The statistical advantage of automatic NLG metrics at the system level

Johnny Tian-Zheng Wei   **Robin Jia**
Department of Computer Science, University of Southern California
{jtwei, robinjia}@usc.edu

Abstract

Estimating the expected output quality of generation systems is central to NLG. This paper qualifies the notion that automatic metrics are not as good as humans in estimating system-level quality. Statistically, humans are unbiased, high variance estimators, while metrics are biased, low variance estimators. We compare these estimators by their error in pairwise prediction (which generation system is better?) using the bootstrap. Measuring this error is complicated: predictions are evaluated against noisy, human predicted labels instead of the ground truth, and metric predictions fluctuate based on the test sets they were calculated on. By applying a bias-variance-noise decomposition, we adjust this error to a noise-free, infinite test set setting. Our analysis compares the adjusted error of metrics to humans and a derived, perfect segment-level annotator, both of which are unbiased estimators dependent on the number of judgments collected. In MT, we identify two settings where metrics outperform humans due to a statistical advantage in variance: when the number of human judgments used is small, and when the quality difference between compared systems is small.[1]

1 Introduction

Automatic metrics are involved in many developmental settings for natural language generation (NLG) systems. In machine translation (MT), metrics like BLEU (Papineni et al., 2002) enable settings where the amount of human effort required would be infeasible, such as architecture or hyperparameter search (Britz et al., 2017). As objective, reproducible quantities, BLEU scores facilitate cross-paper comparisons (Post, 2018). Historically, progress in MT has been attributed to its
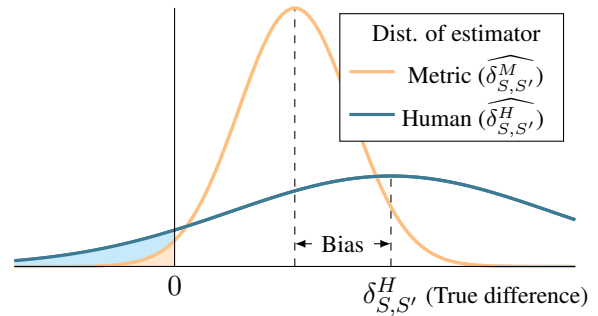

Figure 1: Distribution of estimators for the true difference in system quality $\delta^H_{S,S'}$ between two generation systems (for illustrative purposes). Notation is defined in §2.3. An estimate incurs prediction error if its sign is opposite to the true difference. While humans provide an unbiased estimator of the difference, a biased estimator derived from a metric can have a smaller error probability (shaded areas) due to its lower variance. Evidence supporting the illustration can be found in §5.

use (Callison-Burch et al., 2006). Metrics are an active research area in many NLG subfields, including summarization (Lin, 2004), dialogue (Tao et al., 2018), and image captioning (Anderson et al., 2016), which seek to realize the goal of quick and reliable automatic evaluation.

In all these subfields, the primary goal when conducting evaluation is typically to compare NLG systems. Both human annotators and automatic metrics produce *segment-level* scores, i.e., scores for individual examples, so comparing systems requires aggregating segment-level scores into an overall *system-level* score for each system. Ideally, we would compare systems by their expected human annotator score (an average over infinite human judgments), which we term the true system quality. In practice, we can only estimate this expectation with a sample mean over a finite number of human judgments. Metrics offer a cheaper alternative: we can instead compare systems by their aggregate metric scores on a number of system out-

---

[1] The data and code to reproduce our analyses are available at https://github.com/johntzwei/metric-statistical-advantage.

puts. When comparing systems, we care primarily about how well we estimate the *difference* of their true system qualities, and in particular the sign of this difference (i.e., which system is better), which we term the true pairwise label.

There is a gap in our understanding of *system-level* metrics. To recount a perplexing anecdote, in the most recent edition of the WMT metrics shared task (Mathur et al., 2020b), initial human evaluation disagreed with most metrics on a pairwise prediction of two translation systems. In a manual re-evaluation, the second round results favored the metrics. Our paper offers a statistical explanation for how humans could go "wrong": even if human estimation for the difference in system quality is unbiased, it has high variance. On the other hand, while estimators based on metrics are biased, they have low variance. It is therefore possible for metrics to give a more accurate pairwise prediction than humans when the bias is small (see illustration in Figure 1). Our paper explores this distinction in the following three questions:

(1) *How can we evaluate system-level metrics?* When observing estimator error in terms of pairwise predictions, predictions are evaluated against noisy, human predicted labels rather than the ground truth. In addition, metric predictions fluctuate based on the sample of outputs from the generation system. To disentangle these properties, we examine observed estimator error under a bias-variance-noise decomposition. **Under simulation, we find that the label noise and metric variance account for a small fraction of observed error in both MT and summarization.**

(2) *How good are these metrics?* We compare the errors of metric estimators computed on an infinite number of system outputs, against human estimators with varying amounts of human judgment. We also derive the error of a perfect segment-level annotator (i.e. they provide noiseless/expected human scores for each output), which is also unbiased and judgment dependent. **Empirically, some MT metrics exceed the performance of unbiased estimators with a small number of judgments.**

(3) *What are the limits of system-level evaluation?* The perfect segment-level annotator, as the noiseless human, provides an optimistic estimate for the number of human judgments necessary to achieve a fixed performance. With a power analysis, we can analytically calculate the number of judgments necessary to detect differences between

systems of varying sizes. **When differences in system quality are small, a prohibitively large number of perfect annotator judgments are required to give a correct pairwise prediction.**

2 Formalization

2.1 System-level scores

We will now formalize scoring at the system level, adopting notation from Chaganty et al. (2018). Let $\mathcal{X}$ be a distribution over inputs (e.g. source sentences), and $\mathcal{S}$ be a set of systems (e.g. all translation systems in WMT). Each system $S \in \mathcal{S}$ takes input $x \sim \mathcal{X}$ and returns output $z = S(x)$ (e.g. z is a translation). Let $H(z)$ be a random variable representing a human judgment according to some evaluation prompt (e.g. translation adequacy, from 0-100). A central quantity of interest is the quality of system S, defined as

$$\mu_S^H = \mathbb{E}_{x \sim \mathcal{X}}[H(S(x))] \tag{1}$$

and is not directly observable as it requires infinite human judgment. We can estimate (1) with a finite test set of n examples. Let $x^{(1)}, \ldots, x^{(n)} \overset{\text{i.i.d.}}{\sim} \mathcal{X}$ be a sampled test set and $z^{(1)}, \ldots, z^{(n)}$ be the set of outputs where each $z^{(i)} = S(x^{(i)})$. Human judgments are sampled independently as $y^{(i)} \sim H(z^{(i)})$. The sample mean

$$\widehat{\mu_S^H} = \frac{1}{n}\sum_{i=1}^{n} y^{(i)} \tag{2}$$

is an unbiased estimator of (1). Only (2) is observable, which is a noisy approximation of (1).

A cheaper alternative to estimating the true quality scores is with an estimator based on an automatic metric. Let M (e.g. BERTSCORE) be an automatic metric that takes as input any number of outputs from a system S and produces score

$$\widehat{\mu_S^M} = M(z^{(1)}, \ldots, z^{(n)}) \tag{3}$$

where $\widehat{\mu_S^M}$ is a biased estimator of μ_S^H. As the test set is sampled, the metric score has non-zero variance. Note that while we use the greek letter μ, only some system-level metrics (e.g. ROUGE) are averages of their segment-level counterparts (their score decomposes to $\widehat{\mu_S^M} = \frac{1}{n}\sum_{i=1}^{n} M(z^{(i)})$). Empirically, we find that metrics using other aggregation strategies have convergent properties similar to an average (see Appendix B). We sidestep this by defining the "true" metric score as

$$\mu_S^M = M(z^{(1)}, \ldots, z^{(m)}) \tag{4}$$

for test sets of size m sufficiently large so that this true score is nearly constant.

2.2 Problems in evaluating with correlation

Research in system-level metrics have a tradition of evaluating metric correlation to human judgment with the Pearson correlation coefficient (Reiter, 2018). Formally, these evaluations compare $\widehat{r_M} = \text{Corr}_\mathcal{S}(\widehat{\mu_S^H}, \widehat{\mu_S^M})$ for different metrics M.

Recently, Mathur et al. (2020a) highlights two issues with the use of correlation: First, Pearson's r is neither interpretable nor reflective of system-level metric use in practice. Second, outlier systems (systems with very high/low human/metric scores) can arbitrarily inflate Pearson's r, and outlier systems often exist. Mathur et al. (2020a) propose evaluating metric accuracy in pairwise prediction (can the metric differentiate which generation system is better?) as an alternative that mitigates the issues mentioned above.

We add two points that apply to any measure of metric performance, correlation or pairwise predictions: First, metrics cannot be perfect due to noise in human labels. For instance, while r ranges from $[-1, 1]$, even for the metric that predicts μ_S^H it has $\text{Corr}_\mathcal{S}(\widehat{\mu_S^H}, \mu_S^H) < 1$ due to noise in $\widehat{\mu_S^H}$. It is unclear what is the true upper bound of performance we can expect to achieve. Second, direct measurement of any performance measure on our datasets introduces sample bias (Engstrom et al., 2020). For correlation, $\widehat{r_M}$ could be high because $\widehat{\mu_S^H}$ and $\widehat{\mu_S^M}$ happened to align for this data collection, but a repeat experiment could yield different results. A more holistic view is to give an estimate of average case performance.[2]

The evaluation methodology we derive in §4 addresses the latter points we raise for pairwise predictions and mean squared error (which has direct relationship to the correlation). However, we also believe that pairwise predictions is a step in the right direction, and our discussion continues with pairwise predictions.

2.3 Pairwise predictions

We will now formalize pairwise predictions. For systems $S, S' \in \mathcal{S}$, define the true difference in their system scores as

$$\delta_{S,S'}^H = \mu_S^H - \mu_{S'}^H \qquad (5)$$

and the observed difference as

$$\widehat{\delta_{S,S'}^H} = \widehat{\mu_S^H} - \widehat{\mu_{S'}^H} \qquad (6)$$

and likewise for the differences $\delta_{S,S'}^M$ and $\widehat{\delta_{S,S'}^M}$ w.r.t. to a metric M. In practice, we are interested in the pairwise prediction of S and S' i.e. whether $\delta_{S,S'}^H \overset{?}{>} 0$, given that we have collected human judgments (we observe $\widehat{\delta_{S,S'}^H} \lessgtr 0$), or computed metric scores (we observe $\widehat{\delta_{S,S'}^M} \lessgtr 0$). Refer to Figure 1 for an illustration.

To operationalize the pairwise prediction of S and S', let the true pairwise label

$$\Delta_{S,S'}^H = \text{sign}(\delta_{S,S'}^H) \qquad (7)$$

be defined as the central quantity of interest. Define the human predicted pairwise label as

$$\widehat{\Delta_{S,S'}^H} = \text{sign}(\widehat{\delta_{S,S'}^H}) \qquad (8)$$

and likewise for the true and estimated predictions $\Delta_{S,S'}^M$ and $\widehat{\Delta_{S,S'}^M}$ w.r.t. to a metric M. The 0-1 classification loss for metric M on this example is

$$L(\Delta_{S,S'}^H, \widehat{\Delta_{S,S'}^M}) = \mathbb{I}[\Delta_{S,S'}^H \neq \widehat{\Delta_{S,S'}^M}] \qquad (9)$$

and the pairwise error of an estimator is the loss incurred averaged over all pairwise examples. Ideally, we could calculate the true error of M

$$\text{Err}_{\text{true}}(M) = \mathbb{E}_\mathcal{S}[L(\Delta_{S,S'}^H, \Delta_{S,S'}^M)] \qquad (10)$$

but we can only compute an error of M with noisy human labels and metric scores estimated from finite sized test sets

$$\text{Err}_{\text{obs}}(M) = \mathbb{E}_{\mathcal{X},\mathcal{S}}[L(\widehat{\Delta_{S,S'}^H}, \widehat{\Delta_{S,S'}^M})] \qquad (11)$$

which is typically estimated when we calculate metric pairwise accuracy from our datasets.

3 Datasets

3.1 WMT16-19 metrics shared task

Data. We use the past 4 years of to-English translation data from the WMT metrics shared task (Bojar et al., 2016b, 2017; Ma et al., 2018, 2019).[3] Across all years and language pairs, there are 261 MT systems. Pairs of MT systems are extracted within each year, within each language pair, resulting in 1324 pairwise examples. For each output of an MT system, there are one or more humans judgements and one reference for metric scoring. 1306-5117 outputs were collected for each MT system totaling

---

[2] Pearson's r was not formulated for individual distributions μ_S^H and μ_S^M for each datapoint, so applying the William's test (Graham and Baldwin, 2014) also falls short here.

[3] The WMT20 metrics shared task data was not publicly available at the time of submission.

	Err$_{obs}(\cdot)$	Error components				Err$_{obs}(\cdot)$	Error components		
		c_0Noise	Bias	c_1Var			c_0Noise	Bias	c_1Var
Optimal ($\Delta^{H*}_{S,S'}$)	0.047	0.000	0.000	**0.047**	Optimal ($\Delta^{H*}_{S,S'}$)	0.045	0.000	0.000	**0.045**
Human ($\widehat{\Delta^{H}_{S,S'}}$)	0.065	0.019	0.000	**0.047**	Human ($\widehat{\Delta^{H}_{S,S'}}$)	0.067	0.022	0.000	**0.046**
BERTSCORE	0.102	0.003	**0.086**	0.013	ROUGE	0.296	-0.006	**0.294**	0.008
CHRF	0.124	0.010	**0.105**	0.009	METEOR	0.296	0.004	**0.287**	0.005
BLEURT**	0.128	0.005	**0.108**	0.016	ROUGE-WE	0.317	0.007	**0.301**	0.008
BLEU	0.141	0.008	**0.127**	0.007	BERTSCORE	0.330	-0.004	**0.338**	-0.004
TER	0.184	0.002	**0.173**	0.009	SUPERT***	0.390	0.000	**0.382**	0.008

Table 1: Decomposition of the pairwise error of different metrics (left: WMT, right: SummEval). Highlighted in bold is the largest error component. 10K boostrap trials are conducted for estimation of the expectations (estimation error $< 10^{-3}$). *Denotes an estimator assumed to be unbiased in the simulation. **BLEURT is evaluated only on WMT2019. ***SUPERT is a reference-less metric.

about 1312-5612 judgments, depending on the year and language pair. For ease of interpretation, we always use raw direct assessment judgments which range from 0-100.

Metrics. We evaluate the performance of the three metrics included in SacreBleu (BLEU, TER, chrF; Post, 2018; Koehn et al., 2007). These three have also participated in every year of the metrics task as baselines. In addition, we include two recently developed metrics: BERTSCORE (Zhang et al., 2020) and BLEURT (Sellam et al., 2020). Both metrics are found to effectively utilize contextual embeddings (Devlin et al., 2019), and BLEURT is a learned metric (tuned on data outside of WMT2019). For all metrics, we use the default settings for scoring. Since BLEURT is trained on WMT15-18, we test it only on WMT2019 pairs.

3.2 SummEval

Data. The SummEval dataset (Fabbri et al., 2020) contains 100 outputs from 17 summarization systems. This results in 136 pairwise examples. For each system output, 3 expert judgments, and 11 references for metric scoring. Each summarization is judged in four categories from 0-5: coherence, consistency, fluency, and relevance. To compute system-level human scores for a system, we first average over categories for an aggregate expert score, and then average the aggregated expert scores per system. Metric scores for system outputs were computed against as many references as possible.

Metrics. We evaluate the performance of several metrics that were found to be effective at the system-level in Fabbri et al. (2020). This includes the traditional ROUGE-4 (Lin, 2004) summarization metric, its extension ROUGE-WE (Ng and Abrecht, 2015), and METEOR (Lavie and Agarwal, 2007). In addition, we include two metrics

based on BERT (Devlin et al., 2019). BertScore (Zhang et al., 2020), also present in the WMT analysis, and SUPERT (Gao et al., 2020), which is a reference-less metric for summarization.

4 Decomposing observed metric error

Two sources of variation distinguish the observed pairwise error (11) from the true error in (10) — the noise in the human predicted labels due to finite judgements, and the variance in the metric due to finite test sets. Approximating (11) is straightforward with the bootstrap, but disentangling the error from these two sources of variation requires more care. With the bias-variance-noise decomposition, we can adjust our observed error estimates to the noise-free, infinite test set setting of the true error.

4.1 The bias-variance-noise decomposition

The bias-variance-noise decomposition due to Domingos (2000) decomposes the observed pairwise error in (11) w.r.t. two constant labels for any pairwise example on systems $S, S' \in \mathcal{S}$:
- The *true pairwise label* for this example is

$$\Delta^{H*}_{S,S'} := \arg \min_{y \in \{-1,1\}} \mathbb{E}_{\mathcal{X}}[L(\widehat{\Delta^{H}_{S,S'}}, y)] \quad (12)$$

and the estimator that produces these true labels has, by definition, the lowest observed error. In the decomposition, the human predicted label noise and metric bias is defined relative to the true labels. Assuming the central limit theorem (proof in Appendix A), we actually have $\Delta^{H*}_{S,S'} = \Delta^{H}_{S,S'}$ as defined in eq. (5).

- The *main prediction* of a metric for this ex. is

$$\Delta^{M*}_{S,S'} = \arg \min_{y \in \{-1,1\}} \mathbb{E}_{\mathcal{X}}[L(\widehat{\Delta^{M}_{S,S'}}, y)] \quad (13)$$

and we assume that the metric prediction converges onto the main prediction as the test data

increases for S and S' (empirically validated in Appendix B). In the decomposition, the metric variance is defined relative to the main prediction.

Starting from the loss incurred by M on this pairwise example, the decomposition gives us

$$\mathbb{E}_{\mathcal{X}}[L(\widehat{\Delta^H_{S,S'}}, \widehat{\Delta^M_{S,S'}})] = c_0 \text{Noise}(\widehat{\Delta^H_{S,S'}}) \quad (14)$$
$$+ \text{Bias}(\widehat{\Delta^M_{S,S'}}) + c_1 \text{Var}(\widehat{\Delta^M_{S,S'}})$$

where

- $\text{Noise}(\widehat{\Delta^H_{S,S'}}) = \mathbb{E}[L(\widehat{\Delta^H_{S,S'}}, \Delta^{H*}_{S,S'})]$ where the noise is an irreducible loss incurred by computing pairwise accuracy to the human predicted labels instead of the true labels. Note that this noise term also exactly corresponds to the lowest achievable observable error (see §4.2).

- $\text{Bias}(\widehat{\Delta^M_{S,S'}}) = L(\Delta^{H*}_{S,S'}, \Delta^{M*}_{S,S'})$ where the bias is 0 if the main prediction is correct (w.r.t. to the true label), and 1 otherwise. Note that this term is also the true error of a metric estimator in a noise-free, infinite test set setting. For unbiased estimators this term is zero, as their main prediction matches the true label.

- $\text{Var}(\widehat{\Delta^M_{S,S'}}) = \mathbb{E}[L(\widehat{\Delta^M_{S,S'}}, \Delta^{M*}_{S,S'})]$ where the variance is a likelihood that the estimator deviates from its main prediction under random sampling.

- $c_0 = 2P_{\mathcal{X}}(\widehat{\Delta^M_{S,S'}} = \Delta^{H*}_{S,S'}) - 1$ which means that the influence of label noise on the error becomes small if the estimator prediction are close to random chance. When the estimator gives constant predictions, the sign of c_0 is dependent on the estimator's correctness.

- $c_1 = 1$ if $\widehat{\Delta^M_{S,S'}} = \Delta^{H*}_{S,S'}$ and $c_1 = -1$ otherwise. Variance can both increase and decrease the observed error. If the estimator is unbiased, the variance causes the prediction to from the correct main prediction. On the other hand, for a biased estimator, deviation from its incorrect main prediction occasionally decreases the error.

Unlike the decomposition for mean squared error, the interaction between the c_0 and Var terms only allows the error of two hypothetical settings to be read off directly from the table: when Noise $\to 0$, corresponding to estimator error when computed against the ground truth; or when Noise $+$ Var $\to 0$, when the ground truth is used and metrics have access to an infinite test set for scoring.

4.2 A lower bound for the observed error

By definition the constant estimator that produces the true pairwise labels $\Delta^{H*}_{S,S'}$ (defined in (12)) for each pairwise example has the lowest possible observable error. The observable error of this optimal estimator is exactly $\mathbb{E}[L(\widehat{\Delta^H_{S,S'}}, \Delta^{H*}_{S,S'})] = \text{Noise}(\widehat{\Delta^H_{S,S'}})$. Since this estimator is constant it has no variance, and since it is instantiated by definition it has no bias. Analytically, the observed error of any estimator is lower bounded by $\text{Noise}(\widehat{\Delta^H_{S,S'}})$ and is the agreement of our human predicted labels with the ground truth.

4.3 Best-faith estimation with the bootstrap

Assuming the bootstrap (Efron and Tibshirani, 1993) which is a common procedure in NLP (Dror et al., 2018), we can estimate the expectation quantities in the decomposition. By assuming that sampling with replacement from our datasets approximates real sampling, we can repeatedly simulate the quantity in an expectation. Taking the mean over trials gives the bootstrap estimate of the expectation. We emphasize that this is a regular application of a widely accepted technique—the bootstrap assumption allows us to study problems that would be impossible due to the cost of repeat experiments.

4.4 Results

The following analyses refer to the error components (averaged over all examples) from the simulated decomposition presented in Table 1.

The noise component almost always accounts for a small fraction of the total error. We found this to be counterintuitive—while the lowest observable error (optimal predictions, see §4.2) incur about 5% error on both datasets, the influence of the noise is much smaller than those errors suggest. For the constant c_0 scaling the noise, $c_0 = 0$ if the metric prediction is near random. Since the c_0Noise term on average is small two cases hold true: when humans are uncertain about the example (noise term large) metrics are as well (c_0 term small), and when metrics are certain about the examples (c_0 term large) humans are as well (noise term small). The second case empirically shows studying the sampling distribution of metrics (Koehn, 2004; Berg-Kirkpatrick et al., 2012) is effective, as metric certainty in the difference of system quality often implies human certainty.

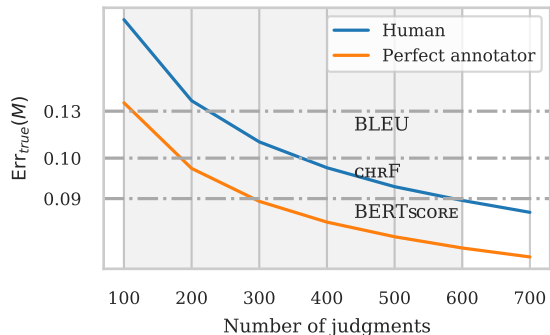**Metric variance introduces little to the pair-**

Figure 2: Comparison of metrics to human and perfect annotator estimators with varying number of judgments in WMT. Errors are adjusted to an idealized setting where true predictions are used for evaluation and metrics are computed on infinite test sets; here metric predictions become constant, so their errors are constant. Shaded in grey is the region where BERTSCORE is superhuman. Results for SummEval are in Appendix D.

wise error, because it is low. Alternatively, metrics stand to gain little from using more test set examples. In MT, dropping both the noise and variance components for the error results in at most a 1 or 2 percent reduction in the observed error (see §9 for the implications in metrics research). Metrics generally have low variance, so at the test set sizes of WMT and SummEval, they are likely to converge to their main predictions.

5   Comparing to the human estimator

In §4, several MT metrics approach the error of the WMT human evaluation. The WMT human evaluation is expensive, using thousands of judgments per translation system. While each human judgment has associated monetary cost, once a large test set is collected, running metrics only incurs computational cost. This section explores this asymmetry, and seeks to understand how much metric predictions are worth, in terms of human judgments.

5.1   Noise-free, variance-free error estimates

We wish to give our best comparison between metrics and unbiased estimators (humans or the perfect annotator). Ideally, metrics would be given their best chance to perform, by using an infinite test set. With the decomposition, we can adjust metric errors estimates to a noise-free and infinite test set setting by taking only their bias component. For human and perfect annotator estimators, we can adjust their errors to a noise-free setting by taking only the variance component. The following

sections compare these adjusted errors.

5.2   Simulating the perfect annotator

While we can estimate the lower bound to the pairwise error for a given dataset (in §4.2), it is achieved by a constant estimator using system-level ground truth. Comparing segment-level metrics against the unbiased "perfect annotator", or the best scorer at the segment-level, is more informative. At the high-level, we can simulate scoring with the perfect annotator at n judgments using the human estimator at $n' > n$ judgments to match the variance of the perfect annotator estimator.

Let's start from the unbiased human estimator $\widehat{\mu_S^H}$ (2). Recall that the estimator is a sample mean, so its variance is $\mathrm{Var}(\widehat{\mu_S^H}) = \mathrm{Var}(H(x))/n$. An insight from Chaganty et al. (2018) gives us the decomposition of the variance of $H(x)$

$$\mathrm{Var}(H(x)) = \mathrm{Var}(\mathbb{E}[H(x)|x]) \qquad (15)$$
$$+ \mathbb{E}[\mathrm{Var}(H(x)|x)]$$

with the law of total variance. In words, the *variance* term can be thought of as the variance of each output sentence's true quality score (some translations produced by S are better than others) and the *expectation* term is the noise introduced by the humans when estimating the quality of a sentence (human scores have mean 0 noise around an output's true quality score).

One intuition is that even if a perfect annotator gives the correct score for each sentence, every time, there is still some unavoidable variance in the estimator due to the variance of the hypothetical quality scores for each output. To formalize this notion, let $P(x) = \mathbb{E}[H(x)|x]$ be the human scoring function of a "perfect annotator", and the estimator $\widehat{\mu_S^P}$ be an empirical mean of n independent samples from $P(x)$ similar to eq. (1). As a sample mean, $\mathrm{Var}(\widehat{\mu_S^P}) = \mathrm{Var}(P(x))/n$. Relating this to (15)

$$\mathrm{Var}(H(x)) = \mathbb{E}[\mathrm{Var}(H(x)|x)] + \mathrm{Var}(P(x)) \quad (16)$$

and while $\mathrm{Var}(P(x))$ is not directly observable, we can calculate $\mathrm{Var}(H(x))$ with the sample variance on all the human judgments, and $\mathbb{E}[\mathrm{Var}(H(x)|x)]$ with a pooled variance over variances from repeat human judgments on the same output sentence.

Our final step considers the efficiency ratio $r = \mathrm{Var}(H(x))/\mathrm{Var}(P(x))$. If we are interested in the perfect annotator estimator at n judgments, the hu-

True mean std. deviation	1	2	3	4	5
25	32489	8124	3612	2032	1301
23	27499	6876	3057	1720	1102
21	22925	5733	2549	1435	919
19	18766	4693	2087	1175	753
17	15024	3757	1671	941	603
15	11697	2926	1301	733	470

Difference in system quality

Table 2: Power analysis for the number of judgments needed to give a pairwise prediction between two systems at .9 accuracy ($\alpha = 0.05, \beta = 0.95$) under ttest assumptions (normality, equal variance) in WMT. WMT ratings are on a 0-100 scale, and the perfect annotator variance in WMT19 was 19.27. Darker cells indicate less feasible experiments, and the colors are set on a log scale. Results for SummEval are in Appendix D.

man estimator at $n' = rn$ judgments has variance

$$\text{Var}(\widehat{\mu_S^H}) = \frac{\text{Var}(H(x))}{rn} \qquad (17)$$

$$= \frac{\text{Var}(P(x))}{n} = \text{Var}(\widehat{\mu_S^P}) \qquad (18)$$

and we invoke the central limit theorem to claim both $\widehat{\mu_S^P}$ and $\widehat{\mu_S^H}$ are normal. This completes our reasoning that for scoring on the system-level, sampling $n' = nr$ human judgments is nearly equivalent to sampling n perfect annotator judgments. See Appendix C for step-by-step derivations for the perfect annotator variance in our datasets.

5.3 Results

The following analyses refer to the comparison of metric estimators to unbiased estimators at varying number of judgments for WMT in Figure 2.

Judgments from the perfect annotator have low variance, like those of professional linguists. While we do not have data from professional linguists, we can qualitatively compare them to the perfect annotator. A growing body of MT literature focuses on professional linguists (Freitag et al., 2020; Mathur et al., 2020b), and there are at least two known properties of their judgments: their judgments have better interannotator agreement (contain less noise), and they are more sensitive to linguistic phenomena. The perfect annotator has no noise, as they assign a constant score to each segment. However, the perfect annotator in WMT is better described as a noiseless crowdworker. With the biases of crowdworkers, the perfect annotator

may not share the sensitivity property, and our use of crowdworkers may be biased w.r.t. professional linguists.

In terms of average pairwise error, MT metrics have an equivalence to a high number of human judgments. Since the error of the human estimator monotonically increases as the number of judgments decrease, each MT metric has a breakeven point. Metrics outperform human estimators using judgments below this threshold. BERTSCORE is as accurate as using a human estimator with 600 judgments per system, or the perfect annotator estimator with 300 judgments, across the WMT dataset. We highlight the statistical advantage in variance many metrics share, and that this advantage offers a *possibility* that metrics can outperform humans, determined by which human estimator the metric is compared against. This is a consequence of the general fact that humans are unbiased, high-variance estimators, and metrics are biased, low-variance estimators, as depicted in Figure 1. For metrics such as BERTSCORE or CHRF, the bias is low as well, which gives it remarkably good error properties.

6 The limits of human evaluation

The perfect annotator provides optimistic figures for human annotation, providing the best performance for a fixed number of judgments, and requiring the least judgments for a fixed performance. In §5, we saw that the perfect annotator is weak at low number of judgments, due to its non-zero variance. In this section we identify another consequence of the perfect annotator's variance, where estimating small differences in system quality is hard.

6.1 Power analysis of the perfect annotator

The performance of an unbiased estimator is dependent on their variance and the effect size it is trying to detect. This section performs a power analysis to determine how much annotator effort is needed to reliably detect the correct pairwise judgment between two systems (Card et al., 2020). To make an optimistic estimate, we assume our annotator variance is close to that of a perfect annotator. We make two assumptions to apply a basic power analysis for the estimation of the difference of system quality between two systems: normality and equal variance across groups. For parameters $\alpha = 0.05$ (false positive rate) and $\beta = 0.95$ (false negative rate), we can analytically compute the number of

judgments needed to ensure our pairwise judgment is at least $\beta(1-\alpha) \approx 90\%$ accurate. Table 2 contains power analyses for different instantiations of annotator variance and effect size.

In WMT, detecting a difference of 1 point requires at least 10K perfect annotator judgments, for different instantiations of its variance. To put this in perspective, the top 5 `zh-en` translation systems in WMT19 differed by less than 3 points (Barrault et al., 2019). Depending on how much is paid per judgment, this cost can quickly become infeasible. Here, the merit of such a task may be argued, as knowing a small difference exists between two systems may not always be productive. From a scientific perspective, many NLG techniques will yield small improvements, and not being able to detect small differences means we will not know whether these techniques are useful.

6.2 Metrics more easily achieve significance

Since metrics tend to have lower variance, metrics often achieve significance in estimating the difference of system qualities, when humans cannot. For instance, BERTSCORE achieves significance in estimating quality differences over half of the pairwise examples where humans do not (see Appendix §E). **In extreme cases, human evaluation is nearly as bad as flipping a coin, but the metric can still offer a consistent prediction between two systems.** When comparing systems similar in quality, practitioners must accept that the number of possible analyses are limited. In ablation studies where similar systems are often compared, metrics may be our only insight into system performance. With white-box metrics such as BLEU, value can be derived from qualitative insight (e.g. systems with high BLEU score have high n-gram overlap with the reference set). In addition, we may qualitatively analyze output statistics not intended to correlate with humans judgment at all (Neubig et al., 2019).

7 Caveats to the analysis

Our analysis assumes that the human judgments are unbiased. In WMT16-19, direct assessment (Graham et al., 2013) was used to elicit judgments from a combination of crowdworkers and researchers. Direct assessment (DA) uses an adequacy evaluation prompt ("Rate how much you agree that the output translation adequately expresses the meaning of the reference translation") and asks contribu-

tors to rate on a 0-100 scale.

The unbiased ground truth is not a fixed goalpost. A number of factors are known to change the eventual ranking of translation systems with human scoring. Employing a different collection methodology, such as human translation edit rate (HTER) of instead of DA, can result in divergent system rankings (Graham et al., 2016). In an earlier edition of WMT, DA judgments were collected with both a grammaticality prompt and an adequacy prompt, corresponding to different system rankings by the respective attribute (Bojar et al., 2016a). Several studies have shown scoring differences between professional linguists and crowdworkers which are due in part to the fact that linguists are more sensitive to linguistic phenomena (Fabbri et al., 2019; Freitag et al., 2019).

The goals of an evaluation should be decided by the practitioner. We do not give suggestions on any particular goals, and practitioners should understand what their application is, and which evaluation is the best approximation (refer to Gatt and Krahmer, 2018). Unfortunately, since the existing data in this domain is limited, our analyses are limited as well. However, the statistical techniques apply to any empirical method. We hope that our analysis inspires others to think about statistical limits in this domain.

8 Pushing the limits of evaluation

To push the limits of what can be evaluated, we need to improve on fundamental aspects of human evaluation. On the human side, we may focus on creating larger effect sizes or reducing noise by adopting new annotation schemes (Läubli et al., 2018; Shapira et al., 2019) or employing professional linguists (Fabbri et al., 2020; Toral et al., 2018). To make the human estimator more efficient, we may consider adaptive data collection techniques to stop data collection early when significance is achieved, in a statistically sound manner (Johari et al., 2017).

Strategies combining human and metric evaluation are also shown to have potential. Variance reduction techniques can be applied to the human estimator by taking advantage of strong metrics (Chaganty et al., 2018). Another bottleneck in human evaluation is in the random sampling of the test set. Metrics could form the basis of an importance sampling procedure to choose test sets that would best differentiate two systems, as a form of

robust evaluation (Chaganty et al., 2017).

On the metric side, if we can reliably estimate metric bias, we can skip human evaluation altogether when the metric is known to be good. Probabilistic reinterpretations of current metrics could be a useful technique for confidence estimation (Keith and O'Connor, 2018). Optimistically, metrics could have provable guarantees, ensuring the correctness of metric decisions (Jia et al., 2019).

9 Best practices for metrics research

We reinterpret problems in evaluating metrics with correlation (§2.2) as a set of guidelines for metrics research. To next year's organizers of the WMT metrics shared task and the broader metrics community we suggest the following: **(1)** Pairwise accuracy has desirable properties as an evaluation measure for metrics. Our bias-variance-noise decomposition shows that the observed pairwise accuracy is very close to the true pairwise accuracy from a noise-free, infinite test set setting (§4.4). We suggest the use of pairwise accuracy as it reflects metric performance well (which may be verified using this analysis). As a normalized form of pairwise accuracy, Kendall's τ is also a suitable measure. **(2)** Since pairwise accuracy is computed against noisy human predictions, on average, it should be impossible for metrics to achieve a perfect accuracy. We suggest providing an upper bound of metric performance (§4.2) to clarify how much improvement is possible for metrics on the dataset.

10 Related work

The fact that a manual evaluation can be weak, and an automatic one can be better is gaining attention in the metrics community. Mathur et al. (2020b) studied a disagreement between crowdworkers and metrics, and a reevaluation favored the metrics over the human prediction. Recently, Freitag et al. (2021) shows that metrics can achieve higher agreement with professional linguists than crowdworkers in judging translation systems. Their results fit into our formalization: if we assume professional linguists are unbiased, the bias and variance properties of metrics combined are superior to those of crowdworkers. Our analysis assumes that crowdworkers are unbiased, where they assume professional linguists are instead.

We wish to highlight several works which inspired the elements of ours: Chaganty et al. (2018) and Hashimoto et al. (2019) formalize metrics as statistical estimators and provide understanding of their statistical properties and limits. In the replication of ImageNet, Engstrom et al. (2020) found that dataset bias accounted for classifier performance differences between the original and the replicated dataset, and provide a decomposition for the sources of error. In automated essay scoring, scorers are often evaluated against noisy human judgment, and Loukina et al. (2020) developed the PRMSE to calculate the MSE between scorer prediction and the true judgment, rather than noisy judgment. Finally, in bioinformatics, Li et al. (2020) derive an upper bound of the R^2 coefficient due to experimental noise when regressing on experiment-derived results.

11 Conclusion

Through rigorous comparison between metrics, humans, and the perfect segment-level annotator, we identify the settings where metrics outperform humans due to a statistical advantage in variance. These results challenge the notion that metrics are always secondary to human evaluation. Instead, we encourage practitioners to understand when human evaluation is weak, and when metrics are necessary. Finally, we hope to provide tools for analysis and future directions for evaluation.

Acknowledgments

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019.

Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Ondrej Bojar, C. Federmann, B. Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016a. Ten years of wmt evaluation campaigns: Lessons learnt. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem"*.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016b. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. Massive exploration of neural machine translation architectures. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1451, Copenhagen, Denmark. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia. Association for Computational Linguistics.

Arun Chaganty, Ashwin Paranjape, Percy Liang, and Christopher D. Manning. 2017. Importance sampling for unbiased on-demand evaluation of knowledge base population. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1038–1048, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pedro M. Domingos. 2000. A unified bias-variance decomposition and its applications. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pages 231–238. Morgan Kaufmann.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Bradley Efron and Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. Springer.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Jacob Steinhardt, and Aleksander Madry. 2020. Identifying statistical bias in dataset replication. *CoRR*, abs/2005.09619.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021.

Experts, errors, and context: A large-scale study of human evaluation for machine translation.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Intell. Res.*, 61:65–170.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4129–4142, Hong Kong, China. Association for Computational Linguistics.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 2017. Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 1517–1525. ACM.

Katherine Keith and Brendan O'Connor. 2018. Uncertainty-aware generative models for inferring document class prevalence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4575–4585, Brussels, Belgium. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Gang Li, Jan Zrimec, Boyang Ji, Jun Geng, Johan Larsbrink, Aleksej Zelezniak, Jens Nielsen, and Martin KM Engqvist. 2020. Performance of regression models as a function of experiment noise.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Anastassia Loukina, Nitin Madnani, Aoife Cahill, Lili Yao, Matthew S. Johnson, Brian Riordan, and Daniel F. McCaffrey. 2020. Using PRMSE to evaluate automated scoring systems in the presence of label noise. In *Proceedings of the Fifteenth Workshop*

on Innovative Use of NLP for Building Educational Applications, pages 18–29, Seattle, WA, USA → Online. Association for Computational Linguistics.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020a. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and OndÅ™ej Bojar. 2020b. Results of the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 682–687, Minneapolis, Minnesota. Association for Computational Linguistics.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 722–729. AAAI Press.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers*, pages 113–123, Belgium, Brussels. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Equivalence between optimal prediction and true system differences

There is a slight difference between the definition of the true difference in (5) which we can alternatively define as

$$\Delta_{S,S'}^{H} = \text{sign}(\mathbb{E}[\widehat{\mu_S^H} - \widehat{\mu_{S'}^H}]) \qquad (19)$$

and the definition of the optimal prediction $\Delta_{S,S'}^{H*}$ in (12), which is positive when

$$P_{\mathcal{X}}(\widehat{\mu_S^H} - \widehat{\mu_{S'}^H} > 0) > \frac{1}{2} \qquad (20)$$

and the two are not immediately equivalent. However, if we assume that the central limit theorem applies (which can be reasonable as our sample means always have $n > 100$) and $X = \widehat{\mu_S^H} - \widehat{\mu_{S'}^H}$ is normal, the CDF of X is

$$F(x) = \phi((x - \mathbb{E}[X])/\text{Var}(X)) \qquad (21)$$

where ϕ is the CDF of the standard normal distribution. Since the standard normal is centered and symmetric, $\phi(x) > 1/2 \iff x > 0$. Together we have

$$F(x) > \frac{1}{2} \iff \mathbb{E}[X] > x \qquad (22)$$

where for $x = 0$ the left and right hand sides are equivalent to (20) and (19), respectively.

B Convergence of metric predictions to the main prediction

A key assumption in interpreting the results from the bias-variance-noise decomposition in §4 is that as system-level metrics have access to more outputs for evaluation, metric predictions converges onto the main prediction.

For many metrics, their system-level score is the mean of their segment-level scores (e.g. BLEURT, BERTSCORE, ROUGE etc.). This is true for all summarization metrics. For these metrics, assuming the central limit theorem allows us to prove that metrics converge to the main prediction, similar to the proof in Appendix A. However, some MT metrics (BLEU, TER, and CHRF) are not simple averages of their segment-level scores, making them harder to analyze.

For system-level metrics that are not simple averages, we analytically observe that their aggregation method is similar to a mean (e.g. BLEU is a macro-average). We empirically verify that as the system-level metric evaluates on more test set outputs, their pairwise predictions converge to the

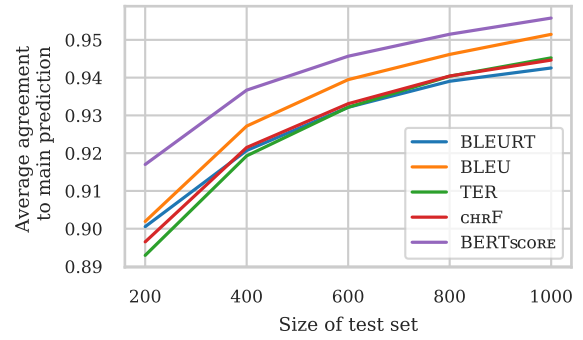main predictions. Refer to Figures 3 and 4.

Figure 3: Average agreement of the main prediction to metric predictions computed from varying test set sizes in WMT. The main predictions were derived from all of our data. Each point was an estimated with 10K bootstrap trials. As the size of the test set increases, we see that the agreement monotonically increases. Note that only BLEURT and BERTSCORE are means of their segment-level scores.
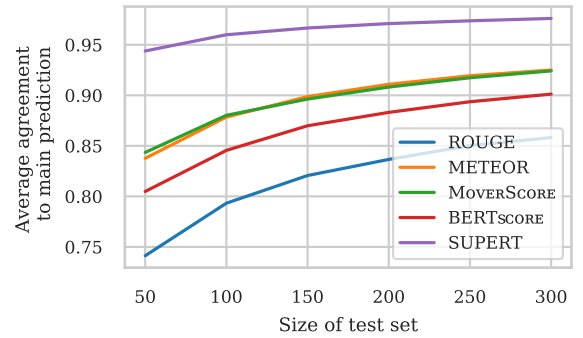

Figure 4: Average agreement of the main prediction to metric predictions evaluated on varying test set sizes in SummEval. The main predictions were derived from all of our data. Each point was an estimated with 10K bootstrap trials. As the size of the test set increases, we see that the agreement monotonically increases. Note that all metrics are means of their segment-level scores.

C Efficiency ratios for the perfect annotator

With repeat human judgments for a given output example, we can estimate the variance of the perfect annotator (or true segment-level score variance) in WMT and SummEval. For WMT, we use only valid judgments (SYSTEM and REPEAT judgments), and discard all attention check judgments (BAD_REF judgments). For SummEval, we use the dataset as is.

In WMT, we analyze all the to-English data grouped by year. We believe this grouping is appropriate because the to-English evaluation is batched together every year. Direct assessment, which WMT uses to collect human judgments (Graham et al., 2013), is a score assigned by crowdworkers to an English translation while referring to an English reference, requiring only monolingual knowledge.

	2016	2017	2018	2019
$\sqrt{\text{Var}(H(x))}$	30.01	29.65	28.21	28.81
$\sqrt{\mathbb{E}[\text{Var}(H(x)|x)]}$	17.53	22.96	19.57	21.42
$\sqrt{\text{Var}(P(x))}$	24.36	18.76	20.33	19.27
$\text{Var}(H(x))/\text{Var}(P(x))$	1.52	2.50	1.93	2.24

Table 3: Step-by-step derivation for the efficiency ratio r (fourth row) of the perfect annotator estimator for WMT16-19 as defined in §4.1. Square roots are taken so that values are in terms of the original units (standard deviations, judgments range from 0-100). These were calculated on to-English data only.

	Expert	Turker
$\sqrt{\text{Var}(H(x))}$	0.717	0.745
$\sqrt{\mathbb{E}[\text{Var}(H(x)|x)]}$	0.293	0.475
$\sqrt{\text{Var}(P(x))}$	0.655	0.574
$\text{Var}(H(x))/\text{Var}(P(x))$	1.201	1.686

Table 4: Step-by-step derivation for the efficiency ratio r (fourth row) of the perfect annotator estimator for SummEval as defined in §4.1. Square roots are taken so that values are in terms of the original units (standard deviations, judgments range from 1-5). Note that there is little agreement between experts and turkers at the system level.

D SummEval analysis results

The main analyses in §5 and §6 are presented for SummEval here. When comparing expert humans to metrics, no metric comes close to expert performance at any number of expert judgments. For the power analysis, small differences are also hard to detect, similar to the findings in WMT. Note that while the perfect expert requires relatively less judgments compared to the perfect crowdworker in WMT, judgments from experts are likely to be much more expensive.
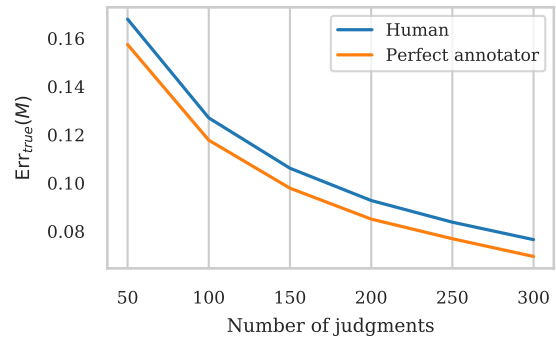

Figure 5: Comparison of metrics to human and perfect annotator estimators with varying number of judgments in SummEval. Errors are adjusted to an idealized setting where true predictions are used for evaluation and metrics are computed on infinite test sets; here metric predictions become constant, so their errors are constant. No metric comes close to expert performance at any number of judgments (ROUGE, the best performing summarization metric, has error 0.221).
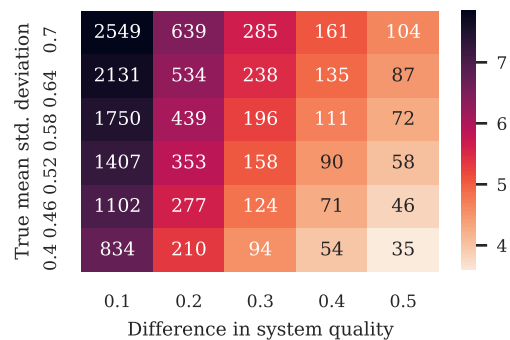

Table 5: Power analysis for the number of judgments needed from the perfect expert to give a pairwise judgment between two systems at .9 accuracy ($\alpha = 0.05, \beta = 0.95$) under ttest assumptions (normality, equal variance) in SummEval. SummEval ratings are on a 1-5 scale, and the true segment quality variance was 0.655. Darker cells indicate less feasible experiments, and the colors are set on a log scale.

E Metric and human significance breakdown

For the pairwise examples in WMT, we provide the co-occurrence of significance for metric and human estimators. Refer to Figures 6 and 7 for analyses on BERTSCORE and BLEURT, respectively.
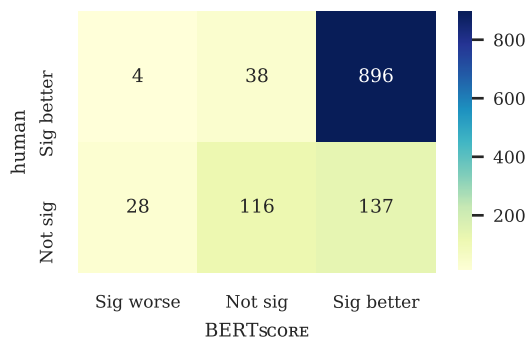

Figure 6: Co-occurrence of BERTSCORE and human significance on pairs in WMT16-19. Pairs are ordered so that the human difference in system quality is always positive. Significance is tested with a one-sided bootstrap resampling test, in the direction of the difference for both humans and metrics with 1K trials at $\alpha = 0.05$. BERTSCORE achieves significance more than half of the time when humans cannot.
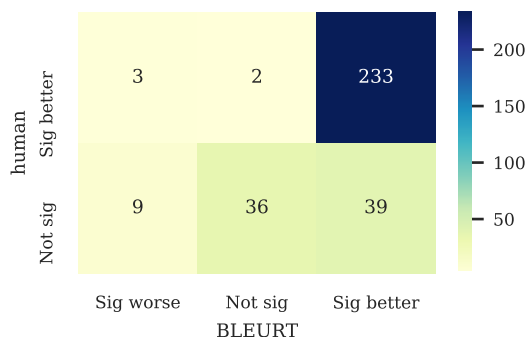

Figure 7: Co-occurrence of BLEURT and human significance on pairs in WMT19. Pairs are ordered so that the human difference in system quality is always positive. Significance is tested with a one-sided bootstrap resampling test, in the direction of the difference for both humans and metrics with 1K trials at $\alpha = 0.05$. BLEURT achieves significance more than half of the time when humans cannot.