Unifying Human and Statistical Evaluation for Natural Language Generation

Tatsunori B. Hashimoto*[1,2] Hugh Zhang*[1] Percy Liang[1,2]
(* equal contribution)
[1]Department of Computer Science [2]Department of Statistics
Stanford University
{thashim,hughz}@stanford.edu pliang@cs.stanford.edu

Abstract

How can we measure whether a natural language generation system produces both high quality and diverse outputs? Human evaluation captures quality but not diversity, as it does not catch models that simply plagiarize from the training set. On the other hand, statistical evaluation (i.e., perplexity) captures diversity but not quality, as models that occasionally emit low quality samples would be insufficiently penalized. In this paper, we propose a unified framework which evaluates both diversity and quality, based on the optimal error rate of predicting whether a sentence is human- or machine-generated. We demonstrate that this error rate can be efficiently estimated by combining human and statistical evaluation, using an evaluation metric which we call HUSE. On summarization and chit-chat dialogue, we show that HUSE detects diversity defects which fool pure human evaluation and that techniques such as annealing for improving quality actually decrease HUSE due to decreases in diversity.

1 Introduction

Generating text is a core part of many NLP tasks such as image captioning (Lin et al., 2014), open-domain dialogue (Sordoni et al., 2015), story generation (Roemmele, 2016), and summarization (Nallapati et al., 2016). However, proper evaluation of natural language generation has proven difficult (Liu et al., 2016; Novikova et al., 2017; Chaganty et al., 2018). A good evaluation metric should not only capture the *quality* of generation, but also the *diversity* of generation, which is especially crucial for open-ended tasks like dialogue or story generation.

Human evaluation, which is often viewed as the gold standard evaluation, captures quality perfectly but fails to capture diversity. As an example, for language modeling, a model that directly plagiarizes sentences from the training set would pass the human quality bar but would have zero generalization ability and thus have inadequate diversity. On the other hand, *statistical evaluation*—i.e., perplexity on a test set—captures diversity, as it ensures a model must have some probability of generating novel sentences, but perplexity provides an inadequate measure of quality (Theis et al., 2015): modifying a perfect model by making it incapable of generating just a single reference sentence results in *infinite* perplexity even though the model has near perfect diversity and sample quality. Automatic metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin and Rey, 2004) capture quality better than perplexity but still correlate poorly with human evaluation, the gold standard for judging sample quality (Novikova et al., 2017; Chaganty et al., 2018); moreover, these metrics do not capture diversity.

We propose defining the ideal evaluation metric as twice the error rate of the *optimal discriminator* that classifies whether a sentence is generated from the reference distribution or from the model (Section 2). If a model generates gibberish (low quality), the optimal discriminator can classify these accurately as coming from the model.

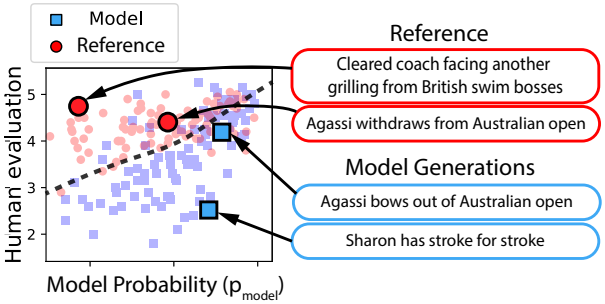

Figure 1. HUSE is twice the classification error of distinguishing reference and generated text represented as (human judgment, p_{model}) pairs. This identifies samples with defects in both quality (Sharon has stroke . . .) and diversity (Cleared coach facing . . .).

If the reference distribution contains sentences the model cannot generate (low diversity), the optimal discriminator can classify these accurately as coming from the reference.

Unfortunately, the optimal discriminator is unavailable. Human discriminators cannot capture diversity effectively, and learned discriminators—e.g., from a Generative Adversarial Network (Goodfellow et al., 2014) or one trained on human judgments (Lowe et al., 2017)—are too unreliable to use as a replacement for human evaluation.

Our key result (Section 3) is that the optimal classifier depends only on two numbers: the probability of a sentence under the model and the probability under the reference distribution. The former can be computed from the model, and the latter can be well approximated by human judgment scores. The resulting two-dimensional space is illustrated in Figure 1. We apply a simple k-nearest neighbor classifier to this space and define Human Unified with Statistical Evaluation (HUSE) as twice the leave-one-out error of this classifier.

We apply HUSE to four natural language generation tasks (Section 5): language modeling, dialogue, story generation, and summarization. First, we show that human evaluation alone is insufficient to discriminate model generations from the references, leading to inflated estimates of model performance. In contrast, HUSE is able to reveal gaps between current models and true human performance. We also show that techniques for improving sample quality such as annealing actually increase distinguishability between the model and reference due to substantial losses in diversity.

2 Optimal Discriminator

Consider a natural language generation task where the model is given a context x (e.g., a dialogue history) drawn from some $p(x)$ and must output a distribution over possible sentences $p_{\text{model}}(y \mid x)$. We define an idealized evaluation metric based on whether p_{model} is close to a *reference distribution* p_{ref}, which is generally human-generated.[1] Specifically, consider a random variable y drawn from either the reference or the model based on an indicator $z \sim \text{Bernoulli}(\frac{1}{2})$:

$$y \mid x, z \sim \begin{cases} p_{\text{ref}}(y \mid x) & \text{if } z = 1 \\ p_{\text{model}}(y \mid x) & \text{if } z = 0. \end{cases} \quad (1)$$

Define L^* to be twice the lowest possible error over any discriminator f that attempts to determine z based on x and y:

$$L^* \stackrel{\text{def}}{=} 2 \inf_f \mathbb{P}[f(x, y) \neq z]. \quad (2)$$

L^* measures similarity between p_{model} and p_{ref}; it is 0 if p_{model} and p_{ref} are disjoint and 1 if they are identical.[2]

Obstacles. Unfortunately, L^* is unattainable because it requires computing the optimal discriminator. In the spirit of the Turing Test, we could consider using the error rate of a human discriminator f_{hum} as a proxy. However, existing human evaluation requests sentences judgments one at a time, resulting in f_{hum} having knowledge of p_{ref} but not p_{model}. Thus, f_{hum} is unable to determine which sentences a model *cannot* generate.

As a concrete example, suppose p_{ref} placed a uniform distribution over some set S. Without knowledge of p_{model} the most sensible discriminator is to predict $z = 1$ (reference) when $y \in S$. This discriminator achieves the same classification error of 0.5 for both the perfect model $p_{\text{model}} = p_{\text{ref}}$ and one which can only return a single $y \in S$. We could try to reveal p_{model} to humans by showing multiple samples simultaneously, but this is expensive – and as we will later see – unnecessary.

Another option is to learn f over an expressive class of functions such as neural networks on data sampled from p_{model} and p_{ref}. This is analogous to learning the discriminator in a Generative adversarial network (GAN) (Goodfellow et al., 2014) or an evaluation metric from human judgments (Lowe et al., 2017). However, as (x, y) are high-dimensional objects, training a good classifier is extremely difficult (and perhaps not significantly easier than solving the original generation problem). Indeed, learned evaluation metrics do not generalize very well (Lowe et al., 2017; Chaganty et al., 2018). Unlike these approaches which seek to replace human evaluation, our focus will instead be on combining human and automatic statistical evaluation to estimate the optimal classifier error.

[1] While some tasks care only for quality and thus only require p_{model} to place mass on *some* high quality y, we demand p_{model} to place mass on *all* high quality y. This diversity is important for open-ended tasks such as dialogue or story generation.

[2] Note that L^* is a linear function of total variational distance: $\|p_{\text{model}} - p_{\text{ref}}\|_{\text{TV}} \stackrel{\text{def}}{=} \sum_{x,y} p(x) |p_{\text{model}}(y \mid x) - p_{\text{ref}}(y \mid x)| = (1 - L^*)$. See Appendix A.1 for details.

3 Human Unified with Statistical Evaluation (HUSE)

Our key result is that the optimal discriminator depends on (x, y) only through a two-dimensional sufficient statistic (Section 3.1), motivating an approximation which we call HUSE (Section 3.2). We finish by showing how the various evaluation metrics relate (Section 3.3).

Firstly, for any feature map ϕ that maps (x, y) to $\phi(x, y) \in \mathbb{R}^d$, define an evaluation score $L(\phi)$ to be twice the error rate of the optimal discriminator that depends on (x, y) through ϕ:

$$L(\phi) \stackrel{\text{def}}{=} 2 \inf_f \mathbb{P}[f(\phi(x, y)) \neq z]. \quad (3)$$

Note that the evaluation score $L(\phi)$ given by a feature map ϕ optimizes over all functions that depend on ϕ (3). Thus, the more information ϕ contains, the lower $L(\phi)$ is. This has two implications: First, any feature map ϕ yields an (optimistic) *upper bound* on L^*, meaning that $L(\phi)$ can detect when a model is poor but cannot certify that it is good. Second, adding features to ϕ can only improve this bound.

3.1 Two features suffice

Let us consider the following two-dimensional feature map:

$$\phi_{\text{opt}}(x, y) \stackrel{\text{def}}{=} [p_{\text{ref}}(y \mid x), p_{\text{model}}(y \mid x)]. \quad (4)$$

From the arguments above, it is clear that $L(\phi_{\text{opt}}) \geq L^*$, but perhaps somewhat surprisingly, we actually have equality:

Proposition 1. *The two-dimensional feature map ϕ_{opt} achieves the optimal discriminator score: $L(\phi_{opt}) = L^*$.*

Proof We compute the true posterior over z given x, y. Since $p(z = 1) = p(z = 0) = \frac{1}{2}$, $p(y \mid x, z = 1) = p_{\text{ref}}(y \mid x)$ and $p(y \mid x, z = 0) = p_{\text{model}}(y \mid x)$, by Bayes' rule:

$$p(z = 1 \mid x, y) = \frac{p_{\text{ref}}(y \mid x)}{p_{\text{ref}}(y \mid x) + p_{\text{model}}(y \mid x)}.$$

The optimal discriminator simply predicts $z = 1$ if $p_{\text{ref}}(y \mid x) > p_{\text{model}}(y \mid x)$ and $z = 0$ otherwise. In other words, the decision line is $\phi_{\text{opt}}(x, y)_1 > \phi_{\text{opt}}(x, y)_2$. □

3.2 HUSE features

While we can directly compute $p_{\text{model}}(y \mid x)$ for many probabilistic models, $p_{\text{ref}}(y \mid x)$ is unattainable, so $L(\phi_{\text{opt}})$ is not computable. However, the *wisdom of the crowds* (Surowiecki, 2004; Ungar et al., 2012) suggests that pooling together the judgments of many humans can often produce surprisingly reliable estimates of real-world probabilities such as $p_{\text{ref}}(y \mid x)$, even if no individual human is particularly reliable. With this motivation, we ask Amazon Mechanical Turk workers to rate a sentence from 1–5 based on how "typical" it is (see Appendix A.2 for more details). We define $\text{HJ}(x, y)$ to be the average response over 20 crowdworkers. Figure 2 shows that for a language modeling task on the Reddit corpus,[3] $\text{HJ}(x, y)$ strongly correlates with the actual log-frequency of y in the corpus. The high correlation suggests that human judgments $\text{HJ}(x, y)$ are a good surrogate for $\log p_{\text{ref}}$.

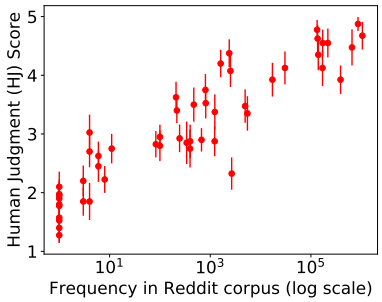

Figure 2. On the Reddit corpus, human judgment (HJ) of the "typicality" of a sentence y correlates strongly ($r = 0.92$) with its frequency in the corpus, suggesting that HJ is a good surrogate for $\log p_{\text{ref}}$. Error bars at the 90% confidence interval.

In addition, we found that rather than using the model probability $p_{\text{model}}(y \mid x)$ directly as a feature, normalizing by sentence length $\text{len}(y)$ yielded lower (tighter) scores, and since any feature map yields a valid upper bound on L^*, we define the following feature map:

$$\phi_{\text{huse}}(x, y) \stackrel{\text{def}}{=} \left[\frac{\log p_{\text{model}}(y \mid x)}{\text{len}(y)}, \text{HJ}(x, y)\right] \quad (5)$$

which is used to define the (population) *HUSE score* $L(\phi_{\text{huse}})$.

3.3 Guarantees derived from HUSE

We now show that the HUSE score satisfies two nice properties: (i) a model with low HUSE score

[3] We used the Reddit corpus due to crowdworker familiarity, corpus size, and short average sentence length, which results in a wide range of sentence frequencies.

must have low optimal classification error (i.e. HUSE can detect low-quality models) and (ii) a model with high HUSE score must have high classification error when using humans as classifiers.

Consider a feature map that only includes human evaluation: $\phi_{\text{hj}}(x, y) \stackrel{\text{def}}{=} [\text{HJ}(x, y)]$. Because ϕ_{huse} also incorporates human evaluation, $L(\phi_{\text{huse}})$ is always tighter than the human discriminator error $L(\phi_{\text{hj}})$:

Proposition 1 (Relationship between HUSE, human evaluation, and optimal scores)**.**

$$L(\phi_{hj}) \geq L(\phi_{huse}) \geq L^*. \tag{6}$$

Furthermore, the main difference between $L(\phi_{\text{huse}})$ and L^* is that the former uses $\text{HJ}(x, y)$ and the latter uses p_{ref}. But as we argued using Figure 2, $\text{HJ}(x, y)$ is strongly correlated with p_{ref}, so we expect $L(\phi_{\text{huse}})$ to be relatively close to L^*.

4 Evaluating models with HUSE

In this section, we show how we can estimate the population HUSE score $L(\phi)$ from finite data (Section 4.1). We then show how HUSE can be decomposed into a score that measures quality (HUSE-Q) and a score that measures diversity (HUSE-D), which allows us to study quality-diversity tradeoffs (Section 4.2).

4.1 Learning a discriminator

For any feature map ϕ, we show how to produce an estimate of $L(\phi)$. Fix n contexts $x_1, \ldots, x_n$. First, we draw n examples $y_1, \ldots, y_n$ from the reference distribution $p_{\text{ref}}(y \mid x)$, which are usually human-generated sentences from a test set. We also draw n examples $y'_1, \ldots, y'_n$ from the model $p_{\text{model}}(y \mid x)$ we wish to evaluate. Next, for each of the $2n$ examples (x, y), we compute the feature map $\phi(x, y)$, which might involve evaluating the model probability $p_{\text{model}}(y \mid x)$ as well as collecting human judgments $\text{HJ}(x, y)$ from crowdworkers.

Finally, we compute the leave-one-out error of a classifier that tries to predict whether a given example (x, y) comes from the reference distribution ($z = 1$) or the model ($z = 0$). The choice of classifier is not important, but we chose k-nearest neighbors because it is simple, requires no training, and can capture arbitrary continuous decision boundaries. Specifically, we set $k = 15$ and define neighbors using L_2 distances over the feature vectors $\phi(x, y)$ scaled componentwise to have unit variance. The overall procedure for computing the estimate $\hat{L}(\phi)$ is formally defined in Algorithm 1.

Algorithm 1 Computing discriminator score

Require: Feature map ϕ
Contexts $x_1, \ldots, x_n$
Reference outputs $y_1, \ldots, y_n$
Model outputs $y'_1, \ldots, y'_n$

1: Construct dataset:

$$\mathcal{D} = \bigcup_{i=1}^{n} \{(\phi(x_i, y_i), 1), (\phi(x_i, y'_i), 0)\}$$

2: $\hat{L}(\phi) \stackrel{\text{def}}{=}$ leave-one-out error of k-NN on $\mathcal{D}$

4.2 Quality-diversity decomposition

We now define the (empirical) *HUSE score* using the feature map ϕ_{huse}:

$$\text{HUSE} \stackrel{\text{def}}{=} \hat{L}(\phi_{\text{huse}}). \tag{7}$$

We define the quality component of HUSE (HUSE-Q) similarly using human judgments alone:

$$\text{HUSE-Q} \stackrel{\text{def}}{=} \hat{L}(\phi_{\text{hj}}). \tag{8}$$

Since humans alone can detect quality defects in a model, any increase in error from removing p_{model} must come from a model's lack of diversity. Therefore, we define the diversity component (HUSE-D) as follows:

$$\text{HUSE-D} \stackrel{\text{def}}{=} 1 + \text{HUSE} - \text{HUSE-Q}, \tag{9}$$

which implies the decomposition $(1 - \text{HUSE-D}) + (1 - \text{HUSE-Q}) = 1 - \text{HUSE}$. As long as the discriminators are non-trivial (obtaining better than chance performance with sufficient data), all scores are contained in $[0, 1]$. Here, $\text{HUSE-D} = 1$ implies that the model suffers no diversity defects, while $\text{HUSE-D} = 0$ indicates that the examples could be discriminated perfectly due to a lack of diversity.

5 Experiments

5.1 Experimental setup

We use HUSE to evaluate three different types of single-sentence natural language generation tasks:

Method	Summarization		Story generation		Chit chat dialogue		LM
	$t = 1.0$	$t = 0.7$	$t = 1.0$	Retrieval	$t = 1.0$	$t = 0.7$	$t = 1.0$
HUSE	**0.53**	0.26	**0.06**	0.00	**0.56**	0.49	**0.86**
HUSE-Q	0.58	**0.92**	0.15	**0.47**	0.56	**0.92**	**0.88**
HUSE-D	**0.95**	0.34	**0.91**	0.53	**1.00**	0.57	**1.02**

Table 1. Performance achieved by the best models on the four tasks, as measured by overall goodness-of-fit (HUSE), sample quality (HUSE-Q) and diversity (HUSE-D). The scale ranges from 0.0 (completely distinguishable) to 1.0 (indistinguishable from reference) where the implied classification error is HUSE/2.

(i) unconditional and high entropy (language modeling); (ii) conditional and high entropy (story generation, chit-chat dialogue); and (iii) conditional and low entropy (summarization).

In more detail, the four tasks are:

- **Summarization**: Giganews story to headline dataset and the pre-trained model from Gehrmann et al. (2018).
- **Story generation**: Last sentence generation for ROC stories (Mostafazadeh et al., 2016) using a standard OpenNMT model with global attention (Klein et al., 2017).
- **Language modeling**: One billion word benchmark language model from Jozefowicz et al. (2016).
- **Chit-chat Dialogue**: Two-turn chit-chat dialogue dataset constructed from Reddit comments and their replies (Appendix A.3). We train a convolutional model from fairseq (Gehring et al., 2017).

For each task, we evaluate three standard schemes for managing diversity-quality tradeoffs. We exclude beam search since HUSE ≈ 0 due to its extreme lack of diversity.

- **Temperature annealing**: for any probabilistic model that generates words sequentially, we sample a word proportional to $p^{1/t}$ where p is the model's distribution and t is the temperature parameter.
- **Retrieval**: using Apache `solr`, we retrieve responses from the training set using the default BM25 similarity metric.
- **Overfitting**: models were trained for triple the number of minibatches necessary to minimize validation loss. This forces the model to aggressively memorize the training set and increase generation quality.

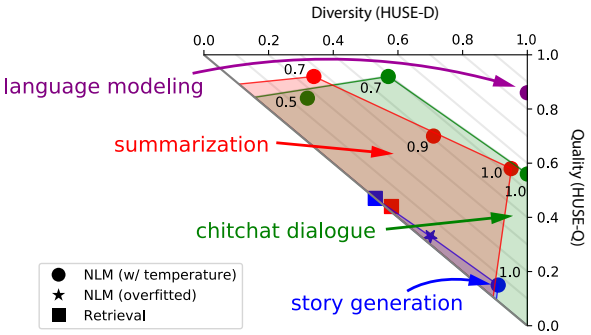

Figure 3. Tradeoffs between HUSE-D and HUSE-Q for various tasks. Closer to the top right is better, and shaded gray lines indicate HUSE. Generation mechanisms can trade-off between diversity and quality but cannot easily increase the underlying model performance (HUSE).

We selected several schemes for each task based on whether we expected them to improve either HUSE-Q or HUSE-D. For cost reasons, we did not comprehensively measure all combinations of tasks and generation strategies, but our evaluations cover the general range of available diversity-quality tradeoffs.

Finally, we collect human judgments HJ(x, y) as per Section 4.1 where we query 20 Amazon Mechanical Turk crowdworkers for typicality ratings on 100 reference and model sentences.

5.2 Overall results

The HUSE scores across the four tasks vary widely. Table 1 shows that language models are nearly indistinguishable, with HUSE $= 0.86$ and implied discriminator error of 43%.

In contrast, both summarization and dialogue are highly distinguishable (HUSE ≈ 0.5) with relatively low quality when sampled from $t = 1.0$. Human evaluation alone (HUSE-Q) would suggest that using temperature annealing to emphasize high-probability outputs substantially improves the model ($t = 0.7$). However, we find that this increase in sample quality comes at the cost of diversity. Examining the achievable HUSE and diversity tradeoffs in Figure 3 shows that mech-

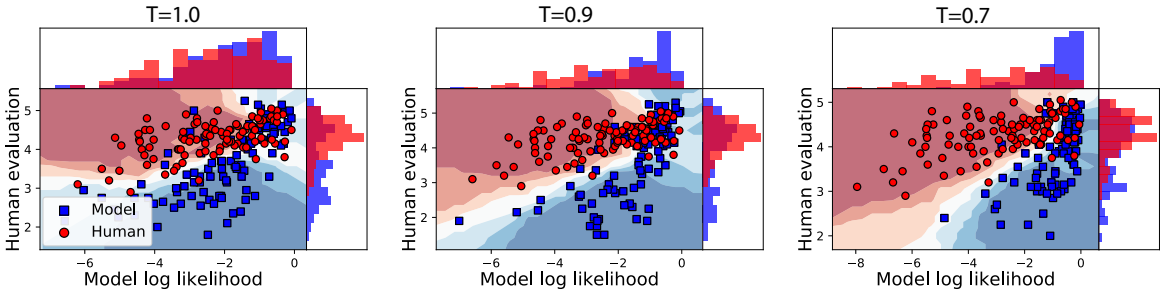

Figure 4. The two dimensional classification problem in Algorithm 1 on the summarization task with different softmax temperatures (three panels). Each point represents a sentence ($\phi_{\text{huse}}(x_i, y_i)$), color is the source of the sentence (z), shading is the classification confidence of the nearest neighbor classifier.

anisms such as annealing which improve sample quality actually degrade HUSE due to severe losses in diversity.

We find that all schemes and models are inadequate for story generation on ROC stories. The original model ($t = 1.0$) is very easily distinguishable by a human (HUSE-Q $= 0.15$), corresponding to a discriminator error of 7%. The retrieval models can improve this to HUSE-Q $= 0.47$, but this comes at the expense of a catastrophic loss of diversity. Even the best scheme we found (overfitting the model) could not avoid this tradeoff.

Finally, we observe that directly sampling the model ($t = 1.0$) is always diverse. This suggests that human evaluation is an appropriate evaluation for systems which are trained to maximize log-likelihood and generate via sampling.

5.3 Understanding HUSE

Since HUSE is estimated from a two-dimensional classification problem, we can directly visualize the classification problem to understand the diversity-quality tradeoffs.

Figure 4 shows $\phi_{\text{huse}}(x_i, y_i)$ for the summarization task on both reference (blue square) and model (red circle) outputs. The shaded areas indicate the decision boundary of the k-nearest neighbor classifier.

At temperature $t = 1.0$, we find that the classification boundary is mostly horizontal, implying that human judgment can distinguish model outputs from references. However, there exists a cluster of sentences with high HJ and high p_{model} which are essentially indistinguishable. Examining the samples in this top-right region reveals that these are news stories with short headlines such as "nadal pulls out of sydney international" which can be reliably generated even at $t = 1.0$.

At lower temperatures of $t = 0.9$ and $t = 0.7$ the boundary shifts towards becoming diagonal. Although the distribution is no longer directly separable on human judgment, the two distributions are quite separable with the inclusion of p_{model}. There is no longer a group of un-identifiable points on the top right, as the model-generated text now receives noticeably higher p_{model} than the references.

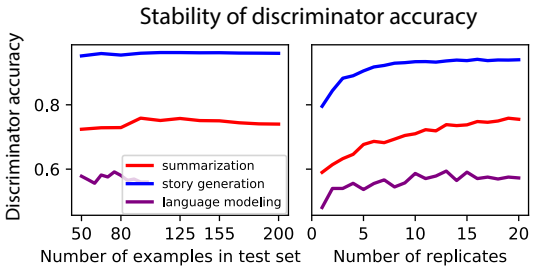

Figure 5. Estimates of HUSE are robust to small test set size, but generally require $\approx$ 20 replicate measurements for HJ.

Using Figure 4, we can identify individual examples which were correctly and incorrectly classified based on $\log p$ and HJ. Table 2 shows examples of both quality failures and diversity failures identified by HUSE. For example, the "Diversity failure" table shows that the summarization model does not understand that "front office" is a way to refer to the president and general manager and thus assigns very low probability to the reference. Improving these models on the diversity failures will require that the model understand these more subtle paraphrases. We can also identify model successes, where the model outputs are truly indistinguishable from the reference, and those in which the reference itself is low quality.

5.4 HUSE stability

Recall that HUSE depends on HJ(x, y) which reduces noise by averaging over crowdworkers. We show that depending on the task, fewer replicate

	Quality failure	$\log p_{\text{model}}$	HJ
Context:	two new vaccines have been shown effective against rotavirus, which is responsible for a half-million infant deaths in poor countries each year, research studies published wednesday said.		
Model	two new vaccines in the poor countries were effective against go-it-alone study says	-2.3	2.6
Reference	new vaccines for key UNKNOWN virus shown effective	-4.0	4.3
	Diversity failure		
Context:	the buffalo bills sacked tom donahoe as president and general manager on wednesday, fulfilling expectations of a shake-up after another failure to make the national football league playoffs.		
Model	bills sack UNKNOWN as president gm and general manager	-0.9	4.3
Reference	nfl 's bills shake up front office	-5.1	4.3
	Model is indistinguishable		
Context:	us veteran and eight-time grand slam winner andre agassi has withdrawn from this month 's australian open due to a nagging ankle injury , his management team announced thursday .		
Model	agassi bows out of australian open after injury	-1.4	5.3
Reference	agassi withdraws from australian open	-0.3	4.9
	Model outperforms reference		
Context:	israeli prime minister ariel sharon was undergoing an emergency operation thursday after suffering a massive stroke.		
Model	sharon undergoing emergency operation	-0.8	4.9
Reference	timeline of sharon era	-4.7	2.9

Table 2. Summarization examples extracted from Figure 4 for the OpenNMT pretrained model ($t = 1.0$) and references.

crowdworkers could be used to estimate HJ.

Figure 5 shows the result of subsampling our original data on 200 sentences and 20 crowdworkers and estimating HUSE. First, we find that using 50 test set examples (Figure 5, left) is sufficient to give accurate estimates of HUSE. Next, we find that the necessary replicate count depends heavily on the task. For easily distinguishable tasks (story generation), 10 replicates suffice, while less distinguishable tasks (summarization) require more than 20 replicates to obtain accurate estimates.

6 Related work

The current state of NLG evaluation. Existing approaches to NLG evaluation use a mix of quality and diversity measures. Out of the 26 NLG papers at ACL 2018, six perform only human evaluation, fourteen measure human evaluation and a diversity metric such as perplexity or n-gram diversity, and six do not even evaluate using human judgments.

While perplexity and n-gram counts can *in principle* evaluate diversity, their practical implementations suffer from serious drawbacks. When human evaluation and perplexity are both evaluated, they are almost always done on separate models – human evaluations are done on beam-searched output, while perplexity is computed on the softmax outputs. This makes it appear as if the models can simultaneously generate high quality outputs while also being diverse, when in fact they can only be one at a time based on whether they sample or run beam search.

On the other hand, n-gram diversity was proposed by (Li et al., 2016) to combat the generic utterance problem where models repeat phrases such as 'I don't know'. Unfortunately, n-gram diversity is computed *across* contexts by counting the number of unique n-grams generated, and so does not measure a model's ability to generate multiple valid utterances at any single context. In particular, a model which can only output a single utterance per context (e.g., via memorization or retrieval) can still have high n-gram diversity as long as the memorized sentences are unique.

Finally, *all* existing diversity measures are computed separately from human evaluation. This results in two incomparable evaluation metrics, which prevent us from reasoning about tradeoffs between diversity and quality. In contrast, HUSE allows us to make precise statements about the cost of diversity because it is a single metric which decomposes into diversity and quality terms.

Related evaluations of diversity. The importance of diverse responses has previously acknowledged for summarization (Nenkova et al., 2007) and information retrieval (Clarke et al.,

2008). Our work differs in considering a *single* evaluation measure that captures quality and diversity applicable to *any* generation task.

Automated metrics based on n-gram overlap such as BLEU, METEOR, ROUGE (Papineni et al., 2002; Lavie and Denkowski, 2009; Lin and Rey, 2004) work well for machine translation but do not generalize well to domains with a diverse spectrum of correct responses. While variants (Sun and Zhou, 2012; Galley et al., 2015; Shima and Mitamura, 2011) have adapted such metrics to high entropy generative environments, they are still significantly inferior to the human judgments they attempt to mimic.

Caccia et al. (2018) recently examined the diversity and quality tradeoffs for different language model architectures on synthetic datasets. However, as their approach relies on measuring log-likelihoods under both the model and reference distributions, it *cannot* be applied to real data where p_{ref} is unavailable. Our main conceptual contribution overcomes this by showing that HJ is an acceptable proxy for p_{ref}.

Sajjadi et al. (2018) also examines diversity and quality (which they call precision and recall) in the context of generative image models. However, they rely on assuming that p_{ref} and p_{model} can be estimated accurately using the Fréchet Inception Distance (FID) (Heusel et al., 2017). HUSE avoids such assumptions and instead directly leverages the human judgments HJ, resulting in a simple and reliable metric more suitable for use as a gold-standard.

Estimating optimal classification error. Evaluating a model by estimating its optimal classification error has been considered by several earlier works (Olsson et al., 2018; Kannan and Vinyals, 2016; Li et al., 2017; Bruni and Fernandez, 2017; Bowman et al., 2016). However, these methods have focused on classifying sentences directly which is quite challenging to do reliably. In fact, existing adversarial evaluation methods do not yet reliably outperform human classification (Kannan and Vinyals, 2016; Bruni and Fernandez, 2017). We propose the use of both human evaluation and model probabilities as part of the adversarial evaluation framework, and demonstrate that the resulting classifier reliably outperforms humans and captures both the sample quality and diversity of a model.

Distributional divergence estimation. Our proposed evaluation metric is closely related to the total variation distance which has been studied extensively in the distribution testing literature. It is known that total variation distance estimates have pessimistic minimax estimation rates in high dimensions (Balakrishnan and Wasserman, 2017). Our work overcomes this by utilizing p_{model} and an estimate of p_{ref}. Other approaches to distributional testing include the maximum mean discrepancy (MMD) and Wasserstein distances, which achieve better rates but at the cost of relying strongly on a kernel and distance metric respectively (Tolstikhin et al., 2016; Singh et al., 2018). Although such divergences are easier to estimate than the total-variation distance, the implied convergence rates are still too slow to be used as a replacement for human evaluation.

7 Discussion

In this paper, we demonstrate that the current gold standard of human evaluation does not penalize under-diverse models. To remedy this, we propose HUSE, a general purpose evaluation which can be applied to any model for which we can calculate p_{model}. HUSE is a upper bound to the optimal classification error of distinguishing reference and model generated text, and *never* does worse than human classification. HUSE leverages both p_{model} and HJ, ensuring that models which have high HUSE are both high-quality and diverse.

In many ways, our work can be seen as an extension to the classic Turing Test (Turing, 1950). Instead of relying on just a human classifier, we approximate the *optimal* classifier by adding additional information about the model through p_{model}.

Finally, our work here focuses on natural language generation as a distribution matching task, where the goal is $p_{\text{ref}} \approx p_{\text{model}}$. However, many settings may have different goals. In machine translation (Bahdanau et al., 2015), for example, a single high-quality output may suffice. Other tasks may have external measures of model quality (task oriented dialogue), or there may be a desire to generate with *super* human quality (story generation). In such cases, HUSE can be extended by constraining the classifier f appropriately for the task. Given that the current state of NLG has not yet approached human level generation, we leave such investigations to future work.

References

D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

S. Balakrishnan and L. Wasserman. 2017. Hypothesis testing for high-dimensional multinomials: A selective review. *arXiv preprint arXiv:1712.06120*.

S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. 2016. Generating sentences from a continuous space. In *Computational Natural Language Learning (CoNLL)*, pages 10–21.

E. Bruni and R. Fernandez. 2017. Adversarial evaluation for open-domain dialogue generation. In *Proceedings of the SIGDIAL 2017 Conference*.

M. Caccia, L. Caccia, W. Fedus, H. Larochelle, J. Pineau, and L. Charlin. 2018. Language gans falling short. *arXiv preprint arXiv:1811.02549*.

A. Chaganty, S. Mussmann, and P. Liang. 2018. The price of debiasing automatic metrics in natural language evaluation. In *Association for Computational Linguistics (ACL)*.

C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *ACM SIGIR*.

M. Galley, C. Brockett, A. Sordoni, Y. Ji, M. Auli, C. Quirk, M. Mitchell, J. Gao, and B. Dolan. 2015. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. *arXiv preprint arXiv:1506.06863*.

J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

S. Gehrmann, Y. Deng, and A. M. Rush. 2018. Bottom-up abstractive summarization. In *Empirical Methods in Natural Language Processing (EMNLP)*.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*.

R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

A. Kannan and O. Vinyals. 2016. Adversarial evaluation of dialogue models. In *NIPS 2016 Workshop on Adversarial Training*.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

A. Lavie and M. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23.

J. Li, M. Galley, C. Brockett, J. Gao, and W. B. Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Human Language Technology and North American Association for Computational Linguistics (HLT/NAACL)*, pages 110–119.

J. Li, W. Monroe, T. Shi, A. Ritter, and D. Jurafsky. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

C. Lin and M. Rey. 2004. Looking for a few good metrics: ROUGE and its evaluation. In *NTCIR Workshop*.

T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll'ar, and C. L. Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755.

C. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Association for Computational Linguistics (ACL)*.

N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, and J. Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *North American Association for Computational Linguistics (NAACL)*.

R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

A. Nenkova, R. J. Passonneau, and K. McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. In *ACM Transactions on Speech and Language Processing*.

J. Novikova, O. Dušek, A. C. Curry, and V. Rieser. 2017. Why we need new evaluation metrics for NLG. In *Empirical Methods in Natural Language Processing (EMNLP)*.

C. Olsson, S. Bhupatiraju, T. Brown, A. Odena, and I. Goodfellow. 2018. Skill rating for generative models. *arXiv preprint arXiv:1808.04888*.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Association for Computational Linguistics (ACL)*.

M. Roemmele. 2016. Writing stories with help from recurrent neural networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

M. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. 2018. Assessing generative models via precision and recall. *arXiv preprint arXiv:1806.00035*.

H. Shima and T. Mitamura. 2011. Diversity-aware evaluation for paraphrase patterns. In *Empirical Methods in Natural Language Processing (EMNLP)*.

S. Singh, A. Uppal, B. Li, C. Li, M. Zaheer, and B. Poczos. 2018. Nonparametric density estimation under adversarial losses. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 246–257.

A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Nie, J. Gao, and B. Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *North American Association for Computational Linguistics (NAACL)*.

H. Sun and M. Zhou. 2012. Joint learning of a dual SMT system for paraphrase generation. In *Association for Computational Linguistics (ACL)*.

J. Surowiecki. 2004. *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. Doubleday and Co.

L. Theis, A. van den Oord, and M. Bethge. 2015. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*.

I. Tolstikhin, B. K. Sriperumbudur, and B. Scholkopf. 2016. Minimax estimation of maximum mean discrepancy with radial kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1930–1938.

A. M. Turing. 1950. Computing machinery and intelligence. *Mind*, 49:433–460.

L. Ungar, B. Mellors, V. Satopää, J. Baron, P. Tetlock, J. Ramos, and S. Swift. 2012. The good judgment project: A large scale test of different methods of combining expert predictions. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

A Appendix

A.1 Relationship between total variation distance and optimal discriminator error

This is a standard result, replicated here for completeness:

Proposition 2. *The total variation distance is related to the optimal discriminator error as follows:* $\|p_{model} - p_{ref}\|_{TV} = (1 - L^*)$.

Proof Fix any x. Define $a_y \stackrel{\text{def}}{=} p_{\text{ref}}(y \mid x)$ and $b_y \stackrel{\text{def}}{=} p_{\text{model}}(y \mid x)$. Let $S \stackrel{\text{def}}{=} \{y : a_y < b_y\}$ be the y where the p_{model} assigns higher probability than p_{ref}, and define $A \stackrel{\text{def}}{=} \sum_{y \in S} a_y$ and $B \stackrel{\text{def}}{=} \sum_{y \in S} b_y$ be the aggregated probabilities. On S, the optimal discriminator should return $z = 0$ (model). This is an error when $z = 1$, which occurs with probability $\frac{1}{2}A$. Analogously, on the complement of S, the error probability (when $z = 0$) is $\frac{1}{2}(1 - B)$. The total contribution to L^* is thus $A + (1 - B)$. The rest follows from algebra:

$$\|p_{\text{model}} - p_{\text{ref}}\|_{\text{TV}} = \frac{1}{2}\|p_{\text{model}} - p_{\text{model}}\|_1 \tag{10}$$

$$= \frac{1}{2}[(B - A) + (1 - A) - (1 - B)] \tag{11}$$

$$= B - A = (1 - L^*) \tag{12}$$

□

A.2 Amazon Mechanical Turk for human judgments

In order to show that HUSE can be reliably estimated even with simple crowdsourcing techniques, we used a single uniform task design where we asked Amazon Mechanical Turk workers to rate the typicality of a sentence from 0-5. We defined 0 as invalid (grammatically or factually incorrect) and 5 as 'very typical'. HJ(x, y) is defined as the average score that crowdworkers assign to a response y given the context x. We did not perform substantial filtering or qualification checks beyond HIT acceptance rate.

We observe that measuring many replicates is sufficient to get low-variance estimates of HJ. For easy classification tasks (such as story generation) we require five to ten replicates, while for hard tasks such as summarization at least twenty replicates are needed (Section 5.4). Manual inspection suggests that up to 20% of the collected data are low-quality but that this noise is uncorrelated with the sentence being rated and outweighed by a larger majority of honest and reasonably accurate data.

A.3 Reddit Dataset

We use a subset of Reddit comments from 2006-2018 scraped from https://pushshift.io/. We construct a dictionary containing the 10000 most popular words and preprocess the dataset by removing deleted posts, out-of-vocabulary tokens, profanity, comments with less than 10 upvotes, and comments with over 400 tokens.

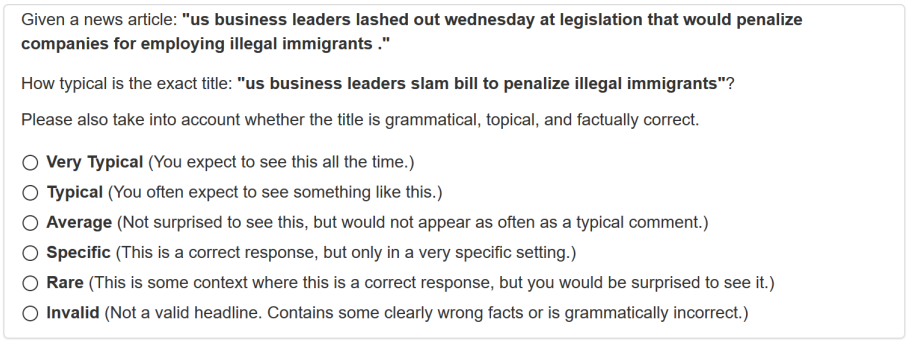

Figure 6: Amazon Mechanical Turk survey design for eliciting HJ in the summarization task.