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 (i) HUSE detects diversity defects which fool pure human evaluation and that (ii) techniques such as annealing for improving quality actually decrease HUSE due to decreased 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 creative, open-ended tasks like dialogue or story generation.

Human evaluation, which is often viewed as the gold standard evaluation, captures quality but fails to capture diversity. As an example, for language
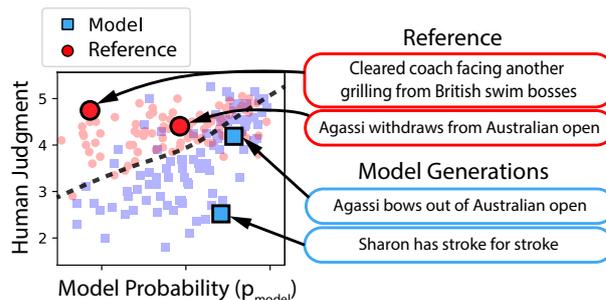

Figure 1: HUSE is twice the classification error of distinguishing reference and generated text based on human judgment scores and model probabilities. HUSE identifies samples with defects in quality (Sharon has stroke . . .) and diversity (Cleared coach facing . . .).

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 reference test set—captures diversity, as it ensures a model must assign reasonable probability to novel sentences, but perplexity provides an inadequate measure of quality (Theis et al., 2015). For example, modifying a perfect model by removing its ability to generate even a single test sentence results in *infinite* perplexity even though the model is still near-perfect. 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 and fail to capture diversity (Novikova et al., 2017; Chaganty et al., 2018).

Existing approaches to combining statistical and human evaluation have been ad-hoc, leading to misleading performance measures. A common approach is to measure diversity through the perplexity of a probabilistic model and quality through human evaluation on beam-searched out-

puts. This gives the illusion that a single model is high-quality and diverse, while the reality is that it shows we can have either a diverse model (when sampling from the distribution used to compute perplexity) *or* a high-quality model (when beam-searching).

In this paper, we define the idealized evaluation metric as twice the error of the *optimal discriminator* for classifying sentences as coming from the reference distribution or the model (Section 2). If a model generates gibberish (low quality), the optimal discriminator can classify these accurately as coming from the model. 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 for rigorous evaluation.

Our key result (Section 3) is based on the observation 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 directly from the model, and we show that 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 in 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, chitchat 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 deficiencies of current models. We also show that common techniques for improving sample quality such as annealing actually increase distinguishability between the model and reference due to losses in diversity.

2 Optimal Discriminator

Consider a natural language generation task where the model is given a context x (e.g., a dialogue his-

tory) drawn from some prior $p(x)$ 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 $z \sim \text{Bernoulli}\left(\frac{1}{2}\right)$:

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

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} instead, often considered the gold standard for evaluation. However, while humans might have knowledge of p_{ref}, they do not have full knowledge of p_{model} and thus would have difficulties determining 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

---

[1] While some tasks only care about quality and thus only require p_{model} to place mass on *some* high quality y, we demand that p_{model} places mass on *all* high quality y as given by p_{ref}. This diversity is important for open-ended tasks such as dialogue or story generation. Also note that p_{ref} need not be the human distribution, or match the training distribution. It can be defined as the distribution given by experts.

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

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 learning 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.

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).

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

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

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^* (2), meaning that $L(\phi)$ might be able 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:

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

From the arguments above, it is clear that $L(\phi_{\text{opt}}) \geq L^*$, but perhaps more 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 boundary is given by $\phi_{\text{opt}}(x, y)_1 > \phi_{\text{opt}}(x, y)_2$. □

More generally, we can obtain this equality with a wider class of ϕ. It will hold exactly for any invertible transformation of ϕ_{opt} (Appendix Corollary 1), and approximately for any ϕ which has high mutual information with ϕ_{opt} (Appendix Theorem 1). This means that we can substitute p_{ref} with noisy, possibly un-normalized estimates and still obtain accurate estimates of L^*.

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 as a way to estimate $p_{\text{ref}}(y \mid x)$. (see Appendix A.3 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 $\log p_{\text{ref}}$.

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. We therefore define the HUSE features as follows:

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

---

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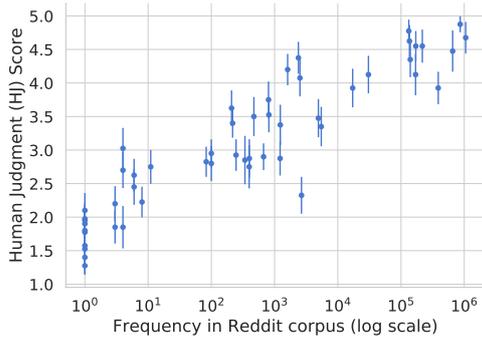
Figure 2: On the Reddit corpus, human judgment (HJ) of the "typicality" of a sentence y correlates strongly ($r = 0.92$) with its frequency in the corpus, suggesting that HJ is a good surrogate for $\log p_{\text{ref}}$. Error bars at the 90% confidence interval.

and define the (population) *HUSE score* as $L(\phi_{\text{huse}})$.

3.3 Guarantees derived from HUSE

We now show that the HUSE score satisfies two nice properties: (i) HUSE does at least as well as human evaluation and (ii) a low HUSE score is *sufficient* to show that a model is far from the reference distribution.

To show (i), 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 (lower) than the human discriminator error $L(\phi_{\text{hj}})$:

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

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

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}, and good approximations to p_{ref} provide approximation guarantees for $L(\phi_{\text{huse}})$ (Appendix Theorem 1).

4 Evaluating models with HUSE

In this section, we show how we can estimate the error rate $L(\phi)$ from finite data (Section 4.1). We then show how the HUSE estimate ($\hat{L}(\phi_{\text{huse}})$) 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 classification problems for HUSE are two-dimensional, which allows us to accurately estimate error rates using a k-nearest neighbors classifier. We opt to use nearest neighbors classifiers as they are simple, require no training, and can asymptotically capture arbitrary continuous decision boundaries. Specifically, we set $k = 16$ 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 Estimating error rates under ϕ
___
Require: Feature map ϕ, number of neighbors k
    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}:

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

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

1692

alone:

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

Since humans 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:

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

which implies the decomposition $(1 - \text{HUSE-D}) + (1 - \text{HUSE-Q}) = 1 - \text{HUSE}$. As long as the discriminators are non-degenerate (obtaining better performance than chance and $\text{HUSE} > \text{HUSE-Q}$), 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: (i) unconditional and high entropy (language modeling); (ii) conditional and high entropy (story generation, chit-chat dialogue); and (iii) conditional and low entropy (summarization). We show that HUSE provides a direct and interpretable measure of diversity on high-entropy tasks, while also serving as a useful model diagnostic on low-entropy ones.

The four tasks along with the datasets and models are as follows:

- **Summarization**: Giganews story to headline dataset and the pre-trained model from Gehrmann et al. (2018). The dataset consists of 3.8 million news story-headline pairs. Examples from this dataset are shown in Table 2.

- **Story generation**: Last sentence generation for ROC stories (Mostafazadeh et al., 2016) consisting of 96,198 examples of partially written four-sentence stories as input, and a single sentence which completes the story as the target. We use a standard OpenNMT model with global attention (Klein et al., 2017).
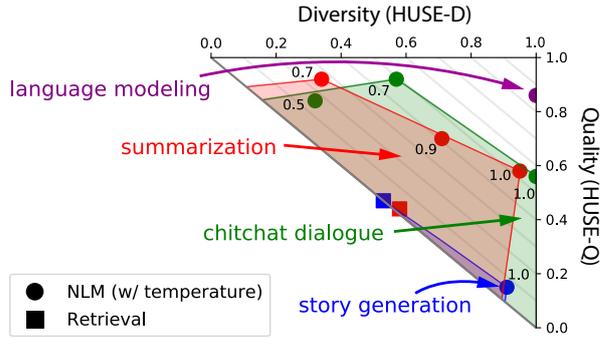
Figure 3: Tradeoffs between HUSE-D and HUSE-Q. Points are models and color indicates task. Neural models (circle) generate using temperature annealing (point labels indicate temperature). Models closer to the top right are superior, and gray diagonal lines indicate equivalent HUSE. A shaded region for a task indicates models which are strictly dominated (worse HUSE with the same HUSE-D–HUSE-Q proportion). Annealing can trade-off between diversity and quality but cannot easily increase the underlying model performance (HUSE).

- **Language modeling**: One billion word benchmark pre-trained language model from Jozefowicz et al. (2016). The task consists of generating a single sentence from the one billion word newswire text distribution.

- **Chit-chat dialogue**: Two-turn chit-chat dialogue dataset consisting of 37.3 million comment-response pairs from Reddit (Appendix A.4). Comments are generally short (5–15 tokens) and cover a single topic (e.g. given "wow how did i not notice that", the response is "you were focusing on other things its understandable"). We train a convolutional model using fairseq (Gehring et al., 2017).

For all the tasks, we train neural models and evaluate their diversity-quality tradeoffs as we change the decoding scheme for generation. Our primary evaluation concerns diversity trade-offs involving **temperature annealing** which is a generation technique applicable to any probabilistic model that generates words sequentially. In temperature annealed models, we sample a word w proportional to $p^{1/t}(w)$ where p is the model probability of w given previous words and t is the temperature parameter. We excluded beam search since it qualitatively behaves similarly to temperature annealing with low temperatures and $\text{HUSE} \approx 0$ due to beam search being extremely

Score	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 for HUSE and HUSE-Q ranges from 0.0 (completely distinguishable from reference) to 1.0 (indistinguishable from reference) where the implied classification error is HUSE/2. HUSE-D may exceed 1.0 with small sample sizes when HUSE-Q > HUSE.

under diverse.

As a non-neural baseline, we also consider retrieval based models based on Apache `solr` on a few tasks. For this approach, we retrieve the single most relevant response from the training set using the BM25 similarity metric on inputs. Such models are known to perform well in tasks with complex outputs such as program generation (Hayati et al., 2018; Hashimoto et al., 2018) and style transfer (Li et al., 2018).

For cost reasons, we did not measure certain combinations of task and generation mechanisms. We did not measure retrieval for chit-chat dialogue, as we observed its outputs were lower quality than a low-temperature neural model. We also did not anneal language models, as the generation quality from the language model was already high, and our goal was to show that they achieved high HUSE. Our set of measurements, while not comprehensive, generally covers the available quality-diversity tradeoffs for conditional tasks.

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 100 model sentences. Since our models generate UNK (unknown and out-of-vocabulary) tokens, we instructed crowdworkers to treat UNK tokens as rare, but appropriate words for the context.

5.2 Overall results

The HUSE scores across the four tasks vary widely. Table 1 shows that single-sentence 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 sug-

gest that using temperature annealing ($t = 0.7$) to emphasize high-probability outputs substantially improves the model (HUSE-Q goes from 0.58 to 0.92 for summarization and 0.56 to 0.92 for dialogue). However, we find that this increase in sample quality comes at the cost of diversity (HUSE-D goes from 0.95 to 0.34 for summarization and 1.0 to 0.57 for dialogue). Examining the achievable HUSE and diversity tradeoffs in Figure 3 shows that mechanisms such as annealing which improve sample quality actually degrade HUSE due to severe losses in diversity.

We find that all generation 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 diversity.

Finally, we observe that directly sampling from the model ($t = 1.0$) is always diverse. This suggests that human evaluation is an appropriate evaluation for generation systems that are directly sampled (rather than beam-searched).

5.3 Model error analysis with HUSE

Since HUSE is estimated from a two-dimensional classification problem, we can directly visualize the classification problem to understand defects in both model quality and diversity.

Figure 4 shows both reference points $\phi_{\text{huse}}(x_i, y_i)$ (blue squares) and model points $\phi_{\text{huse}}(x_i, y_i')$ (red circles) for the summarization task. The shaded areas indicate the decision boundary of the 16-nearest neighbor classifier.

At temperature $t = 1.0$, we find that the classification boundary is mostly horizontal, implying that human judgment alone can distinguish model outputs from references. There is a cluster of sentences with high HJ and high p_{model} which are es-
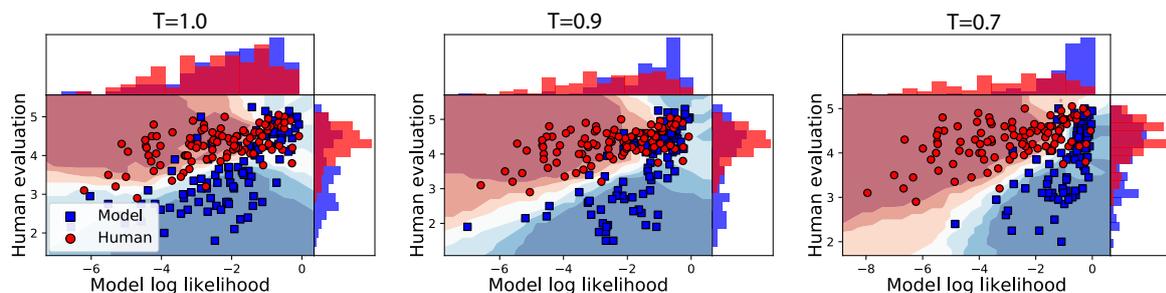
Figure 4: The two-dimensional classification problem in Algorithm 1 on the summarization task with different softmax temperatures (three panels). Each point represents a reference sentence $\phi_{\text{huse}}(x_i, y_i)$ or model-generated sentence $\phi_{\text{huse}}(x_i, y_i')$. The color denotes the source of the sentence (z), shading is the classification confidence of the nearest neighbor classifier.

sentially 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 $t = 1.0$. However, the model frequently generates low quality samples that can easily be distinguished such as "two new vaccines in the poor countries were effective against go-it-alone study says" (Table 2).

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 clearly separable with the inclusion of p_{model}.

Using Figure 4, we can identify individual examples which were correctly and incorrectly classified based on p_{model} 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 ($t = 0.7$) has an extremely low probability of generating some reference sentences ("NFL's bills shake up front office") and is thus under-diverse. Closer examination of the model shows that the probability of generating "front office" is low, since it is an unusual way to refer to the president and general manager. Improving these models on the diversity failures will require that the model understand more subtle paraphrases. We can also identify model successes, where the model outputs are indistinguishable from the reference in terms of quality ("Agassi bows out of Australian Open after injury"), and the model assigns high probability to the reference ("Agassi withdraws from Australian Open").
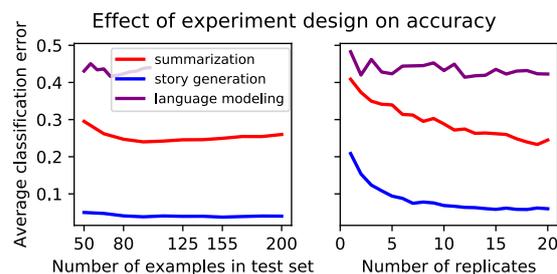

Figure 5: Estimates of HUSE are robust to small test set size, but generally require ≈ 20 crowdworker measurements for each example.

5.4 HUSE stability

Since HUSE depends on human crowdworker annotations, one might ask if it is possible to reduce either the number of annotated examples, or number of distinct crowdworkers for each example. We show that for low-quality models, substantially fewer annotations are needed.

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

6 Related work

The current state of NLG evaluation. Existing approaches to NLG evaluation use a hodgepodge mix of quality and diversity measures. Out of the 26 NLG papers at ACL 2018, six perform only hu-

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 <UNK> 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 <UNK> 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

Table 2: Example reference and model outputs (capitalization added for readability) corresponding to Figure 4 (summarization task) that were shown to crowdworkers (left column). Crowdworkers were shown samples from the model (including the <UNK> token) and returned human judgments (right column). Using human judgments and the model probability, we can identify several types of failures. Quality failures are examples that are classified by human judgment. Diversity failures are examples that are classified by model probabilities. Finally some examples are not easily classified, as they have similar human judgment and model probability scores.

man evaluation, fourteen measure human evaluation and a diversity metric such as perplexity or n-gram diversity, and six do not 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 identify models with 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 only outputs a single memorized utterance per context

(e.g., via memorization or retrieval) can still have high n-gram diversity as long as the memorized sentences differ across contexts.

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 tradeoffs between model quality and 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 been 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 human judgments, 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. 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, but these approaches require knowledge of a ground truth metric or kernel space (Tolstikhin et al., 2016; Singh et al., 2018). Although such divergences are easier to estimate than the total variation distance from samples, the implied convergence rates are still too slow to be practically useful.

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 strategy which can be applied to any model for which we can calculate a model's sampling probabilities. HUSE is an upper bound on the optimal classification error of distinguishing reference and model-generated text, and *never* does worse than human classification. HUSE leverages both model probabilities and human judgments, ensuring that models which do well on the metric are both high-quality and diverse.

Our work can be viewed as a "superhuman version" of the classic Turing Test (Turing, 1950). Instead of relying on just a human classifier, we approximate the *optimal* classifier, which can utilize information about the model in addition to the reference. We also modify the classification problem and seek to identify whether a sample comes from a (potentially superhuman) reference distribution, rather than the human distribution. These two changes lead to tractable, rigorous estimators which can quantify tradeoffs between model quality and diversity on a wide range of generation tasks.

Reproducibility. All code, data, and experiments are available on the CodaLab platform at https://worksheets.codalab.org/worksheets/0x88644b5ee189402eb19d39d721d1005c.

References

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. Feder and N. Merhav. 1994. Relations between entropy and error probability. *IEEE Transactions on Information Theory*, 40:259–266.

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)*.

T. Hashimoto, K. Guu, Y. Oren, and P. Liang. 2018. A retrieve-and-edit framework for predicting structured outputs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

S. A. Hayati, R. Olivier, P. Avvaru, P. Yin, A. Tomasic, and G. Neubig. 2018. Retrieval-based neural code generation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

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, R. Jia, H. He, and P. Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *North American Association for Computational Linguistics (NAACL)*.

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:

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

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

$$= B - A = (1 - L^*). \quad (12)$$

$\square$

A.2 Approximation error from ϕ features

Theorem 1. *Let L^* and $L(\phi)$ be the optimal classification error and optimal error under feature map ϕ respectively. Then,*

$$L^* \leq L(\phi) \leq L^* + 2(1 - 2^{-I})$$

where $I \stackrel{\text{def}}{=} I(Z_{opt}; \phi_{opt}(X, Y) \mid \phi(X, Y))$ is the conditional mutual information in bits and Z_{opt} is the prediction of the optimal classifier.

Proof The lower bound falls out of the definition of L^*. To prove the upper bound, a variant of the entropy lower bound by Feder and Merhav (Feder and Merhav, 1994) shows that the error rate for predicting Z_{opt}, via the optimal $f(\phi(X, Y))$ follows

$$P(f(\phi(X, Y)) \neq Z_{opt})$$
$$\leq 1 - 2^{I(Z_{opt}; \phi(X,Y)) - H(Z_{opt})}. \quad (13)$$

Now expand the mutual information using the chain rule

$$I(Z_{opt}; \phi(X, Y)) = I(Z_{opt}; \phi_{opt}(X, Y), \phi(X, Y))$$
$$- I(Z_{opt}; \phi_{opt}(X, Y) \mid \phi(X, Y))$$
$$= -I(Z_{opt}; \phi_{opt}(X, Y) \mid \phi(X, Y)) + H(Z_{opt}).$$

The last line follows from the fact that Z_{opt} is a deterministic function of ϕ_{opt} (Proposition 1). Substituting this into the inequality gives the bound,

$$P(f(\phi(X, Y)) \neq Z_{opt}) \leq 1 - 2^{-I}$$

with $I = I(Z_{opt}; \phi_{opt}(X, Y) \mid \phi(X, Y))$.

Finally, note that Z_{opt} incurs $L^*/2$ error, and we disagree with Z_{opt} at most a $P(f(\phi(X, Y)) \neq Z_{opt})$ fraction of time. Assuming that we get every one of these disagreements wrong gives an upper bound of $L^*/2 + P(f(\phi(X, Y)) \neq Z_{opt})$ on $L(\phi)/2$. $\square$

A straightforward corollary is that whenever ϕ is an invertible function of ϕ_{opt}, the conditional mutual information is zero, and therefore the above inequalities become an equality.

Corollary 1. *Whenever ϕ is an invertible function of ϕ_{opt}, $L(\phi) = L^*$.*

A.3 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 (HIT Approval rate greater than 95 percent and number of HITs approved greater than 50 and location is USA). We constructed each HIT to be 25 examples, and paid one dollar per HIT.

We observe that measuring many replicates is sufficient to get low-variance estimates of HJ. For classification tasks where the model is straightforward to identify from references (such as story generation) we require five to ten replicates, while

Given a news article: **"us business leaders lashed out wednesday at legislation that would penalize companies for employing illegal immigrants ."**

How typical is the exact title: **"us business leaders slam bill to penalize illegal immigrants"**?

Please also take into account whether the title is grammatical, topical, and factually correct.

○ **Very Typical** (You expect to see this all the time.)
○ **Typical** (You often expect to see something like this.)
○ **Average** (Not surprised to see this, but would not appear as often as a typical comment.)
○ **Specific** (This is a correct response, but only in a very specific setting.)
○ **Rare** (This is some context where this is a correct response, but you would be surprised to see it.)
○ **Invalid** (Not a valid headline. Contains some clearly wrong facts or is grammatically incorrect.)

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

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. Even if the data quality is low, HUSE is still a valid upper bound (i.e. models with low HUSE are guaranteed to be distinguishable from humans). Thus the models which we identify as having low-HUSE are reliably distinguishable regardless of the crowdworker quality.

A.4 Reddit Dataset

We use a subset of Reddit comments from 2006-2018 scraped from https://pushshift.io/. We construct a dictionary containing the 10,000 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.