BLEU Neighbors:
A Reference-less Approach to Automatic Evaluation

Kawin Ethayarajh
Stanford University
`kawin@stanford.edu`

Dorsa Sadigh
ILIAD Lab, Stanford University
`dorsa@cs.stanford.edu`

Abstract

Evaluation is a bottleneck in the development of natural language generation (NLG) models. Automatic metrics such as BLEU rely on references, but for tasks such as open-ended generation, there are no references to draw upon. Although language diversity can be estimated using statistical measures such as perplexity, measuring language *quality* requires human evaluation. However, because human evaluation at scale is slow and expensive, it is used sparingly; it cannot be used to rapidly iterate on NLG models, in the way BLEU is used for machine translation. To this end, we propose *BLEU Neighbors*, a nearest neighbors model for estimating language quality by using the BLEU score as a kernel function. On existing datasets for chitchat dialogue and open-ended sentence generation, we find that – on average – the quality estimation from a BLEU Neighbors model has a lower mean squared error and higher Spearman correlation with the ground truth than individual human annotators. Despite its simplicity, BLEU Neighbors even outperforms state-of-the-art models on automatically grading essays, including models that have access to a gold-standard reference essay.

1 Introduction

Despite recent advances on many natural language generation (NLG) tasks – including open-ended generation, chitchat dialogue, and abstractive summarization – evaluation remains a challenge. Automatic metrics such as BLEU rely on references, but for many NLG tasks, there is no single correct answer. In dialogue, the space of acceptable responses to a given prompt is often very large, yet most datasets only provide a few gold-standard references (Serban et al., 2015). In open-ended generation, where text is generated freely by a language model, there are no references at all; statistical measures such as perplexity capture language
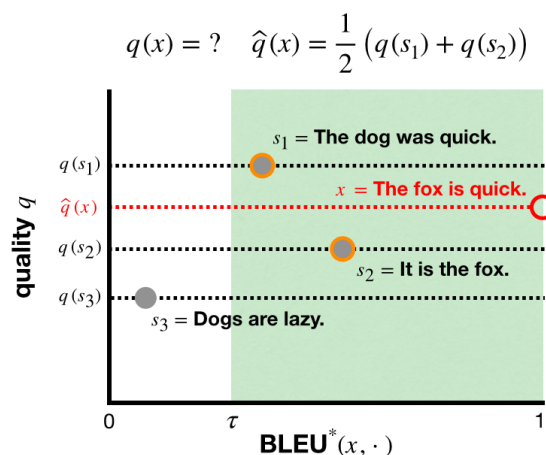
$$q(x) = ? \quad \widehat{q}(x) = \frac{1}{2}\big(q(s_1) + q(s_2)\big)$$

Figure 1: We want to score a sentence x given training examples $S = \{s_1, s_2, s_3\}$ with known quality scores $\{q(s_1), q(s_2), q(s_3)\}$. BLEU Neighbors works as follows: calculate BLEU*$(x, \cdot)$, a variant of the BLEU-4 score, for each s; ignore those below $\tau = 0.08$; take the average score of those that remain to predict $\widehat{q}(x)$.

diversity but not language quality (Hashimoto et al., 2019). These limitations necessitate human evaluation. However, because human evaluation at scale is slow and expensive, it is used sparingly; it cannot be used to rapidly iterate on NLG models, in the way BLEU is used for machine translation.

Prior work on automating reference-less evaluation has largely been limited in scope. Heuristic-based evaluation was found to be effective for grammatical error correction, but the methods used were problem-specific and cannot be extended to other tasks (Napoles et al., 2016; Choshen and Abend, 2018; Asano et al., 2017). Using the log-odds from a language model, Kann et al. (2018) made automatic judgments of sentence-level fluency that correlated moderately well with human judgment, but this captured only one facet of language quality. Approaches that were broader in scope found less success: although ADEM, an RNN trained to

score dialogue responses, was initially thought to correlate well with human judgment (Lowe et al., 2017), it was later found to generalize poorly, placing outsized influence on factors such as response length (Lowe, 2019).

Can we come up with a fast and simple method for reference-less evaluation of language quality, analogous to BLEU for machine translation? Note that our goal here is *not* to supplant human evaluation, but to complement it: as long as the method's predictions correlate moderately well with the ground-truth quality scores, it can be used to speed up NLG model development. Our desiderata are then as follows: simplicity, speed, and a moderately strong correlation with the ground truth. To this end, we propose *BLEU Neighbors*, a new approach to reference-less automatic evaluation.

Our approach is a nearest neighbors model that predicts language quality by using BLEU as a kernel function. We start with training examples S, where each sentence $s \in S$ has a ground-truth quality score $q(s)$. Note that these examples are not references – we do not expect the NLG model being evaluated to generate any sentence in S. In fact, S contains sentences of varying quality, including incoherent sentences with low quality scores. Given a test sentence x, we use the BLEU score to identify its neighbors in the training data: $\{s \mid \text{BLEU}^*(x, s) > \tau, s \in S\}$, where τ is a similarity threshold. Then we simply take the mean of the neighbors' known quality scores to estimate $\widehat{q}(x)$, the quality of x. Consider the test sentence $x = $ *'The fox is quick'*. As seen in Figure 1, it overlaps with $s_1 = $ *'The dog was quick.'* and $s_2 = $ *'It is the fox.'* but not with $s_3 = $ *'Dogs are lazy.'*. Therefore, we estimate $\widehat{q}(x)$ as the mean of $q(s_1)$ and $q(s_2)$ but not $q(s_3)$.

We test BLEU Neighbors on the datasets from HUSE (Hashimoto et al., 2019), where each sentence's ground-truth quality score is the average over 20 human judgments. On the dialogue and open-ended generation datasets, we find that – on average – the BLEU Neighbors model has a lower mean squared error (MSE) and higher Spearman correlation with the ground truth than individual annotators. The premise of our method is that past approaches to reference-less evaluation fell short because they were too ambitious – if a given test sentence is not sufficiently similar to any training example, no prediction should be made at all. Although we sacrifice some coverage in or-

der to make more accurate estimates, this sacrifice is modest: BLEU Neighbors makes predictions for 41% to 99% of sentences from the HUSE datasets. Our method is also data-efficient – none of HUSE datasets have over 400 training examples.

Our method is weakest on evaluating summaries; this is unsurprising, given that summary quality is conditioned on the source text, which the method ignores. In contrast, BLEU Neighbors is surprisingly effective at automatically grading essays, achieving a new state-of-the-art and even beating out models that have access to a gold-standard reference essay. These findings suggest that despite its simplicity, our approach has broad applicability. Although BLEU Neighbors does not measure language *diversity*, it is sufficient for it to measure quality alone. The former is easier to estimate (e.g., perplexity) and can be combined with the BLEU Neighbors score in a hybrid metric (Hashimoto et al., 2019). We conclude by providing some practical advice, such as how to prevent NLG models from explicitly optimizing for a high BLEU Neighbors score without generating high-quality output.

2 Related Work

2.1 Reference-based Evaluation

BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) are the *de facto* canonical metrics of reference-based automatic evaluation. Given a candidate sentence x and a reference sentence s, each metric assigns a score $q(x, s) \in [0, 1]$ based on how well x overlaps with s. Where the metrics differ is in how they define this overlap. Letting $\ell_n(\cdot)$ denote the list of n-grams, the n-gram precision P_n and recall R_n can be defined as follows:

$$
\begin{aligned}
P_n(x, s) &= \frac{1}{|\ell_n(x)|} \sum_{g \in \ell_n(x)} \mathbb{1}[g \in \ell_n(s)] \\
R_n(x, s) &= \frac{1}{|\ell_n(s)|} \sum_{g \in \ell_n(s)} \mathbb{1}[g \in \ell_n(x)]
\end{aligned}
\tag{1}
$$

BLEU The BLEU score for (x, s) is the geometric mean of the n-gram precision P_n up to a chosen n (typically, $n = 4$). BLEU also implements clipping, such that each n-gram $g \in \ell_n(x)$ can be matched at most once. It also includes a brevity penalty to penalize shorter candidates.

METEOR The METEOR score takes the harmonic mean of P_1 and R_1, with greater weight

placed on R_1. It is laxer than BLEU, allowing words in x and s to match, for example, if they are synonyms or share the same stem (Banerjee and Lavie, 2005). Instead of looking at higher order n-grams, METEOR tries to align the tokens in x and s and penalizes alignments that are not contiguous.

ROUGE-L ROUGE-L, the variant of ROUGE we discuss in this work, measures the overlap between x and s as the size of their longest common subsequence $LCS(x, s)$. Specifically, it calculates $LCS(x, s)/P_1(x, s)$ and $LCS(x, s)/R_1(x, s)$ and takes their harmonic mean.

Although there have been advances in reference-based automatic evaluation – such as BEER (Stanojević and Sima'an, 2014) and RUSE (Vedantam et al., 2015), among others (Shimanaka et al., 2018; Ma et al., 2017; Lo et al., 2018; Zhao et al., 2019) – BLEU and METEOR are still widely used for machine translation; ROUGE, for summarization (Liu et al., 2016). This is partially because some of the newer methods are learned metrics that do not generalize well to new domains (Chaganty et al., 2018). Moreover, most do not enjoy the incumbent status that BLEU, ROUGE, and METEOR have. To our knowledge, the current state-of-the-art in reference-based evaluation metrics is BERTScore (Zhang et al., 2019), which uses BERT embeddings (Devlin et al., 2019) to compute similarity at the token-level before aggregating the similarities using importance-weighting. As it is state-of-the-art for reference-based evaluation, it is the only non-canonical metric we consider as a kernel function.

2.2 Reference-less Evaluation

Compared to reference-based evaluation, little work has been done on automating reference-less evaluation. The most successful approaches have been task-specific: heuristic-based evaluation was found to be effective for grammatical error correction (Napoles et al., 2016; Choshen and Abend, 2018; Asano et al., 2017). However, those heuristics cannot be extended to other tasks. Kann et al. (2018) proposed two metrics for judging the fluency of a sentence: sentence-level log-odds ratio (SLOR) and a Wordpiece-based variant named WP-SLOR. Although the latter correlates moderately well (Pearson's $r > 0.40$) with human judgment, it should be noted that sentence-level fluency is only one facet of language quality – a sentence may be probable according to a language model while making little sense to a human.

Approaches that were broader in scope were less successful. ADEM, an RNN trained to score dialogue responses, was initially thought to correlate well with human judgment (Lowe et al., 2017). However, the authors later found that it generalized poorly (Lowe, 2019), placing outsized influence on factors such as response length. It was also found to be vulnerable to adversarial examples (Sai et al., 2019). In any case, ADEM was not a purely reference-less method – it still required a gold-standard reference as input. Rather, its key insight was that the space of acceptable responses is much larger than the handful of gold-standard references provided in dialogue datasets, and that this should be considered when estimating quality.

3 BLEU Neighbors

Given a candidate sentence x, training examples S, and ground-truth quality scores $\{q(s) \mid s \in S\}$, we want to estimate $\widehat{q}(x)$, the language quality of x. How can we do so in a fast and simple manner such that our predictions correlate well with the ground truth? We propose a nearest neighbors model that uses a variant of the BLEU score called BLEU* as the kernel function. Once the neighbors of x have been identified, we take the mean of their known quality scores as $\widehat{q}(x)$.

Definition 3.1. The *non-unigram BLEU-4 score* is a variant of the BLEU-4 score that ignores unigram precision. Where $\beta = \exp\left(\min\left(0, 1 - \frac{|\ell_1(s)|}{|\ell_1(x)|}\right)\right)$ is the brevity penalty and P_i is defined in (1),

$$\text{BLEU}^*(x, s) = \beta \cdot \prod_{i=2}^{4} P_i(x, s)^{1/3} \qquad (2)$$

BLEU Neighbors uses this variant of BLEU as the kernel function. We ignore the unigram precision P_1 because we are not comparing candidates and their direct references, but rather candidates and training examples. It is not uncommon for two random sentences to have stopwords in common, in which case a non-zero P_1 is unexceptional. We validated this empirically as well, finding that ignoring P_1 improves correlation with the ground-truth.

Definition 3.2. Given a candidate sentence x, training examples S, and a similarity threshold $\tau \in [0, 1]$, the *BLEU neighbors* of x are

$$\mathcal{N} = \{s \in S \mid \text{BLEU}^*(x, s) \geq \tau\}$$

To ensure that the quality estimate is stable, we require that $\mathcal{N}$ have a minimum size of $a \in \mathbb{Z}^+$.

Conversely, a candidate sentence that overlaps with many training examples in S likely does so because it contains many common n-grams. This complicates evaluation: since BLEU* does not weigh n-grams by their frequency, an abundance of common n-grams – such as *"on the"* or *"it is"*, for example – can exaggerate the similarity between the candidate and a training example. In this scenario, it is best that no prediction be made at all. Since $\mathcal{N} \subseteq \mathcal{S}$, let $b \in [0, 1]$ denote the largest fraction of S that $\mathcal{N}$ can contain. We express b as a fraction of the training set size $|S|$ because if S is very large, it would not be uncommon for even sentences with rare n-grams to have matches in S.

When $\mathcal{N}$ meets the aforementioned size constraints, the BLEU Neighbors estimate of x's quality is the average of its neighbors' quality scores:

$$\widehat{q}(x) = \begin{cases} \dfrac{1}{|\mathcal{N}|} \sum_{s \in \mathcal{N}} q(s) & a \leq |\mathcal{N}| \leq b|S| \\ \text{undefined} & \text{otherwise} \end{cases} \tag{3}$$

In other words, $\mathcal{N} \subseteq S$ comprises all the training examples that are sufficiently similar to the candidate with respect to BLEU*. If there are fewer than a examples or more than $b|S|$ examples in $\mathcal{N}$, then no prediction is made; otherwise, the estimate $\widehat{q}(x)$ is the average quality of the examples in $\mathcal{N}$.

Although τ, a, b are parameters to be set, we find that $\tau = 0.08, a = 5, b = 0.66$ are near-optimal for all tasks (see section 5.2). This universality allows BLEU Neighbors to be used out-of-the-box, without hyperparameter tuning. Note that S should only be used to train the evaluator (i.e., BLEU Neighbors). The generator (i.e., the NLG model being evaluated) should not have access to S; otherwise, it could optimize for a high BLEU Neighbors score by including n-grams that only belong to examples with a high ground-truth quality, thus artificially inflating the quality estimates.

Definition 3.3. Given a set of candidates $\mathcal{X}$ to be evaluated, the *coverage* of $\mathcal{X}$ is the proportion of candidates for which $\widehat{q}(x)$ is defined.

This is a key distinction between BLEU Neighbors and prior approaches to reference-less evaluation: our approach does not necessarily make a prediction for all candidates. This is by design – as mentioned earlier, we surmise that past approaches fell short because they were too ambitious, trying

to score sentences that simply could not be scored. There is a trade-off between coverage and prediction error, with greater coverage generally coming at the cost of greater prediction error.

4 Experiments

NLG Tasks We test BLEU Neighbors on evaluating sentences from the following NLG tasks: chitchat dialogue, open-ended sentence generation (from a language model), and abstractive summarization. Hashimoto et al. (2019) provided a dataset for each of these tasks, which we collectively refer to as the HUSE datasets. We ignore the story generation dataset in that work because the machine-generated examples are far from human quality and can thus be trivially assigned a low quality score.

Each dataset contains a mixture of machine- and human-generated sentences, in roughly equal proportion. Each sentence in the HUSE datasets was judged by 20 human annotators, who assigned it a label based on its typicality. These labels map to an integer score from 0 to 5. We divide the raw judgment by 5 to bound it in $[0, 1]$ and then take the mean across all 20 annotators, which we treat as the ground-truth language quality $q(s)$ for each sentence s. Because these datasets are small, we use leave-one-out prediction. That is, given a candidate sentence from a particular HUSE dataset, we treat the remaining $n - 1$ sentences as S.

Grading Essays We also test our model on automatically grading essays from the ASAP-SAS dataset[1]. Although each essay is a multi-sentence paragraph, we did not adapt our model in any way. Each essay's quality score is an integer from 0 to 3, which we divide by 3 to bound in $[0, 1]$. This normalization is done for the sake of consistency. Because there are distinct training and test sets, we draw the training examples from the training data and the candidates to be evaluated from the test data. The ASAP-SAS data is also broken down by topic. The current state-of-the-art model only evaluates on topic #3 – specifically, on essays from topic #3 that contain 5 to 15 sentences (Clark et al., 2019). Therefore, to allow for a fair comparison, we also draw test sentences from this subset.

Threshold Settings Unless otherwise stated, for all HUSE datasets, we use $\tau = 0.08, a = 5, b = 0.66$. These settings were chosen to maximize the Spearman correlation with the ground-truth quality

---

[1] https://www.kaggle.com/c/asap-sas

	Dialogue			Open-ended Generation			Summarization		
	MSE	ρ	Coverage	MSE	ρ	Coverage	MSE	ρ	Coverage
Human (best)	0.0208	0.878	1.00	0.0177	0.861	1.00	0.0200	0.921	1.00
Human (average)	0.0807	0.456	1.00	0.0719	0.472	1.00	0.0802	0.405	1.00
BLEU Neighbors	**0.0164**	**0.470***	0.76	0.0204	**0.575***	0.41	0.0213	**0.325***	0.99
ROUGE Neighbors	0.0197	0.342*	0.86	**0.0174**	0.077	0.47	0.0226	0.245*	0.97
METEOR Neighbors	0.0165	0.382*	0.47	0.0209	0.395	0.22	**0.0180**	0.240	0.12
BERTScore Neighbors	0.0229	0.150*	0.89	0.0192	0.566*	0.32	0.0223	0.225	0.53

Table 1: The mean squared error (MSE) and Spearman's ρ of language quality predictions $\widehat{q}(\cdot)$ with respect to the ground truth $q(\cdot)$. The lowest MSE and highest ρ across all models is in bold and * signifies $p < 0.01$. For all tasks, BLEU Neighbors achieves a higher Spearman's ρ than its ROUGE, METEOR, and BERTScore counterparts. For dialogue and open-ended generation, it even has a lower MSE and higher ρ than human annotators on average.
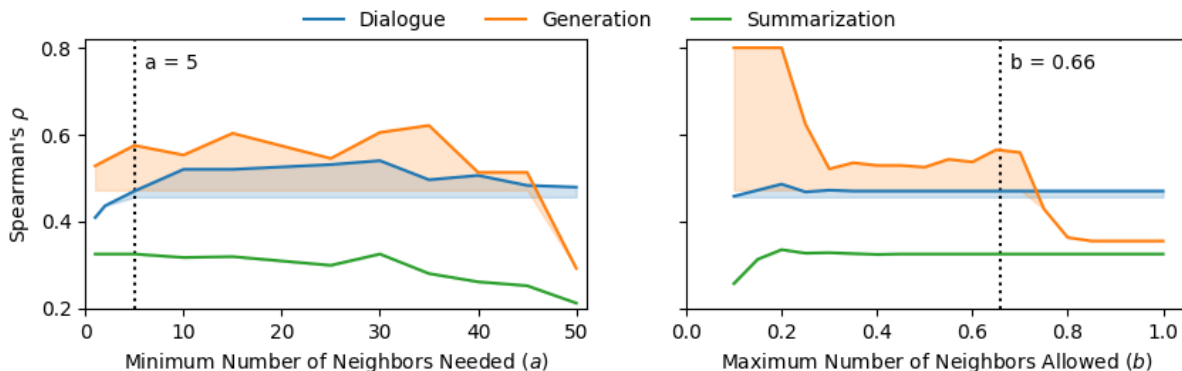

Figure 2: Spearman's ρ between BLEU Neighbors estimates $\widehat{q}(\cdot)$ and the ground-truth quality $q(\cdot)$ as each evidence threshold changes, while the other is held constant at $a = 5$ or $b = 0.66$. a is the minimum number of neighbors needed; b is the maximum allowed (as a fraction of the training set). For all tasks, increasing a improves correlation, up to a point. Only the correlation for open-ended generation is sensitive to changes in b, which decreases as b increases. The shaded area for each task indicates above-human performance (on average).

while retaining at least 40% coverage. The same settings were used for essay grading, except with no upper bound on $|\mathcal{N}|$ (i.e., $b = 1$), since each essay is a multi-sentence text that has some overlap with most essays in the training data. In section 5.2, we show how a, b can be adjusted to trade off some performance for greater coverage (and vice-versa).

Other Kernel Functions In addition to using a variant of the BLEU score as the kernel function, we try other automatic metrics, including ROUGE, METEOR, and BERTScore (Zhang et al., 2019). As with BLEU, a single value of τ for each metric works universally well: 0.06 (for ROUGE); 0.18 (for METEOR); 0.10 (for BERTScore).

5 Results

5.1 BLEU Neighbors vs. Humans

In Table 1, using mean squared error (MSE) and the Spearman correlation, we compare the language quality predictions $\widehat{q}(\cdot)$ made by our various mod-

els with the ground-truth quality $q(\cdot)$. Because the ground-truth quality is the mean over 20 annotator judgments, we provide the performance of the best human annotator and the average performance across all individual annotators. Note that not all annotators scored all the examples: the average MSE and ρ we report in Table 1 is the average over what each annotator obtained on their respective subset of the data. We find that there is a significant gap between the best- and average-case, both in terms of MSE and Spearman's ρ. For example, on the summarization task, the MSE and Spearman's ρ of the best human annotator is 4x and 2x better than those of annotators on average.

Spearman Correlation As shown in the second section of Table 1, for all tasks, we find that BLEU Neighbors has a higher Spearman correlation with the ground truth than its ROUGE, METEOR, and BERTScore counterparts. For open-ended generation and dialogue, it even outperforms the average-case human annotator. Only on evaluating sum-
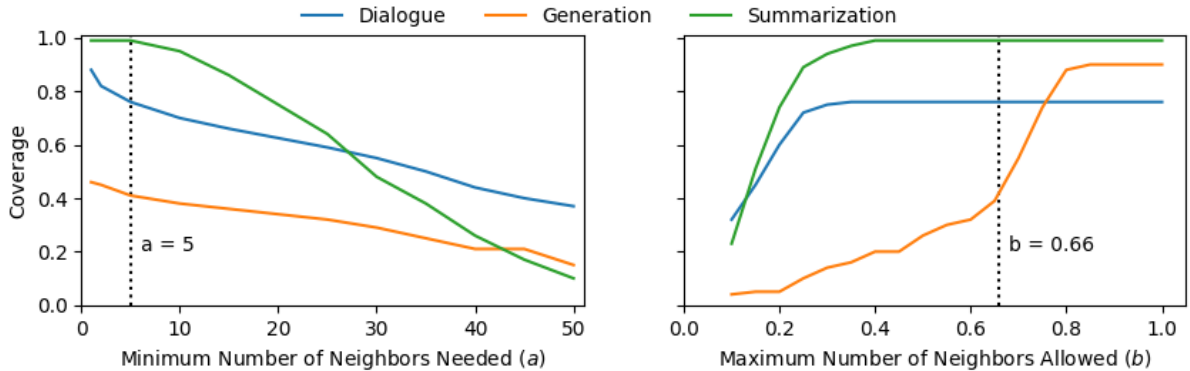
Figure 3: The coverage (i.e., the fraction of sentences for which BLEU Neighbors makes predictions) as each evidence threshold changes while the other is held constant at $a = 5$ or $b = 0.66$. a is the minimum number of neighbors needed; b is the maximum number allowed (as a fraction of the training set). For all tasks, coverage falls as a increases and b decreases (i.e., as the range for the acceptable number of neighbors gets smaller).

maries does the average-case annotator beat all nearest neighbors models with respect to Spearman's ρ; this is unsurprising, given that summary quality is strongly conditioned on the source text, which these models ignore. Despite the impressive performance of BLEU Neighbors, it should be noted that it is still well behind the *best* human annotator for each task.

Mean Squared Error Although BLEU Neighbors achieves a much higher Spearman correlation with the ground-truth quality than its counterparts, the model that achieves the lowest MSE varies across tasks. How can we reconcile these observations? We find that the variance of the ground-truth quality is quite small for all datasets. By just predicting the mean of $q(\cdot)$ for all candidates, we can get an MSE for each task that is only slightly higher than the best annotator's. Models that obtain the lowest MSE while also having a low Spearman's ρ are thus making low-variance estimates close to the mean that do not correlate well with the ground truth. Also, annotators of the HUSE datasets assigned discrete scores (Hashimoto et al., 2019), while $q(\cdot)$, being an average over those scores, is continuous. This is conducive to human annotators having a higher MSE than the models.

5.2 Varying the Evidence Thresholds

Recall that BLEU Neighbors has two evidence thresholds: a, the minimum number of neighbors needed to make a prediction, and b, the maximum number of neighbors allowed (as a fraction of the training set S). In Figure 2, we plot the Spearman correlation between predictions $\widehat{q}(\cdot)$ and the

ground truth $q(\cdot)$ as each threshold changes, while the other is held constant at the default setting ($a = 5, b = 0.66$). In Figure 3, we plot the change in coverage as the thresholds change.

Spearman Correlation The correlation for all tasks is sensitive to changes in a, with the correlation peaking at $a = 30$ or $a = 35$ before declining. This is intuitive: increasing the amount of evidence required yields more robust predictions, but sentences that meet the stringent requirement of having at least $a \geq 35$ neighbors likely have many common n-grams, making them harder to score. While performance on all tasks is sensitive to a, only performance on open-ended generation is sensitive to b, with the correlation decreasing as b increases (i.e., as we loosen the upper bound on the number of neighbors). This suggests that sentences in the dialogue and summarization datasets do not have many neighbors to begin with, which is why tightening the upper bound has little effect. Sentences in the open-ended generation data, on the other hand, seem to have many more neighbors on average, resulting in ρ being inversely related to b. The two sudden drops in Spearman's ρ for open-ended generation – at approximately $b = 0.2$ and $b = 0.7$ – suggests that the distribution of $|\mathcal{N}|$, the number of neighbors, is multi-modal.

Coverage within a Model The higher a is and the lower b is, the more candidates we reject for having too few or too many neighbors. In Figure 3, the coverage falls linearly as a increases but rises linearly before plateauing as b increases. The plateau is indicative of no candidate sentence having that many neighbors to begin with.

45

Source Task	Target Task		
	$\rightarrow$ D	$\rightarrow$ G	$\rightarrow$ S
Dialogue (D) $\rightarrow$	**0.470**	0.206	0.032
Generation (G) $\rightarrow$	0.310	**0.575**	-0.070
Summarization (S) $\rightarrow$	0.276	0.095	**0.325**

Table 2: BLEU Neighbors performance when the training and test examples are sourced from different tasks. For example, the intersection of $G \rightarrow$ and $\rightarrow D$ means that training examples from open-ended generation are used to score dialogue data. In this setup, a moderate Spearman's ρ can still be achieved on the dialogue data.

Coverage across Models   Holding constant the evidence thresholds a and b, we see in Table 1 that coverage *across* different models is unrelated to MSE and Spearman's ρ. For all models, τ is set to minimize the MSE and maximize Spearman's ρ. However, models with the lowest MSE or highest ρ on a given task are not necessarily the most selective (i.e., those with the lowest coverage). BLEU Neighbors, which has the highest correlation on all tasks, has a coverage of 41%, 76%, and 99% on open-ended generation, dialogue, and summarization respectively. In other words, the trade-off between coverage and prediction error exists *within* a model – as a function of parameters a and b – but not *across* different types of models.

Performance vs. Coverage   As seen in Figures 2 and 3, there is a trade-off when choosing a and b. Higher a and lower b result in better performance (i.e., greater correlation with the ground truth), but they also decrease coverage. Recall the default settings: $a = 5, b = 0.66$. Even though correlation on most tasks peaks at $a = 30$ or $a = 35$, we choose $a = 5$ as the default because we want to keep the coverage as high as possible. By choosing $a > 1$, however, we still see some benefit from requiring a minimum number of neighbors. We choose $b = 0.66$ because it is near the end of a plateau past which performance on open-ended generation data drops precipitously. In other words, the default settings of a, b are near Pareto-optimal, maximizing coverage while outperforming human annotators on average. Some performance can be traded off for additional coverage by picking a different point (a, b) on the Pareto frontier.

5.3   Low-Hanging or High-Hanging Fruit?

Does BLEU Neighbors only make predictions for sentences that humans consider easy to score (i.e., low-hanging fruit)? Let $\mathcal{A}_i$ denote the set of all

sentences for task i and $\mathcal{B}_i \subseteq \mathcal{A}_i$ denote the subset of those sentences for which BLEU Neighbors makes predictions. We can answer this question by comparing the average MSE of human annotators on $\mathcal{A}_i$ with their average MSE on $\mathcal{B}_i$, which we will denote as $\overline{\text{MSE}}(\mathcal{A}_i)$ and $\overline{\text{MSE}}(\mathcal{B}_i)$ respectively. We cannot use the Spearman correlation for comparison because not every annotator scored every sentence; recall that the statistics reported in Table 1 are computed over each annotator's performance on their subset of the data.

If our model were only scoring the easy-to-score sentences, then we would expect $\overline{\text{MSE}}(\mathcal{A}_i)$ to be significantly larger than $\overline{\text{MSE}}(\mathcal{B}_i)$. However, for both summarization and open-ended generation, we find that there is no statistically significant difference between these means at any level. Only on the dialogue dataset could this theory partially explain the success of our model: $\overline{\text{MSE}}(\mathcal{B}_{\text{dialogue}})$ is 15.6% lower than $\overline{\text{MSE}}(\mathcal{A}_{\text{dialogue}})$ and this difference is significant at $p < 0.01$. However, the average-case annotator MSE on the subset of the dialogue data scored by ROUGE Neighbors is only 2×10^{-4} higher than $\overline{\text{MSE}}(\mathcal{B}_{\text{dialogue}})$, yet ROUGE Neighbors performs far worse than its BLEU counterpart (see Table 1). This implies that the success of BLEU Neighbors is much more than simply picking the right sentences to score.

5.4   Cross-Task Performance

In Table 2, we report the Spearman's ρ for BLEU Neighbors when the training and test examples are drawn from different tasks. Of all the tasks, performance on dialogue is the most robust: regardless of which task is used to source the training data, it is possible to achieve a moderately strong correlation ($\rho > 0.27$), albeit with lower coverage. Performance on summarization drops to near zero in this setup – this is unsurprising, given that summary quality is strongly conditioned on the source text, which is ignored. For open-ended generation, it is still possible to achieve a weak correlation ($\rho > 0.09$) with this setup. Curiously, the coverage for open-ended generation actually improves when the training data is sourced from a different task, so it may be possible to adjust parameters a, b to trade off some coverage for a higher correlation.

5.5   How much Training Data is Needed?

In Figure 4, we plot the performance of BLEU Neighbors on the HUSE datasets for different amounts of training data. This is simulated by
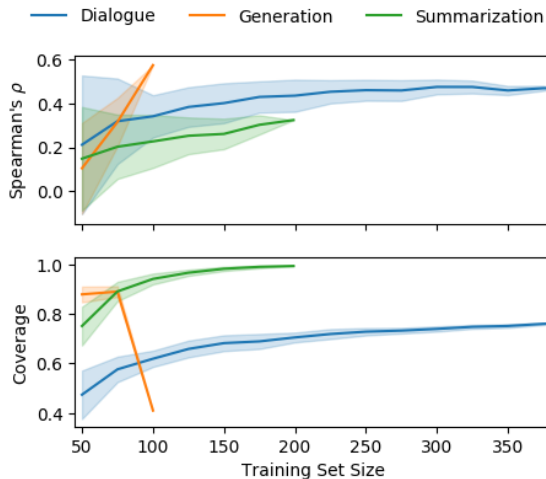
Figure 4: BLEU Neighbors performance when only a random subset of the training data is used. With more data, both coverage and the Spearman correlation with the ground truth improve, albeit with diminishing returns. The shaded area denotes one standard deviation (i.e., variation in performance across random samples).

Model	Train Topic	ρ	Coverage
ROUGE-L	3	0.441	1.0
WMS (ELMo)	3	0.443	1.0
SMS (ELMo)	3	0.451	1.0
S+WMS (ELMo)	3	0.490	1.0
	1	0.301	1.0
	2	0.392	1.0
	4	-0.128	1.0
	5	0.466*	1.0
BLEU Neighbors	6	0.383	1.0
	7	0.305	1.0
	8	**0.500***	1.0
	9	0.375	1.0
	10	0.437*	1.0
	$\widetilde{3}$	0.465*	1.0

Table 3: Spearman's ρ between predicted essay quality and the ground truth, where the test essays are from Topic #3 and * denotes $p < 0.01$. When using essays from Topic #8 as the training data, BLEU Neighbors is state-of-the-art, even beating out models with access to a gold-standard reference essay for Topic #3.

drawing a random subset of the data with n examples, doing hold-one-out prediction with $n-1$ examples, and then taking the mean performance over 20 such runs. We find that BLEU Neighbors is surprisingly robust on all tasks, with 75 training examples being sufficient to achieve a Spearman's $\rho > 0.30$ on dialogue and open-ended generation while retaining above 60% coverage. As more data is used, both coverage and the Spearman correlation improve, though there are diminishing returns. Unsurprisingly, the variation in performance across random subsets also drops as more data is used.

5.6 Automated Essay Grading

In Table 3, we report the performance on the essay grading task described in Section 4, where the goal is to score essays from Topic #3 of the ASAP-SAS dataset. Unlike the NLG tasks, every test example here is a multi-sentence paragraph, which makes scoring more difficult: ten random sentences may be high-quality on their own while making little sense when put together. The difficulty of this task is compounded by the fact that the ground-truth quality of each essay is based on a gold-standard reference for Topic #3. Since BLEU Neighbors does not use references, it is at a disadvantage compared to approaches that do, such as ROUGE-L. Excluding ROUGE-L, all the models we list in Table 3 are optimal transport methods that leverage text embeddings (Clark et al., 2019).

Despite not being given the gold-standard reference, when BLEU Neighbors is trained with sample essays from Topic #8, it achieves a new state-of-the-art: a Spearman's ρ of 0.500 between its predicted scores and the ground-truth quality judgments. However, due to the small amount of test data, this improvement over the state-of-the-art is not statistically significant at $p < 0.01$ when using a Williams test. Still, its coverage is 100%, meaning that it makes predictions for all of the test essays. As seen in the second half of Table 3, the performance of the model depends strongly on which topic the training data is sourced from. This is unsurprising, given that some topics are more related to #3 than others. Some topics (e.g., #4) are so different from the test topic that its training examples are of no use, leading to very poor quality estimates. When we use essays from all topics but topic #3 as the training data – denoted in Table 3 as $\widetilde{3}$ – we still outperform most of the past approaches.

6 Limitations and Future Work

Quality + Diversity Although the BLEU Neighbors model does not measure language diversity, this is by design. Consider that if an NLG model were ideal, even the optimal discriminator could not tell whether its outputs were human- or machine-generated. Hashimoto et al. (2019) proved that such an optimal discriminator would only need two statistics, a measure of language diversity (e.g., perplexity) and a measure of lan-

guage quality. The former is trivial to compute – it is the latter that is cost- and time-intensive, and which we thus try to automate using BLEU Neighbors. These two measures can be combined using a metric such as HUSE (Hashimoto et al., 2019), meaning that it is sufficient for our model to predict quality alone. The next step would be to use such a hybrid metric in rapidly evaluating NLG models during development.

Preventing "Hacks" How can we prevent the NLG model being evaluated from "hacking" a BLEU Neighbors model so as to receive inflated quality estimates for all its outputs? As mentioned in section 3, one way to prevent this is to use disjoint training sets for the NLG model and BLEU Neighbors, so that the former has no idea what the latter considers a high-quality candidate. Additionally, it would help to have a large set of training examples for BLEU Neighbors and then subsample it during each evaluation instance, as that would discourage NLG models from generating n-grams that just so happen to occur in one or two high-quality examples in the training data. Moreover, BLEU Neighbors is intended to speed up NLG model development – not supplant humans – so any attempts to inflate quality estimates during development would have poor long-term outcomes.

Metric Learning The success of BLEU Neighbors can largely be ascribed to it using BLEU*, a variant of the BLEU-4 score, as the kernel function in sentence space. Despite its simplicity, BLEU* works surprisingly well. There is likely a more convoluted variant of BLEU-4 that works even better for this purpose – one that excludes stopwords, one that places greater weight on rarer n-grams, etc. Instead of specifying a kernel function, it may also be possible to learn one. For example, instead of representing each sentence as a sequence of words, one could transform it into a sentence embedding and then learn a kernel function as a metric in the embedding space. This is one direction of future work.

Diverse Datasets Although BLEU Neighbors performs well in our experiments, because of the small size of the datasets we use, not all results are statistically significant. One limitation of the HUSE datasets in particular is that, as mentioned earlier, the annotators scored different subsets of the data. In order to more faithfully compare our method against human annotators, we need larger

datasets from a more diverse array of tasks, where every example is scored by every annotator.

Because of the leave-one-out paradigm we use on the HUSE datasets, the test examples were scored in part with the help of scored training examples that were generated by the same model. Table 2 shows that cross-task performance is generally poor, with the exception of dialogue data. Would the performance still be poor if we used model-generated training examples from the same task but used a different model to generate them? This is a possibility that should be explored. It is also unclear what exactly is driving the success of BLEU Neighbors. For example, if it is exploiting annotation artefacts, then its success would be far less impressive (Gururangan et al., 2018). Understanding these possible failure cases is an important direction for future work. Developing a theoretical understanding of BLEU Neighbors – as has been done with static word embeddings, for example (Levy and Goldberg, 2014; Ethayarajh et al., 2019a,b; Ethayarajh, 2019) – would be ideal.

7 Conclusion

The absence of a reference-less evaluation metric for language quality has been an impediment to developing NLG models. To address this problem, we proposed *BLEU Neighbors*, a nearest neighbors model that leverages the BLEU score as a kernel function in sentence space. Our simple approach worked surprisingly well: it outperformed human annotators – on average – in predicting the quality of dialogue and open-ended generation data. We also found BLEU Neighbors to be state-of-the-art on automatically grading essays, even beating out models that had access to a gold-standard reference essay. Moreover, our model is fast, data-efficient, and easy-to-use; it has only two hyperparameters and those have settings that work universally well, across various tasks. Still, BLEU Neighbors is intended to complement, not supplant, human evaluation – its speed, simplicity, and ease of use makes it ideal for rapidly iterating on NLG models long before any human evaluation is done.

References

Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. 2017. Reference-based metrics can be replaced with reference-less metrics in evaluating grammatical error correction systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–348.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

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

Leshem Choshen and Omri Abend. 2018. Reference-less measure of faithfulness for grammatical error correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 124–129.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760.

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

Kawin Ethayarajh. 2019. Rotate king to get queen: Word relationships as orthogonal transformations in embedding space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3494–3499.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019a. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019b. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

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

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Chi-kiu Lo, Michel Simard, Darlene Stewart, Samuel Larkin, Cyril Goutte, and Patrick Littell. 2018. Accurate semantic textual similarity for cleaning noisy parallel corpora using semantic machine translation evaluation metric: The nrc supervised submissions to the parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 908–916.

Ryan Lowe. 2019. Introducing retrospectives: 'real talk' for your past papers.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126.

Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. 2017. Blend: a novel combined mt metric based on direct assessment—casict-dcu submission to wmt17 metrics task. In *Proceedings of the second conference on machine translation*, pages 598–603.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2115.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Ananya B Sai, Mithun Das Gupta, Mitesh M Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating adem: A deeper look at scoring dialogue responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6220–6227.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758.

Miloš Stanojević and Khalil Sima'an. 2014. Beer: Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578.