On the Relation between Quality-Diversity Evaluation and Distribution-Fitting Goal in Text Generation

Jianing Li [1 2] **Yanyan Lan** [1 2] **Jiafeng Guo** [1 2] **Xueqi Cheng** [1 2]

Abstract

The goal of text generation models is to fit the underlying real probability distribution of text. For performance evaluation, quality and diversity metrics are usually applied. However, it is still not clear to what extend can the quality-diversity evaluation reflect the distribution-fitting goal. In this paper, we try to reveal such relation in a theoretical approach. We prove that under certain conditions, a linear combination of quality and diversity constitutes a divergence metric between the generated distribution and the real distribution. We also show that the commonly used BLEU/Self-BLEU metric pair fails to match any divergence metric, thus propose CR/NRR as a substitute for quality/diversity metric pair.

1. Introduction

Text generation is an essential task for many NLP applications, such as machine writing (Zhang et al., 2017a), machine translation (Bahdanau et al., 2014), image captioning (Rennie et al., 2017) and dialogue system (Li et al., 2017). Text generation models work by either explicitly modeling the probability distribution of text (Mikolov et al., 2010; Yu et al., 2017), or implicitly learning a generator which maps noise data to text (Zhang et al., 2017b; Chen et al., 2018). Both approaches aim at generating text with the same distribution of given text data.

To achieve the distribution-fitting goal, divergence metrics are usually applied as the training objective for text generation models, which take minimal value 0 if and only if the model distribution exactly recover the real text distribution. Typical choices include the Kullback-Leibler divergence by maximum likelihood estimation (MLE) (Mikolov et al., 2010), and Jensen-Shannon divergence or Wasserstein distance by adversarial training (Yu et al., 2017; Gulrajani et al., 2017). However during evaluation, divergence-based metrics fails to distinguish two under-fitting cases from each other: the low-quality case that generate unrealistic text, and the low-diversity case that generates dull and repeated text. As such, quality and diversity metrics are introduces to help the model diagnosis, such as BLEU (Papineni et al., 2002) and Self-BLEU (Zhu et al., 2018). High generation quality requires the model to generate realistic samples, i.e. generated samples are free of grammatical or logical errors. High generation diversity requires the model to generate diverse samples, i.e. generated samples are less likely to be duplicate and contain diverse unique patterns.

Despite popular application of quality-diversity metrics in evaluation of text generation models (Chen et al., 2018; Lu et al., 2018b; Fedus et al., 2018; Alihosseini et al., 2019), the relationship between such evaluation and the distribution-fitting goal is still not clear. It seems to be a tacit consensus in recent works that a model with both higher quality and higher diversity also better fit the real text distribution (Caccia et al., 2018; Li et al., 2019; d'Autume et al., 2019). However, such assumption is yet to be verified. This is critical since a potential inequivalence may result in misleading evaluation conclusions. In this paper, we try to answer this question under the unconditional text generation setting by a theoretical approach.

To bridge the gap between distribution-fitting goal and quality-diversity evaluation, we require the optimal solutions from divergence minimization to be consistent with that of quality-diversity maximization. As such, we first give a general definition of quality and diversity. Then, we study a Multi-Objective Programming (MOP) problem which maximizes quality and diversity simultaneously. We prove there exists a family of Pareto-optimal solutions for this MOP problem, i.e. solutions which cannot be outperformed in terms of both quality and diversity. Then we prove the real distribution belongs to this Pareto-optimal family if and only if quality-diversity metrics are used in pairs with strong restrictions. Under such condition, a linear combination of quality and diversity constitutes a divergence metric between the generated distribution and the real distribution.

[1]CAS Key Laboratory of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China [2]University of Chinese Academy of Sciences, Beijing, China. Correspondence to: Yanyan Lan <lanyanyan@ict.ac.cn>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

For quality-diversity metrics used in practice, we show that the widely applied BLEU/Self-BLEU metric pair fails to match any divergence metric. This is highlighted by a counter-intuitive observation that real text samples are significantly outperformed by manually constructed models over both BLEU and Self-BLEU. Therefore, we further propose Coverage Rate (CR) and Negative Repetition Rate (NRR) as substitute based on above theoretical analysis. Experiments show that CR/NRR act well as quality/diveristy metrics respectively, while a linear combination of CR/NRR acts well as divergence metric.

2. Related Work

To evaluate the performance of text generation models, many evaluation metrics are designed from different perspectives. Early neural text generation models use Perplexity (PPL) to show how well a language model fit the training data (Mikolov et al., 2010). This is a divergence-based metric, and is still adopted in recent works (Fedus et al., 2018; Lu et al., 2018a; Subramanian et al., 2018). Calculation of PPL may be intractable for implicit models, so other divergence-based metrics are also practical choices, such as Kernel Density Estimation (Zhang et al., 2017b), Word Mover Distance (Lu et al., 2018a), MS-Jaccard (Alihosseini et al., 2019), and Frechet Distance (Semeniuta et al., 2018; Alihosseini et al., 2019; d'Autume et al., 2019). However, divergence metrics provide limited information for model diagnosis, and may not correlate well with task performance (Chen et al., 1998; Fedus et al., 2018). Therefore, the quality and diversity of generated text are further considered as complementary metrics, which are also practical requirements in real applications (Zhang et al., 2018; Hashimoto et al., 2019; Gao et al., 2019).

For quality metrics, the evaluation is closely related to the ground truth distribution. Yu et al. (2017) propose to use Negative Log-Likelihood where the real distribution is known in advance, which measures the average log-probability of generated samples over the real distribution. If the real distribution is not explicitly given, BLEU (Papineni et al., 2002) and ROUGE (Lin & Och, 2004) are usually applied, which measure the n-gram overlap between generated samples and a set of reference ground truth samples. For diversity metrics, the evaluation is performed within the model itself. Li et al. (2015) proposed Distinct-n as diversity metric, which calculates the ratio of unique n-grams in generated samples. Zhu et al. (2018) proposed Self-BLEU, which is similar to BLEU but use generated samples as reference set.

There was a time in the past that only quality metrics are applied for evaluation, such as in works of SeqGAN (Yu et al., 2017), RankGAN (Lin et al., 2017), and LeakGAN (Guo et al., 2017). However after an observation of the quality-diversity tradeoff problem, Zhu et al. (2018) suggest to use a hybrid of both quality and diversity metrics, such as BLEU and Self-BLEU. This suggestion is widely adopted by many analytical works (Lu et al., 2018b; Caccia et al., 2018; Semeniuta et al., 2018; Alihosseini et al., 2019), as well as newly proposed methods, such as FM-GAN (Chen et al., 2018), DDR (Li et al., 2019), and ScratchGAN (d'Autume et al., 2019). Despite the prevailing application of quality-diversity evaluation, its relationship with divergence metrics remains unclear, which poses great uncertainty for evaluation conclusions. Our work will help to build bridges between quality-diversity and divergence, and provide guidance for choosing appropriate quality-diversity metrics.

3. Definition of Quality and Diversity

Currently there is no unified definition for quality and diversity in text generation, which brings great challenges for further theoretical studies. In fact, it is not easy to define a general form of quality and diversity due to various understandings of these two aspects. Thus before moving on to further analysis, we first try to give a general form of quality and diversity in a mathematical view, though it may not be comprehensive enough to cover all possible understandings.

3.1. A General Form of Quality and Diversity

Text data is usually discrete, so we make the following notations. Assume the vocabulary size is $|V|$, and the maximum length is L, then the distribution of text data can be described by a categorical distribution with size $N = |V|^L$. We denote the real distribution and the generated model distribution as $P(x) = (P_1, P_2, \cdots, P_N)$ and $Q(x) = (Q_1, Q_2, \cdots, Q_N)$, respectively.

In general, the *Quality* of a text generation model measures how likely the generated text are to be realistic text in human's view. Since the value of real probability $P(x)$ can be viewed as reflecting the realistic degree of a text x, the expectation of some function over $P(x)$ could be used to quantify quality. For example, in works of Yu et al. (2017) and Nie et al. (2018), *Log-Likelihood (LL)* is used as the quality metric, where $LL(Q; P) = \mathbb{E}_{x\sim Q} \log P(x)$. Following this idea, we propose a general form of quality, i.e., $U(Q; P) = \mathbb{E}_{x\sim Q} f_u[P(x)]$, where f_u is a function over $P(x)$.

Similarly, the *Diversity* of a text generation model measures how much difference there are among generated texts. From the viewpoint of information, *Shannon-Entropy (SE)* of $Q(x)$ can be used as a natural diversity metric, where $SE(Q) = -\mathbb{E}_{x\sim Q} \log Q(x)$. From another understanding view, a text x should be less likely to be generated again if the diversity is high. This idea has been adopted in biology to evaluate the diversity of biocoeno-

sis, named as the *Simpson's Diversity Index (SDI)*, where $SDI(Q) = 1 - \mathbb{E}_{x\sim Q}Q(x)$. Summarizing these two different understandings, we obtain a general form of diversity, i.e. $V(Q) = -\mathbb{E}_{x\sim Q} f_v[Q(x)]$.

To this end, we propose a general form of quality and diversity metrics as follows:

$$U(Q) = U(Q;P) = \mathbb{E}_{x\sim Q} f_u[P(x)] = \sum_{i=1}^{N} Q_i \cdot f(P_i),$$

$$V(Q) = -\mathbb{E}_{x\sim Q} f_v[Q(x)] = \sum_{i=1}^{N} g(Q_i),$$

where $f_u(x)$ is denoted as $f(x)$ and $-x \cdot f_v(x)$ as $g(x)$.

3.2. The Rationality of Quality and Diversity

To guarantee U and V are rational quality and diversity metrics, we need to discuss about the conditions of f and g. Without loss of generality, we first assume that f is differentiable and g is twice differentiable. Further, the following requirements are necessary for rational quality and diversity:

1. Generating more samples with higher real probability yields higher overall quality;
2. Distributing the probability more equally yields higher overall diversity.

Mathematically, these two requirements can be formalized as the following two properties:

1. If $P_i > P_j$, then for $Q' = (Q_1, \dots, Q_i + \epsilon, \dots, Q_j - \epsilon, \dots)$, there is $U(Q') > U(Q)$ for any $\epsilon \in (0, Q_j)$.

2. If $Q_i \geq Q_j$, then for $Q' = (Q_1, \dots, Q_i + \epsilon, \dots, Q_j - \epsilon, \dots)$, there is $V(Q') < V(Q)$ for any $\epsilon \in (0, Q_j)$.

Then we can obtain the conditions of f and g by the following theorem:

Theorem 1. *The following conditions are both sufficient and necessary to satisfy the properties 1-2: For any x_1, x_2 s.t. $x_1 > x_2 > 0$ and $x_1 + x_2 \leq 1$, we have $f(x_1) > f(x_2)$ and $g'(x_1) < g'(x_2)$.*

According to Theorem 1, it is necessary for $f(x)$ to be strictly monotonically increasing and $g(x)$ to be strictly concave for $x \in (0, \frac{1}{2})$. For simplicity, we only consider the cases where such properties hold for $x \in (0, 1)$, thus get a sufficient condition:

1. $f(x)$ is strictly monotonically increasing for $x \in (0, 1)$;
2. $g(x)$ is strictly concave for $x \in (0, 1)$.

Under this condition, we can see that a model with highest quality will distribute all its density to text with highest real probability, and a model with highest diversity will be uniform, which are consistent with human understandings.

4. Analysis of Quality-Diversity Evaluation

In this section, we show how and to what extent can the quality-diversity evaluation reflect the distribution-fitting goal. The key idea is to solve the Multi-Objective Programming (MOP) problem which tries to maximize quality and diversity simultaneously. We give the structure of all the Pareto-optima of this MOP problem, which constitutes the Pareto-frontier. Then we prove the ground truth distribution lies in this frontier if and only if f and g are paired according to a given rule. Under such condition, a linear combination of quality and diversity constitutes a divergence metric, which means the quality-diversity evaluation is sufficient to reflect the distribution-fitting goal.

4.1. The MOP Problem

We consider the following MOP problem:

$$\max_{Q} \ (U(Q), V(Q))$$
$$s.t. \sum_{i=1}^{N} Q_i = 1$$
$$\forall i, \ Q_i \geq 0$$

The goal is to maximize both quality and diversity, while keeping Q a legal distribution. The optimal solutions of a MOP problem are called Pareto-optima, which means no other solution can beat them consistently over all objectives.

We give definitions of the terminologies of Pareto-optimality below:

Definition 1. *For two distributions Q and Q', if one of the following conditions are satisfied, we say that Q is dominated by Q'.*

1. *$U(Q') > U(Q)$ and $V(Q') \geq V(Q)$;*
2. *$U(Q') \geq U(Q)$ and $V(Q') > V(Q)$.*

A solution Q is called a Pareto-optimum if it is not dominated by any Q'. The set containing all the Pareto-optima is called the Pareto-frontier.

Intuitively, a Pareto-optimum is a solution that there is no distribution can achieve both higher quality and higher diversity than it. And all the Pareto-optima constitutes the Pareto-frontier. The Pareto-frontier may collapse into one solution which leads to a global optimum, e.g. if P is uniform, the unique optimal solution would be $Q^* = P$. However

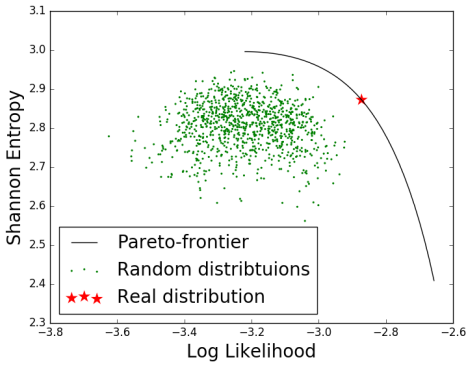


Figure 1. Illustration of the Pareto-frontier of LL-SE metric pair on a toy categorical distribution, which contains 20 categories and probabilities are sampled from uniform distribution with normalization.

it is often the case where the objectives in MOP problem cannot reach their optima consistently, which results in a family of optimal solutions. Therefore, the structure of the Pareto-frontier under a non-uniform P is what we care about.

4.2. The Pareto-frontier

We show the structure of the Pareto-frontier by giving the following theorem:

Theorem 2. *For a distribution Q, if P is not uniform, then:*

(1) The following condition is both sufficient and necessary for Q to be a Pareto-optimum: there exist real value $w \leq 0$ and b that for any $i = 1, \ldots, N$, there is

$$Q_i = \hat{g}'^{-1}[w \cdot f(P_i) + b],$$

where

$$\hat{g}'^{-1}(x) = \begin{cases} g'^{-1}(x) & \text{if } x < g'(0), \\ 0 & \text{if } x \geq g'(0), \end{cases}$$

(2) b is correspondent to w, i.e. b is fixed once w is fixed. If $f(x) < 0$ for all $x \in [0, 1]$, then b is strictly monotonically increasing w.r.t. w. If $f(x) > 0$ for all $x \in [0, 1]$, then b is strictly monotonically decreasing w.r.t. w.

(3) Denote a Pareto-optimum Q as $Q(w)$, then for any $w_1 < w_2$: if $w_1, w_2 \in [B, 0]$, there is $Q(w_1) \neq Q(w_2)$ and $U(Q(w_1)) > U(Q(w_2)), V(Q(w_1)) < V(Q(w_2))$; if $w_1, w_2 \in (-\infty, B]$, there is $Q(w_1) = Q(w_2)$; where $B = \frac{g'(\frac{1}{M}) - g'(0)}{f(P_{m_1}) - f(P_{m_2})}$, and $P_{m_1} = \max_i P_i$, $P_{m_2} = \max_{P_i \neq P_{m_1}} P_i$, $M = \#\{i | P_i = P_{m_1}\}$, # denotes the cardinality of a set.

According to Theorem 2, different ws lead to different distributions, so we can change w from 0 to B and get a family of optimal solutions with different quality and diversity. As such, for a non-uniform P, the Pareto-frontier is a family of distributions.

We can see quality and diversity act as a tradeoff if we want to maximize them at the same time. Since all distributions in the Pareto-frontier are Pareto-optima, trying to improve one metric for an optimum will lead to another optimum at most, thus inevitably causing another metric to drop. This result provides support for the quality-diversity tradeoff problem observed in previous works (Zhu et al., 2018; Caccia et al., 2018).

We show the result of Theorem 2 here on a special case. We pair Log-Likelihood (LL) with Shannon-Entropy (SE), the corresponding Pareto-optima can be written as

$$Q_i = \frac{P_i^\beta}{Z},\ Z = \sum_{i=1}^{N} P_i^\beta,\ \beta \geq 0,$$

we have $w = -\beta$, and $b = 1 + \log Z$. These Pareto-optima are formerly used as quality-diversity tradeoff solutions by Li et al. (2019).

An illustration of the Pareto-frontier on a toy distribution is shown in Figure 1. We can see that quality and diversity are negatively correlated for solutions in the Pareto-frontier. Note that the ground truth distribution lies exactly on the frontier in this LL-SE case, which can be checked by setting $\beta = 1$. We will then show this is the key to the relation between quality-diversity metrics and divergence metrics.

4.3. Relationship with Divergence

To bridge the gap between the distribution-fitting goal and quality-diversity evaluation, it is necessary for the optimal solutions from divergence minimization to be consistent with that from quality-diversity maximization. Since $Q = P$ is the optimal solution with minimum divergence and the above Pareto-frontier is the set of optimal solutions with maximal quality and diversity, we require $Q = P$ to be in the Pareto-frontier. Theoretical results are shown in the following Theorem:

Theorem 3. *The following condition is both sufficient and necessary for $Q = P$ to be a Pareto-optimum for any P: there exist $w_0 \leq 0$ and b_0 that*

$$g(x) = w_0 \int_0^x f(u)\mathrm{d}u + b_0 x.$$

If the above condition is satisfied, then $Q = P$ corresponds to a Pareto-optimum with $w = w_0$ and $b = b_0$, and it is the only distribution that maximize $\Psi(Q) = \alpha U(Q) + (1 - \alpha)V(Q)$ with $\alpha = \frac{w_0}{w_0 - 1} \in [0, 1)$, and $D(P||Q) = \Psi(P) - \Psi(Q)$ becomes a divergence metric.

We find that if quality and diversity metrics are carefully chosen, namely g is the integral of an affine transformation

of f, we can get a divergence metric by a linear combination of these two metrics.

The LL-SE case satisfies the condition in Theorem 3. Under this special case, there is $\Psi(Q) = \frac{1}{2}\mathrm{LL}(Q) + \frac{1}{2}\mathrm{SE}(Q)$, and

$$D(P||Q) = \frac{1}{2}\sum_{i=1}^{N} Q_i \cdot \log\frac{Q_i}{P_i},$$

which is exactly the Reverse KL divergence if the constant $\frac{1}{2}$ is ignored. This linearly combined divergence metric can be viewed as a tangent line of the Pareto-frontier curve in Figure 1, and the real distribution is the tangent point.

Since such condition is also necessary, the real distribution is unlikely to be a Pareto-optima if we use casually chosen metrics. This means, there would be one distribution achieving both higher quality and higher diversity than the ground truth, which is implausible. Therefore, if the condition in Theorem 3 is not satisfied, it would be unlikely to measure the divergence using a combination of quality and diversity.

Now we can conclude that, it is sufficient to reflect the distribution-fitting goal by a hybrid of quality-diversity evaluation. However, specific metrics should be chosen carefully, in order to avoid the potential violation of such property. Suppose such property is violated severely, featured by a huge gap between the ground truth distribution and the Pareto-frontier, then a model which perfectly fits the real distribution would be significantly outperformed by another model over both quality and diversity, resulting in misleading conclusions.

Therefore in the next section, we will examine the existence of the gap for quality-diversity metrics used in practice, and provide suggestions on the choice of quality-diversity metrics.

5. Options for Quality-Diversity Metrics

It is yet to be examined that whether existing quality-diversity metrics are sufficient to reflect the distribution-fitting goal. For metrics satisfying our defined general form in Section 3.1, conclusions can be drawn directly by applying Theorem 3. For example, the Log-likelihood (LL) is widely used as quality metric, which is correspondent to NLL-oracle (Yu et al., 2017) and Reverse PPL (Subramanian et al., 2018). As proved above, LL satisfies the condition in Theorem 3 if it's paired with Shannon Entropy (SE). Consequently, it is safe to use LL-SE together as in the work of Alihosseini et al. (2019).

However for most scenarios with real text data, the calculation is intractable for the general form of quality-diversity in Section 3.1 as the ground truth distribution is unknown, including the LL-SE pair. Practical metrics (e.g. BLEU and Self-BLEU) thus usually fall out of this framework, and Theorem 3 cannot be applied directly. In order to make a judgement on such metrics, we suggest to consider the compatibility between divergence and quality-diversity metric pair. We say a pair of quality-diversity metrics is *divergence-compatible* if the real distribution is a Pareto-optimum under the MOP problem maximizing both metrics. Such compatibility is a necessary condition for the existence of a corresponding divergence metric which is strictly monotonically decreasing w.r.t. both quality and diversity.

5.1. BLEU and Self-BLEU

BLEU (Papineni et al., 2002) and Self-BLEU (Zhu et al., 2018) are common metrics for quality and diversity evaluation, respectively. Intuitively, BLEU measures the n-gram overlap between a candidate set of generated text and a reference set of real text, while Self-BLEU is the average BLEU score of each generated text with other candidates as reference. High BLEU score means that n-grams in generated text are more likely to appear in real text, thus BLEU can be used as quality metric. Similarly, high Self-BLEU score means that generated text are similar to each other in terms of n-gram, thus Negative Self-BLEU (NSBLEU as abbreviation) can be used as diversity metric.

The expression of BLEU on a candidate set C is:

$$\mathrm{BLEU} = BP \cdot exp(\frac{1}{M}\sum_{n=1}^{M}\log p_n),$$
$$p_n = \frac{\sum_{c\in C}\sum_{gram_n\in c} Count_{clip}(gram_n)}{\sum_{c'\in C}\sum_{gram'_n\in c'} Count(gram'_n)},$$

where BP is the Brevity Penalty which penalizes short sentences, and M denotes the maximum n-gram order. p_n is a precision term, which measures the proportion of grams in the candidate set that also appear in the reference set. BLEU is the geometric mean of p_n for all $n \le M$, multiplied by a penalty term.

The expression of BLEU does not seem to satisfy the general form of quality/diversity defined in Section 3.1. However on some special case, the general form is still satisfied, upon which we show some symptoms indicating the incompatibility of BLEU-NSBLEU. Assume the lengths of text are all 1, so that $M = 1$ and $BP \equiv 1$. In this case, BLEU contains only one term, i.e. $\mathrm{BLEU} = p_1$. Then for candidate set C and reference set R, the expectation of BLEU and NSBLEU over generated distribution Q and real distribution P would be

$$\mathop{\mathbb{E}}_{C\sim Q, R\sim P}\mathrm{BLEU}(C,R) = \sum_{i=1}^{N} Q_i\cdot[1-(1-P_i)^{|R|}],$$

$$\mathop{\mathbb{E}}_{C\sim Q}\mathrm{NSBLEU}(C) = -\sum_{i=1}^{N} Q_i\cdot[1-(1-Q_i)^{|C|-1}].$$

Such expressions satisfy the general form with

$$f(x) = 1 - (1-x)^{|R|}, \quad g(x) = -x + x \cdot (1-x)^{|C|-1}.$$

The condition in Theorem 3 would be satisfied if and only if $|R| = 1$ and $|C| = 2$, which becomes $f(x) = x$ and $g(x) = -x^2$. However, the size of reference set $|R|$ is usually far more than 1, under which cases the BLEU-NSBLEU metric pair would be divergence-incompatible.

Though above analysis is done on a special case, such results imply a potential incompatibility for general BLEU-NSBLEU metric pairs. We will confirm this incompatibility by an empirical approach in Section 6.

5.2. The Proposed Metric Pair

To avoid possible misleading conclusions in practice, we suggest to use diversity-compatible quality-diversity metric pair.

Since the real probability $P(x)$ is required in $U(Q;P)$ under the general form in Section 3.1, calculation of most quality metrics are intractable on real text data. The only exception is the case with $f(x) = x$, paired with $g(x) = -x^2$. The linearity of f can avoid the explicit form of $P(x)$ by sampling from real data, i.e. $U(Q) = \mathbb{E}_{x\in P}\, Q(x)$. We name the corresponding quality metric as *Coverage Rate (CR)*, and diversity metric as *Negative Repetition Rate (NRR)*. Even so, we observe a large variance while estimating CR and NRR on real text data. This is mainly because of the extremely large space of text of $N = |V|^L$. Therefore, estimations of CR/NRR are highly inaccurate in the text space.

We thus suggest to calculate CR-NRR in n-gram space rather than in text space. Derive the n-gram distribution Q_g and P_g from text distribution Q and P, so that

$$\mathrm{CR}_n(Q;P) = \sum_{gram_n \in S_n} Q_g(gram_n) \cdot P_g(gram_n),$$
$$\mathrm{NRR}_n(Q) = -\sum_{gram_n \in S_n} Q_g^2(gram_n),$$

where S_n denotes the set of all possible n-grams. In practice, Q_g and P_g can be estimated by the empirical distribution, i.e. count the number of target n-grams and divide by the total number. Note that if calculated by the longest n-gram with $n = L$, CR_n and NRR_n would exactly recover the original CR and NRR metric in text space, thus can be viewed as a generalized form. In the rest of this paper, we use *CR-NRR* as a default notation in the n-gram space unless explicitly stated.

In the n-grams space, calculation of metric pairs with other f/g functions also becomes possible. However, metrics such as LL-SE suffer from another smoothing problem on real text data, i.e. their values go to infinity if some n-grams do not appear in candidate set or reference set. Therefore, we still suggest to use CR-NRR as a first choice.

Though there is a conversion from the text space to the n-gram space, CR/NRR can still reflect quality/diversity. The CR_n metric measures the average probability for an n-gram in candidate set to appear in the reference set, thus is an indicator of quality. Similarly, NRR_n measures the average probability for an n-gram to appear again in two consecutive sampling processes over the candidate set, thus is an indicator of diversity.

We then check the divergence-compatibility of CR-NRR evaluation. Firstly, CR-NRR is divergence-compatible w.r.t. distributions in the n-gram space, according to Theorem 3. We name the corresponding divergence metric as *CR-NRR Divergence (CND)*, where

$$\Psi_n(Q) = \frac{2}{3}\mathrm{CR}_n(Q;P) + \frac{1}{3}\mathrm{NRR}_n(Q),$$

and

$$\begin{aligned}\mathrm{CND}_n(Q;P) &= 3 \cdot [\Psi_n(P) - \Psi_n(Q)] \\ &= \sum_{gram_n \in S_n} [Q_g(gram_n) - P_g(gram_n)]^2.\end{aligned}$$

Secondly, CR-NRR is also divergence-compatible w.r.t. distributions in the text space. Assume $Q = P$ is dominated by Q' under CR-NRR evaluation, which means $Q_g = P_g$ would also be dominated by Q'_g. This cause contradiction with the compatibility in n-gram space, so the compatibility in text space also holds.

In addition to the divergence-compatibility property, CR-NRR is also easy to acquire. It does not require the explicit value of $P(x)$ or $Q(x)$, thus can be applied on implicit models similarly to BLEU-NSBLEU. Moreover, the time complexity of CR-NRR algorithm is $O(m+n)$, which is much lower than BLEU-NSBLEU with $O(m \cdot (m+n))$, where m and n denote the size of candidate and reference set respectively. To conclude, we suggest to use CR-NRR in n-gram space for quality-diversity evaluation, instead of BLEU-NSBLEU.

6. Experiments

In this section, we perform compatibility analysis of BLEU-NSBLEU, compared with CR-NRR on both synthetic data and real text data. We show that BLEU-NSBLEU is significantly divergence-incompatible, by observing a phenomenon that ground truth text data are clearly outperformed over both BLEU and NSBLEU by some manually constructed model. We also show that CR/NRR are representative for quality/diversity evaluation respectively, while CND is representative for divergence evaluation.

Table 1. Lower-bound of QDisc and DRate w.r.t. BLEU-NSBLEU on synthetic data with different σs.

Metrics	$\sigma = 0.5$		$\sigma = 1.0$		$\sigma = 2.0$	
	QDisc	DRate(%)	QDisc	DRate(%)	QDisc	DRate(%)
BS-1	0.01287	2.55	0.01509	3.29	0.01063	3.15
BS-2	0.02384	9.41	0.01699	4.27	0.01146	1.71
BS-3	2.090×10^{-8}	<0.01	6.045×10^{-6}	0.19	3.878×10^{-4}	0.05

To measure the degree of incompatibility, we calculate the Quality Discrepancy (QDisc) and Discrepancy Rate (DRate):

$$\text{QDisc} = \max_{Q} U(Q) - U(P),\ s.t.\ V(Q) \geq V(P),$$

$$\text{DRate} = \frac{\text{QDisc}}{\max_Q U(Q) - U(Q')},\ Q' = \underset{Q}{argmax}\, V(Q).$$

Intuitively, we try to find a model with best quality while its diversity is no lower than that of real distribution. Then QDisc measures the difference between this model and the real distribution in terms of quality. DRate measures the ratio between QDisc and the total range of quality for all Pareto-optima. A metric pair is divergence-compatible if and only if $\text{QDisc} = 0$.

6.1. Experiments on Synthetic Data

We first run experiments on synthetic data rather than real text data, in order to get the precise values of all metrics. Under this setting, the information of generated distribution Q and real distribution P are explicitly given in advance, thus eliminates the possible variance from sampling. The synthetic data are texts with length L using a pseudo vocabulary V. We construct the real distribution using an oracle LSTM model as in SeqGAN (Yu et al., 2017), whose weights are randomly sampled from a gaussian distribution with $\mu = 0$. Different standard deviation σs are applied to get several synthetic real distributions with different levels of entropy, i.e. distribution with smaller σ is more flat and of higher entropy, and distribution with larger σ is more sharp and of lower entropy.

Calculation of QDisc and DRate can be achieved by a simple binary-search algorithm if the exact form of Pareto-frontier is known. However for BLEU-NSBLEU metric pair, the frontier is unknown since Theorem 2 cannot be applied in this case. Consequently, we opt to used an optimization-based method for the estimation of QDisc. We try to solve the following optimization problem using stochastic gradient descent (SGD) with momentum:

$$Q^* = \underset{Q}{\text{argmax}}\ U(Q) - \lambda \cdot \max(0, V(P) - V(Q)),$$

where λ is a penalty term to discourage the case where divergence is lower than real distribution P. We set $\lambda = 2.0$ in our experiments. So that $\text{QDisc} = U(Q^*) - U(P)$, and the denominator in DRate is also calculated through such optimization-based method.

For BLEU metric with candidate set size m and reference set size n, the expectation can be directly calculated by

$$\underset{C\sim Q, R\sim P}{\mathbb{E}} \text{BLEU}(C, R) = \sum_{C\in V^{L\cdot m}, R\in V^{L\cdot n}} \prod_{i=1}^{m} Q(C_i) \cdot \prod_{j=1}^{n} P(R_j) \cdot \text{BLEU}(C, R).$$

The time complexity (number of terms) of such calculation is $O(|V|^{L\cdot(m+n)})$. This is intolerable for above optimization problem even in text space of normal size. As a result, we set $|V| = 4, L = 3, m = 1, n = 2$, and apply SGD under the Tensorflow framework[1].

We use CN-n and BS-n as abbreviation for CR-NRR and BLEU-NSBLEU with n-gram, respectively. We report the QDisc and DRate of BLEU-NSBLEU in Table 1. Note that the reported QDisc values are corresponding lower bounds, since the optimization-based method does not guarantee a global optimum. These non-zero QDisc values provide a clear support for the incompatibility of BLEU-NSBLEU. We can also see that such discrepancy is significant on some cases, e.g. $\text{QDisc} > 0.02$ and $\text{DRate} = 9.41\%$ for BS-2 on data with $\sigma = 0.5$. A QDisc value of 0.02 means that, we cannot surely claim that a model is better than another when the quality gap is below 0.02, which is already a clear gap for BLEU. We also run similar experiments for CR-NRR. However, no positive lower bound is observed, which is in accordance with our theory.

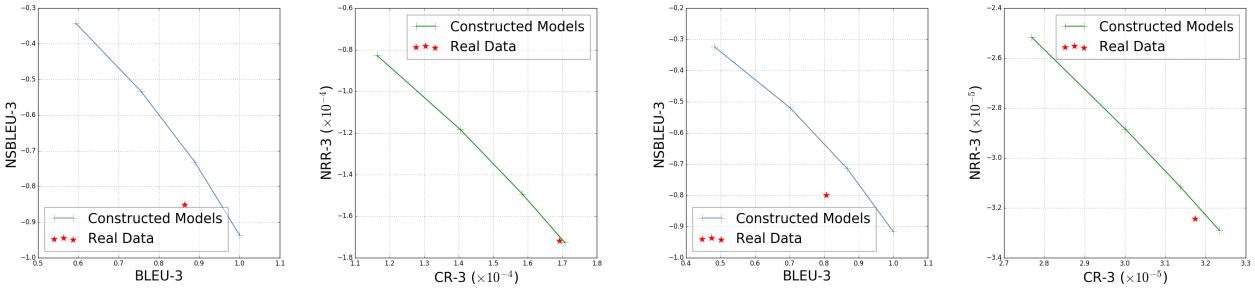
6.2. Experiments on Real Text Data

Significance of quality discrepancy varies on different cases, thus we care about the discrepancies on real text data. We use two public datasets, MSCOCO Image Caption dataset (Chen et al., 2015) and EMNLP2017 WMT News dataset[2]. We use 50,000 sentences as candidate set and another 50,000 as reference set for each dataset [3].

[1] Slight increase of any parameter will consume intolerably more time, and is not necessary for the conclusions.

[2] http://statmt.org/wmt17/translation-task.html

[3] See supplementary material for detailed configurations.



(a) MSCOCO dataset (b) WMT dataset

Figure 2. Evaluation of BLEU-NSBLEU and CR-NRR on real text data. Test data are random text from reference set, mixed with noise with a proportion of $\epsilon = [0.0, 0.2, 0.4, 0.6]$ from right to left.

Table 2. Estimation of QDisc, DRate, Self-Ratio, and Ref-Ratio on real text data.

Metrics	MSCOCO				WMT			
	QDisc	DRate(%)	Self-Ratio	Ref-Ratio	QDisc	DRate(%)	Self-Ratio	Ref-Ratio
BS-2	0.032	3.2	0.034	0.314	0.034	3.4	0.036	0.26
BS-3	0.090	9.0	0.104	0.814	0.117	11.7	0.145	0.88
BS-4	0.162	16.2	0.219	1.46	0.211	21.1	0.339	1.59
CN-2	0.75×10^{-6}	0.013	0.0005	0.006	3.69×10^{-7}	0.016	0.0008	0.025
CN-3	1.07×10^{-6}	0.079	0.0063	0.087	3.45×10^{-7}	0.098	0.0109	0.358
CN-4	1.15×10^{-6}	0.163	0.0247	0.421	3.12×10^{-7}	0.220	0.0525	2.092

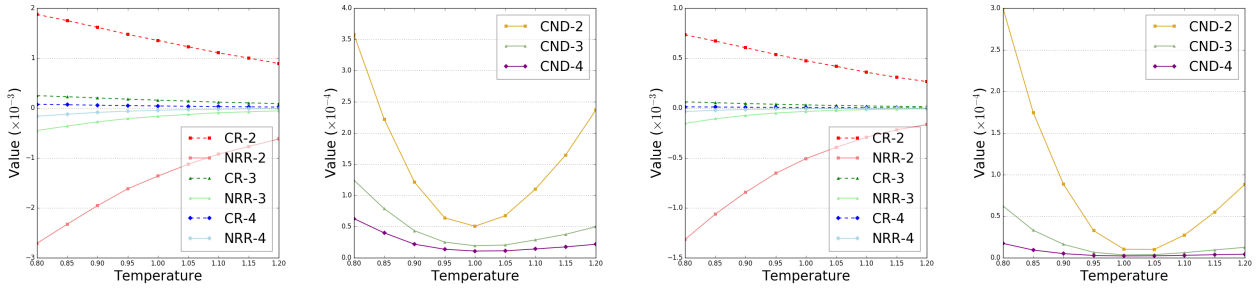
To provide an estimation of QDisc and DRate, we manually construct a family of strong models. We mix the empirical distribution $\tilde{P}$ with truncated uniform distribution M under different proportions, i.e. $Q = (1-\epsilon)\cdot\tilde{P} + \epsilon\cdot M$. During text generation, a random text from reference set is sampled with probability $1-\epsilon$, otherwise a text with random tokens of length L' is constructed with probability ϵ. We try both $L' = 5$ and $L' = L$, and report the case with larger QDisc value.

We estimate QDisc by a linear interpolation between two closest points on the curve w.r.t. quality of real data. For the denominator of DRate in BLEU-NSBLEU, we use 1.0 directly, since $\text{BLEU} = 1$ is reached for highest quality with $\epsilon = 0.0$, and $\text{BLEU} \approx 0$ for highest diversity with $\epsilon = 1.0$. For CR-NRR, CR goes to 0 when diversity is maximized with $\epsilon = 1.0$. As for the maximal value of CR, we estimate it by using a single reference sentence as candidate and select the one with maximal CR value.

For a clearer view of the significance of quality discrepancy, we introduce two additional metrics: Self-Ratio and Ref-Ratio. Self-Ratio calculates the ratio between QDisc and the quality of candidate set. Ref-Ratio calculates the ratio between QDisc and the quality difference of $\epsilon = 0.0$ and $\epsilon = 0.2$. The evaluation results of BLEU-NSBLEU and CR-NRR with 3-gram under $L' = 5$ are shown in Figure 2.

We can see that real data stays close to the CR-NRR curve, while a much larger gap is observed between real data and the BLEU-NSBLEU curve. We give the values of QDisc, DRate, Self-Ratio, and Ref-Ratio in Table 2. BLEU-NSBLEU shows a significant incompatibility, by QDisc values ranging from 0.032 to 0.211. Such huge discrepancy in BLEU is unbearable in real applications, e.g. we cannot claim a model is better than another even if it achieves higher NSBLEU and significantly higher BLEU. As a result, we suggest not to use BLEU-NSBLEU in order to avoid misleading conclusions. CR-NRR also shows a small positive discrepancy, this is due to the inevitable difference between the empirical distributions of candidate set and reference set. However, discrepancy caused by such distribution difference is generally much smaller than BLEU-NSBLEU. We also observe that DRate grows quickly as n-gram becomes longer for CR-NRR, thus we suggest to use CR-NRR with short n-gram such as CN-2 or CN-3.

Next we show how CR/NRR/CND behave on real text data. We apply temperature sweep on an RNN-based language model (RNNLM) pre-trained by maximum likelihood estimation, which is a quick way to get a family of models with quality-diversity tradeoff according to works of Caccia et al. (2018). The RNNLM consists of an embedding layer, an LSTM layer, and a fully-connected output layer. The embedding dimension and number of hidden nodes are all



(a) MSCOCO dataset (b) WMT dataset

Figure 3. Evaluation of CR-NRR and CND on real text data. Test data are generated by temperature-sweep on pre-trained RNNLMs.

set to 128. We train the model using Adam (Kingma & Ba, 2014) optimizer with learning rate 0.001 by 30 epochs. As temperature t grows, the model becomes more close to uniform, so that quality decreases and diversity increases, and minimal divergence is taken near $t = 1.0$. Results are shown in Figure 3, where we can see CR/NRR/CND are representative for quality/diversity/divergence respectively, which clearly fit our expectations. Therefore, we suggest to use CR-NRR for quality-diversity evaluation.

7. Discussion

Our above conclusions are mainly drawn under the unconditional text generation setting, however, quality-diversity evaluation is also getting great attentions under conditional text generation settings, such as dialogue system (Vijayakumar et al., 2016), machine translation (Shen et al., 2019) and image captioning (Ippolito et al., 2019). In this section, we give a brief discussion about quality-diversity evaluation under conditional text generation settings.

Due to different formalization of quality and diversity metrics, our conclusions cannot be directly transferred to conditional text generation settings. Under these settings, the quality of text x under condition c is still defined as monotonically increasing w.r.t. the real conditional probability $P(x|c)$. So that the overall quality metric becomes the expectation of text quality over x and c, which is the case for BLEU. Meanwhile, diversity metrics have two different understandings. One is defined as the average diversity of conditional model distribution $Q(x|c)$ under different c, such as Pairwise-BLEU (Shen et al., 2019). The other is define as the diversity of marginal model distribution $Q(x) = \sum_c P(c)Q(x|c)$, such as Distinct (Li et al., 2015). Formalization of both quality and diversity metrics depart from ours in Section 3.1, and may result in different conclusions, thus require further separate analysis. Though such analyses are not covered here, our work provides a paradigm for future theoretical analysis, including metric definition, Pareto-optimality analysis, and divergence-compatibility judgement.

Another difference lies in the point of view of task goal. While the goal of unconditional text generation is to design models that better fit the text distribution, in conditional text generation however, better human evaluation results are viewed as final goal in most cases. Therefore in these cases, the main focus would be designing metrics that better reflect human evaluation as well as designing training objectives that achieve better evaluation. It is also anticipated that whether human evaluation is compatible with divergence. We regard these as our future work.

8. Conclusion

In this paper, we give theoretical analysis of the relation between quality-diversity evaluation and distribution-fitting goal. We show that when using properly paired quality-diversity metrics, i.e. $g(x)$ is the integral of an affine transformation of $f(x)$, a linear combination of quality and diversity constitutes a divergence metric between the generated distribution and the real distribution. For metrics used in practice, we show the commonly used BLEU and Self-BLEU metric pair fails to reflect the distribution-fitting goal. For a substitute, we suggest to use CR-NRR instead as quality-diversity metric pair.

Acknowledgement

This work was supported by Beijing Academy of Artificial Intelligence (BAAI) under Grants No. BAAI2019ZD0306, and BAAI2020ZJ0303, the National Natural Science Foundation of China (NSFC) under Grants No. 61722211, 61773362, 61872338, 61902381, and 61906180, the Youth Innovation Promotion Association CAS under Grants No. 20144310, and 2016102, the National Key RD Program of China under Grants No. 2016QY02D0405, the Lenovo-CAS Joint Lab Youth Scientist Project.

References

Alihosseini, D., Montahaei, E., and Baghshah, M. S. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 90–98, 2019.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

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

Chen, L., Dai, S., Tao, C., Zhang, H., Gan, Z., Shen, D., Zhang, Y., Wang, G., Zhang, R., and Carin, L. Adversarial text generation via feature-mover's distance. In *Advances in Neural Information Processing Systems*, pp. 4666–4677, 2018.

Chen, S. F., Beeferman, D., and Rosenfeld, R. Evaluation metrics for language models. In *DARPA Broadcast News Transcription and Understanding Workshop*, pp. 275–280. Citeseer, 1998.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

d'Autume, C. d. M., Rosca, M., Rae, J., and Mohamed, S. Training language gans from scratch. *arXiv preprint arXiv:1905.09922*, 2019.

Fedus, W., Goodfellow, I., and Dai, A. M. Maskgan: Better text generation via filling in the _. *arXiv preprint arXiv:1801.07736*, 2018.

Gao, X., Lee, S., Zhang, Y., Brockett, C., Galley, M., Gao, J., and Dolan, B. Jointly optimizing diversity and relevance in neural response generation. *arXiv preprint arXiv:1902.11205*, 2019.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pp. 5767–5777, 2017.

Guo, J., Lu, S., Cai, H., Zhang, W., Yu, Y., and Wang, J. Long text generation via adversarial training with leaked information. *arXiv preprint arXiv:1709.08624*, 2017.

Hashimoto, T. B., Zhang, H., and Liang, P. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*, 2019.

Ippolito, D., Kriz, R., Sedoc, J., Kustikova, M., and Callisonburch, C. Comparison of diverse decoding methods from conditional language models. pp. 3752–3762, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

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

Li, J., Lan, Y., Guo, J., Xu, J., and Cheng, X. Differentiated distribution recovery for neural text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6682–6689, 2019.

Lin, C.-Y. and Och, F. Looking for a few good metrics: Rouge and its evaluation. In *Ntcir Workshop*, 2004.

Lin, K., Li, D., He, X., Zhang, Z., and Sun, M.-T. Adversarial ranking for language generation. In *Advances in Neural Information Processing Systems*, pp. 3155–3165, 2017.

Lu, S., Yu, L., Zhang, W., and Yu, Y. Cot: Cooperative training for generative modeling of discrete data. *arXiv preprint arXiv:1804.03782*, 2018a.

Lu, S., Zhu, Y., Zhang, W., Wang, J., and Yu, Y. Neural text generation: Past, present and beyond. *arXiv preprint arXiv:1803.07133*, 2018b.

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., and Khudanpur, S. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

Nie, W., Narodytska, N., and Patel, A. Relgan: Relational generative adversarial networks for text generation. 2018.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. Self-critical sequence training for image captioning. In *CVPR*, volume 1, pp. 3, 2017.

Semeniuta, S., Severyn, A., and Gelly, S. On accurate evaluation of gans for language generation. *arXiv preprint arXiv:1806.04936*, 2018.

Shen, T., Ott, M., Auli, M., and Ranzato, M. Mixture models for diverse machine translation: Tricks of the trade. *arXiv: Computation and Language*, 2019.

Subramanian, S., Mudumba, S. R., Sordoni, A., Trischler, A., Courville, A. C., and Pal, C. Towards text generation with adversarially learned neural outlines. In *Advances in Neural Information Processing Systems*, pp. 7551–7563, 2018.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D. J., and Batra, D. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv: Artificial Intelligence*, 2016.

Yu, L., Zhang, W., Wang, J., and Yu, Y. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pp. 2852–2858, 2017.

Zhang, H., Lan, Y., Guo, J., Xu, J., and Cheng, X. Tailored sequence to sequence models to different conversation scenarios. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1479–1488, 2018.

Zhang, J., Feng, Y., Wang, D., Wang, Y., Abel, A., Zhang, S., and Zhang, A. Flexible and creative chinese poetry generation using neural memory. *arXiv preprint arXiv:1705.03773*, 2017a.

Zhang, Y., Gan, Z., Fan, K., Chen, Z., Henao, R., Shen, D., and Carin, L. Adversarial feature matching for text generation. *arXiv preprint arXiv:1706.03850*, 2017b.

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., and Yu, Y. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1097–1100. ACM, 2018.