Supplementary Materials

A Preliminaries

Before starting the proofs, we first introduce some preliminaries on the constrained convex optimization problem. Assume $f(x)$, $c_i(x)$, and $h_j(x)$ are continuous differentiable function define on $\mathbb{R}^n$, consider the constrained convex optimization problem defined as follows:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ s.t. \quad & c_i(x) \leq 0, \quad i = 1, 2, \cdots, k \\ & h_j(x) = 0, \quad j = 1, 2, \cdots, l \end{aligned} \tag{1}$$

The optimal solutions for above problem are given by the Lagrange Multiplier approach , as shown in the following theorem:

Theorem 1. *Assume $f(x)$ and $c_i(x)$ are convex, $h_j(x)$ are affine, and c_i are strictly feasible (there exists one x satisfying $c_i(x) < 0$ for all i). Define the Lagrange function as:*

$$L(x, \alpha, \beta) = f(x) + \sum_{i=1}^{k} \alpha_i c_i(x) + \sum_{j=1}^{l} \beta_j h_j(x),$$

where $\alpha \geq 0$. Then the the following conditions are both sufficient and necessary for x to be a solution in problem 1.

$$\begin{aligned} \nabla_x L(x^*, \alpha^*, \beta^*) &= 0 \\ \nabla_\alpha L(x^*, \alpha^*, \beta^*) &= 0 \\ \nabla_\beta L(x^*, \alpha^*, \beta^*) &= 0 \\ \alpha_i^* c_i(x^*) &= 0, \quad i = 1, 2, \cdots, k \\ c_i(x^*) &\leq 0, \quad i = 1, 2, \cdots, k \\ \alpha_i^* &\geq 0, \quad i = 1, 2, \cdots, k \\ h_j(x^*) &= 0, \quad j = 1, 2, \cdots, k \end{aligned} \tag{2}$$

The conditions in Equation 2 are called the Karush-Kuhn-Tucker(KKT) conditions.

B Proof of Theorem 1

For property 1, from $U(Q') - U(Q) = \epsilon f(P_i) - \epsilon f(P_j) = \epsilon[f(P_i) - f(P_j)] > 0$, we get $f(P_i) > f(P_j)$. We then get the conclusion by setting $x_1 = P_i$ and $x_2 = P_j$.

For property 2, $V(Q') - V(Q) = [g(Q_i + \epsilon) + g(Q_j - \epsilon)] - [g(Q_i) + g(Q_j)] < 0$ is true for any $Q_i > Q_j$. Denote $C = Q_i + Q_j$ and $r(x) = g(x) + g(C - x)$, then we have $V(Q') - V(Q) = r(Q_i + \epsilon) - r(Q_i) < 0$ for any Q_i, ϵ. Since $0 < Q_i < Q_i + \epsilon < 1$, we need $r'(x) < 0$ for $x \in (0, 1)$. Then, since $r'(x) = g'(x) - g'(C - x) < 0$ is true for any $0 < C - x < x < 1$. Set $x_1 = C - x$ and $x_2 = x$ and we get $g'(x_1) < g'(x_2)$ for any $x_1 > x_2 > 0$ and $x_1 + x_2 = Q_i + Q_j \leq 1$.

C Lemmas

We give two lemmas to support the proof of Theorem 2 and Theorem 3.

C.1 Lemma 1

Lemma 1. *If Q is a Pareto-optimum, then the following conditions are satisfied: if $P_i > P_j$, then $Q_i \geq Q_j$; if $P_i = P_j$, then $Q_i = Q_j$.*

If $P_i > P_j$, assume $Q_i < Q_j$, we can construct Q' where $Q'_k = Q_k$ for all $k \neq i, j$ and $Q'_i = Q_j, Q'_j = Q_i$. As such, $V(Q') = V(Q)$ but $U(Q') - U(Q) = (Q_j - Q_i)[f(P_i) - f(P_j)] > 0$. This means Q is dominated by Q', which conflicts with the fact that Q is a Pareto-optimum. So $Q_i \geq Q_j$.

If $P_i = P_j$, assume $Q_i \neq Q_j$, and we can further assume $Q_i > Q_j$. Again we construct Q' where $Q'_k = Q_k$ for all $k \neq i, j$ and $Q'_i = Q'_j = \frac{Q_i + Q_j}{2}$. Surely we have $U(Q') = U(Q)$, and $V(Q') - V(Q) = 2g(\frac{Q_i+Q_j}{2}) - g(Q_i) - g(Q_j)$. Since g is strictly concave, we have $V(Q') - V(Q) > 0$, which means Q is dominated by Q'. This causes confliction, so $Q_i = Q_j$.

C.2 Lemma 2

Lemma 2. *Assume $\alpha \in [0, 1)$ and $\Psi(Q) = \alpha U(Q) + (1-\alpha)V(Q)$, then the distribution Q that maximize $\Psi(Q)$ satisfies $Q_i = \hat{g}'^{-1}[w \cdot f(P_i) + b]$, and $w = \frac{\alpha}{\alpha - 1}$.*

Define the optimization problem as follows:

$$\min_{Q} -\alpha \cdot U(Q) - (1-\alpha)V(Q)$$

$$s.t.\ 1 - \sum_{i=1}^{N} Q_i = 0$$

$$\forall i,\ -Q_i \leq 0$$

Again we first check that the prerequisites in KKT are all satisfied. $-U(Q)$ is linear and $-V(Q)$ is convex w.r.t. Q; $1 - \sum_{i=1}^{N} Q_i$ is affine w.r.t. Q; since all Q_i can be positive, so the inequalities are all strictly feasible.

The Lagrange function is:

$$L(Q_i, \lambda, \xi_i) = -\alpha \sum_{i=1}^{N} Q_i f(P_i) - (1-\alpha) \sum_{i=1}^{N} g(Q_i) + \lambda(1 - \sum_{i=1}^{N} Q_i) - \sum_{i=1}^{N} \xi_i Q_i, \quad \xi \geq 0$$

Apply KKT and we get the following conditions for a optimal solution:

$$\forall i,\ \frac{\partial L}{\partial Q_i} = -\alpha f(P_i) - (1-\alpha) g'(Q_i) - \lambda - \xi_i = 0,$$

$$\forall i,\ -\xi_i Q_i = 0$$

For $Q_i \neq 0$, there is $\xi_i = 0$, so

$$Q_i = g'^{-1}[\frac{\alpha}{\alpha - 1} f(P_i) + \frac{\lambda}{\alpha - 1}];$$

for $Q_i = 0$, there is $\xi_i > 0$, so

$$\frac{\alpha}{\alpha - 1} f(P_i) + \frac{\lambda}{\alpha - 1} > g'(0).$$

Denote $w = \frac{\alpha}{\alpha - 1}$ and $b = \frac{\lambda}{\alpha - 1}$ and combine the two cases together, we get:

$$Q_i = \hat{g}'^{-1}[w \cdot f(P_i) + b], \quad w \leq 0,$$

The above derivation is both sufficient and necessary, so we finished the proof.

D Proof of Theorem 2

We give the proofs for three conclusions individually.

D.1 Conclusion 1

Here we only consider the case with $U(Q) \neq \max_Q U(Q)$, and the case where $U(Q) = \max_Q U(Q)$ will be incorporated into conclusion 3. We try to find a distribution Q' with the highest diversity while quality is not lower than Q. Define a convex optimization problem as follows:

$$\begin{aligned} &\min_{Q'} -V(Q') \\ s.t.\ &U(Q) - U(Q') \leq 0 \\ &1 - \sum_{i=1}^{N} Q'_i = 0 \\ &\forall i,\ -Q'_i \leq 0 \end{aligned}$$

For Q to be a Pareto-optimum, it's necessary for $Q' = Q$ to be a solution of above problem. Thus we try to solve this problem next.

We first check that the prerequisites in KKT are all satisfied. $-V(Q')$ is convex w.r.t. Q'; $1-\sum_{i=1}^{N} Q'_i$ is affine w.r.t. Q'; $U(Q) - U(Q')$ and $-Q'_i$ are convex(linear) w.r.t Q'; since all Q'_i can be positive and $U(Q) \neq \max_Q U(Q)$, so the inequalities are all strictly feasible.

The Lagrange function is:

$$L(Q'_i, \lambda, \eta, \xi_i) = -\sum_{i=1}^{N} g(Q'_i) + \lambda(1 - \sum_{i=1}^{N} Q'_i) + \eta \sum_{i=1}^{N} (Q_i - Q'_i) f(P_i) - \sum_{i=1}^{N} \xi_i Q'_i, \quad \eta, \xi \geq 0$$

Apply KKT and we get the following conditions for a optimal solution:

$$\begin{aligned} \forall i,\ \frac{\partial L}{\partial Q'_i} = -g'(Q'_i) - \lambda - \eta f(P_i) - \xi_i = 0, \\ \eta[U(Q) - U(Q')] = 0, \\ \forall i,\ -\xi_i Q'_i = 0 \end{aligned}$$

Since we need $Q' = Q$ to be a solution, so

$$\begin{aligned} \forall i,\ -g'(Q_i) - \lambda - \eta f(P_i) - \xi_i = 0, \\ \forall i,\ -\xi_i Q_i = 0 \end{aligned}$$

For $Q_i \neq 0$, there is $\xi_i = 0$, so $Q_i = g'^{-1}[-\eta f(P_i) - \lambda]$; for $Q_i = 0$, there is $\xi_i > 0$, so $-\eta f(P_i) - \lambda > g'(0)$. Denote $w = -\eta$ and $b = -\lambda$ and combine the two cases together, we get:

$$Q_i = \hat{g}'^{-1}[w \cdot f(P_i) + b], \quad w \leq 0,$$

where

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

Now we get a necessary condition for Q to be a Pareto-optimum. To make it sufficient, we still require that for any two distributions satisfying this form, no one could dominate another. This property can be proved by combining conclusion 2 and 3.

D.2 Conclusion 2

We separate the proof into two parts: (1) b is correspondent to w; (2) the monotonicity of b w.r.t. w.

(1) The sum of all Q_i should be 1. Denote

$$T(w, b) = \sum_{i=1}^{N} \hat{g}'^{-1}[w \cdot f(P_i) + b].$$

Since $g'(x)$ is strictly monotonically decreasing, so $T(w, b)$ is monotonically non-increasing w.r.t. b. If $T(w, b) > 0$, there would be a term which is strictly monotonically decreasing w.r.t. b, under which condition $T(w, b)$ is strictly monotonically decreasing w.r.t. b. Also, $T(w, b)$ is continuous w.r.t. b since g'^{-1} is continuous. When

$$b = g'(0) - w \cdot f(\max_i P_i),$$

there is

$$w \cdot f(P_i) + b \geq w \cdot f(\max_i P_i) + b = g'(0),$$

so $T(w, b) = 0$; when

$$b = g'(\frac{1}{N}) - w \cdot f(\min_i P_i),$$

there is

$$w \cdot f(P_i) + b \leq w \cdot f(\min_i P_i) + b = g'(\frac{1}{N}),$$

so $T(w, b) \geq 1$. From above analysis, the value of T can reach 0 or be greater than 1. So combining the monotonicity of T, there exists and only one b that satisfies $T(w, b) = 1$, leading to a rational distribution.

(2) Define $T(w, b) = \sum_{i=1}^{N} \hat{g}'^{-1}[w \cdot f(P_i) + b(w)]$ as above. Since $T(w, b)$ represents the total probability of a distribution, so there should be $T(w, b) \equiv 1$, thus $\frac{\mathrm{d}T}{\mathrm{d}w} = 0$.

$$\frac{\mathrm{d}T}{\mathrm{d}w} = \sum_{i \in S} \frac{f(P_i) + b'(w)}{g''\{g'^{-1}[w \cdot f(P_i) + b(w)]\}},$$

where $S = \{i | w \cdot f(P_i) + b(w) < g'(0)\}$. By the condition $\frac{\mathrm{d}T}{\mathrm{d}w} = 0$, we get

$$b'(w) = -\frac{\sum_{i \in S} \frac{f(P_i)}{g''\{g'^{-1}[w \cdot f(P_i) + b(w)]\}}}{\sum_{i \in S} \frac{1}{g''\{g'^{-1}[w \cdot f(P_i) + b(w)]\}}}.$$

Since $g''(x) < 0$, so if $f(x) < 0$ for all $x \in [0, 1]$, we can get $b'(w) > 0$, thus b is strictly monotonically increasing w.r.t. w. Similarly, if $f(x) > 0$ for all $x \in [0, 1]$, we can get $b'(w) < 0$, thus b is strictly monotonically decreasing w.r.t. w.

D.3 Conclusion 3

We also separate the proof into two parts: (1) the uniqueness of $Q(w)$; (2) the monotonicity of U and V w.r.t. w.

(1) Since P is not uniform, so we can denote B, P_{m_1}, P_{m_2} as they are in the theorem. According to Lemma 1, since P_{m_1} is the largest one, so the corresponding Q_{m_1} is also the largest one, which means

$$Q_{m_1} = \hat{g}'^{-1}[w \cdot f(P_{m_1}) + b] > 0.$$

Thus we get

$$w \cdot f(P_{m_1}) + b < g'(0).$$

At the same time, because we can get $Q_i = Q_{m_1}$ if $P_i = P_{m_1}$, so we can sum up all the largest Q_i and get

$$M \cdot Q_{m_1} \leq \sum_{i=1}^{N} Q_i = 1,$$

we can get

$$w \cdot f(P_{m_1}) + b \geq g'(\frac{1}{M}). \tag{3}$$

Consider the case where $w \geq B$, we first prove that $w \cdot f(P_{m_2}) + b \leq g'(0)$. Assume

$$w \cdot f(P_{m_2}) + b > g'(0), \tag{4}$$

then $Q_{m_2} = 0$, and there is $Q_i = 0$ for any i satisfying $P_i \leq P_{m_2}$. As a result, there should be $Q_i = \frac{1}{M}$ for all i satisfying $P_i = P_{m_1}$, which means

$$w \cdot f(P_{m_1}) + b = g'(\frac{1}{M}). \tag{5}$$

Subtract Equation 5 by Equation 4, we get

$$w \cdot [f(P_{m_1}) - f(P_{m_2})] < g'(\frac{1}{M}) - g'(0),$$

so

$$w < \frac{g'(\frac{1}{M}) - g'(0)}{f(P_{m_1}) - f(P_{m_2})} = B.$$

This contradict with the fact that $w \geq B$. Thus we have $w \cdot f(P_{m_2}) + b \leq g'(0)$.

Combining the above conclusions, for any $w_1, w_2 \in [B, 0]$, assume $Q(w_1) = Q(w_2)$, then

$$w_1 \cdot f(P_{m_1}) + b_1 = w_2 \cdot f(P_{m_1}) + b_2,$$
$$w_1 \cdot f(P_{m_2}) + b_1 = w_2 \cdot f(P_{m_2}) + b_2.$$

As $P_{m_1} \neq P_{m_2}$, so $w_1 = w_2$, causing contradiction. Thus we have $Q(w_1) \neq Q(w_2)$.

For any $w \leq B$, assume

$$w \cdot f(P_{m_2}) + b < g'(0). \tag{6}$$

By subtracting Equation 3 and Equation 6, we get

$$w \cdot [f(P_{m_1}) - f(P_{m_2})] > g'(\frac{1}{M}) - g'(0),$$

so

$$w > \frac{g'(\frac{1}{M}) - g'(0)}{f(P_{m_1}) - f(P_{m_2})} = B.$$

This causes contradiction, so the above assumption does not hold. Thus we have $w \cdot f(P_{m_2}) + b \geq g'(0)$, which means $Q_{m_2} = 0$. Borrowing the proof above, we know that $Q_i = \frac{1}{M}$ for all i satisfying $P_i = P_{m_1}$. This is a trivial Pareto-optimal case where $U(Q) = \max_Q U(Q)$. Now we know the distribution Q is fixed and does not change as w changes, so for any $w_1, w_2 \leq B$, there is $Q(w_1) = Q(w_2)$.

(2) For the expression of Q_i, since f and g' are both continuous and monotonic, so it is easy to know that Q_i is continuous w.r.t. w, then $U(Q(w))$ and $V(Q(w))$ are both continuous w.r.t. w. We just need to prove the monotonicity.

Assume $B \leq w_1 < w_2 \leq 0$, the goal is to prove that $U(Q(w_1)) > U(Q(w_2))$ and $V(Q(w_1)) < V(Q(w_2))$. According to Lemma 2, w_1 and w_2 have their corresponding $\alpha_1 = \frac{w_1}{w_1 - 1}$ and $\alpha_2 = \frac{w_2}{w_2 - 1}$, and $\alpha_1 > \alpha_2$. Since $Q(w)$ is the optimal solution for problem $\alpha U(Q) + (1-\alpha)V(Q)$, and $Q(w_1)$ is different with $Q(w_2)$, so the following inequalities hold:

$$\alpha_1 U(Q(w_1)) + (1-\alpha_1)V(Q(w_1)) > \alpha_1 U(Q(w_2)) + (1-\alpha_1)V(Q(w_2)),$$
$$\alpha_2 U(Q(w_1)) + (1-\alpha_2)V(Q(w_1)) < \alpha_2 U(Q(w_2)) + (1-\alpha_2)V(Q(w_2)).$$

Subtracting the first equation by the second one, we get

$$(\alpha_1 - \alpha_2)[(U(Q(w_1)) - U(Q(w_2))) - (V(Q(w_1)) - V(Q(w_2)))] > 0.$$

As $\alpha_1 > \alpha_2$, so

$$U(Q(w_1)) - U(Q(w_2)) > V(Q(w_1)) - V(Q(w_2)).$$

Because $Q(w_1)$ and $Q(w_2)$ are both Pareto-optima, there quality and diversity should satisfy one of the following: $U(Q(w_1)) > U(Q(w_2)), V(Q(w_1)) < V(Q(w_2))$ or $U(Q(w_1)) < U(Q(w_2)), V(Q(w_1)) > V(Q(w_2))$. With the derived restriction $U(Q(w_1)) - U(Q(w_2)) > V(Q(w_1)) - V(Q(w_2))$, we know the first one holds, that is $U(Q(w_1)) > U(Q(w_2))$ and $V(Q(w_1)) < V(Q(w_2))$.

E Proof of Theorem 3

The requirement that $Q = P$ being a Pareto-optimum is equivalent to the following condition: for any P, there exist $w_0 \leq 0$ and b_0 that for any i, there is

$$P_i = \hat{g}'^{-1}[w_0 \cdot f(P_i) + b_0].$$

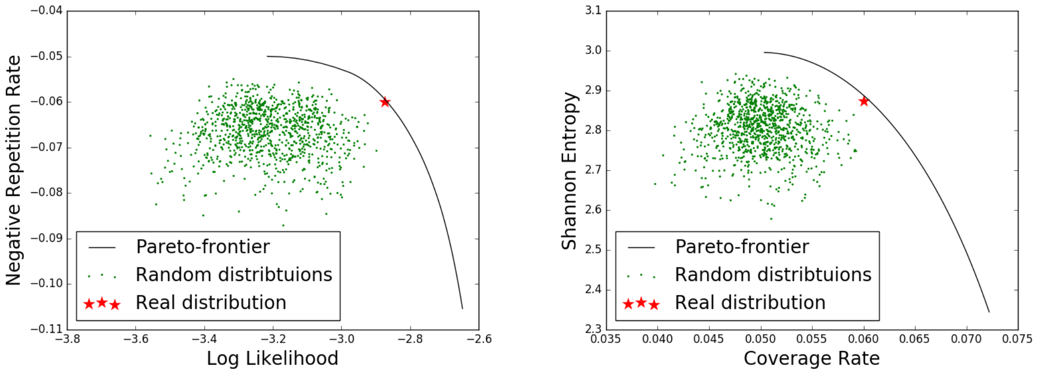

Figure 1: Illustration of the Pareto-frontier on a random toy categorical distribution with size 20. The ground truth distribution is under the frontier curve. **Left:** Pair LL with NRR. **Right:** Pair CR with SE.

This means, for any $P_i > 0$, there is $w_0 \cdot f(P_i) + b_0 = g'(P_i)$. Since f and g' are both continuous, so

$$w_0 \cdot f(0) + b_0 - g'(0) = \lim_{P_i \to 0} w_0 \cdot f(P_i) + b_0 - g'(P_i) = 0.$$

We can see $w_0 \cdot f(P_i) + b_0 = g'(P_i)$ is also true for $P_i = 0$. By solving this differential equation, we get

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

Here b_0 can be any value because $P_i = g'^{-1}[w_0 \cdot f(P_i) + b_0]$ always lead to a plausible distribution P. Under this condition, we know that $Q = P$ is the only distribution that maximize $\Psi(Q) = \alpha U(Q) + (1-\alpha)V(Q)$ where $\alpha = \frac{w_0}{w_0 - 1}$ according to Lemma 2. With above conclusions, it is easy to check that $D(P||Q) = \Psi(P) - \Psi(Q) \geq 0$ and $D(P||Q) = 0$ if and only if $Q = P$, thus $D(P||Q)$ is a divergence metric.

F Pareto-frontier with Mismatched Metrics

We show in Figure 1 that the point $Q = P$ is under the Pareto-frontier curve when quality and diversity metrics are not matched, i.e. the condition in Theorem 3 is not satisfied. We use the same toy dataset, but pair LL with NRR and CR with SE. Note that there is always a gap between the star and the curve, indicating that the real distribution lies on neither of the two Pareto-frontiers.

G Additional Information for Experiments

G.1 Experiments on Synthetic Data

The probabilities of synthetic ground truth distributions are shown in Figure 2. We use different standard deviations to get different kind of distributions. Distribution with $\sigma = 0.5$ is more flat and of higher entropy, and distribution with $\sigma = 2.0$ is more sharp and of lower entropy.

We show the training curve of the optimization process used on synthetic data in Figure 3. Learning rates are adjusted according to each process, so as to find a best distribution. Points are neglected if $V(Q) \leq V(P)$ or $U(Q) \leq U(P)$, i.e. they fail to dominate the ground truth distribution.

We show the correlation between CR/NRR/CND and quality/diversity/divergence on synthetic data, respectively. We use the well-defined Pareto-frontier under LL-SE in text space as target models, i.e. $Q_i \propto P_i^{\beta}$. As β decreases, the corresponding Pareto-optimum becomes more close to uniform distribution, so that quality decreases and diversity increases according to Theorem 2, and minimal divergence is taken when $\beta = 1$ according to Theorem 3. We plot the curves of BLEU-NSBLEU, CR-NRR, and CND in Figure 4. We can see CR/NRR/CND can properly reflect quality/diversity/divergence, respectively.

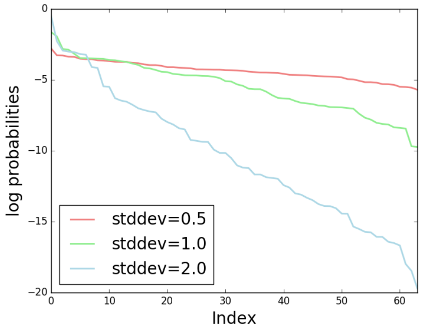

Figure 2: The log-probabilities of three synthetic ground truth distributions used in our experiments, shown in descending order.

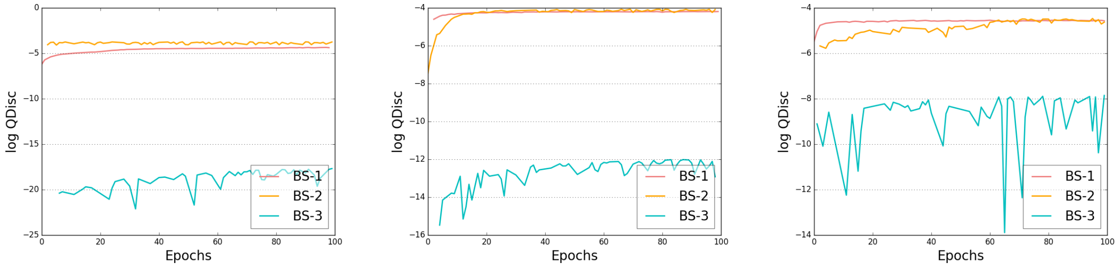

Figure 3: The optimization curve of Quality-Discrepancy for BLEU-NSBLEU metric pair on synthetic data with different standard deviations, $\sigma = 0.5, 1.0, 2.0$ from left to right.

G.2 Experiments on Real Text Data

For MSCOCO dataset, we remove words with frequency lower than 20, as well as sentences containing them. The vocabulary size is 5,473, and maximum text length is 32. Sentences longer than 32 are also removed, and we get a total number of 530,093 sentences. We randomly sample 50,000 sentences as candidate set, 50,000 sentences as reference set, and another 200,000 sentences for training data of the RNNLM.

For WMT dataset, we use the Europarl-v7 part. We remove words with frequency lower than 400, as well as sentences containing them. The vocabulary size is 6,655, and maximum text length is 50. Sentences longer than 50 or shorter than 20 are also removed, and we get a total number of 475,662 sentences. We again randomly sample 50,000 sentences as candidate set, 50,000 sentences as reference set, and another 200,000 sentences for training data of the RNNLM.

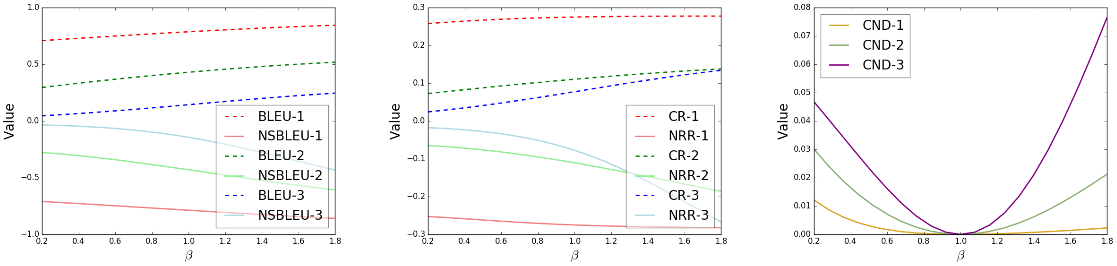

Figure 4: Evaluation of BLEU-NSBLEU, CR-NRR, and CND on synthetic data with $\sigma = 1.0$. Test models are Pareto-optima parameterized by β under LL-SE metric pair.