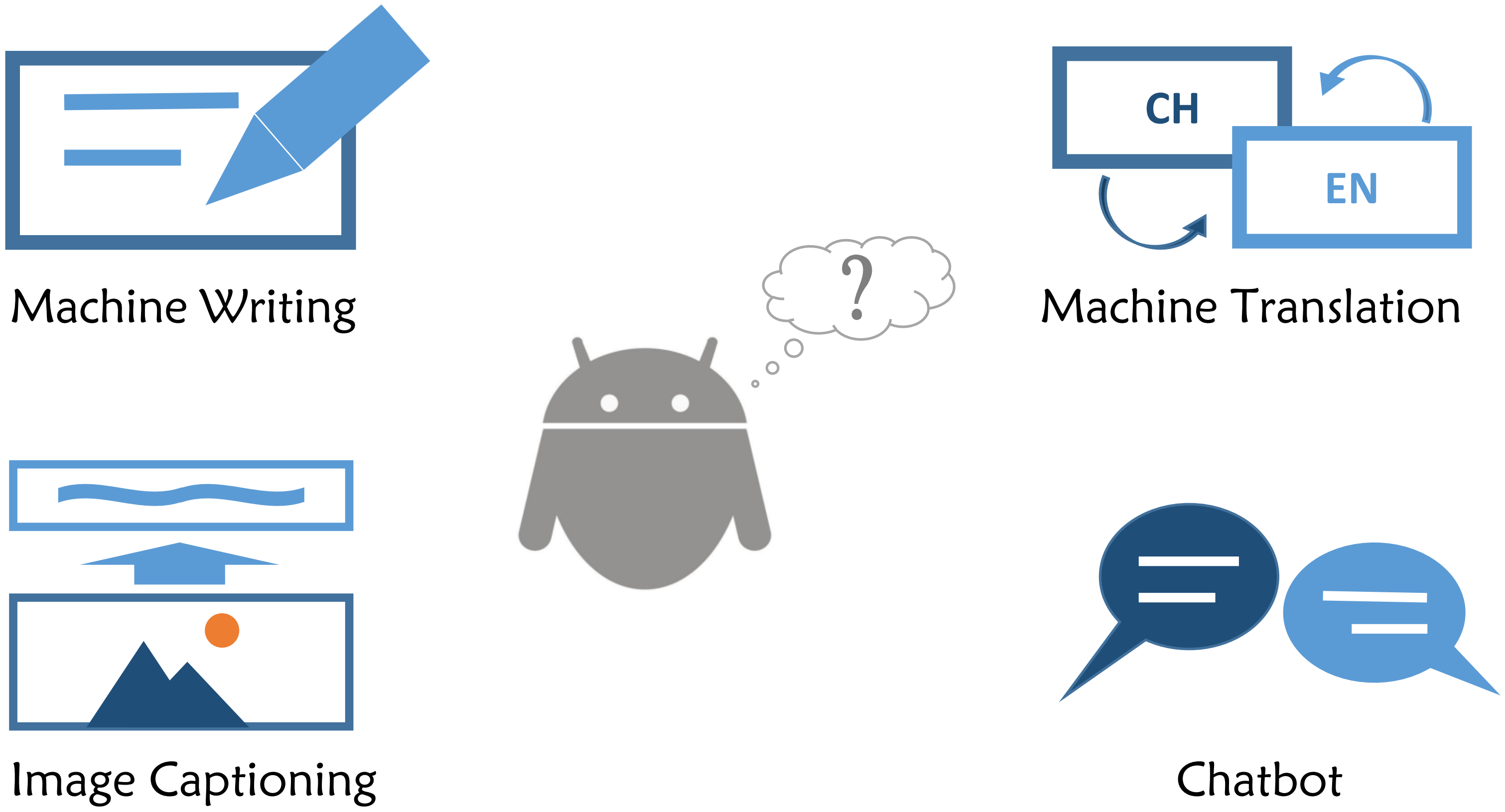
Exploring the Pareto-Optimality between Quality and Diversity in Text Generation Models

Jianing Li, Yanyan Lan, Jiafeng Guo, Xueqi Cheng

Content

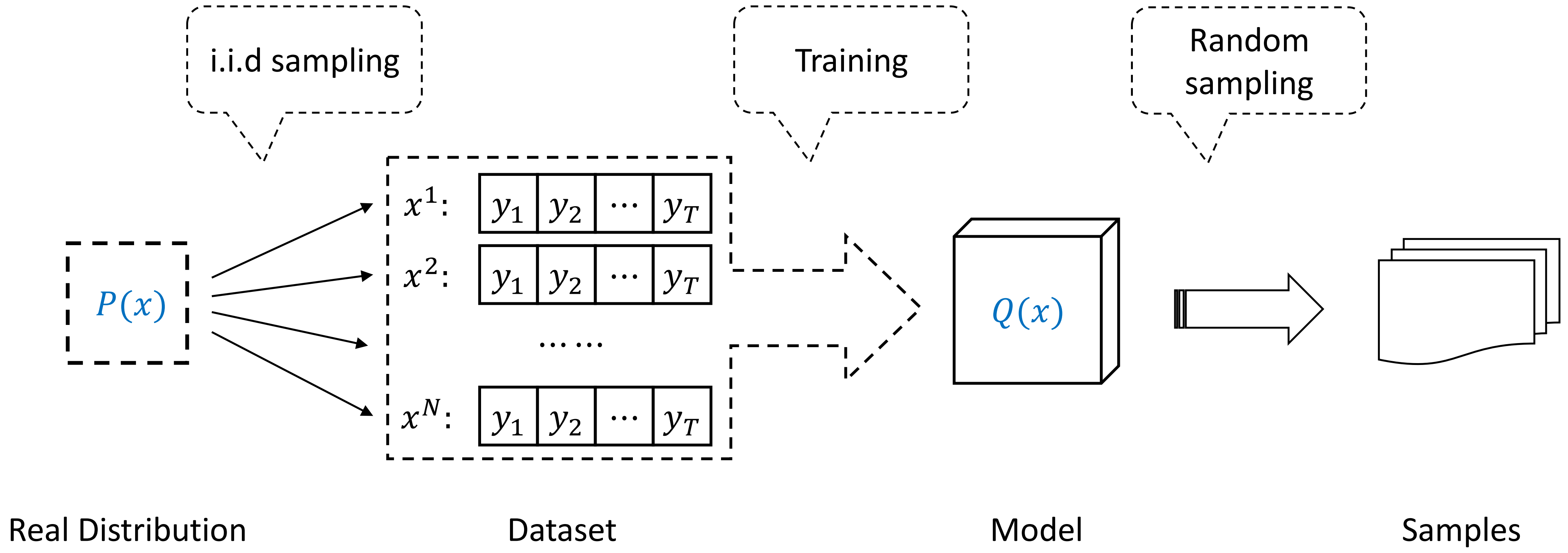
- Background
- Problem & Formalization
- Theoretical Analysis
- Conclusion

Background – Text Generation Tasks



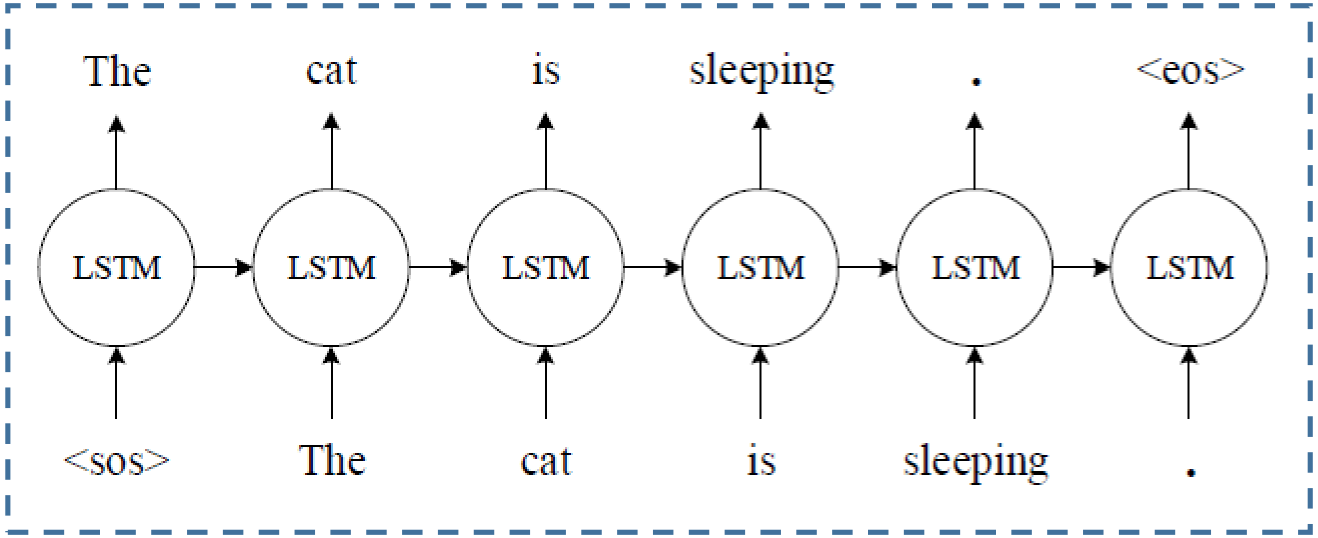
Background – Unconditional Text Generation

- Given a text dataset, build a model $Q(x)$ for text generation.



Background – RNN-based Language Model

- Most widely used method for text generation



Probability Decomposition:

$$x := Y_{1:T}$$

$$Q(Y_{1:T}) = \prod_{i=1}^{T} Q(y_t | Y_{1:t-1})$$

- Training by Maximum Likelihood Estimation (MLE)

$$\max_{Q} \mathbb{E}_{x \sim P} \log Q(x) \Longleftrightarrow \min_{Q} D_{KL}(P||Q) \Longrightarrow Q^* = P$$

Background – Evaluation

- Divergence: How close is it between distribution Q and P
  - Perplexity(PPL): How likely are validation data to be generated by the model

- Quality: How likely are generated samples to be real ones
  - NLL-oracle: How likely are generated samples to appear in the real distribution
  - BLEU: N-gram overlap between generated samples and validation data

- Diversity: How much difference are there between generated samples
  - Distinct: Percentage of unique N-grams within generated sample
  - Self-BLEU: N-gram overlap within generated sample

Problem & Formularization

Problem

- What is the relationship between quality and diversity?
  - Quality and diversity seem to be a trade-off in practice, but there is no theoretical explanation.

- Can quality and diversity be used to evaluate divergence?
  - Calculation of divergence may not be tractable, so quality and diversity are usually used instead in practice. It is not clear if they are sufficient.

- How to balance the quality and diversity in practice?
  - Some tasks focus more on higher quality (or diversity). There is currently no method that maximize one metric while keeping another above a threshold.

Formularization – Notations

- Given the real distribution P and model distribution Q
  - $P = (P_1, \cdots P_N)$
  - $Q = (Q_1, \cdots Q_N)$

- Quality metric: $U(Q)$

- Diversity metric: $V(Q)$

Formularization – Special cases

- LL-SE metric:
  - Log-likelihood with oracle: $U(Q) = \mathbb{E}_{x\sim Q} \log P(x) = \sum_{i=1}^{N} Q_i \log P_i$
  - Shannon Entropy: $V(Q) = -\mathbb{E}_{x\sim Q} \log Q(x) = -\sum_{i=1}^{N} Q_i \log Q_i$

- CR-NRR metric:
  - Coverage Rate: $U(Q) = \mathbb{E}_{x\sim \mathrm{Q}} P(x) = \sum_{i=1}^{N} Q_i \cdot P_i$
  - Negative Repeat Rate: $V(Q) = -\mathbb{E}_{x\sim Q} Q(x) = -\sum_{i=1}^{N} Q_i^2$

Formularization – General

- Quality metric:
  - $U(Q) = U(Q; P) = \sum_{i=1}^{N} Q_i \cdot f(P_i)$
  - Generating more samples with higher real probability yields higher overall quality
  - $\Rightarrow$ $f(x)$ is strictly monotonically increasing w.r.t x

- Diversity metric:
  - $\mathrm{V}(Q) = \sum_{i=1}^{N} g(Q_i)$
  - Distribute the probability more equally yields higher overall diversity
  - $\Rightarrow$ $g(x)$ is strictly concave w.r.t x

Formularization – MOP

- The Multi-Objective Programming problem

$$\max_{Q} \left(U(Q), V(Q)\right)$$

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

- The relationship between quality and diversity lies in the solution of this MOP

Theoretical Analysis

What is the relationship between quality and diversity?

$$\max_{Q} \left(U(Q), V(Q) \right)$$

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

Analysis – Pareto Optimality

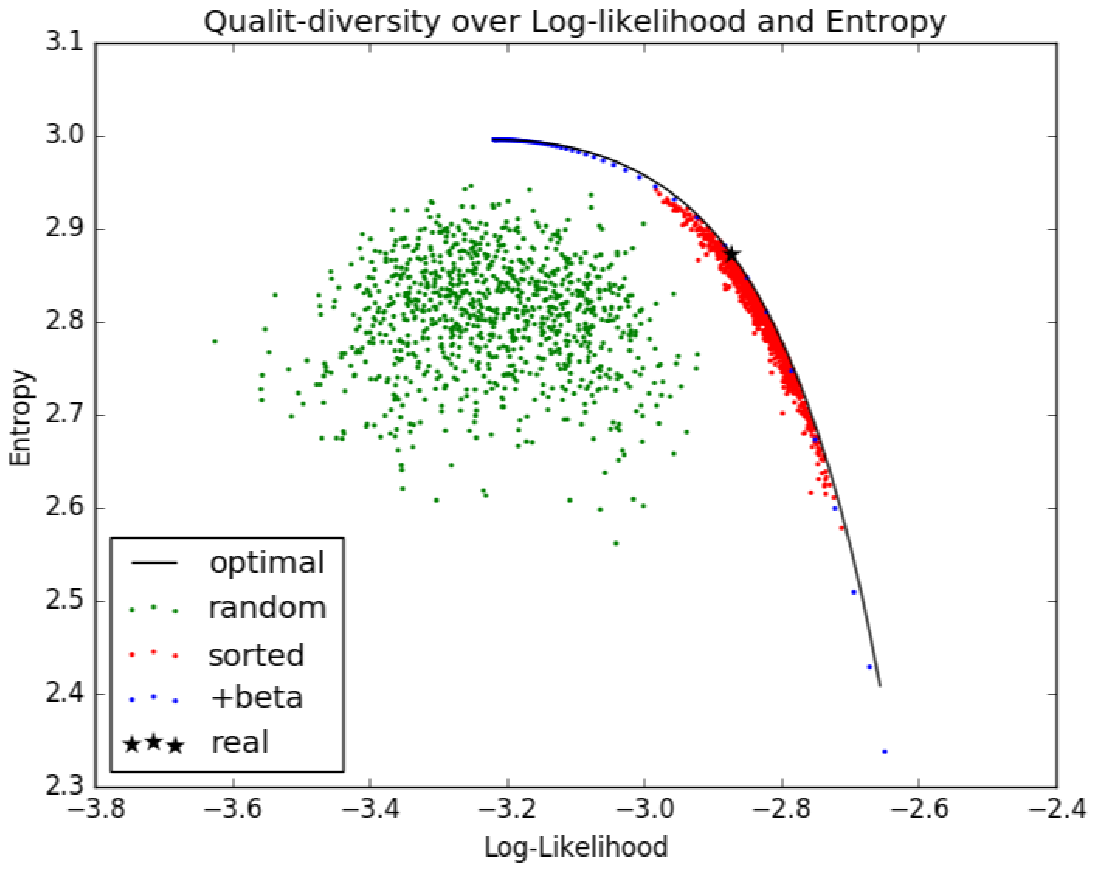
$$\max_{Q} \left(U(Q), V(Q)\right)$$

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

- Pareto-optimum
  - A solution in MOP that no other solution can outperform it over all objectives
- Pareto-frontier
  - The set containing all the Pareto-optima

Analysis – Case Study

- Each point represents a distribution (model)
- Black: Pareto-frontier
- Star: Real distribution P
- Green: Random distributions
- Red: Sorted random distributions
- Blue: Optima from other metrics



Analysis – Pareto Optimum

- Properties of the Pareto-optima
  - The model distribution is order-preserving
  - The model distribution satisfies the following form:

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

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

$$\boxed{U(Q) = U(Q;P) = \sum_{i=1}^{N} Q_i f(P_i), \quad V(Q) = \sum_{i=1}^{N} g(Q_i).}$$

Analysis – Case Study

- LL-SE metric [$f(x) = \log x$, $g(x) = -x \log x$]:

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

$$\boxed{w = -\beta, \ b = 1 + \log Z.}$$

- CR-NRR metric [$f(x) = x$, $g(x) = -x^2$]:

$$Q_i = \frac{\max(P_i + \gamma, 0)}{Z}, \ Z = \sum_{i=1}^{N} \max(P_i + \gamma, 0), \ \gamma > -\max_i P_i,$$

$$\boxed{w = -\tfrac{2}{Z}, b = -\tfrac{2\gamma}{Z}.}$$

Analysis – Trade-off

- The Pareto-frontier contains infinite elements
  - b is determined by w
  - For any $B \leq w \leq 0$, w leads to a different distribution
  - As w grows, the quality $U(Q)$ decreases and the diversity $V(Q)$ increases
  - Each Pareto-optimum corresponds to a solution which maximizes

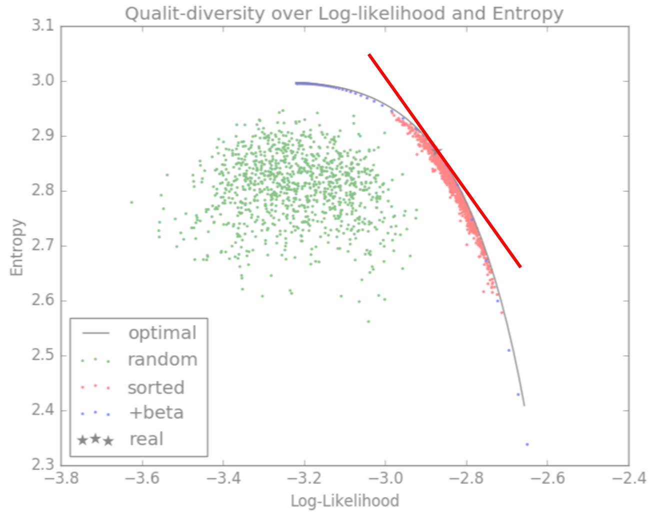
$$W(Q) = \alpha U(Q) + (1 - \alpha) V(Q), \qquad \alpha = \frac{w}{w-1}$$

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

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

There is a trade-off between quality and diversity!

Can quality and diversity be used to evaluate divergence?



Analysis – Relationship with Divergence

- The following condition is both sufficient and necessary for $Q = P$ to be in the Pareto-frontier

$$g(x) = w_0 \int f(x)\mathrm{d}x + b_0 x, \quad w_0 \leq 0.$$

LL-SE: $f(x) = \log x$, $g(x) = -x \log x$

CR-NRR: $f(x) = x$, $g(x) = -x^2$

- If so, then $D(P||Q) = W(P) - W(Q)$ is a divergence metric
  - $W(Q) = \alpha U(Q) + (1-\alpha)V(Q)$, $\alpha = \frac{w_0}{w_0 - 1}$
  - LL-SE: $D(P||Q) = \frac{1}{2}\sum_{i=1}^{N} Q_i \cdot \log \frac{Q_i}{P_i}$
  - CR-NRR: $D(P||Q) = \frac{1}{3}\sum_{i=1}^{N}(P_i - Q_i)^2$

Quality and diversity metrics should be chosen carefully to recover the divergence

How to balance the quality and diversity in practice?

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

Analysis – Algorithm

- We can reach a Pareto-optimum by using the following training objective:

$$\min_Q \mathbb{E}_{x\sim P}\, h[Q(x)],$$

$$h(x) = \int \frac{c}{f^{-1}\left[\frac{g'(x)-b}{w}\right]} \mathrm{d}x, \quad c > 0.$$

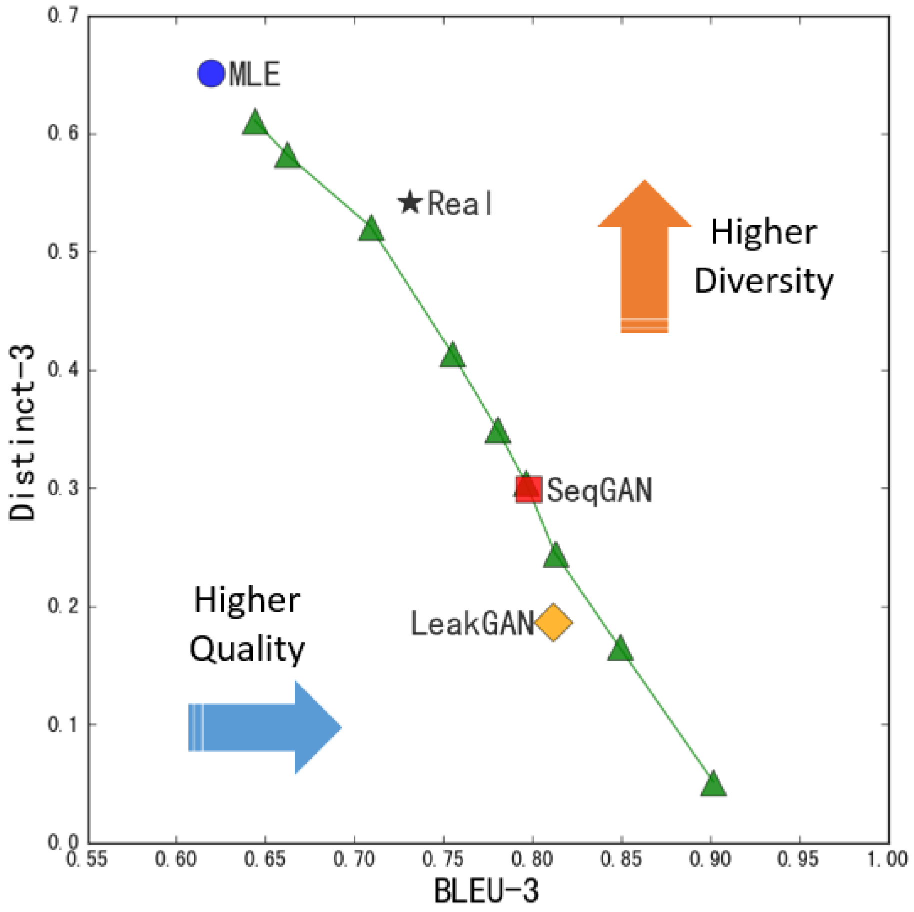
MLE:

$$\boxed{\max_Q \mathbb{E}_{x\sim P}\ \log Q(x)}$$

- Two cases feasible to be used in practice:
  - $f(x) = \log x$ $\quad h(x,w,b,c) = h_1(x,w) \cdot h_2(w,b) \cdot c$
  - $f(x) = x^a$ $\quad h(x,w,b,c) = h_3(x,b) \cdot h_4(w,b) \cdot c$

Analysis – Case Study

- Each point represents a distribution (model)
- Green: Proposed method (LL-SE)
- Star: Real distribution P



Conclusion

Conclusion

- We prove in theory that quality and diversity act as trade-off under unconditional text generation settings.

- We show that quality and diversity can recover the divergence with a linear combination, only if the metrics are carefully chosen.

- We derived a algorithm that can control the degree of quality-diversity trade-off.

Thank you! Q&A

Jianing Li

lijianing@.ict.ac.cn