Differentiated Distribution Recovery for Neural Text Generation

Jianing Li[1,2], Yanyan Lan[1,2,3], Jiafeng Guo[1,2], JunXu[1,2], Xueqi Cheng[1,2]

1. CAS Key Lab of Network Data Science and Technology,

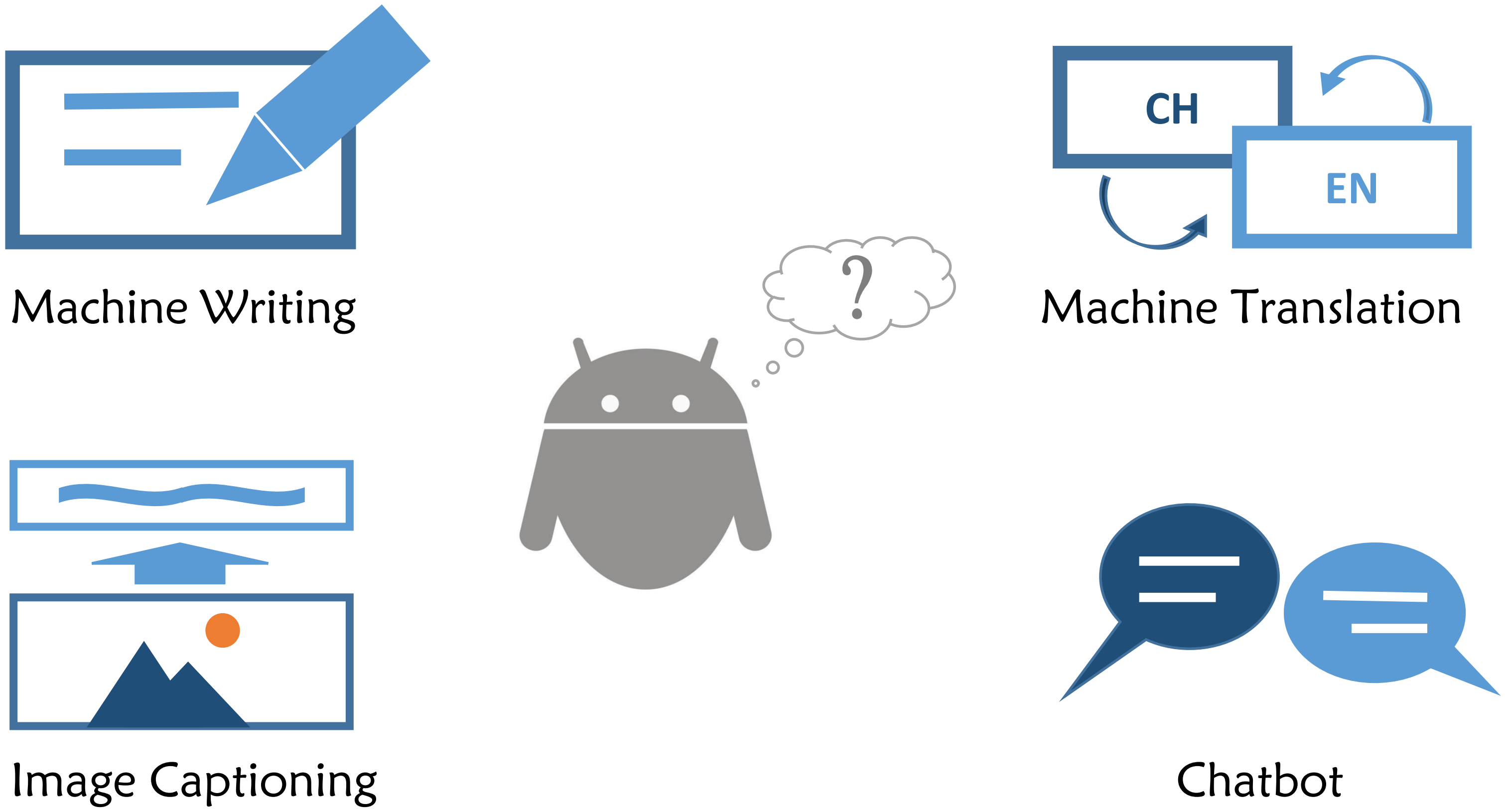
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

2. University of Chinese Academy of Sciences, Beijing, China

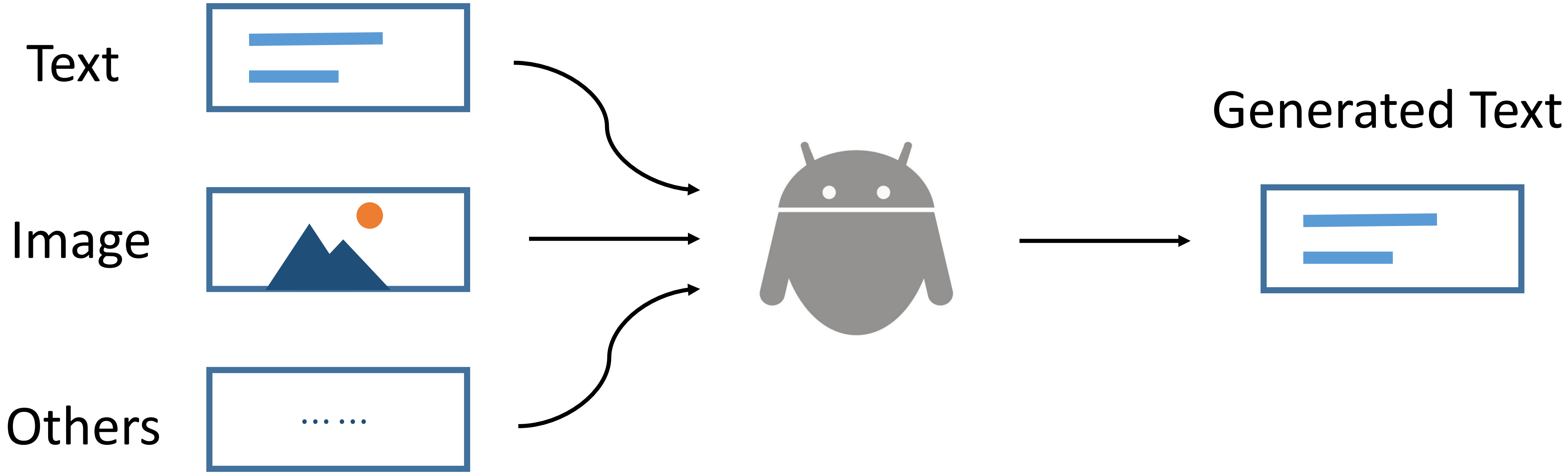
3. Department of Statistics, University of California, Berkeley

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

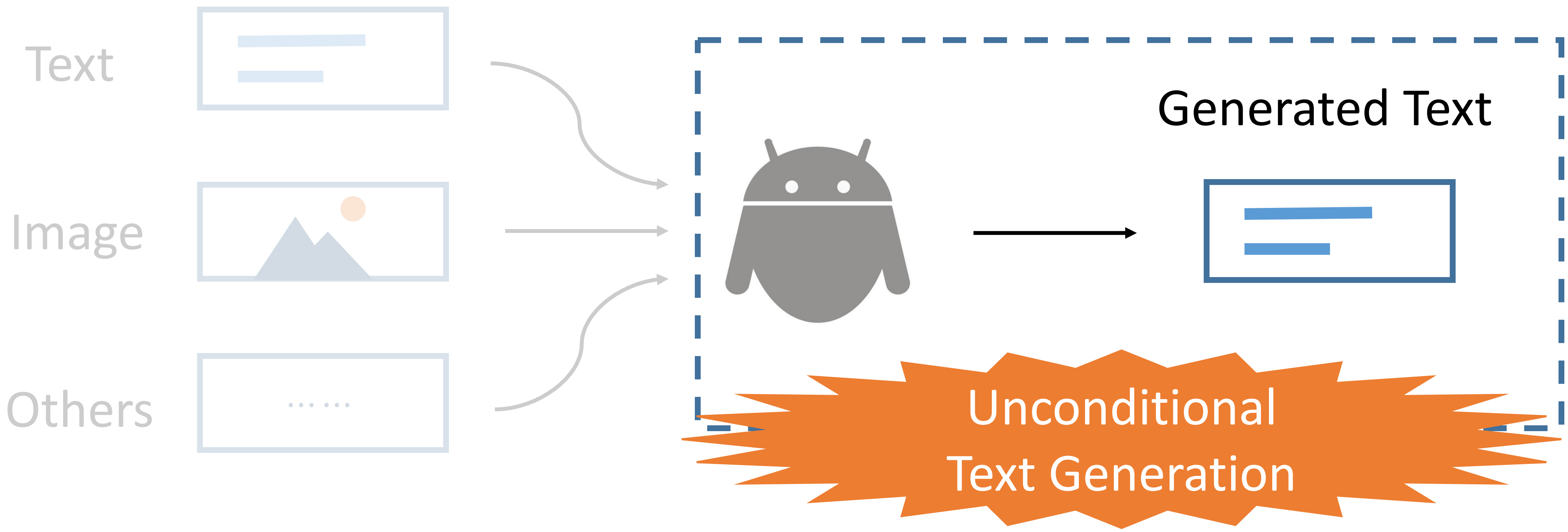
Background – Text Generation Tasks



Background – Text Generation Tasks

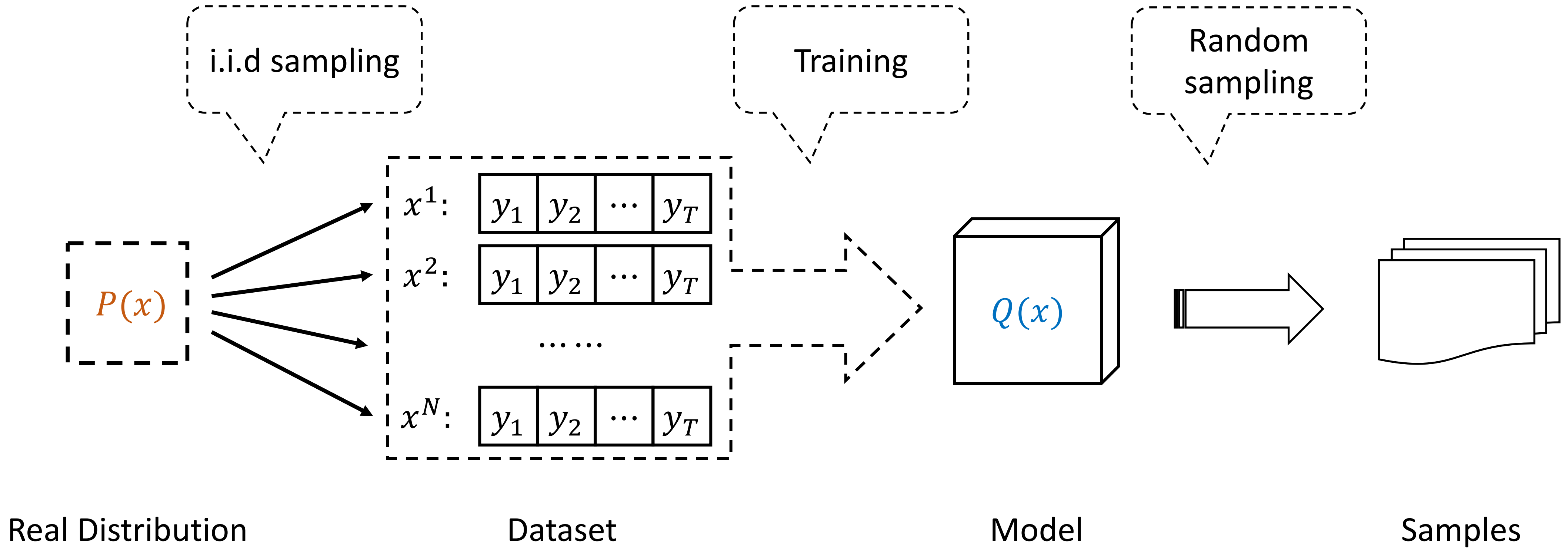


Background – Text Generation Tasks



Task – Unconditional Text Generation

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



Task – Unconditional Text Generation

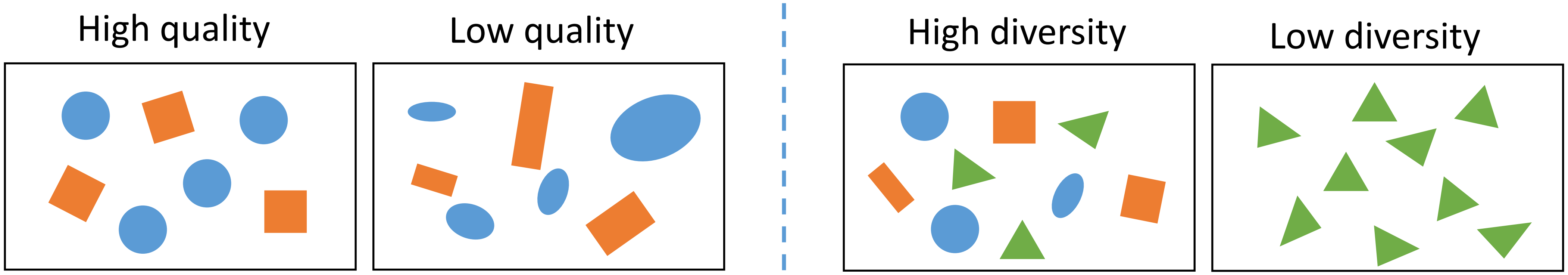
- Evaluation Metrics

  Quality: Generated text should contain less grammatical and logical errors.

  > Primary! We do not accept samples with errors.

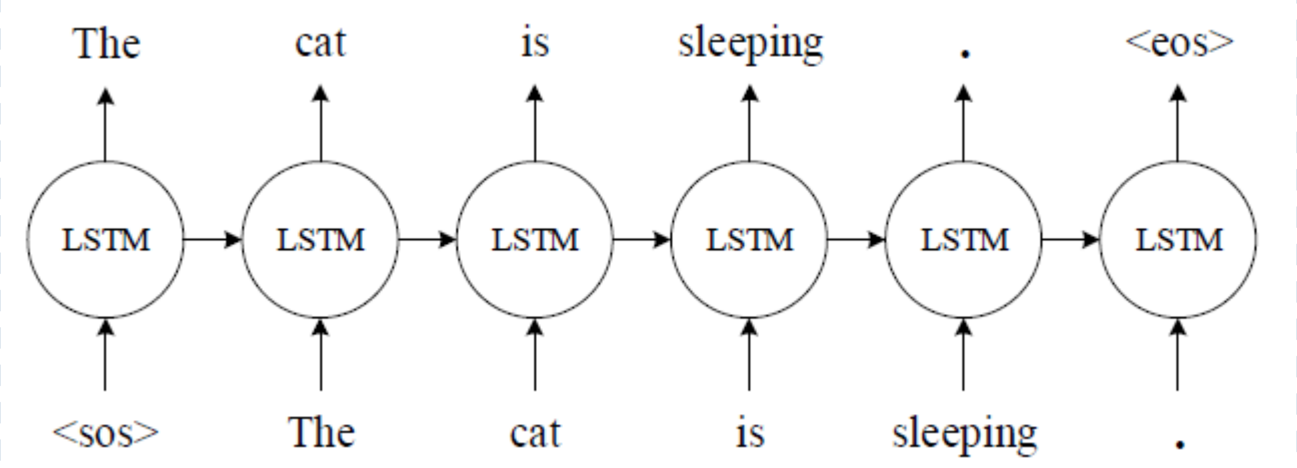
  Diversity: Generated text should not share similar words and structures.

  > Secondary. Similar samples are acceptable to some extent.



Baseline – RNN-based Language Model

- Use Recurrent Neural Networks (RNNs) to model sequential data



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)

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

Precise Distribution Recovery

Baseline – RNN-based Language Model

- RNNLM achieves only 49% Turing Test pass rate on MSCOCO dataset.
- Precise distribution recovery is sensitive to noises and rare patterns in training data.

Some common errors in MSCOCO dataset

A man is talking a picture of himself in the mirror.
A cat laying on top of a table trying to sleep.
A cat staring out a window at snow covered tree's.

Influence of noises should be minimized

How to neglect the impact of bad training data ?

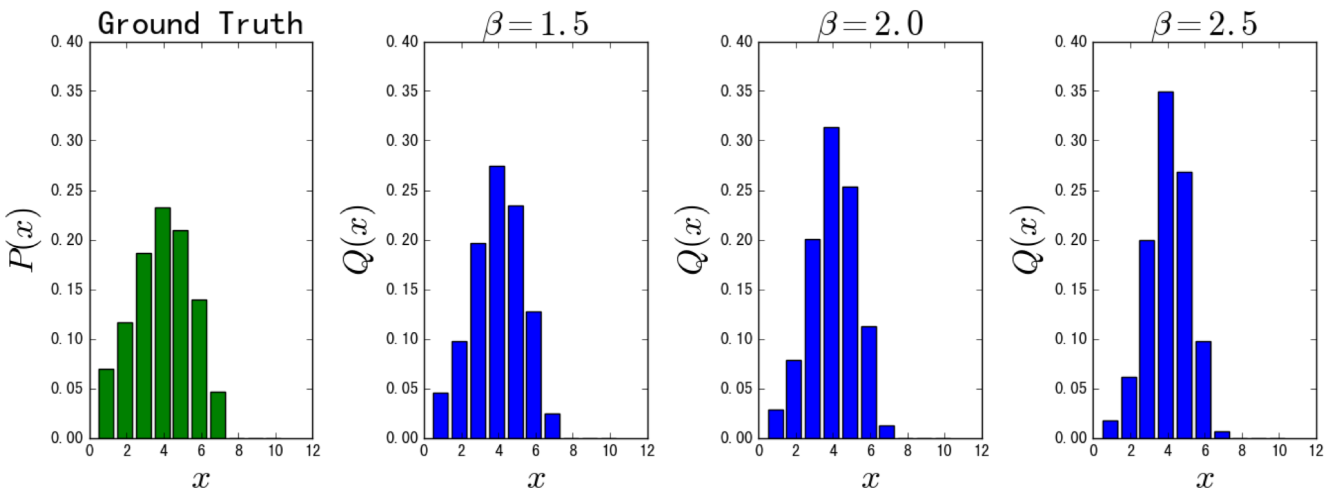
Use different strategy for different data !

Method – Differentiated Distribution Recovery

Assumption: Samples with lower probability under real distribution $P(x)$ are more likely to be bad samples.

Idea: Instead of making the model $Q(x)$ to precisely recover $P(x)$, we encourage samples with high probability and discourage samples with low probability.

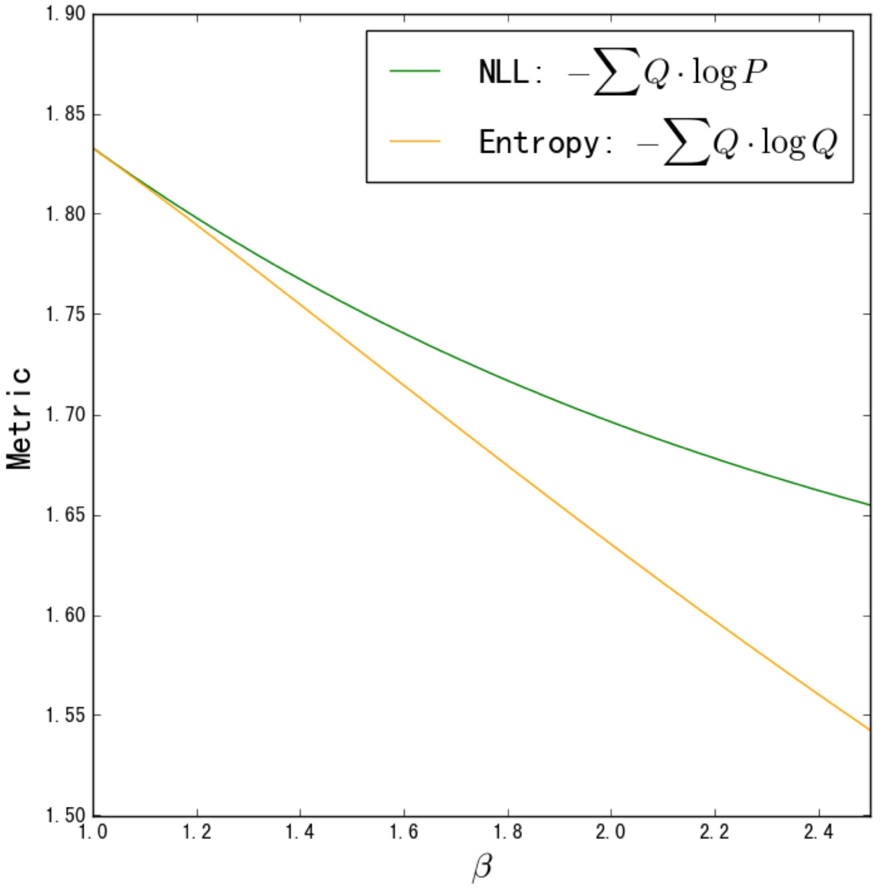
Differentiated Distribution Recovery (DDR):



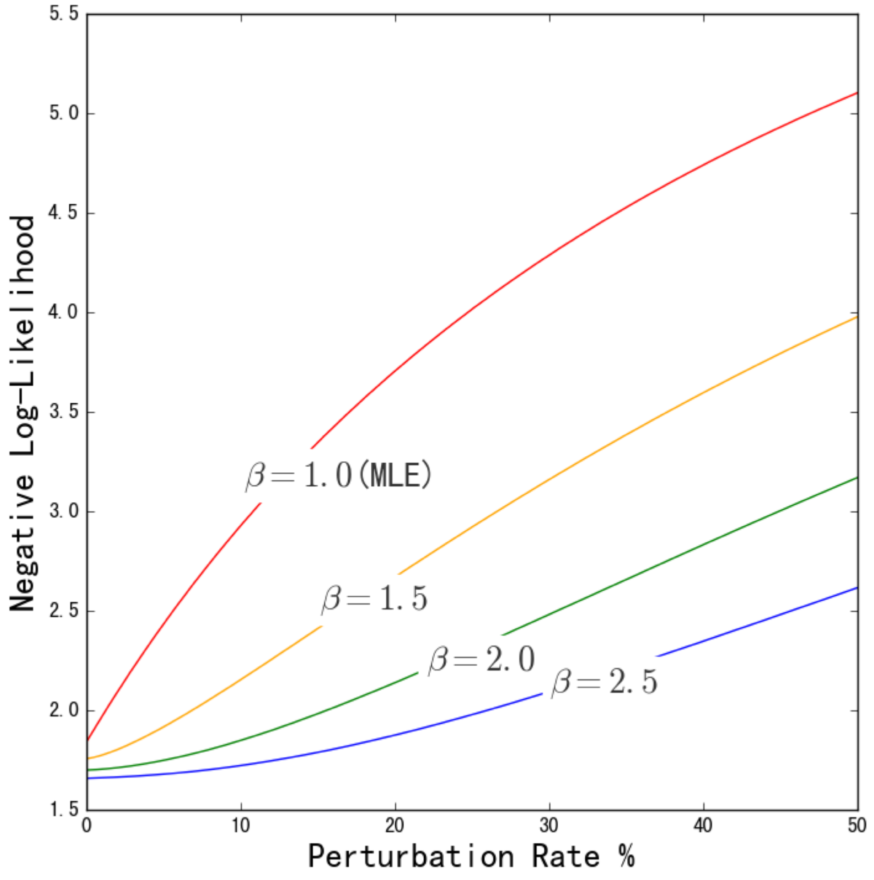
$$Q^*(x) \propto P(x)^\beta$$
$$\beta > 1$$

Method – Differentiated Distribution Recovery

Larger β leads to higher quality but lower diversity



Larger β is more robust to noises



How can we achieve Differentiated Distribution Recovery in practice?

$$\boxed{Q^*(x) \propto P(x)^{\beta}} \Rightarrow \boxed{Q^*(x) = \frac{P(x)^{\beta}}{\sum_x P(x)^{\beta}}}$$

Method – Implementation of DDR

Theorem

Let P and Q be two discrete distributions. With an objective defined as

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

$$f(Q(x);\alpha) = \alpha \cdot Q(x)^{\frac{1}{\alpha}} - \alpha, \qquad \alpha > 1,$$

The optimal Q with respect to the objective can be written as:

$$Q^*(x) = \frac{P(x)^{\beta}}{\sum_x P(x)^{\beta}}, \qquad \beta = \frac{\alpha}{\alpha - 1}$$

Method – Implementation of DDR

Proof

This is a constrained optimization problem:

$$\max_{Q} \mathbb{E}_{x\sim P}\, f[Q(x)],$$
$$f(Q(x);\alpha) = \alpha \cdot Q(x)^{\frac{1}{\alpha}} - \alpha$$
$$s.t. \sum_x Q(x) = 1\,, 0 \le Q(x) \le 1$$

The Lagrange function is:

$$L(Q(x),\lambda,\gamma,\eta) =$$
$$\sum_x P(x) \cdot f[Q(x)] + \lambda[1 - \sum_x Q(x)] - \gamma Q(x) + \eta[Q(x) - 1]$$

Let the first derivative be zero:

$$P(x) \cdot f'[Q^*(x)] = constant,$$
$$\sum_x Q^*(x) = 1,$$
$$0 \le Q^*(x) \le 1$$

We get the optimal $Q(x)$:

$$Q^*(x) = \frac{P(x)^\beta}{\sum_x P(x)^\beta}, \qquad \beta = \frac{\alpha}{\alpha - 1}$$

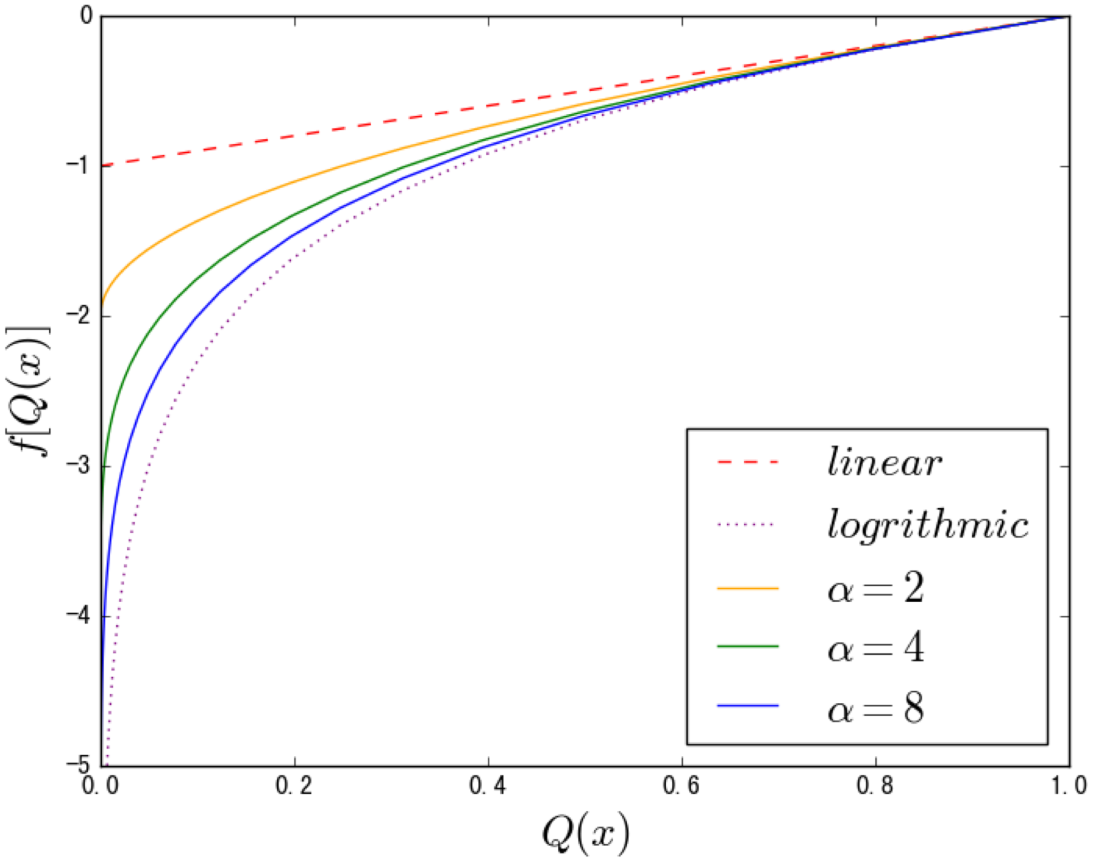
Method – Implementation of DDR

- DDR can be realized by using the following objective function

$$\max_{Q} \mathbb{E}_{x\sim P}\, f[Q(x)],$$
$$f(Q(x);\alpha) = \alpha \cdot Q(x)^{\frac{1}{\alpha}} - \alpha,$$
$$\alpha = \frac{\beta}{\beta - 1} > 1$$

Function f changes from linear to logarithm as α grows

$$f(Q(x);1) = Q(x) - 1,$$
$$\lim_{\alpha\to\infty} f(Q(x);\alpha) = \ln Q(x)$$



Method – Implementation of DDR

Loss function

$$\mathcal{L}(\mathcal{D};\alpha) = -\frac{1}{N}\sum_{i=1}^{N} \alpha \cdot Q\left(Y_{1:T}^{i}\right)^{\frac{1}{\alpha}}$$

$$= -\frac{\alpha}{N}\sum_{i=1}^{N}\prod_{t=1}^{T} Q\left(y_t^i \middle| Y_{1:t-1}^i\right)^{\frac{1}{\alpha}}$$

$$= -\frac{\alpha}{N}\sum_{i=1}^{N}\exp\left\{\frac{1}{\alpha}\sum_{t=1}^{T}\log Q\left(y_t^i \middle| Y_{1:t-1}^i\right)\right\}$$

Easy to implement

No extra training overhead

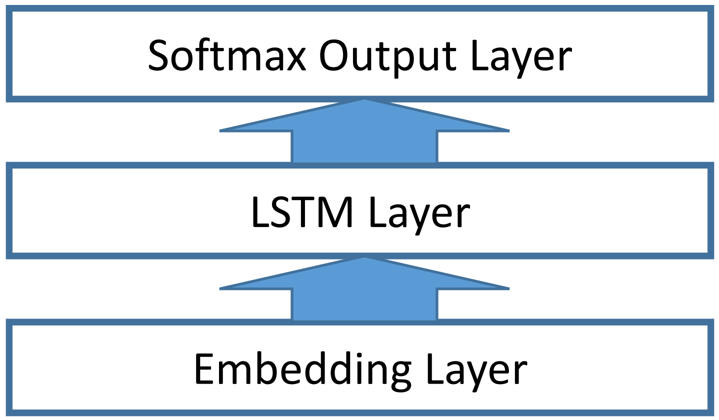
Comparison with Related works

RNNLM (Baseline)	GANs & RL methods	DDR (Ours)
• Low generation quality • Fast training speed • Hard to control the tradeoff between quality and diversity	• Improved generation quality • Low training speed • Hard to control the tradeoff between quality and diversity	• Improved generation quality • Fast training speed • Easy to control the tradeoff between quality and diversity

Experiments – Settings

- Datasets
  - Synthetic data
  - MSCOCO Image Caption dataset
  - EMNLP2017 WMT News dataset
- Baselines
  - RNNLM
  - SeqGAN (Yu et al. 2017)
  - LeakGAN (Guo et al. 2017)

General model architecture

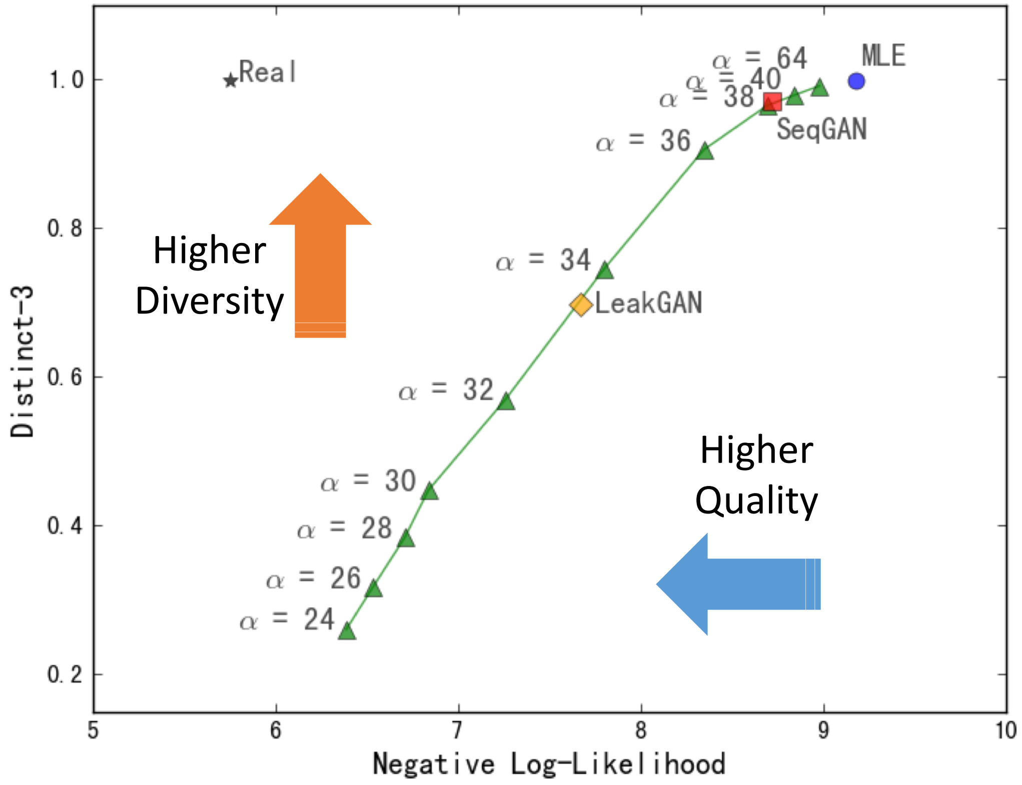


Experiments – Synthetic Data

- Use an oracle model (Yu et al. 2017) to generate data.
- Attributes
  - #Training data: 10000
  - Sequence length: 20
  - Vocabulary size: 5000
- Evaluation metrics:

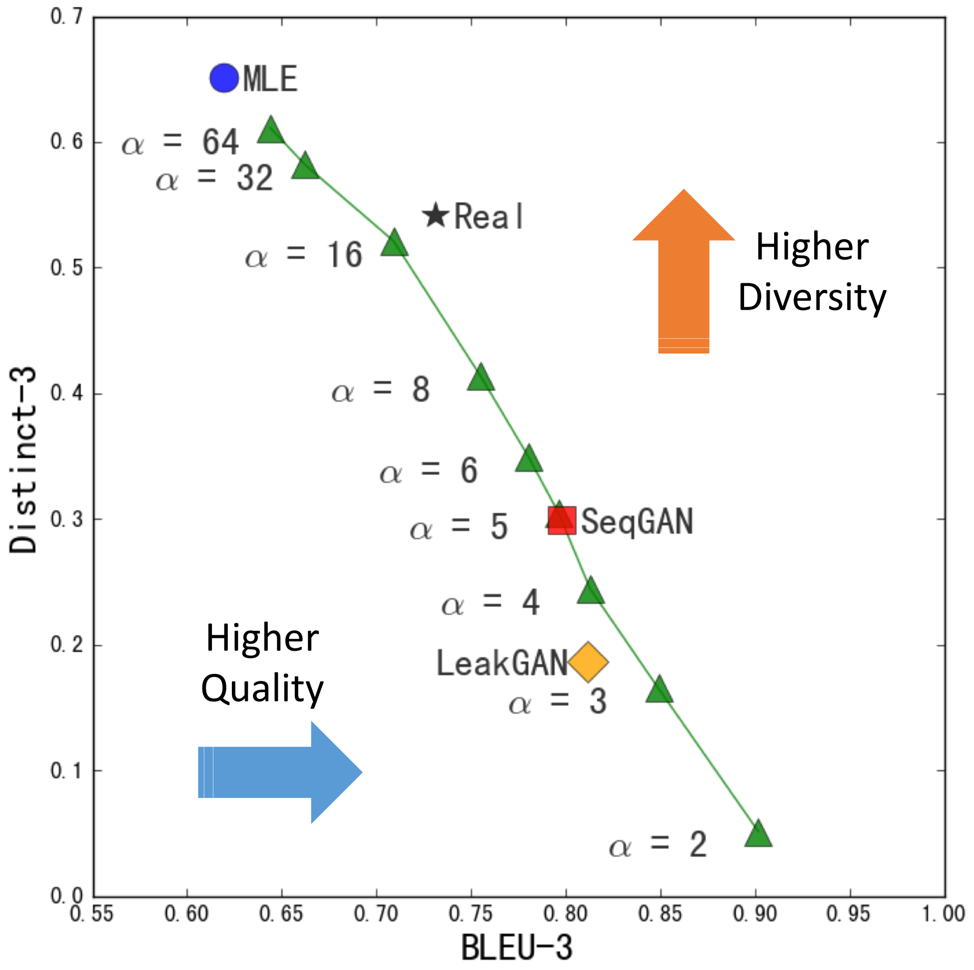
$$NLL = -\mathbb{E}_{Y_{1:T} \sim Q}\left[\sum_{t=1}^{T} \log G_{oracle}(y_t | Y_{1:t-1})\right]$$

$$Distinct\_n = \frac{\#\ \text{Unique}\ n\_grams}{\#\ Total\ n\_grams}$$

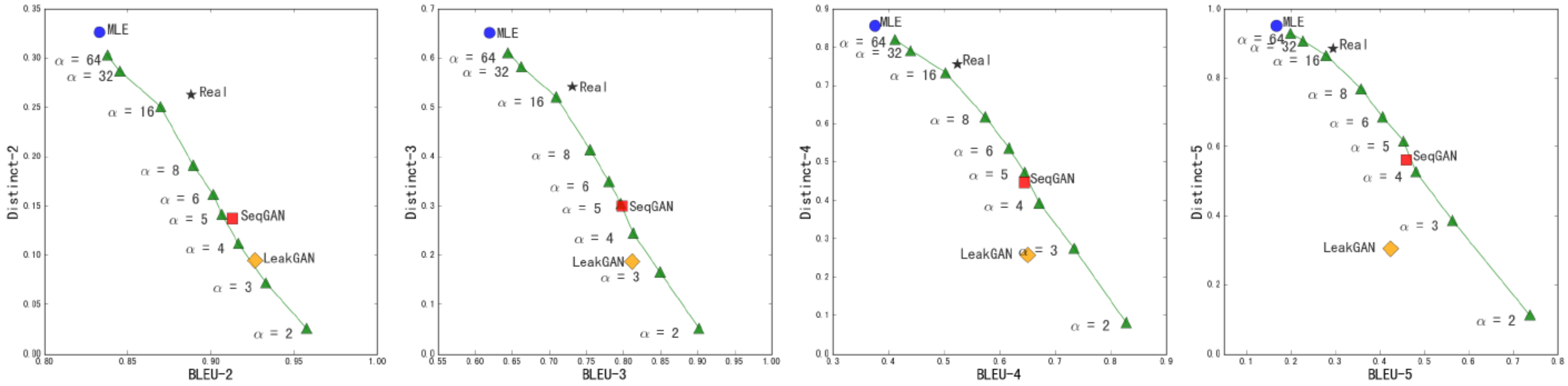


Experiments – Real World Datasets

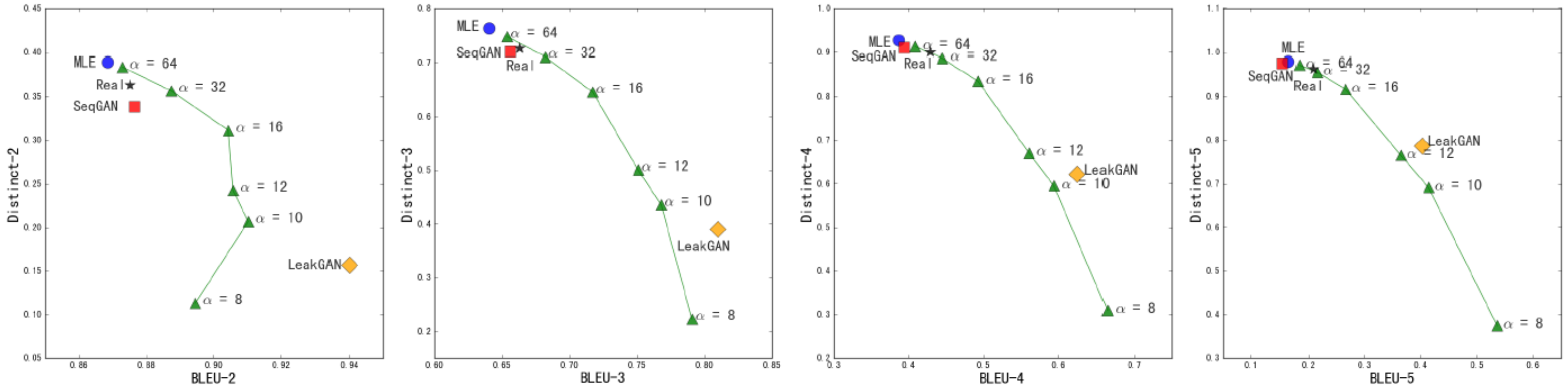
- MSCOCO Dataset
  - #Training/Test data: 80000/5000
  - Sequence length: 32
  - Vocabulary size: 4840
- WMT Dataset
  - #Training/Test data: 200000/10000
  - Sequence length: 50
  - Vocabulary size: 6655
- Evaluation Metrics:
  - BLEU-(2-5)
  - Distinct-(2-5)



Experiments – Real World Datasets



(a) MSCOCO dataset



(b) WMT dataset

Experiments – Generated Samples (MSCOCO)

Method	Generated samples
Real data	A cat stuck in a car with a slightly opened window . Bicycles , cars and a trash can in a garage . A lady talking a self portrait in a fancy bathroom . A man standing in a white kitchen with his arms folded .
MLE	Two young children playing a video game on the Nintendo Wii . Two pancakes on a white paper plate with sauce on the plate . A suitcase with vanilla and yellow markings on top of it . Birds flying on a stone bench next to the tree .
SeqGAN	A group of people standing on top of a snow covered mountain . Two people standing next to each other in the dirt . A brown horse standing next to a white fence on the beach . A cow standing in a grassy area near a body of water .
LeakGAN	A bicycle is locked to a fence by a truck . The interior of a bathroom with a long mirror and partially tiled walls . A small bathroom has toilet , medicine cabinet , and small sink . A woman riding a bicycle down a street in front of shops .

Method	Generated samples
DDR($\alpha=64$)	A woman wearing tennis gear holding a racket and her racquet . A dog sitting on a chair in front of a birthday cake . A bald man lays on a bed in the yellow floral pot . A girl is flying a kite in the sky into the airport .
DDR($\alpha=8$)	A man is sitting in a chair with a white cat . A guy is jumping in the air with a skateboard . A tall giraffe standing on top of a lush green field . Two women pose in front of a very tall building .
DDR($\alpha=2$)	A couple of young men playing a game of baseball . A couple of zebra standing on top of a lush green field . A red stop sign sitting on the side of a road . A man hitting a tennis ball with a tennis racquet .

Experiments – Human Turing Test

- Sample 50 sentences from each model, and mix all sentences together.
- 10 Ph.D students are invited to give scores individually for all samples.
- A sample get +1 score if one think it is possibly written by a human, otherwise get +0 score.

Method	Turing Test Score
Ground Truth	0.772
MLE	0.490
SeqGAN	0.706
LeakGAN	0.758
DDR($\alpha=64$)	0.586
DDR($\alpha=8$)	0.692
DDR($\alpha=2$)	**0.932**

Generation quality is significantly improved with DDR

Experiments – Robustness Test

- Add 10% random noises to MSCOCO dataset.
- See how many bad sentences are generated by models.
- A sample is regarded as bad if its BLEU-2 score is lower than 0.001

Method	Bad Samples %
MLE	8.0
SeqGAN	2.2
LeakGAN	**0.0**
DDR($\alpha=64$)	0.4
DDR($\alpha=32$)	0.1
DDR($\alpha=16$)	**0.0**
DDR($\alpha=8$)	**0.0**
DDR($\alpha=2$)	**0.0**

Model become more robust with DDR

Conclusion

- **Differentiated Distribution Recovery (DDR)** is an efficient way to promote generation quality for unconditional neural text generation.
- **DDR** provides an flexible control between generation quality and diversity through a hyper-parameter α.
- **DDR** makes the model more robust against noises in training data.

Thank you! Q&A

Jianing Li lijianing@software.ict.ac.cn