Deep Adversarial Learning for NLP

William Wang
UC SANTA BARBARA

Sameer Singh
UC Irvine

Slides: **http://tiny.cc/adversarial**

With contributions from Jiwei Li.

Agenda

- Introduction, Background, and GANs (William, 90 mins)
- Adversarial Examples and Rules (Sameer, 75 mins)
- Conclusion and Question Answering (Sameer and William, 15 mins)

Slides: **http://tiny.cc/adversarial**

Outline

- Background of the Tutorial
- Introduction: Adversarial Learning in NLP
- Adversarial Generation
- A Case Study of GANs in Dialogue Systems

Rise of Adversarial Learning in NLP

- Through a simple ACL anthology search, we found that in 2018, there were 20+ times more papers mentioning “adversarial”, comparing to 2016.
- Meanwhile, the growth of all accepted papers is 1.39 times during this period.
- But if you went to CVPR 2018 in Salt Lake City, there were more than 100 papers on adversarial learning (approximately 1/3 of all adv. learning papers in NLP).

Questions I’d like to Discuss

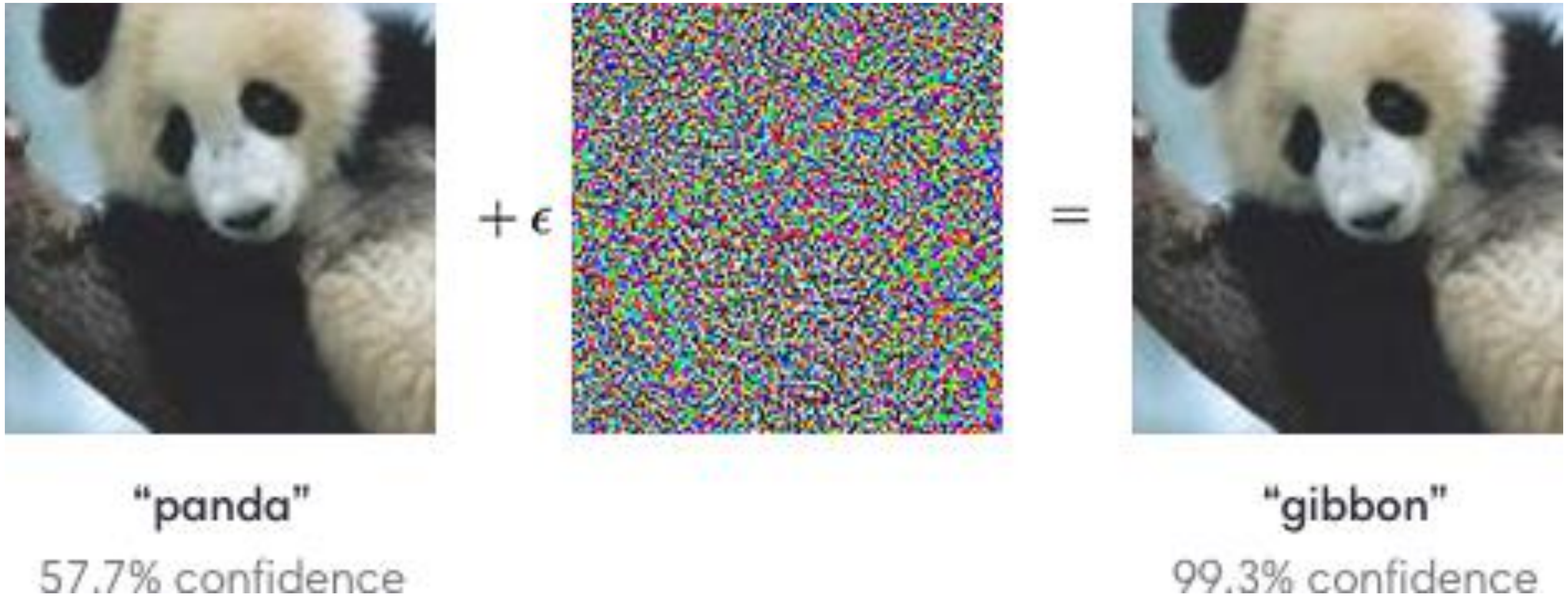
- What are the subareas of deep adversarial learning in NLP?
- How do we understand adversarial learning?
- What are some success stories?
- What are the pitfalls that we need to avoid?

Opportunities in Adversarial Learning

- Adversarial learning is an interdisciplinary research area, and it is closely related to, but limited to the following fields of study:
  - Machine Learning
  - Computer Vision
  - Natural Language Processing
  - Computer Security
  - Game Theory
  - Economics

Adversarial Attack in ML, Vision, & Security

- Goodfellow et al., (2015)



Physical-World Adversarial Attack / Examples (Eykholt et al., CVPR 2018)

Success of Adversarial Learning

CycleGAN (Zhu et al., 2017)

Failure Cases

CycleGAN (Zhu et al., 2017)

Success of Adversarial Learning

GauGAN (Park et al., 2019)

Deep Adversarial Learning in NLP

- There were some successes of GANs in NLP, but not so much comparing to Vision.
- The scope of Deep Adversarial Learning in NLP includes:
  - Adversarial Examples, Attacks, and Rules
  - Adversarial Training (w. Noise)
  - Adversarial Generation
  - Various other usages in ranking, denoising, & domain adaptation.

UCSB

Outline

- Background of the Tutorial
- Introduction: Adversarial Learning in NLP
- Adversarial Generation
- A Case Study of GANs in Dialogue Systems

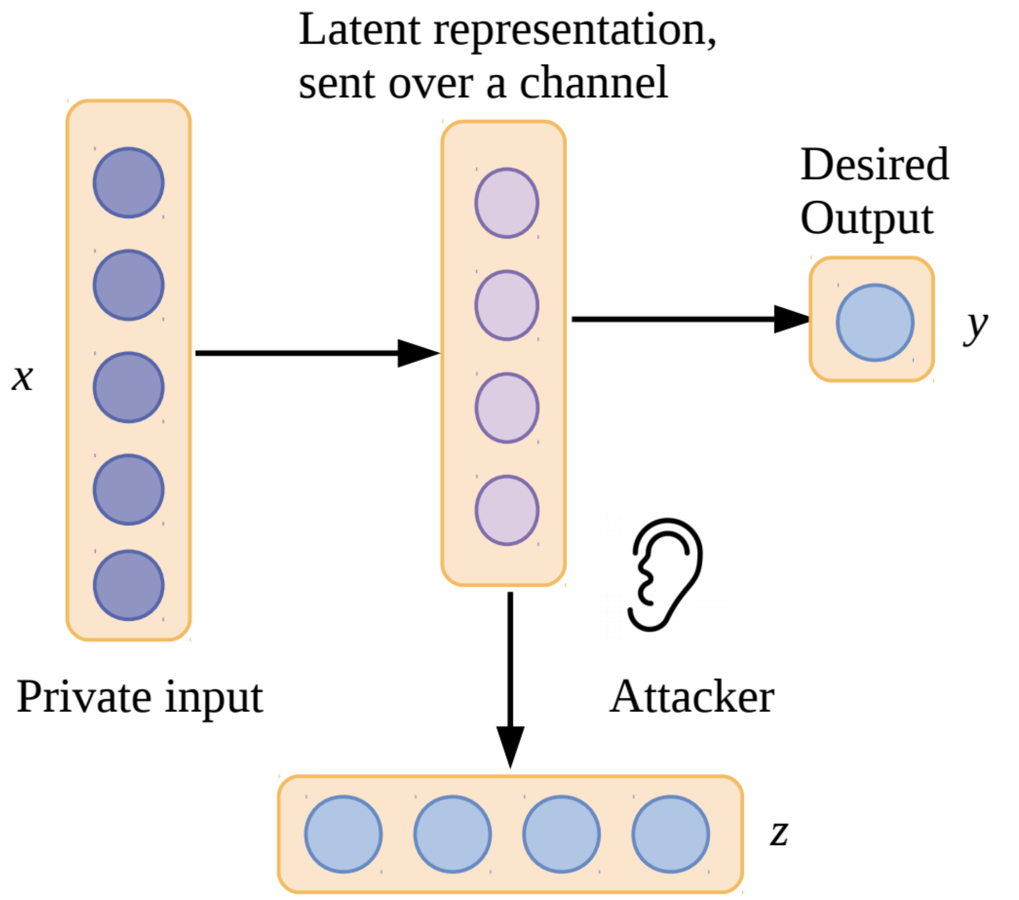
Adversarial Examples

- One of the more popular areas of adversarial learning in NLP.
- E.g., Alzantot et al., EMNLP 2018

Original Text Prediction: **Entailment** (Confidence = 86%)
Premise: *A runner wearing purple strives for the finish line.* **Hypothesis:** *A **runner** wants to head for the finish line.*
Adversarial Text Prediction: **Contradiction** (Confidence = 43%)
Premise: *A runner wearing purple strives for the finish line.* **Hypothesis:** *A **racer** wants to head for the finish line.*

Adversarial Attacks (Coavoux et al., EMNLP 2018)

The main classifier predicts a label y from a text x, the attacker tries to recover some private information z contained in x from the latent representation used by the main classifier.



Adversarial Training

- Main idea:
  - Adding noise, randomness, or adversarial loss in optimization.

- Goal: make the trained model more robust.

Adversarial Training: A Simple Example

- Adversarial Training for Relation Extraction
  - Wu, Bamman, Russell (EMNLP 2017).

- Task: Relation Classification.

- Interpretation: Regularization in the Feature Space.

Adversarial Training for Relation Extraction

$$
\begin{aligned}
L_{\text{adv}}(X;\theta) &= L(X + e_{\text{adv}};\theta), \text{ where} \\
e_{\text{adv}} &= \underset{\|e\| \leq \epsilon}{\arg\max}\, L(X + e;\hat{\theta})
\end{aligned}
$$

$$
e_{\text{adv}} = \epsilon g / \|g\|, \text{ where } g = \nabla_V L(X;\hat{\theta}).
$$

Wu, Bamman, Russell (EMNLP 2017).

Adversarial Training for Relation Extraction

Recall	0.1	0.2	0.3	0.4	AUC
PCNN	0.667	0.572	0.476	0.392	0.329
PCNN-Adv	0.717	0.589	0.511	0.407	0.356
RNN	0.668	0.586	0.524	0.442	0.351
RNN-Adv	**0.728**	**0.646**	**0.553**	**0.481**	**0.382**

Wu, Bamman, Russell (EMNLP 2017).

Outline

- Background of the Tutorial
- Introduction: Adversarial Learning in NLP
- Adversarial Generation
- A Case Study of GANs in Dialogue Systems

GANs (Goodfellow et al., 2014)

- Two competing neural networks: generator & discriminator

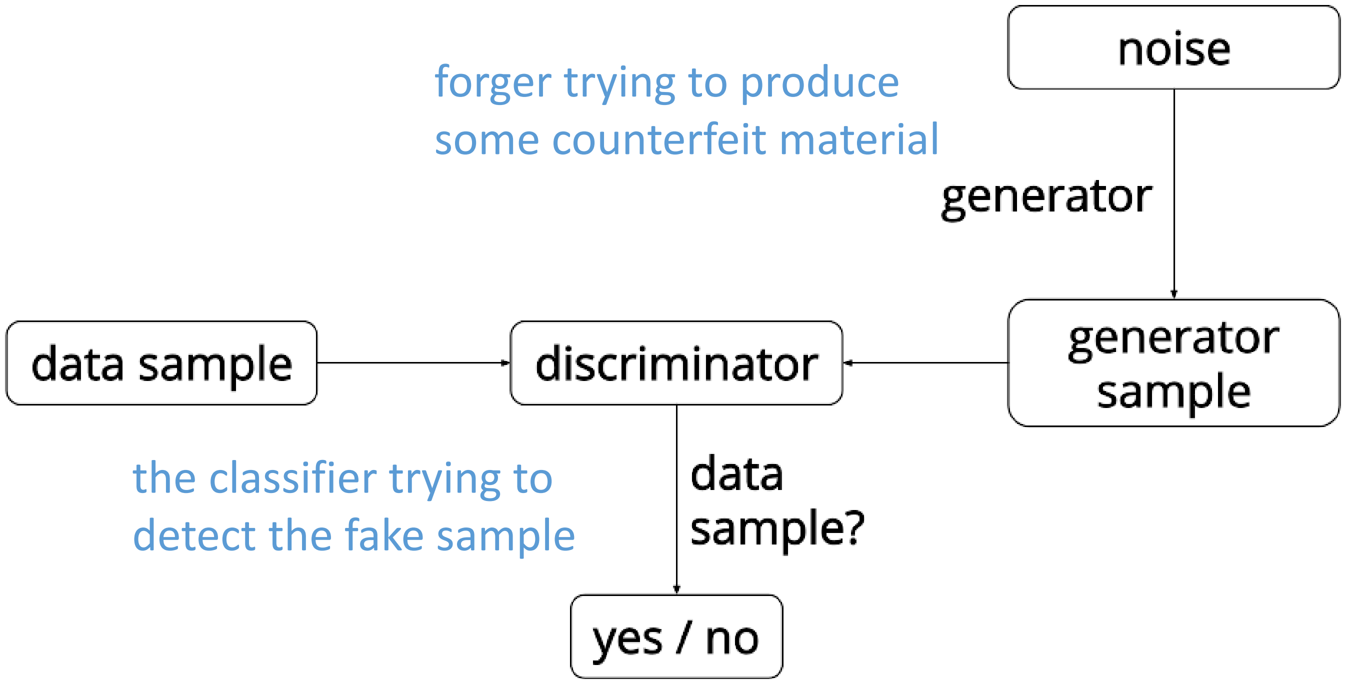


Image: https://ishmaelbelghazi.github.io/ALI/

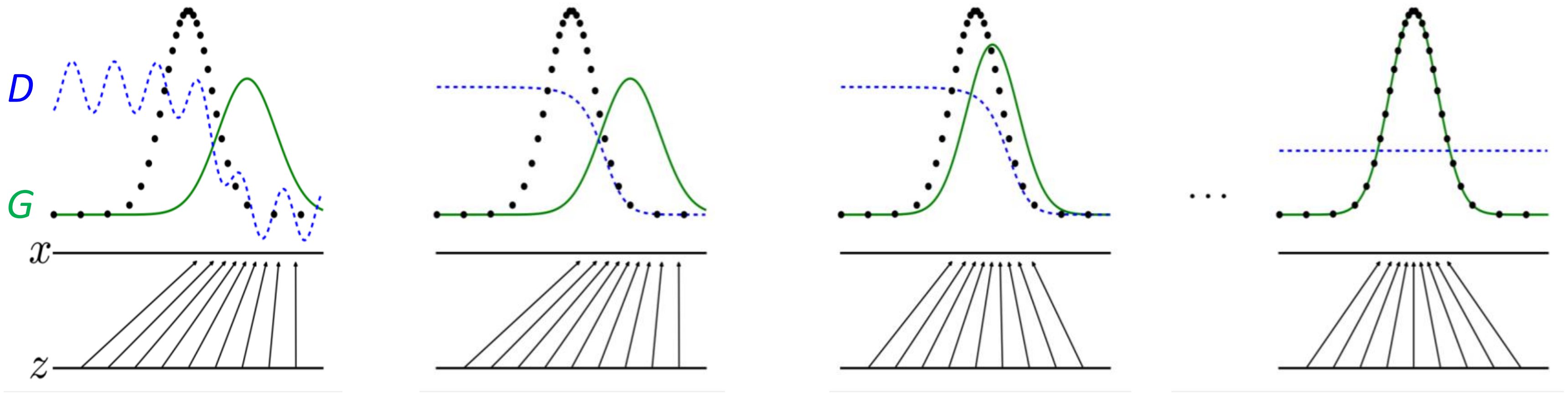
GAN Objective

$$\min_G \max_D V(D, G)$$

$$= \mathbb{E}_{q(\mathbf{x})}[\log(D(\mathbf{x}))] + \mathbb{E}_{p(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$

$$= \int q(\mathbf{x}) \log(D(\mathbf{x})) d\mathbf{x} + \iint p(\mathbf{z}) p(\mathbf{x} \mid \mathbf{z}) \log(1 - D(\mathbf{x})) d\mathbf{x} d\mathbf{z}$$

$D(x)$: the probability that x came from the data rather than generator



GAN Training Algorithm

Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator, k, is a hyperparameter. We used $k = 1$, the least expensive option, in our experiments.

for number of training iterations **do**

Discriminator

  for k steps **do**

  - Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
  - Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
  - Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right].$$

  end for

Generator

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

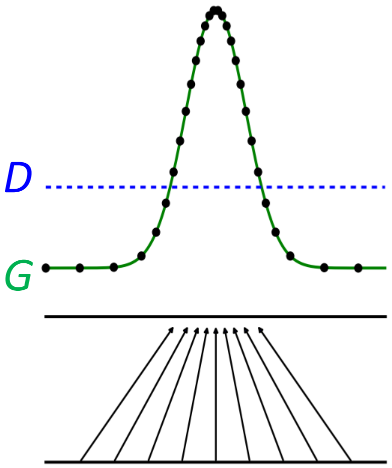
end for

The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.

GAN Equilibrium

- Global optimality
  - Discriminator

$$D^*(\mathbf{x}) = \frac{q(\mathbf{x})}{q(\mathbf{x}) + p(\mathbf{x})}$$

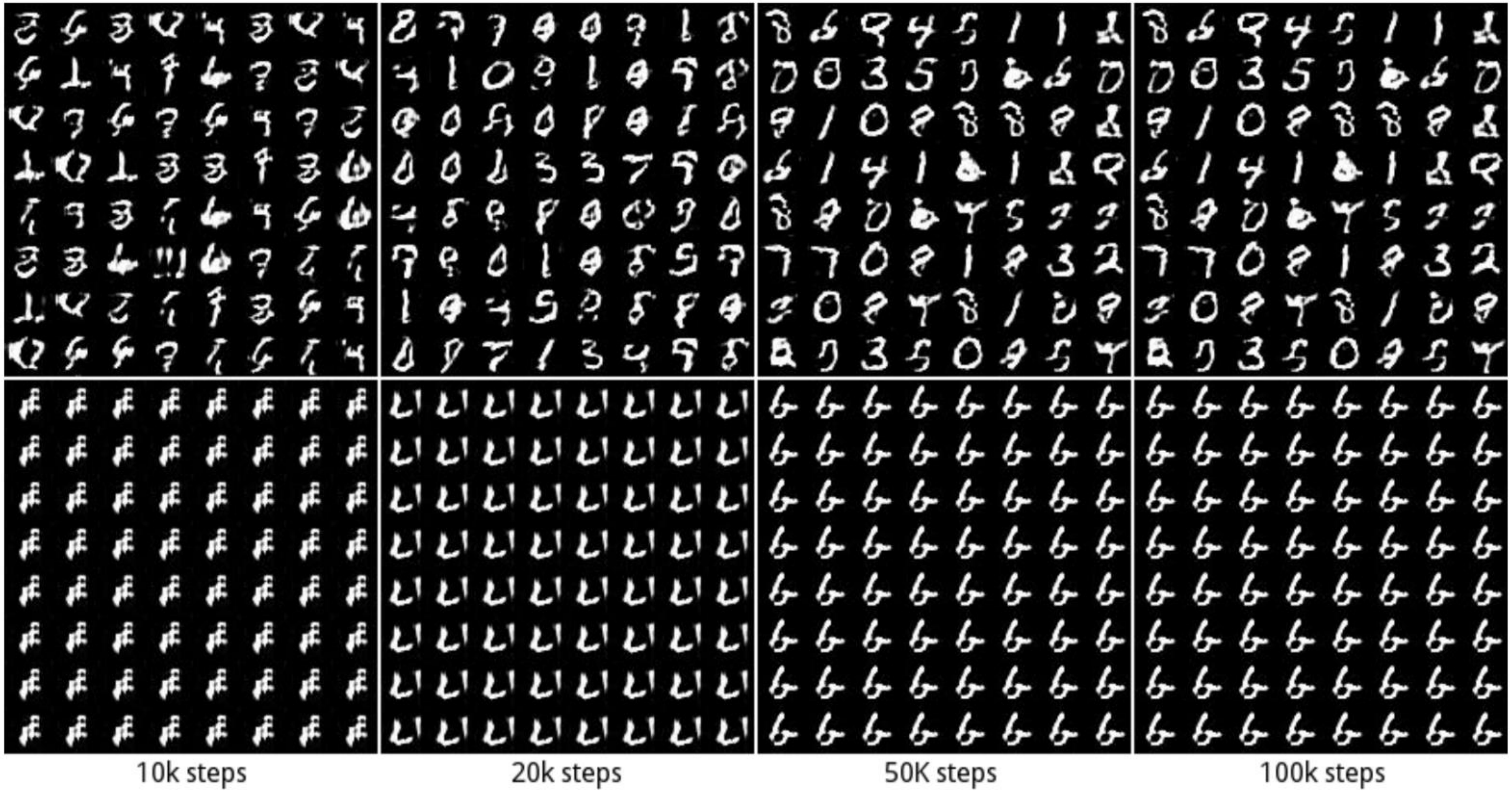


  - Generator

$$G^*(\mathbf{z}) \text{ s.t. } p(\mathbf{z}) = q(\mathbf{x})$$

Major Issues of GANs

- Mode Collapse (unable to produce diverse samples)



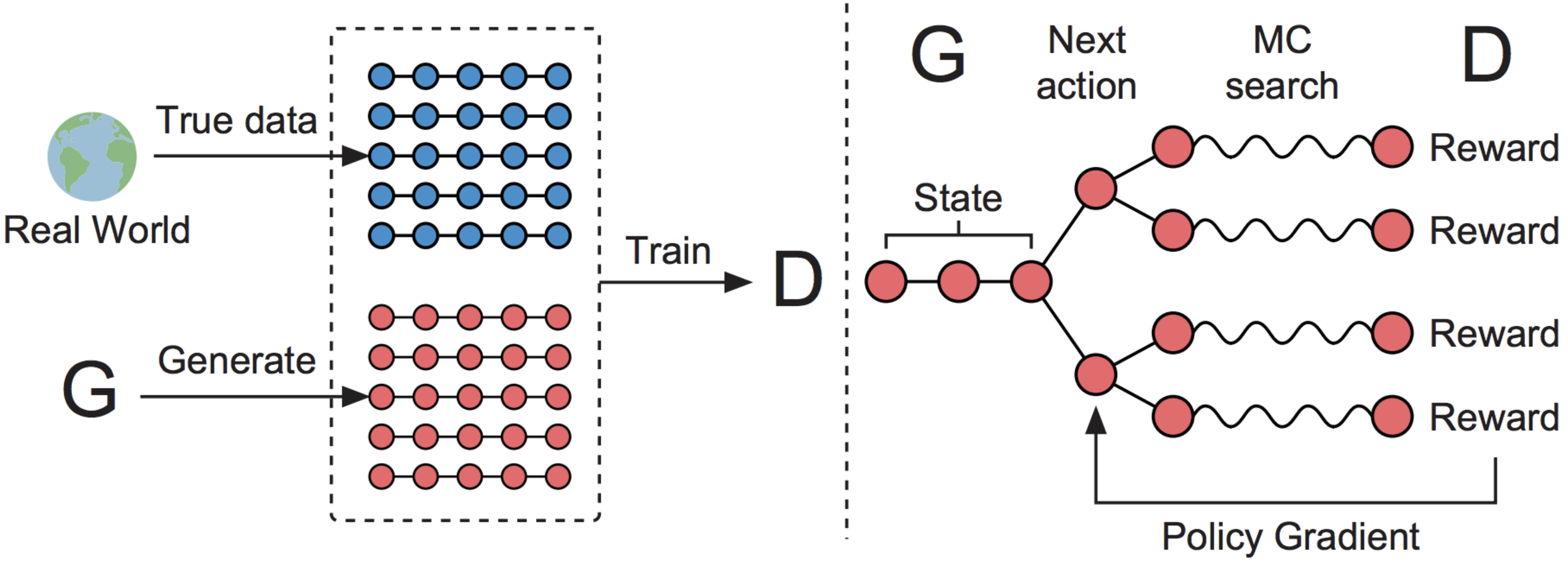
Major Issues of GANs in NLP

- Often you need to pre-train the generator and discriminator w. MLE
  - But how much?
- Unstable Adversarial Training
  - We are dealing with two networks / learners / agents
  - Should we update them at the same rate?
- The discriminator might overpower the generator.
- With many possible combinations of model choice for generator and discriminator networks in NLP, it could be worse.

Major Issues of GANs in NLP

- GANs were originally designed for images
  - You cannot back-propagate through the generated X
- Image is continuous, but text is discrete (DR-GAN, Tran et al., CVPR 2017).

SeqGAN: policy gradient for generating sequences
(Yu et al., 2017)



Training Language GANs from Scratch

- New Google DeepMind arxiv paper (de Masson d'Autume et al., 2019)
    - Claims no MLE pre-trainings are needed.
    - Uses per time-stamp dense rewards.
    - Yet to be peer-reviewed and tested.

Why shouldn’t NLP give up on GAN?

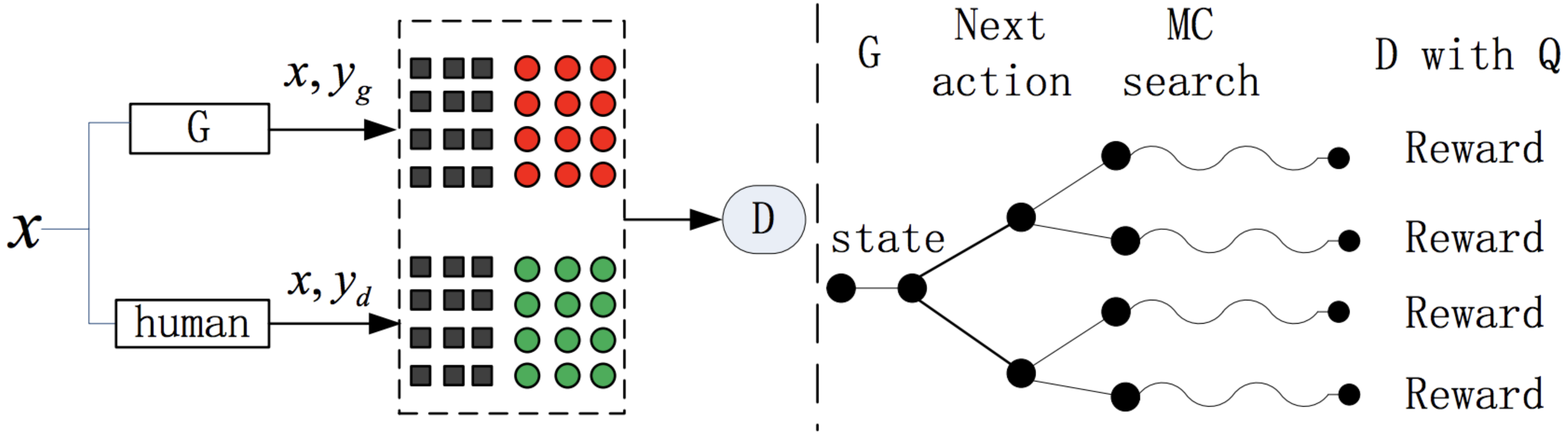
- It’s unsupervised learning.
- Many potential applications of GANs in NLP.
- The discriminator is often learning a metric.
- It can also be interpreted as self-supervised learning (especially with dense rewards).

Applications of Adversarial Learning in NLP

- Social Media (Wang et al., 2018a; Carton et al., 2018)
- Contrastive Estimation (Cai and Wang, 2018; Bose et al., 2018)
- Domain Adaptation (Kim et al., 2017; Alam et al., 2018; Zou et al., 2018; Chen and Cardie, 2018; Tran and Nguyen, 2018; Cao et al., 2018; Li et al., 2018b)
- Data Cleaning (Elazar and Goldberg, 2018; Shah et al., 2018; Ryu et al., 2018; Zellers et al., 2018)
- Information extraction (Qin et al., 2018; Hong et al., 2018; Wang et al., 2018b; Shi et al., 2018a; Bekoulis et al., 2018)
- Information retrieval (Li and Cheng, 2018)
- Another 18 papers on Adversarial Learning at NAACL 2019!

GANs for Machine Translation

- Yang et al., NAACL 2018
- Wu et al., ACML 2018



SentiGAN (Wang and Wan, IJCAI 2018)

Idea: use a mixture of generators and a multi-class discriminator.

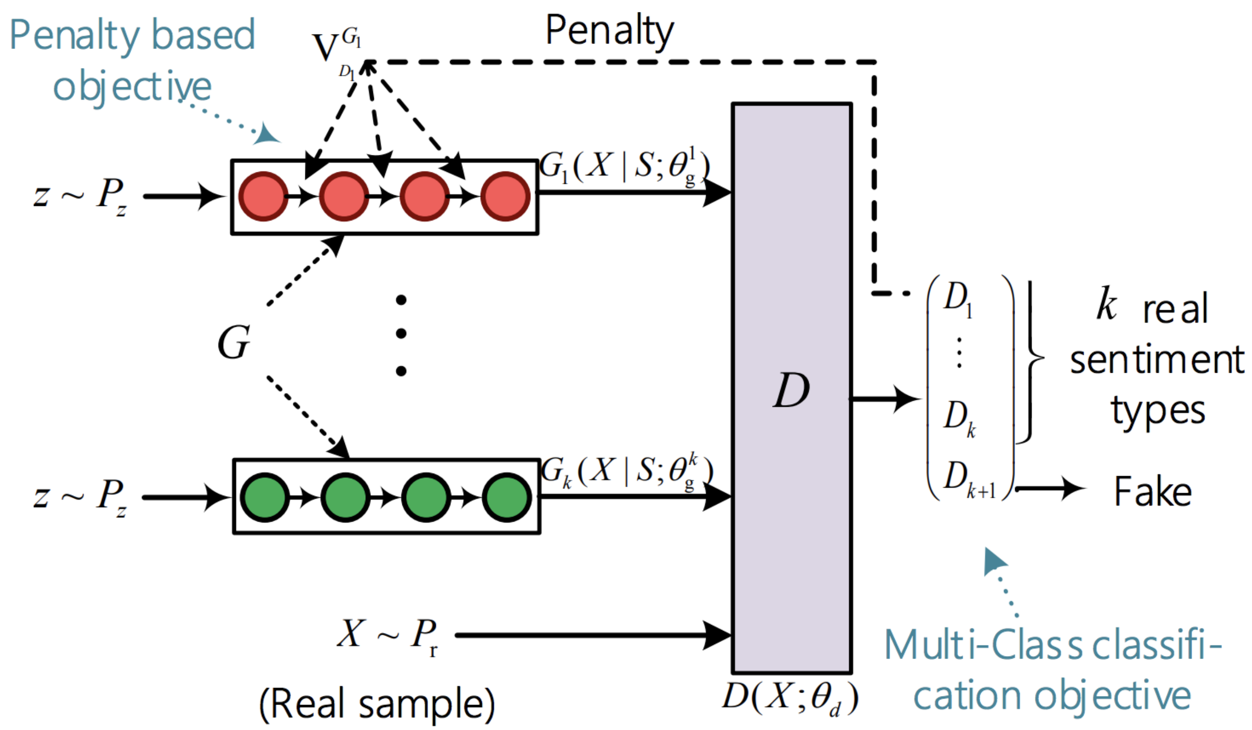
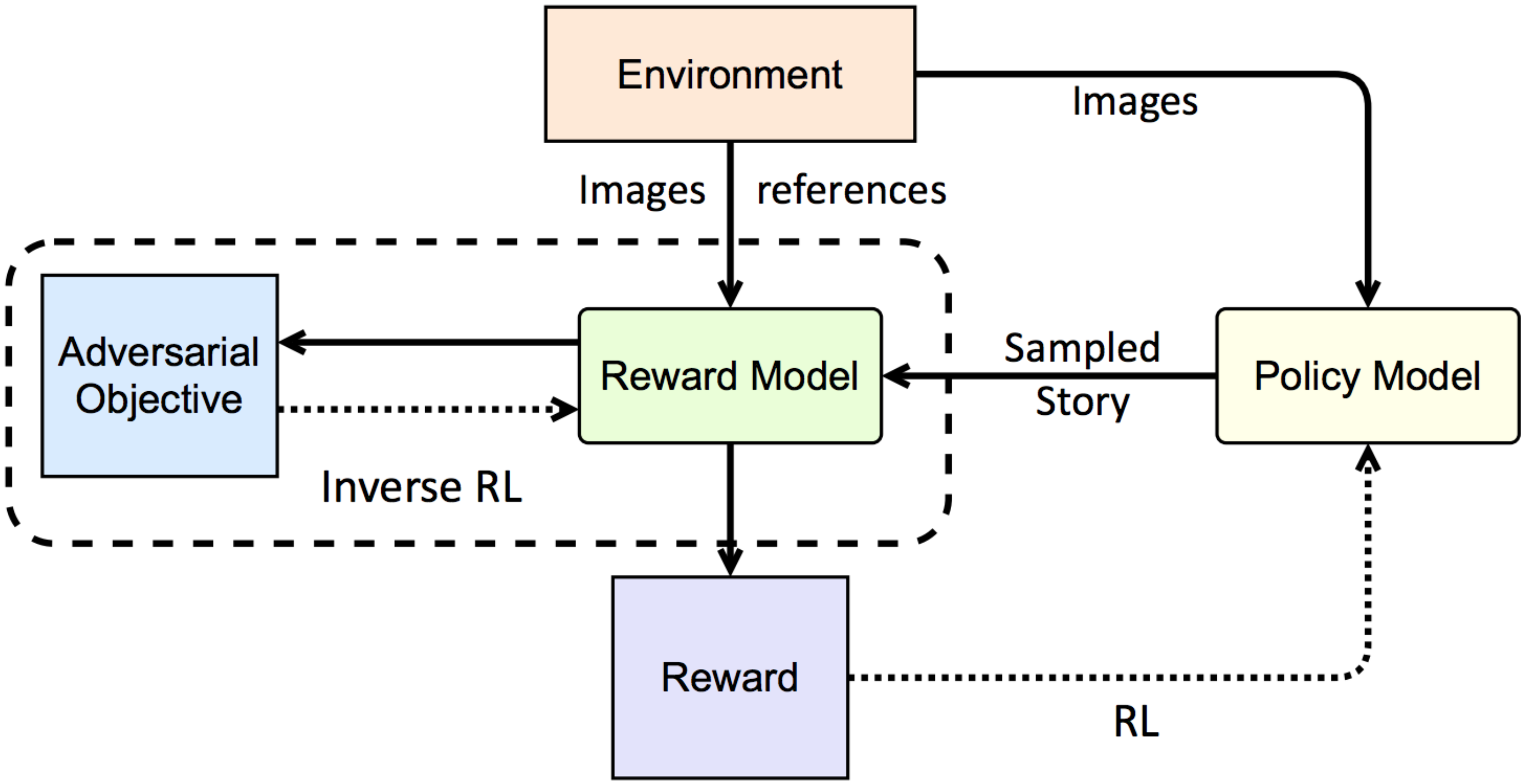


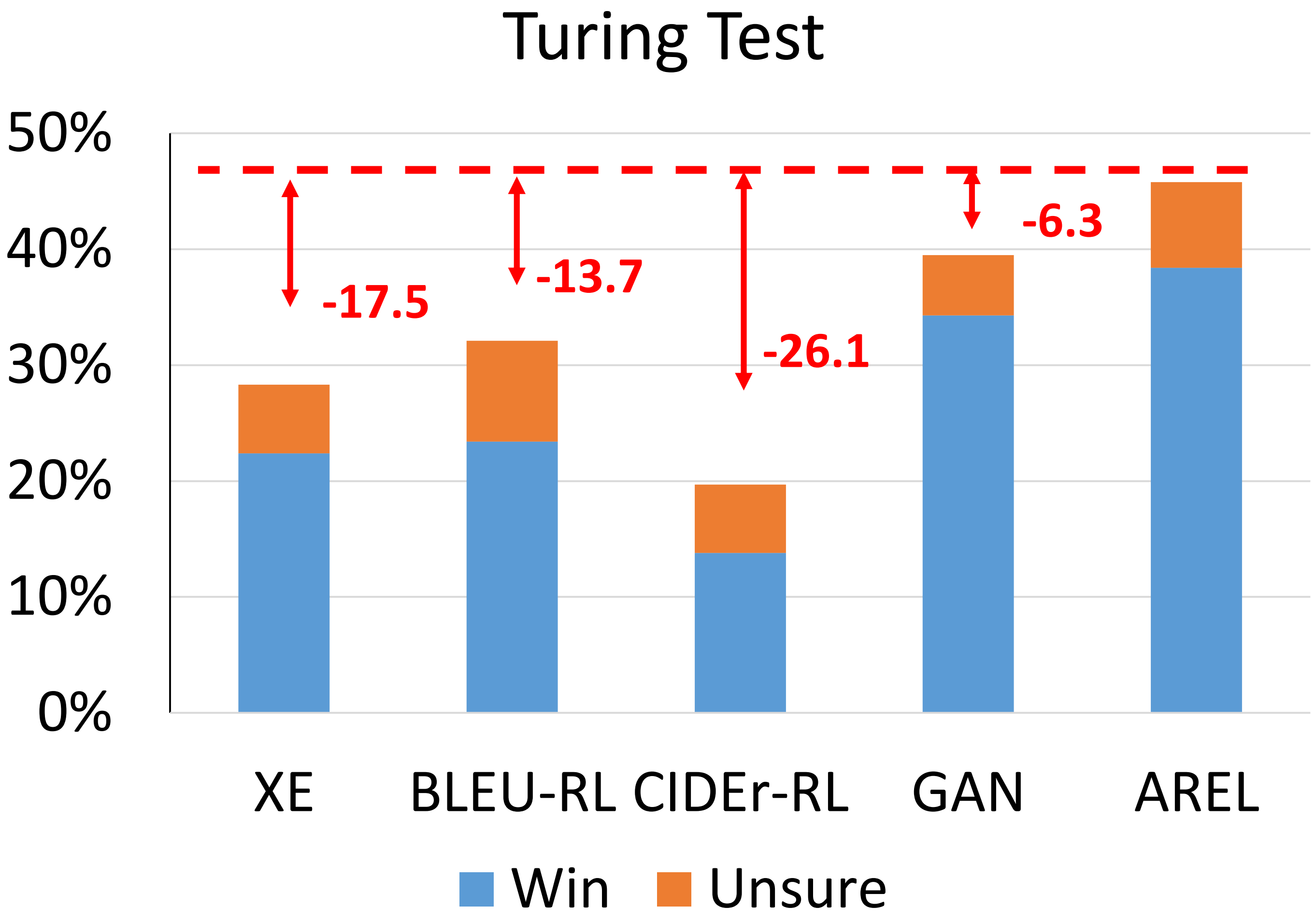
Figure 1: The framework of SentiGAN with k generators and one multi-class discriminator.

No Metrics Are Perfect: Adversarial Reward Learning (Wang, Chen et al., ACL 2018)

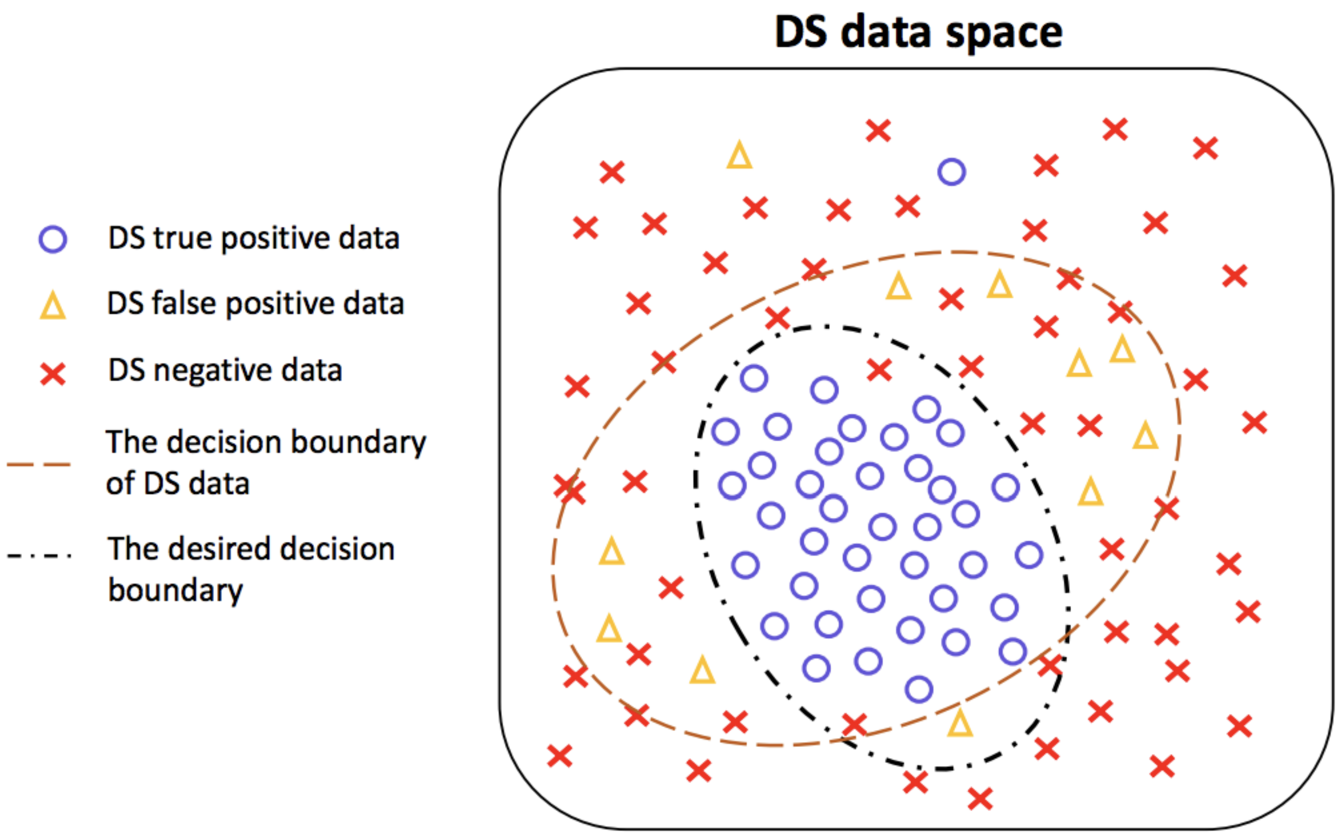


AREL Storytelling Evaluation

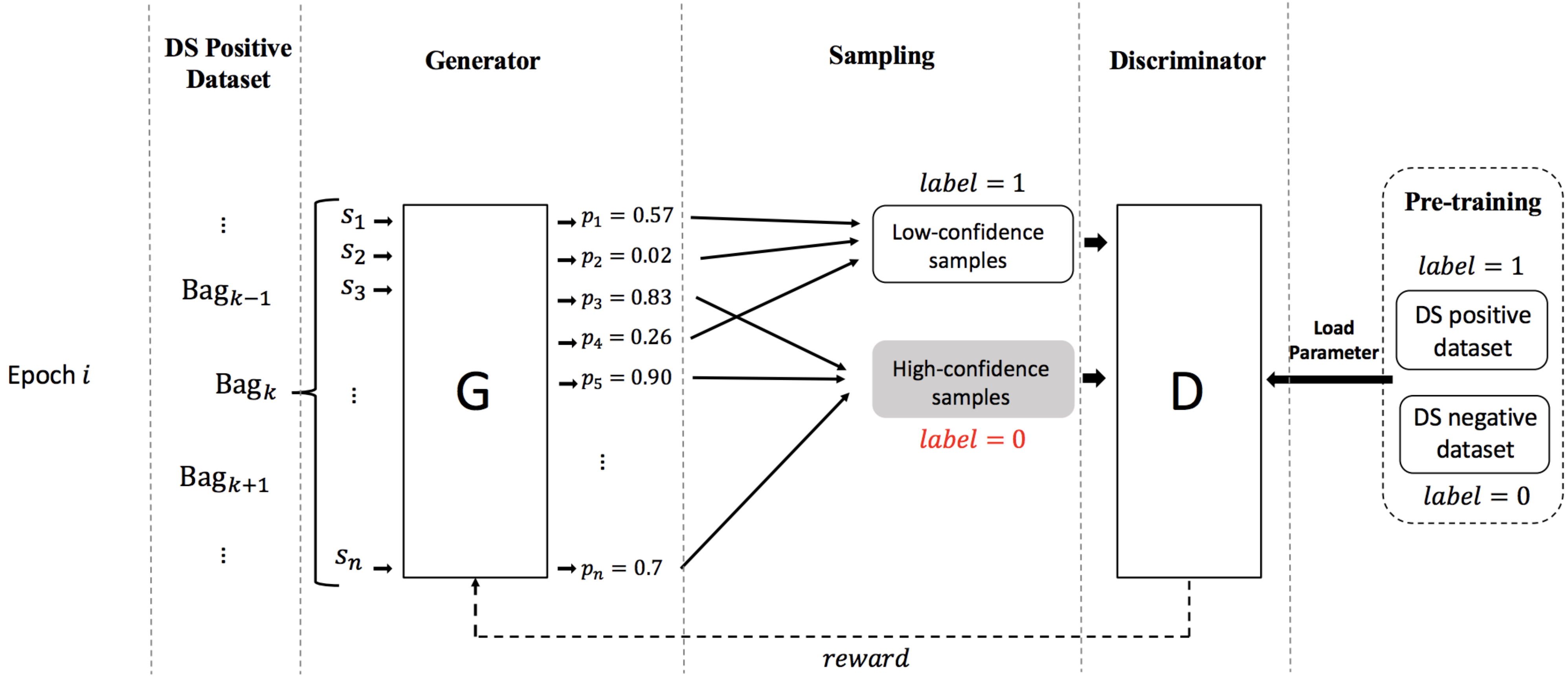
- Dataset: VIST (Huang et al., 2016).



DSGAN: Adversarial Learning for Distant Supervision IE (Qin et al., ACL 2018)

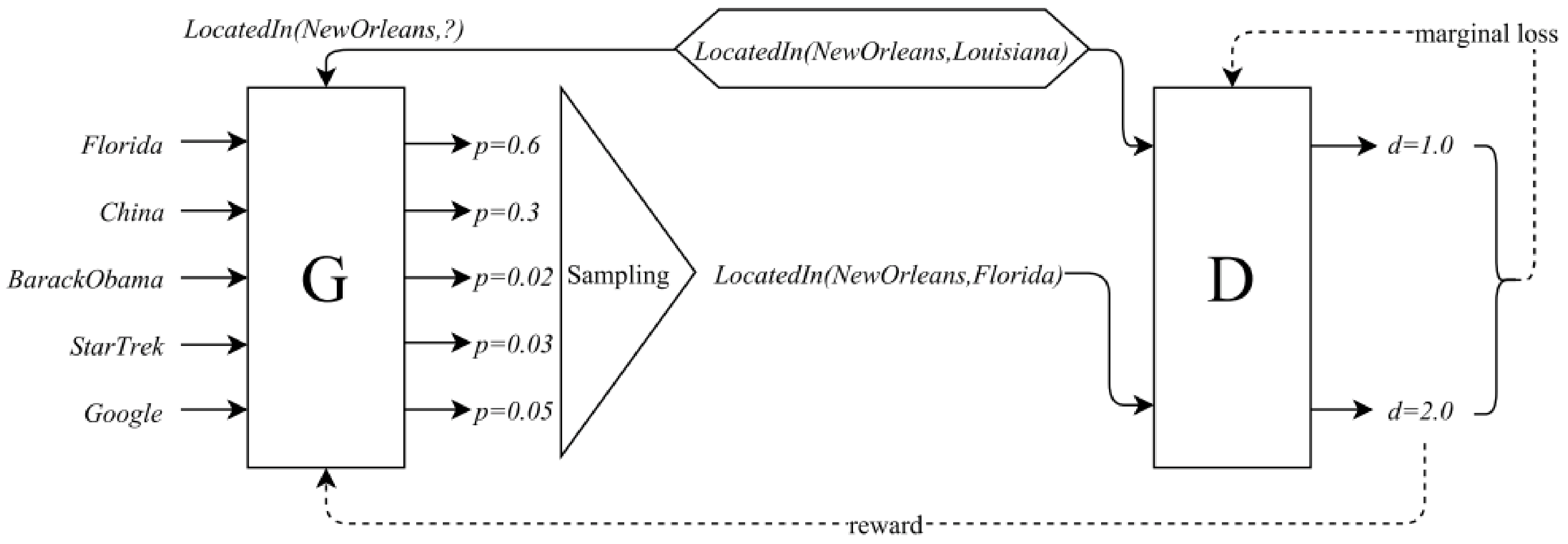


DSGAN: Adversarial Learning for Distant Supervision IE (Qin et al., ACL 2018)



KBGAN: Learning to Generate High-Quality Negative Examples (Cai and Wang, NAACL 2018)

Idea: use adversarial learning to iteratively learn better negative examples.



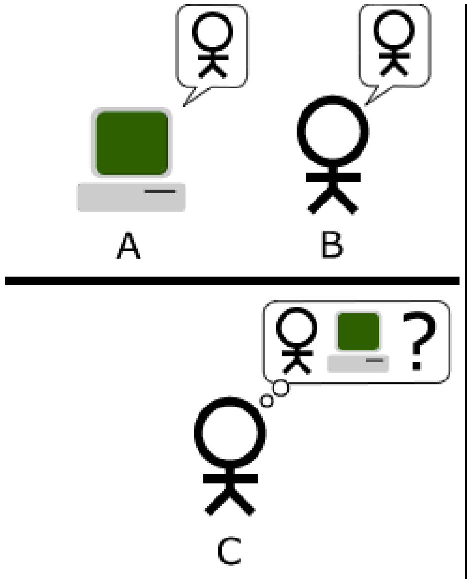
Outline

- Background of the Tutorial
- Introduction: Adversarial Learning in NLP
- Understanding Adversarial Learning
- Adversarial Generation
- A Case Study of GANs in Dialogue Systems

What Should Rewards for Good Dialogue Be Like ?

Reward for Good Dialogue

Turing Test



JI3

Reward for Good Dialogue

I’m 25.

How old are you ?

I don’t know what you are talking about

A human evaluator/ judge

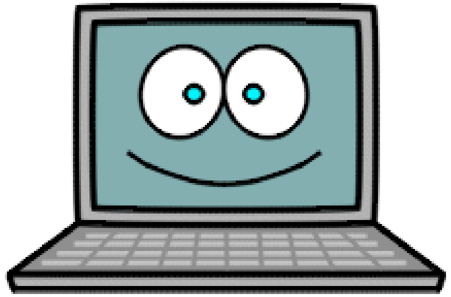
JI3

Reward for Good Dialogue

I’m 25.

How old are you ?

I don’t know what you are talking about



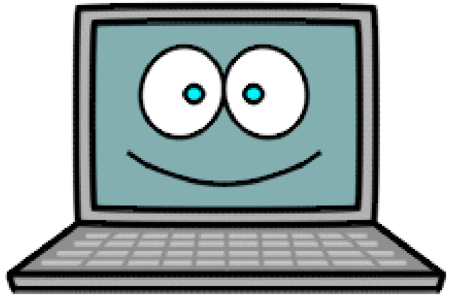
JI3

Reward for Good Dialogue

P= 90% human generated

I’m 25.

How old are you ?

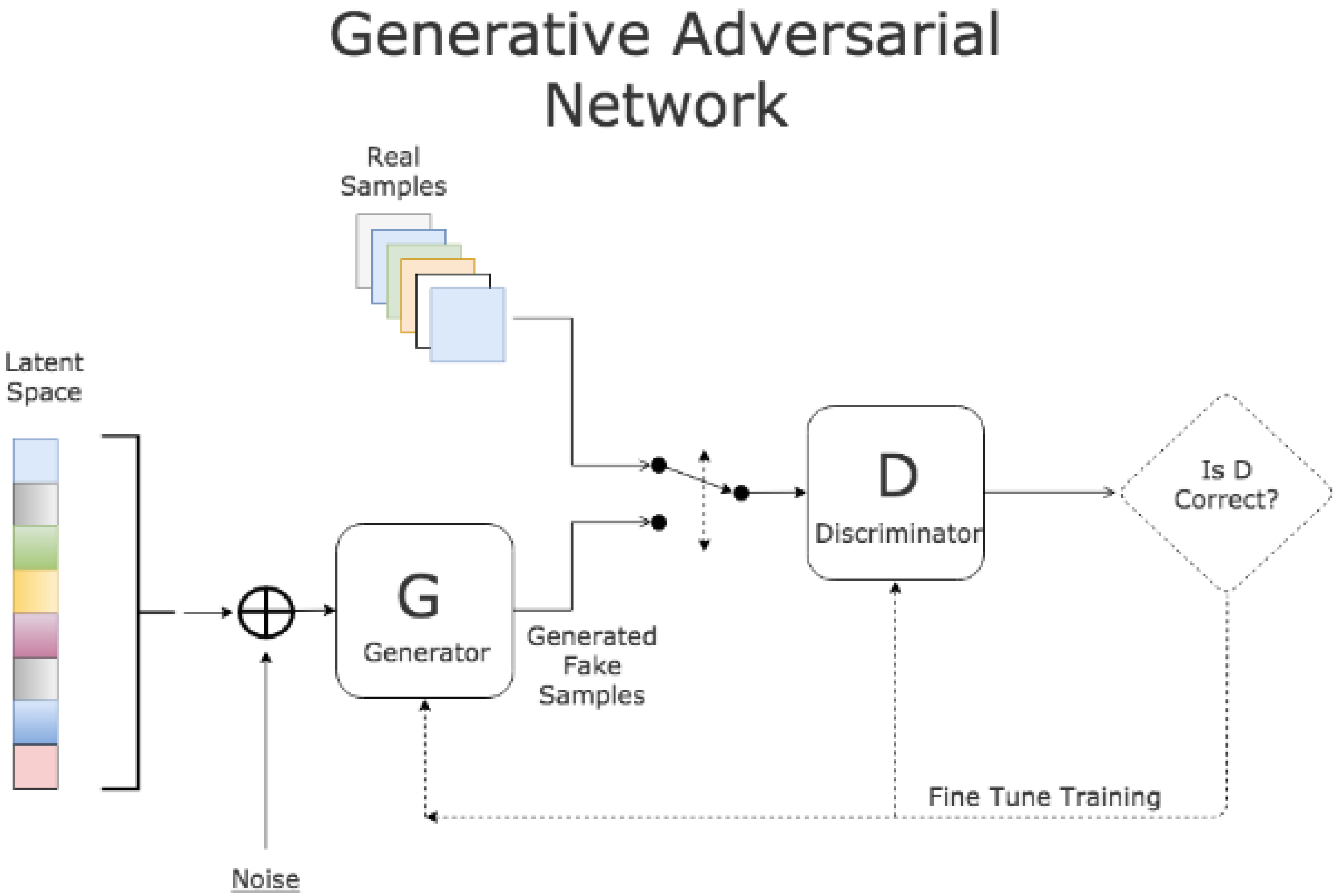


I don’t know what you are talking about

P= 10% human generated

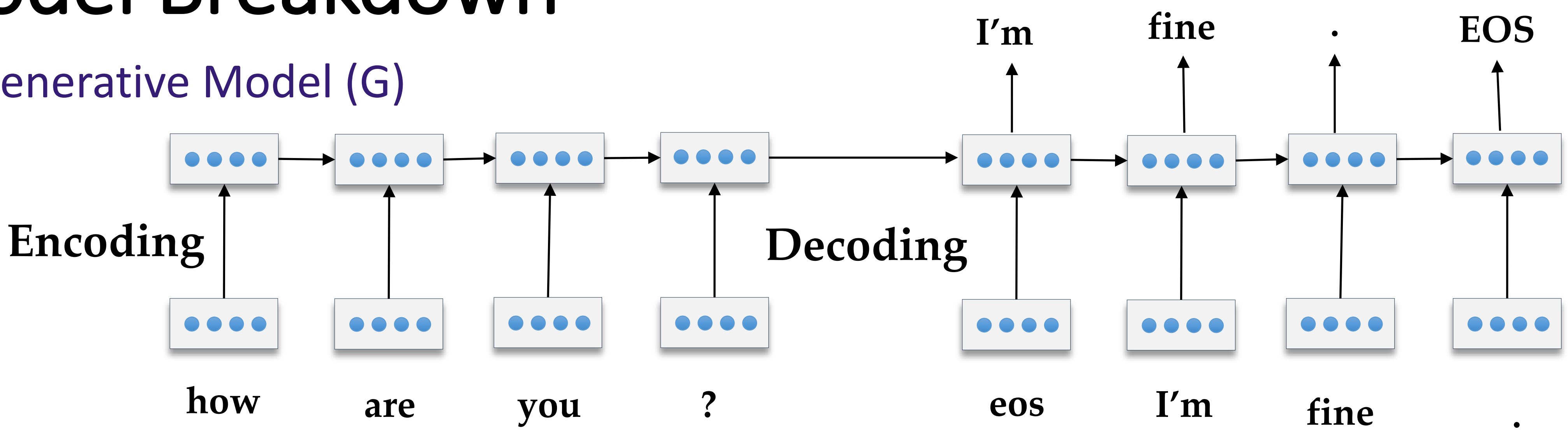
JI3 JI4

Adversarial Learning in Image Generation (Goodfellow et al., 2014)



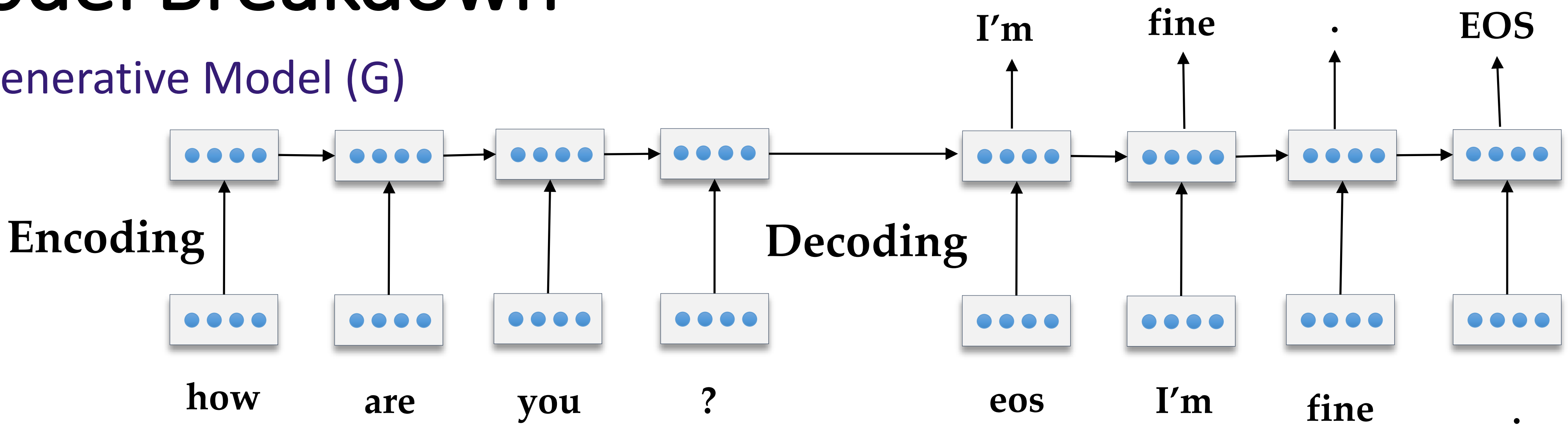
Model Breakdown

Generative Model (G)

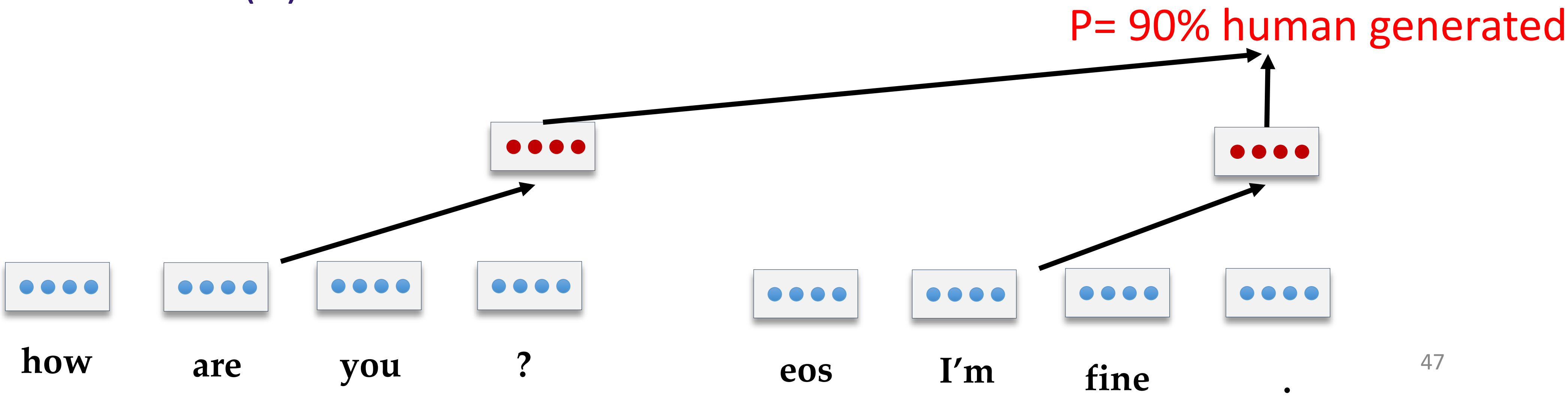


Model Breakdown

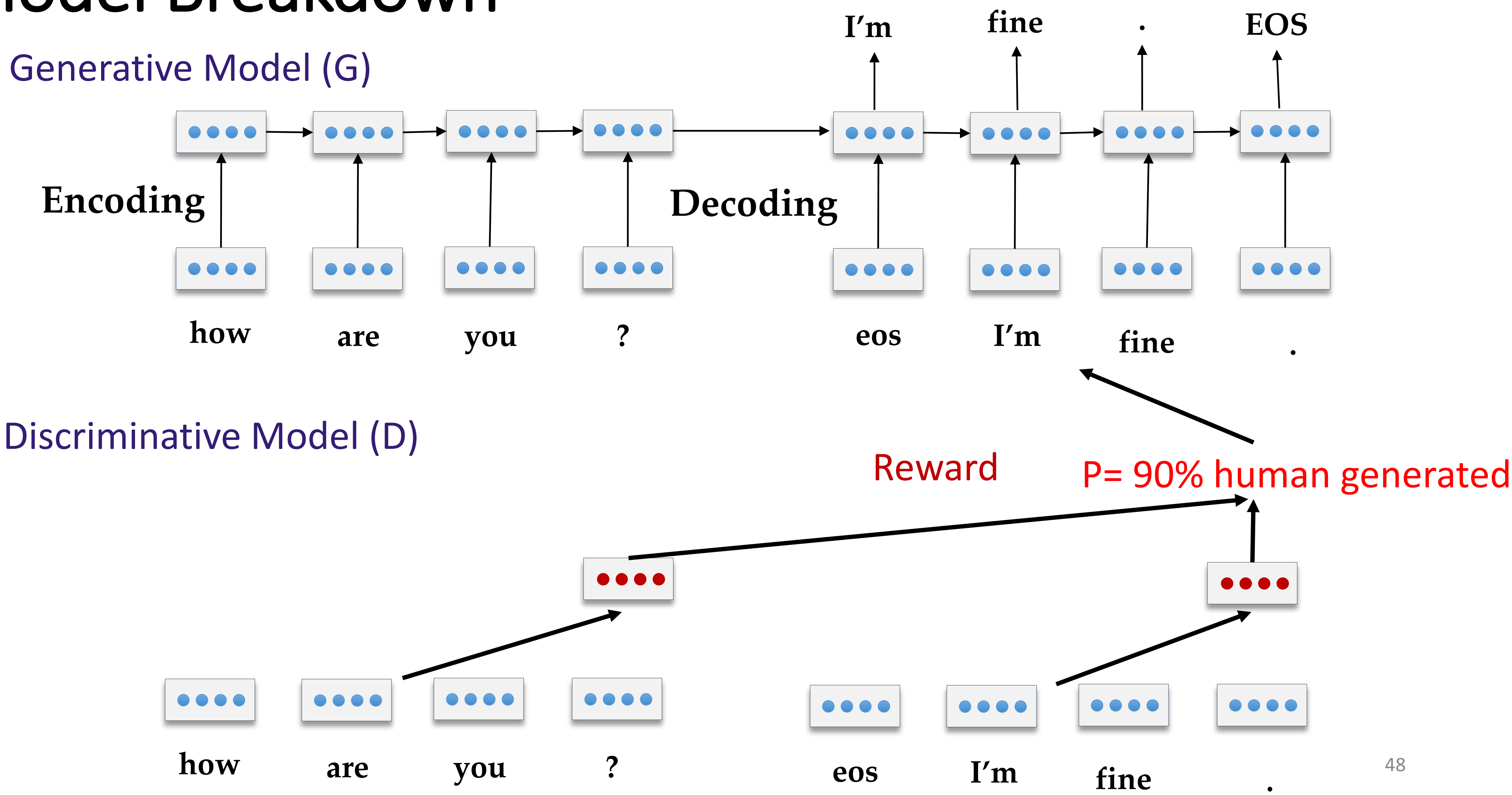
Generative Model (G)



Discriminative Model (D)

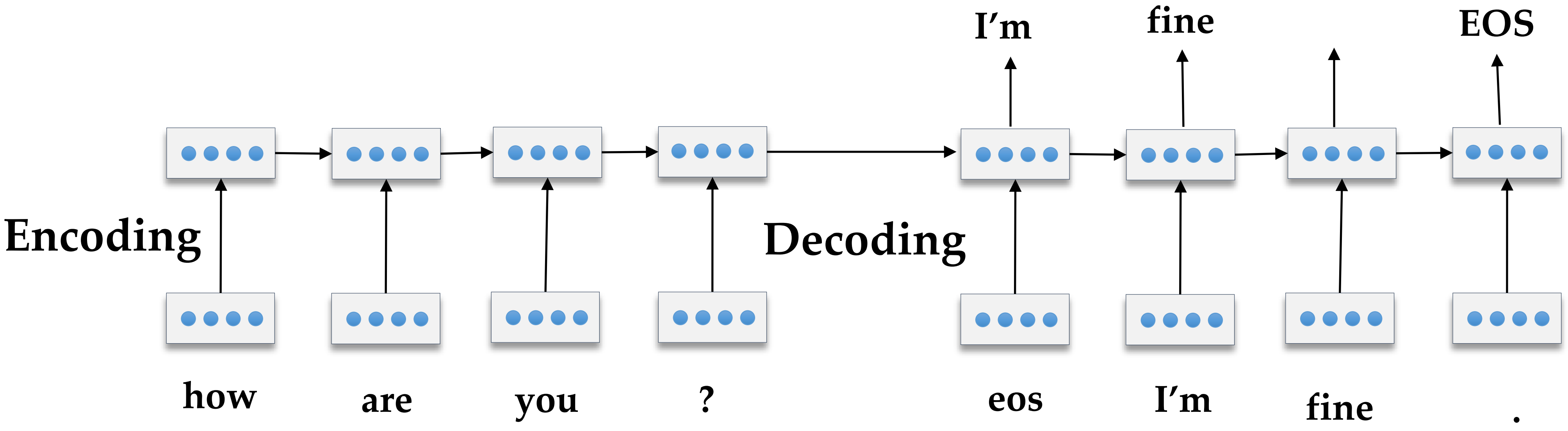


Model Breakdown



Policy Gradient

Generative Model (G)



REINFORCE Algorithm (William,1992)

$$J = E[R(y)]$$

Adversarial Learning for Neural Dialogue Generation

For number of training iterations **do**

. **For** i=1,D-steps **do**

. Sample (X,Y) from real data

. Sample $\hat{Y} \sim G(\cdot|X)$

. Update D using (X, Y) as positive examples and $(X, \hat{Y})$ as negative examples.

. **End**

For i=1,G-steps **do**

Sample (X,Y) from real data

Sample $\hat{Y} \sim G(\cdot|X)$

Compute Reward r for $(X, \hat{Y})$ using D.

Update G on $(X, \hat{Y})$ using reward r

Teacher-Forcing: Update G on (X, Y)

End

End

Update the Discriminator

Update the Generator

The discriminator forces the generator to produce correct responses

Human Evaluation

Setting	adver-win	adver-lose	tie
single-turn	0.62	0.18	0.20
multi-turn	0.72	0.10	0.18

The previous RL model only perform
better on multi-turn conversations

Results: Adversarial Learning Improves Response Generation

Human Evaluator

vs a vanilla generation model

Adversarial Win	Adversarial Lose	Tie
62%	18%	20%

Sample response

Tell me ... how long have you had this falling sickness ?

System	Response

Sample response

Tell me ... how long have you had this falling sickness ?

System	Response
Vanilla-Seq2Seq	I don’t know what you are talking about.

Sample response

Tell me ... how long have you had this falling sickness ?

System	Response
Vanilla-Seq2Seq	I don't know what you are talking about.
Mutual Information	I'm not a doctor.

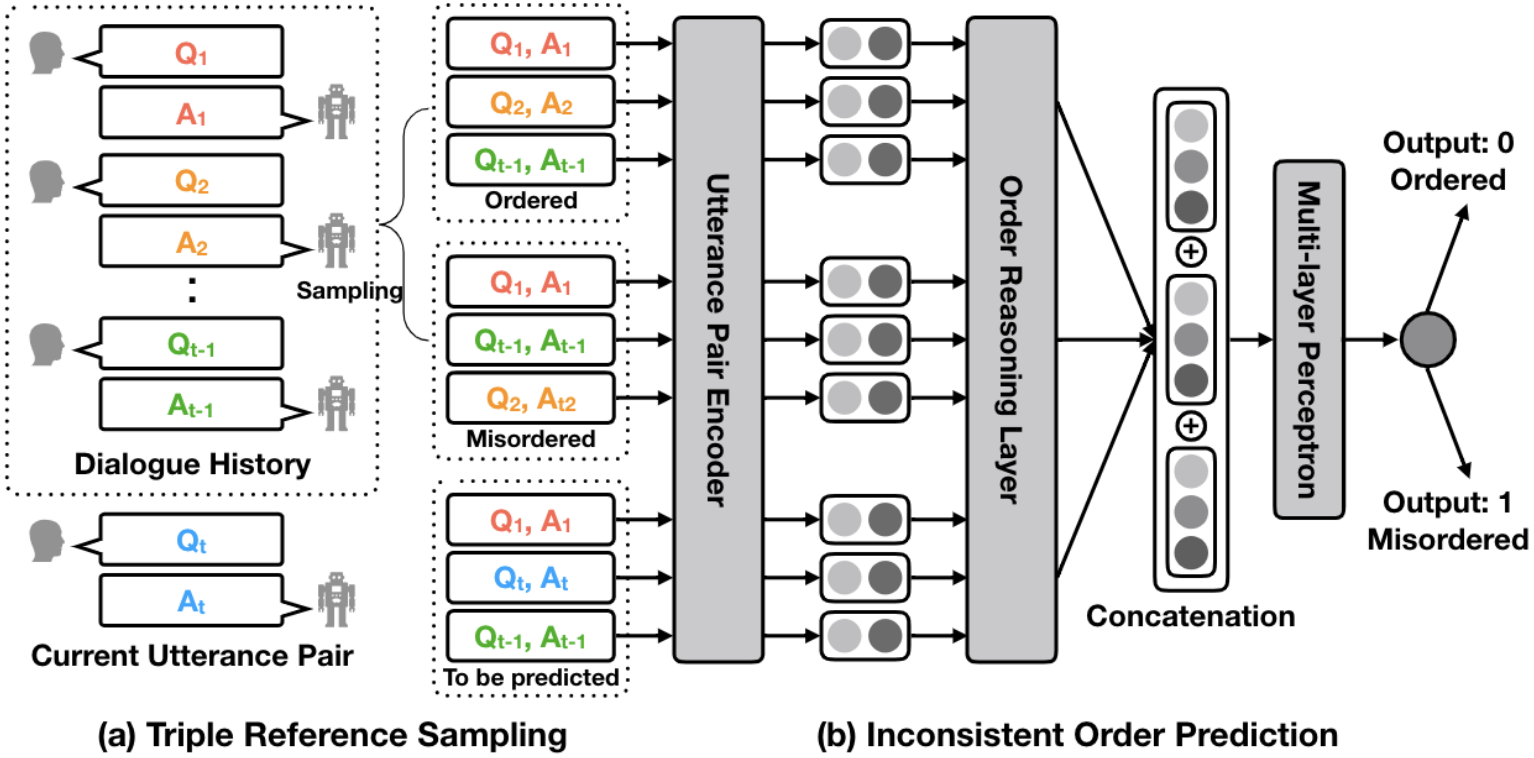
Sample response

Tell me ... how long have you had this falling sickness ?

System	Response
Vanilla-Seq2Seq	I don't know what you are talking about.
Mutual Information	I'm not a doctor.
Adversarial Learning	A few months, I guess.

Self-Supervised Learning meets Adversarial Learning

- Self-Supervised Dialog Learning (Wu et al., ACL 2019)
- Use of SSL to learn dialogue structure (sequence ordering).



Self-Supervised Learning meets Adversarial Learning

- Self-Supervised Dialog Learning (Wu et al., ACL 2019)
- Use of SSN to learn dialogue structure (sequence ordering).
- REGS: Li et al., (2017) AEL: Xu et al., (2017)

Win	REGS	AEL	$\mathcal{SSN}$
Single-turn Percentage	.095	.192	**.713**
Multi-turn Percentage	.025	.171	**.804**

Conclusion

- Deep adversarial learning is a new, diverse, and inter-disciplinary research area, and it is highly related to many subareas in NLP.
- GANs have obtained particular strong results in Vision, but yet there are both challenges and opportunities in GANs for NLP.
- In a case study, we show that adversarial learning for dialogue has obtained promising results.
- There are plenty of opportunities ahead of us with the current advances of representation learning, reinforcement learning, and self-supervised learning techniques in NLP.

UCSB Postdoctoral Scientist Opportunities

- Please talk to me at NAACL, or email william@cs.ucsb.edu.

Thank you!

- Now we will take an 30 mins break.

UCSB