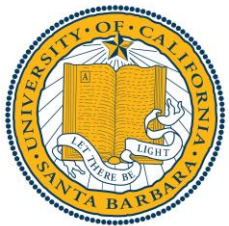


Deep Adversarial Learning for NLP



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

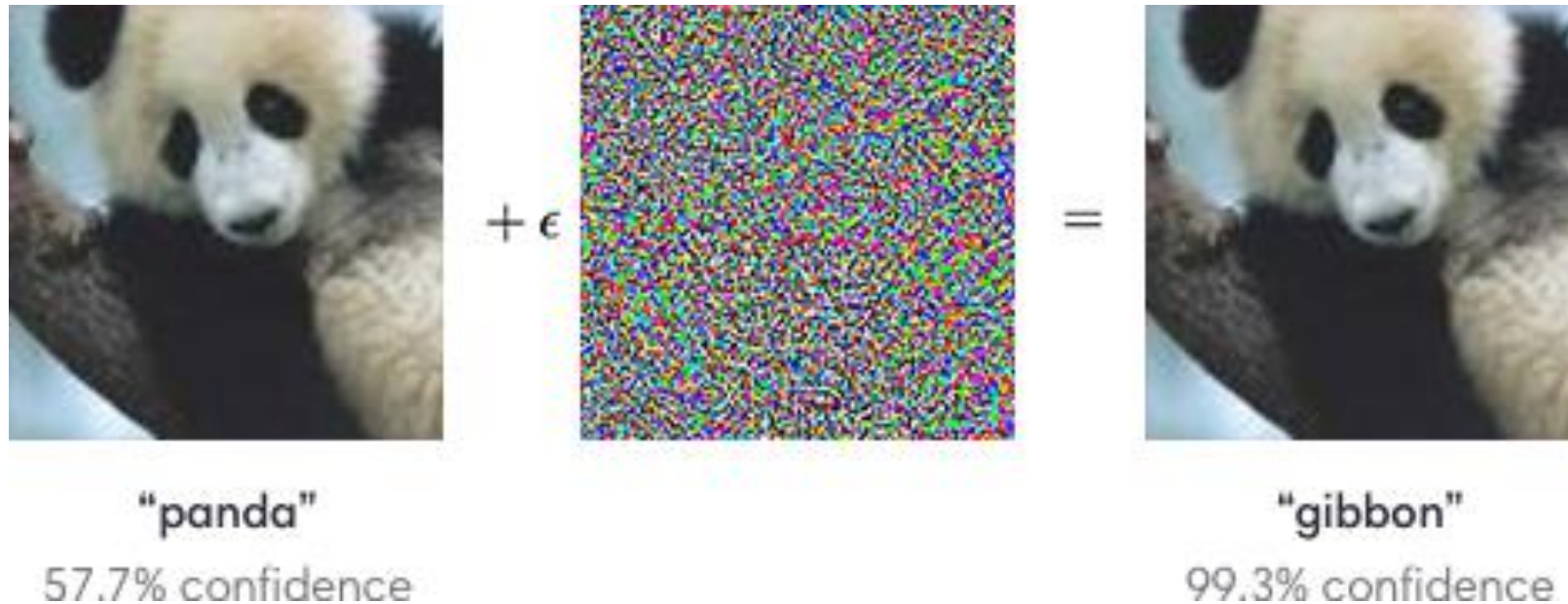
- What are the subareas of deep adversarial learning in NLP?
- How do we understand adversarial learning?
- What are some success stories?
- What are the pitfalls that we need to avoid?

Opportunities in Adversarial Learning

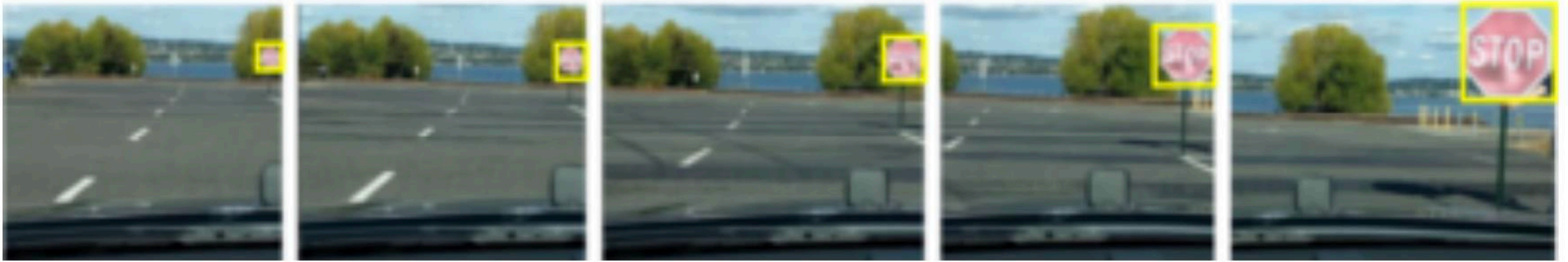
- Adversarial learning is an interdisciplinary research area, and it is closely related to, but limited to the following fields of study:
 - Machine Learning
 - Computer Vision
 - Natural Language Processing
 - Computer Security
 - Game Theory
 - Economics

Adversarial Attack in ML, Vision, & Security

- Goodfellow et al., (2015)



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



Success of Adversarial Learning



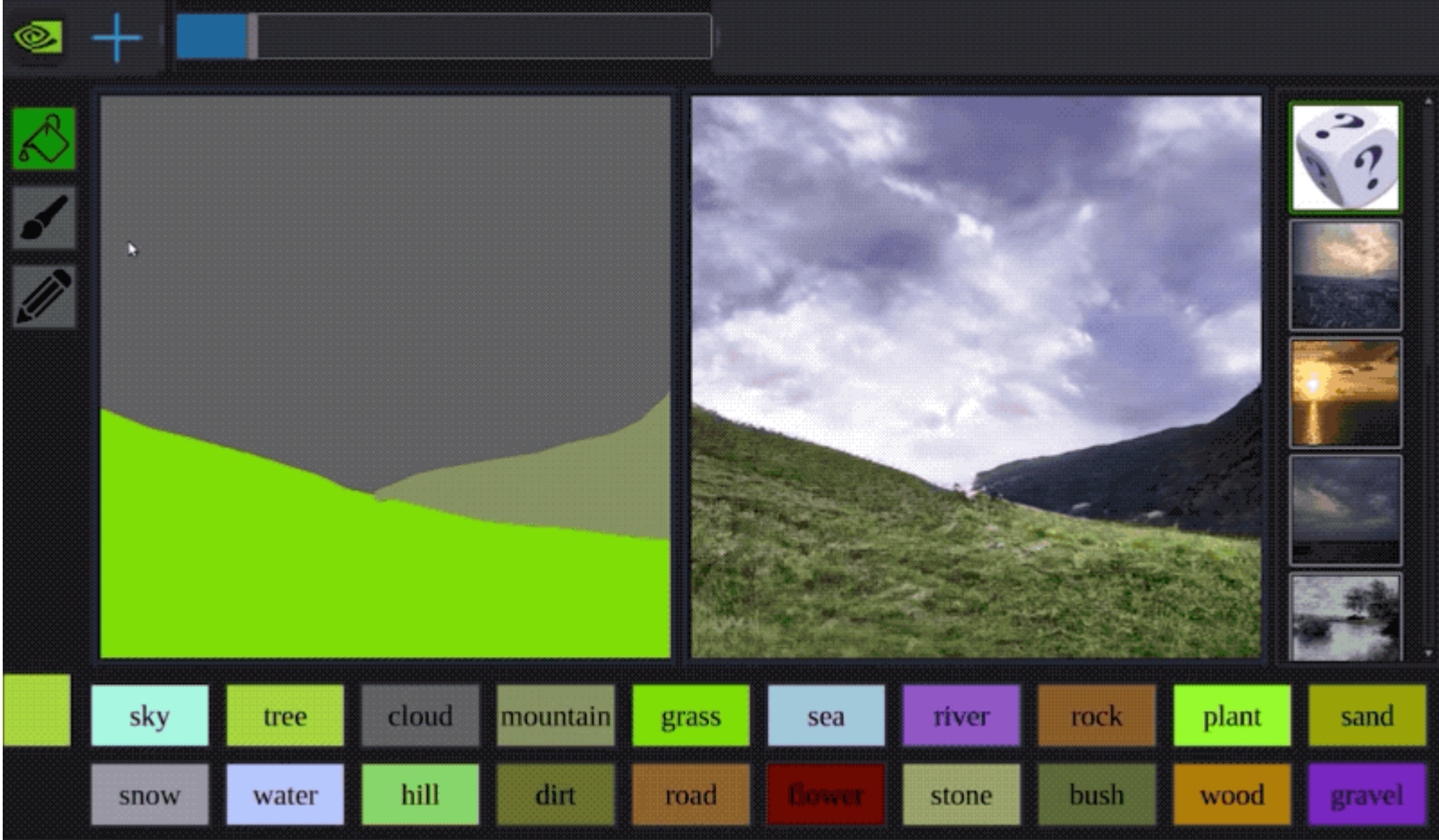
CycleGAN (Zhu et al., 2017)

Failure Cases



CycleGAN (Zhu et al., 2017)

Success of Adversarial Learning



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.

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.*

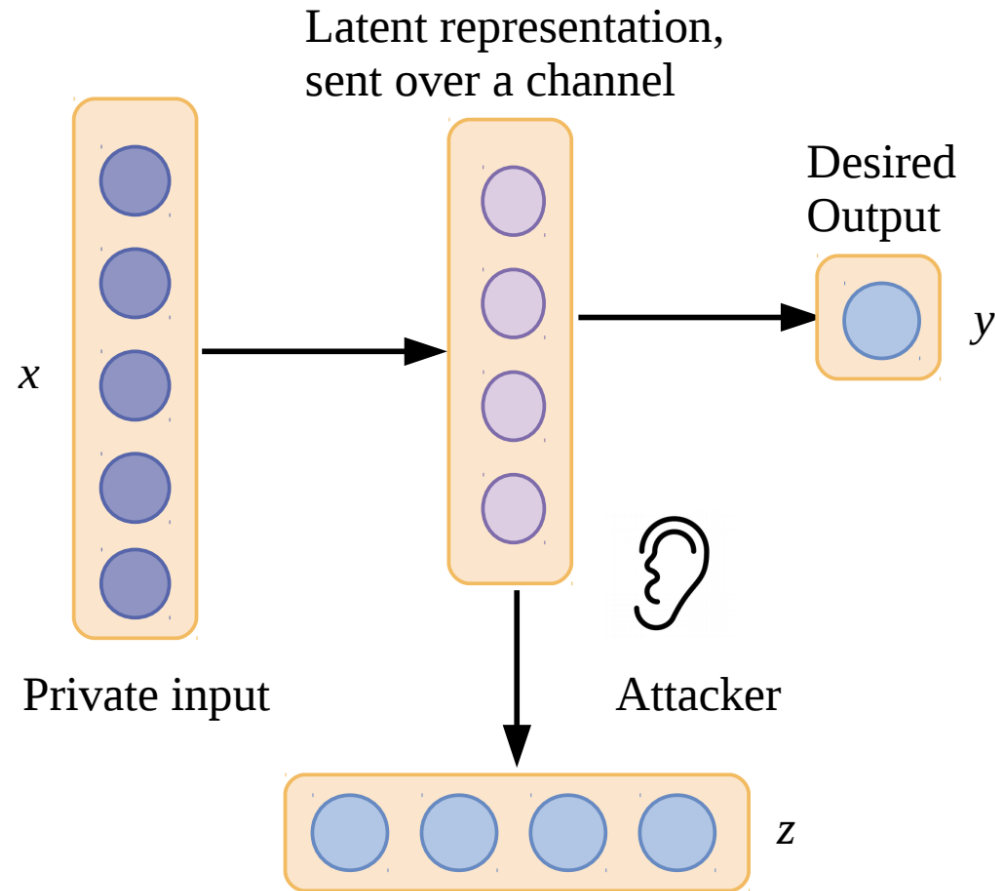
Adversarial Text Prediction: **Contradiction** (Confidence = 43%)

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

$$L_{\text{adv}}(X; \theta) = L(X + e_{\text{adv}}; \theta), \text{ where}$$
$$e_{\text{adv}} = \arg \max_{\|e\| \leq \epsilon} L(X + e; \hat{\theta})$$

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

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

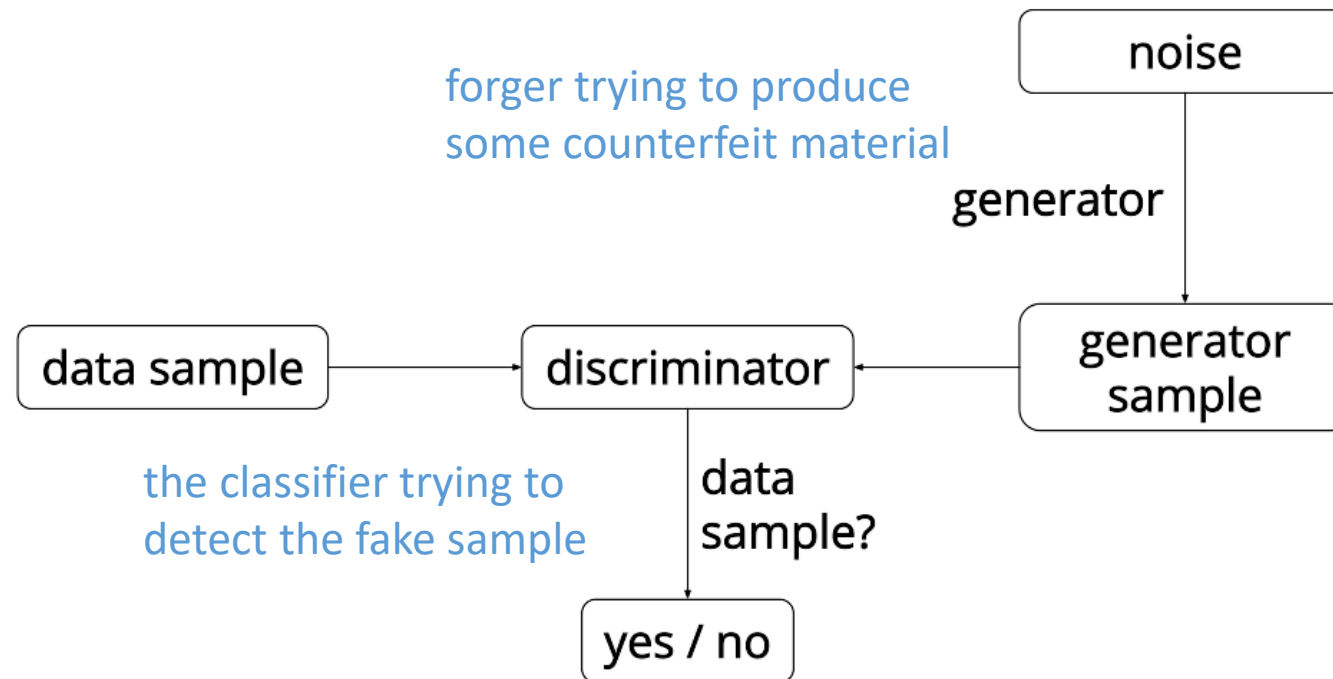
Wu, Bamman, Russell (EMNLP 2017).

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



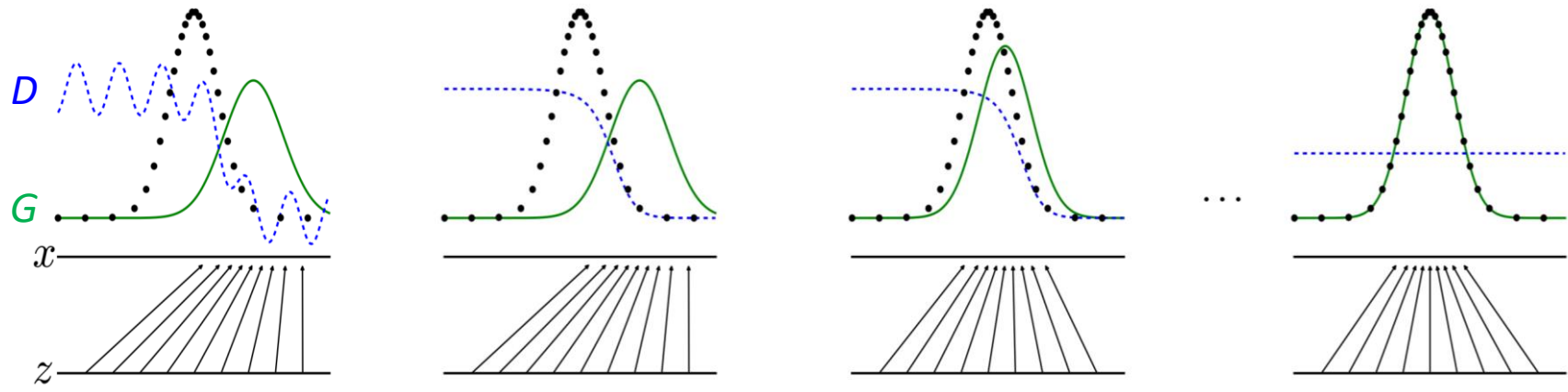
GAN Objective

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

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

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

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



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**

for k steps **do**

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

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

end for

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

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

end for

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

Discriminator

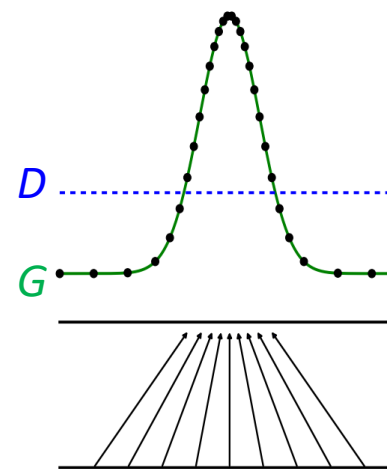
Generator

GAN Equilibrium

- Global optimality
 - Discriminator
- Generator

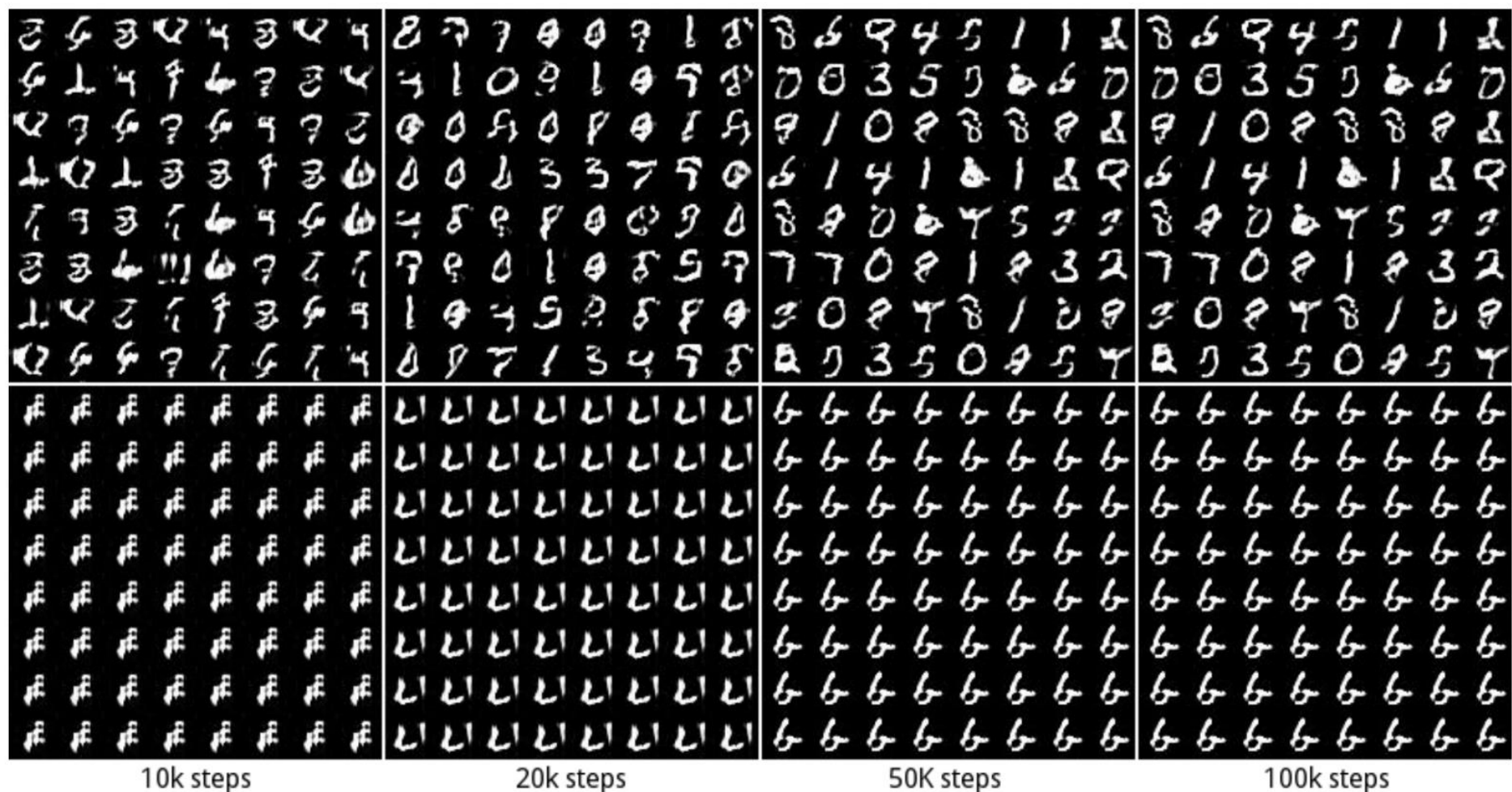
$$D^*(\mathbf{x}) = \frac{q(\mathbf{x})}{q(\mathbf{x}) + p(\mathbf{x})}$$

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



Major Issues of GANs

- Mode Collapse (unable to produce diverse samples)



Major Issues of GANs in NLP

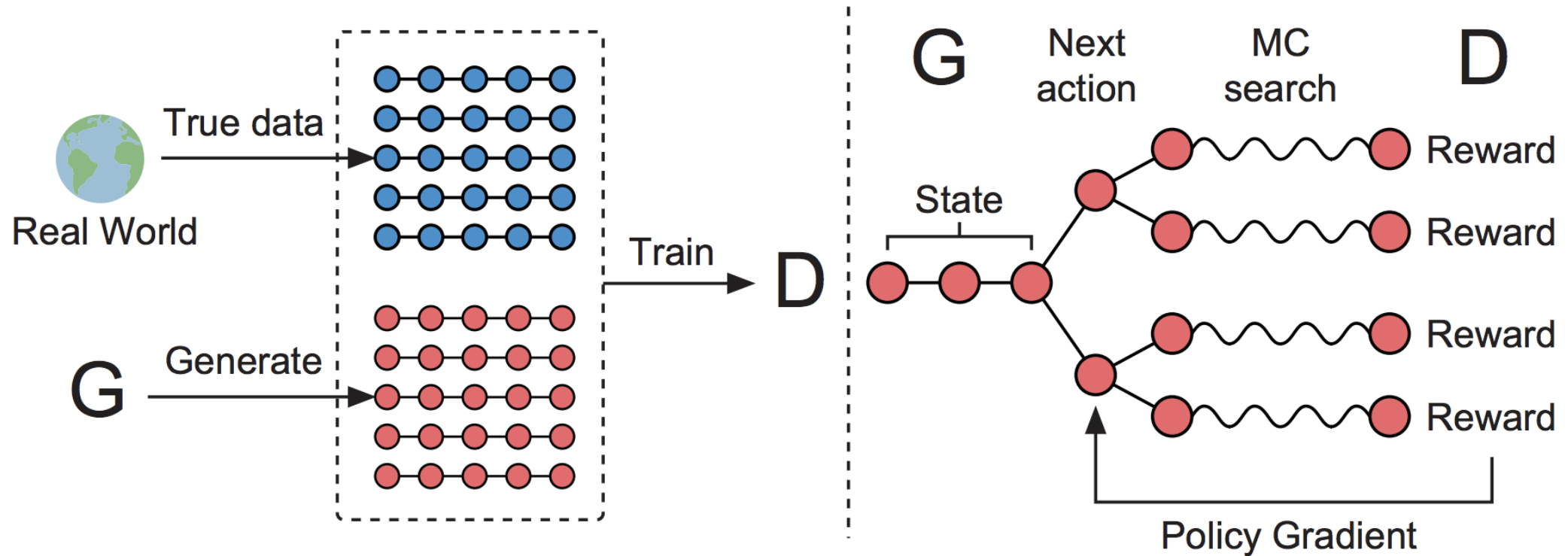
- Often you need to pre-train the generator and discriminator w. MLE
 - But how much?
- Unstable Adversarial Training
 - We are dealing with two networks / learners / agents
 - Should we update them at the same rate?
- The discriminator might overpower the generator.
- With many possible combinations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

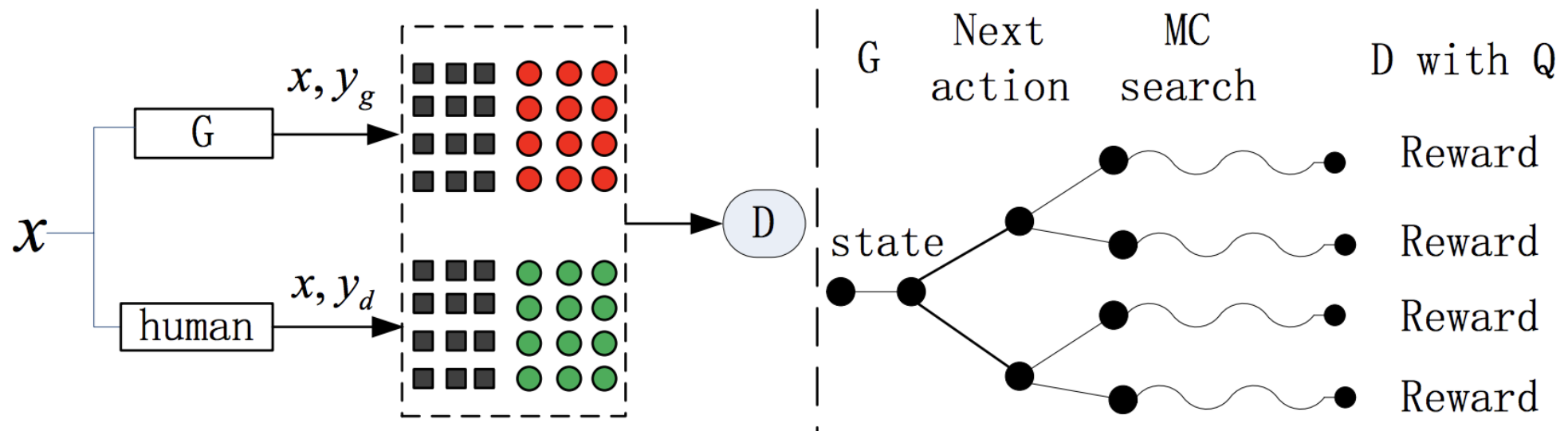
- It's unsupervised learning.
- Many potential applications of GANs in NLP.
- The discriminator is often learning a metric.
- It can also be interpreted as self-supervised learning (especially with dense rewards).

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.

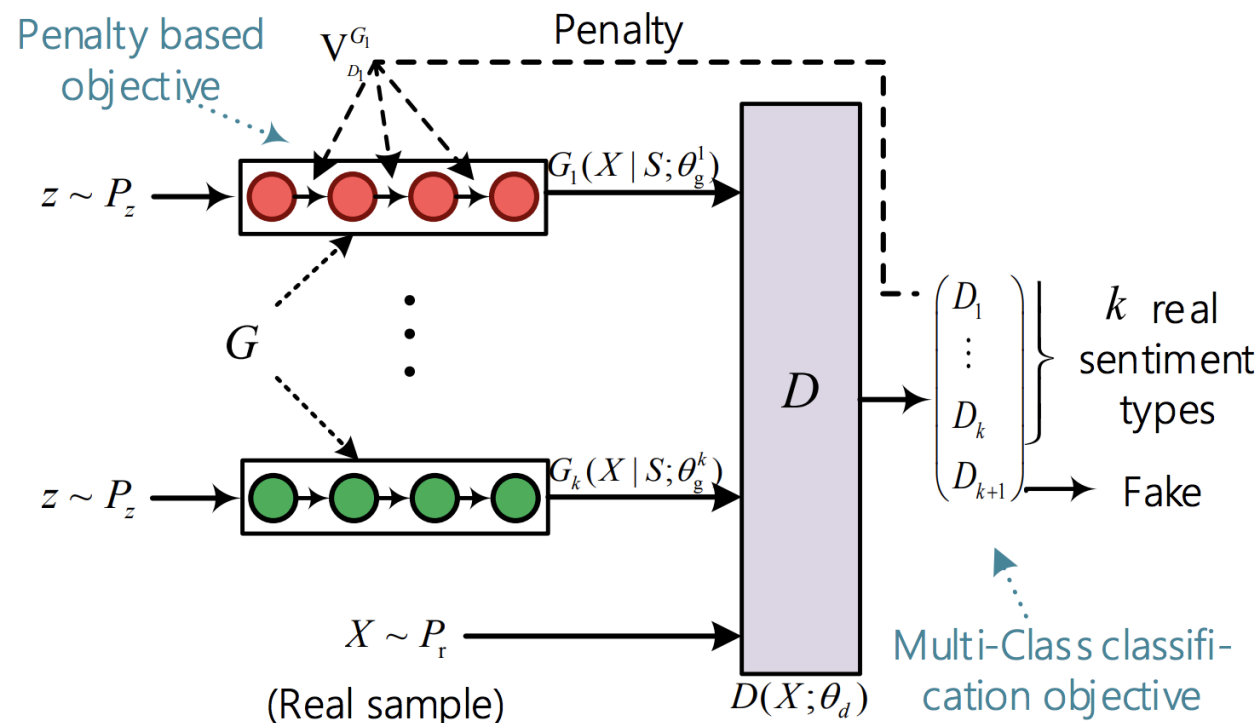
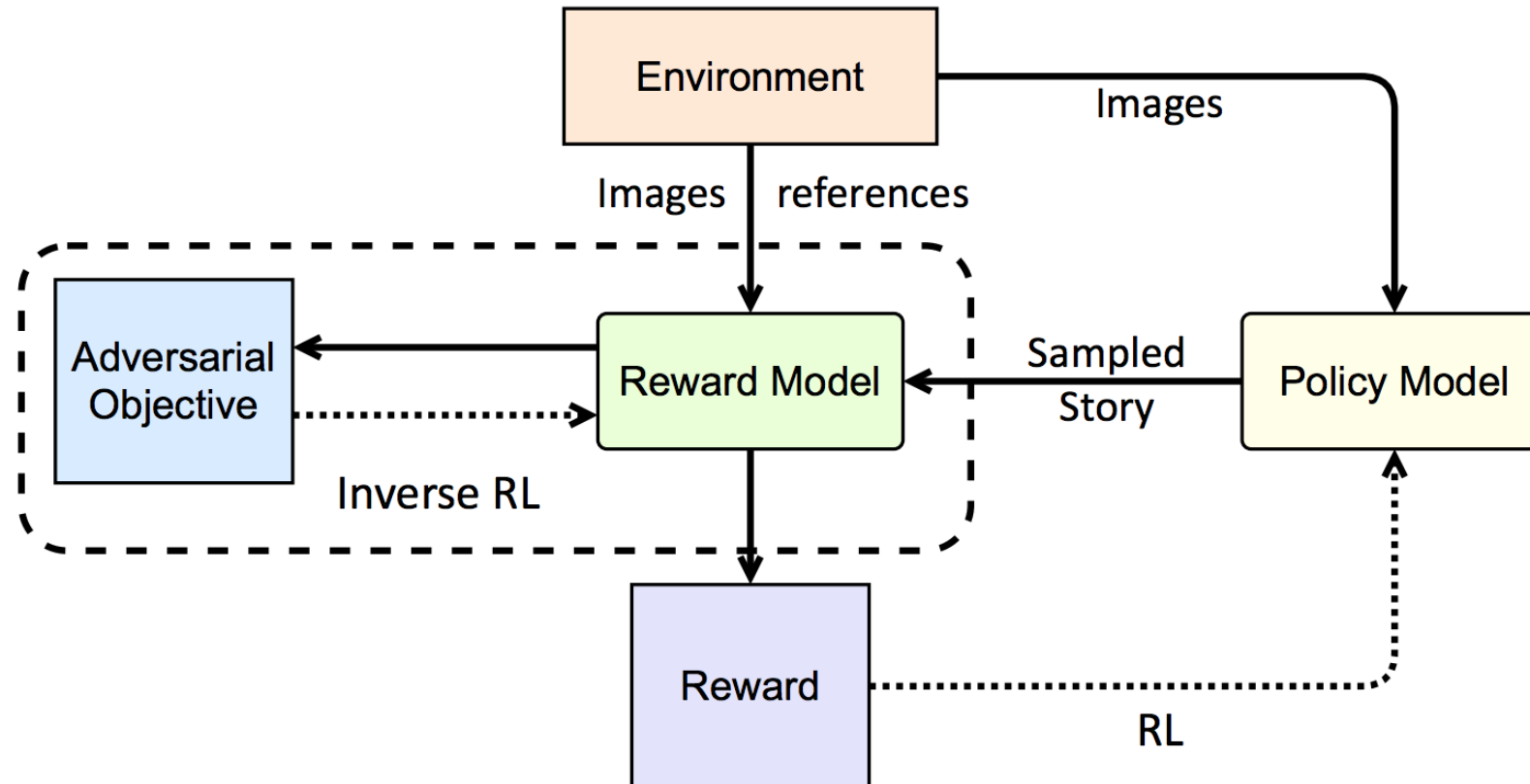


Figure 1: The framework of SentiGAN with k generators and one multi-class discriminator.

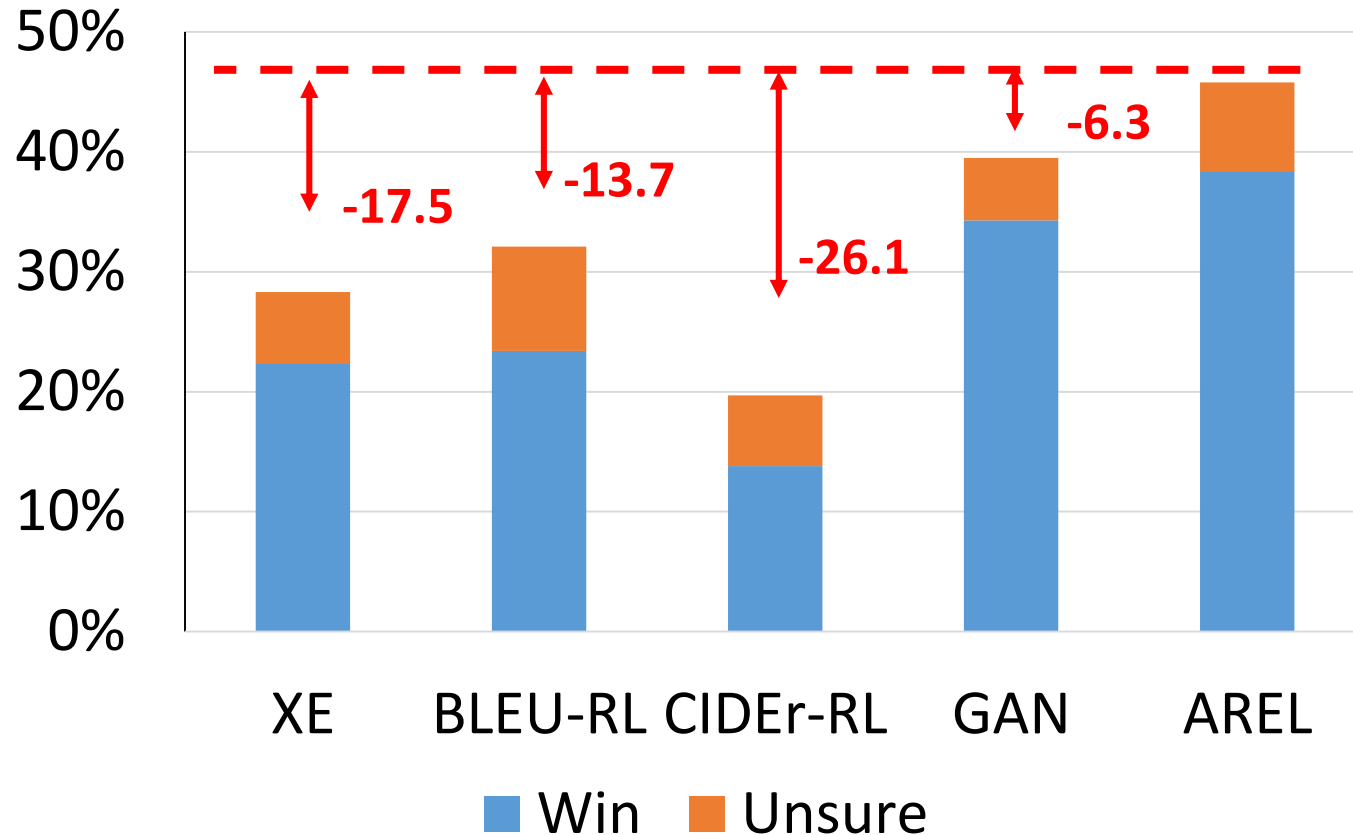
No Metrics Are Perfect: Adversarial Reward Learning (Wang, Chen et al., ACL 2018)



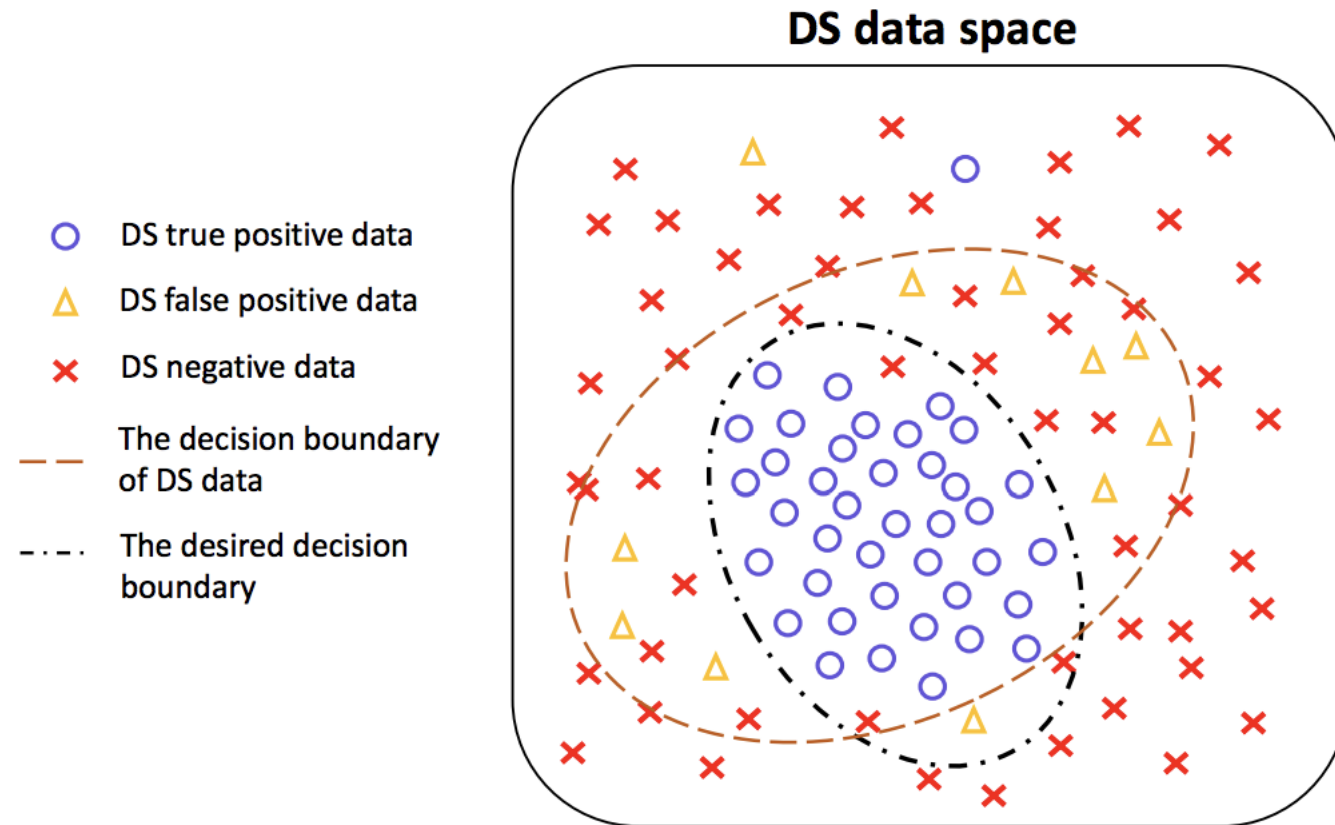
AREL Storytelling Evaluation

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

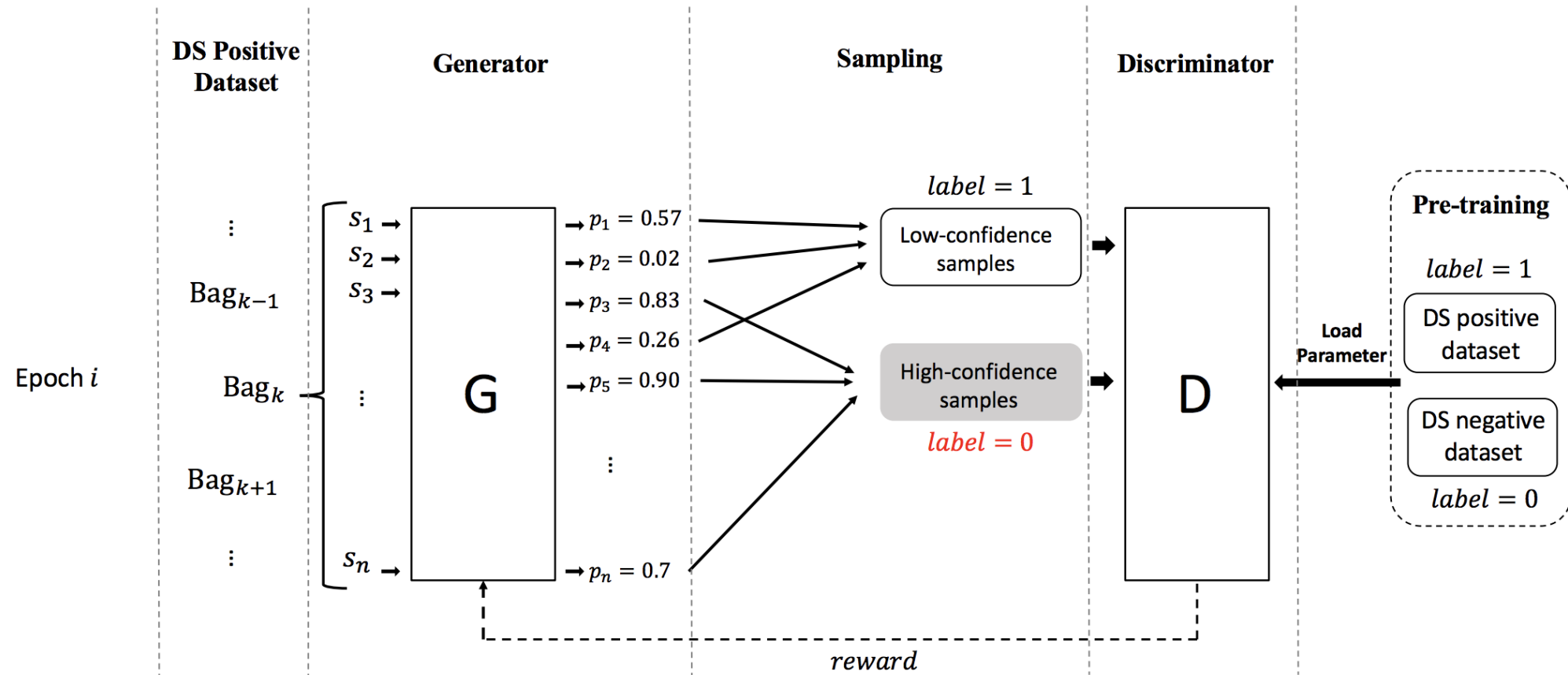
Turing Test



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

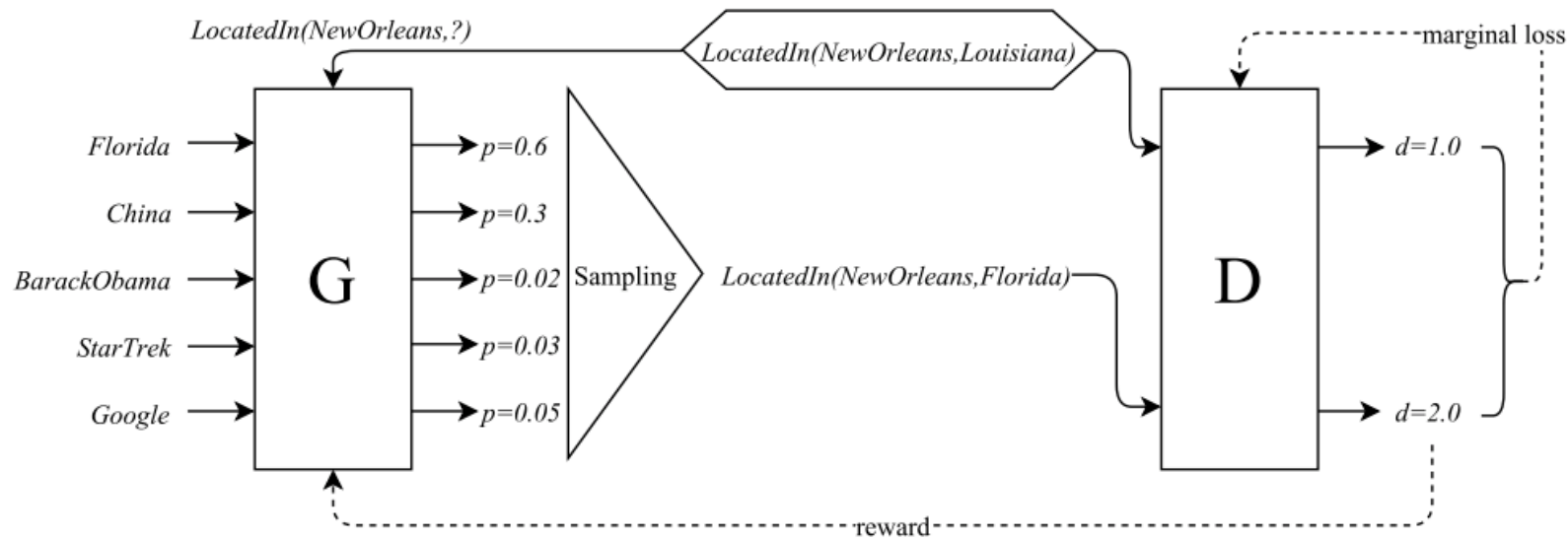


DSGAN: Adversarial Learning for Distant Supervision IE (Qin et al., ACL 2018)



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

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



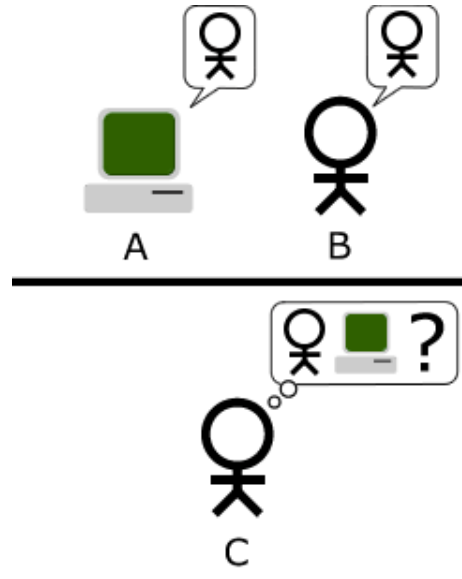
Outline

- 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



Reward for Good Dialogue

How old are you ?

I'm 25.

I don't know what you are talking about

A human evaluator/ judge



Reward for Good Dialogue

How old are you ?



I'm 25.



I don't know what you are talking about



Reward for Good Dialogue

How old are you ?



P= 90% human generated

I'm 25.

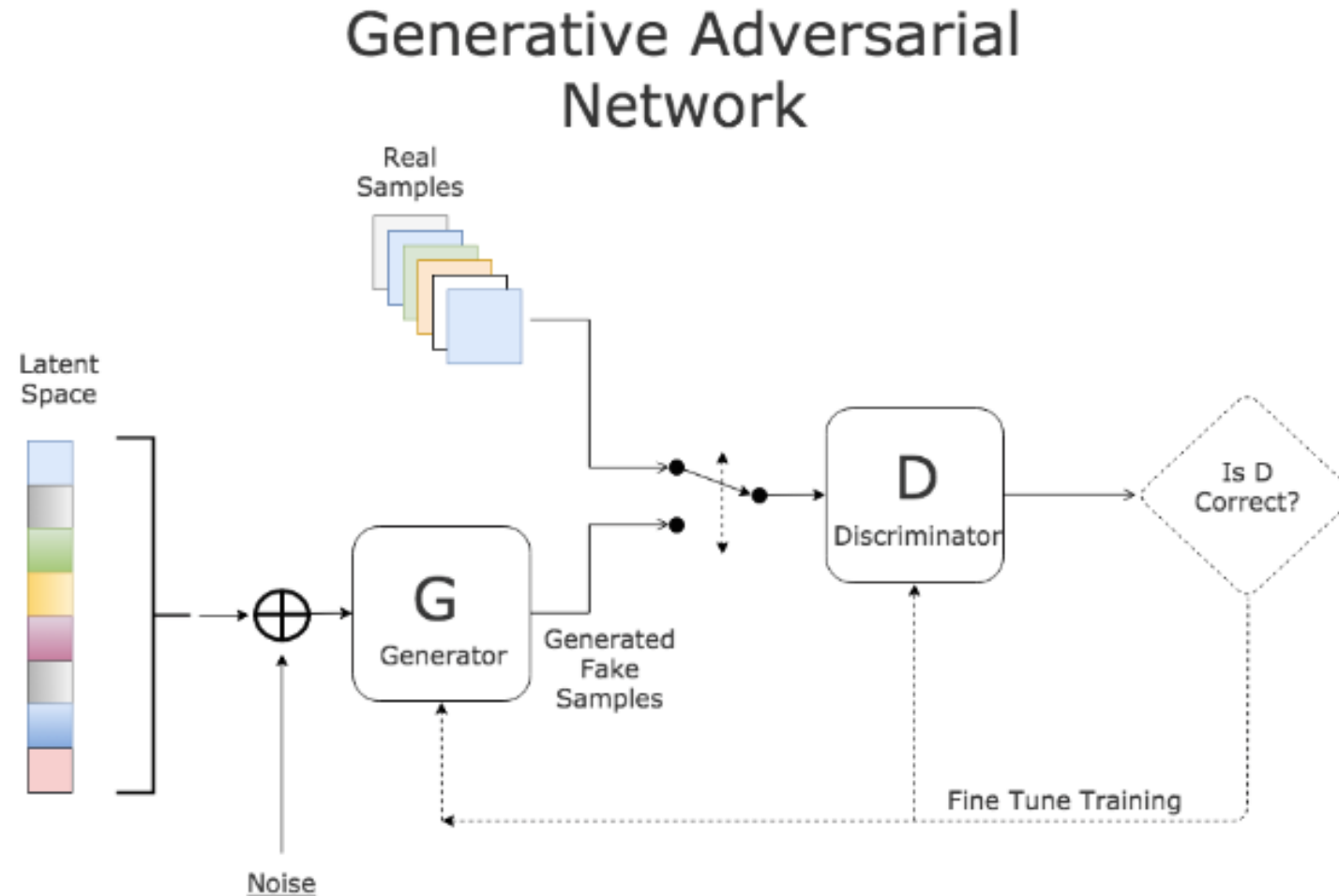


I don't know what you are talking about

P= 10% human generated

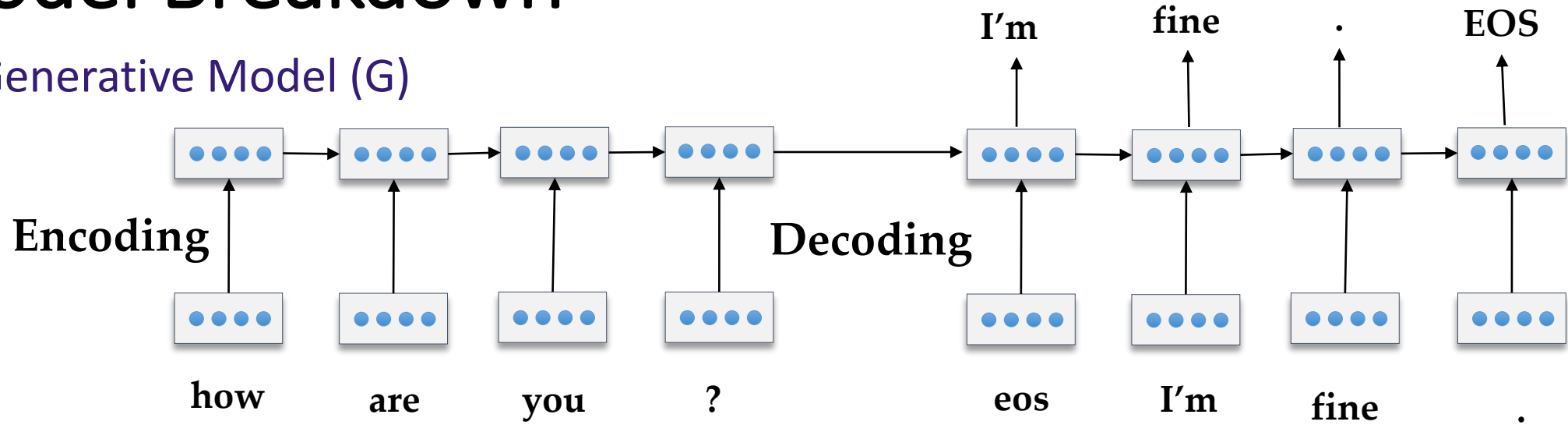


Adversarial Learning in Image Generation (Goodfellow et al., 2014)



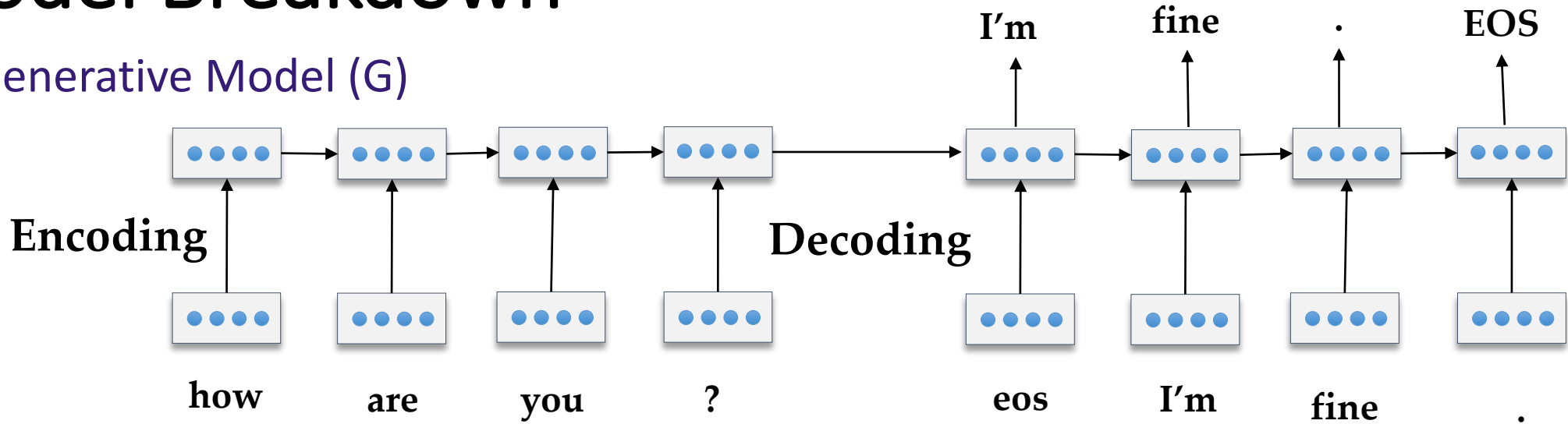
Model Breakdown

Generative Model (G)

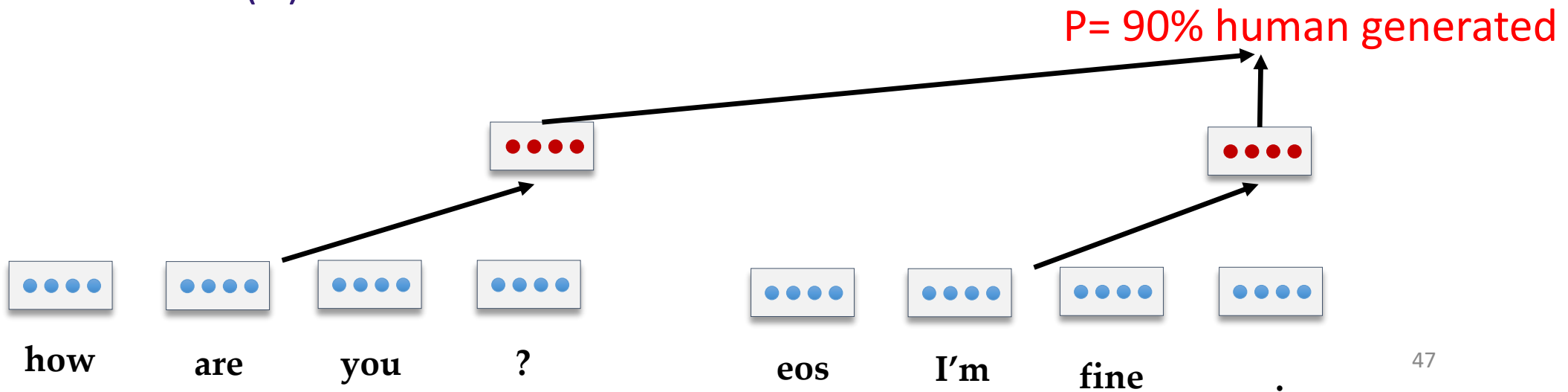


Model Breakdown

Generative Model (G)

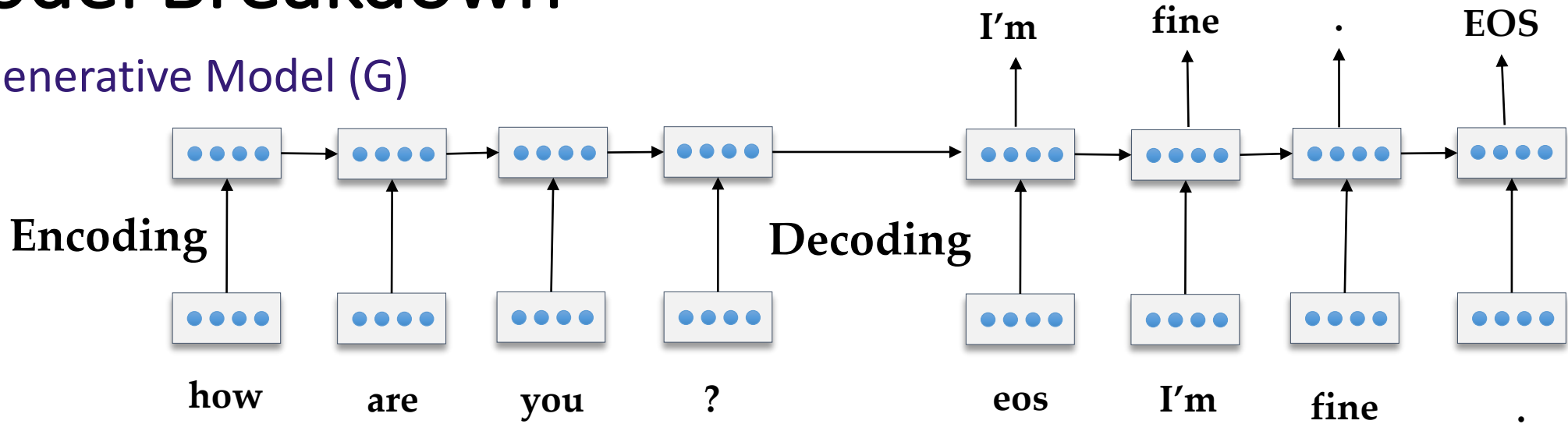


Discriminative Model (D)

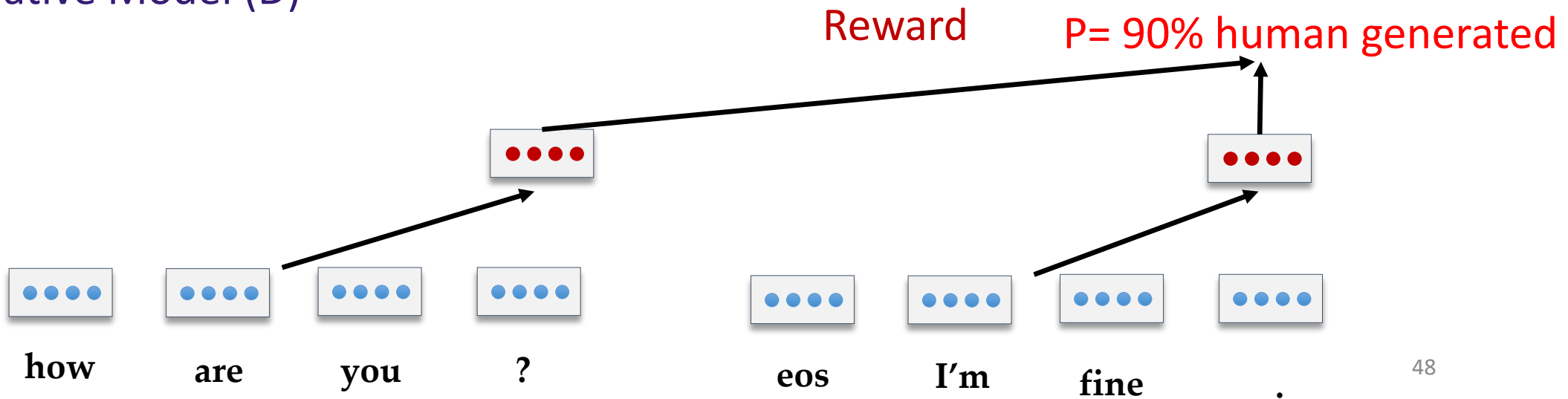


Model Breakdown

Generative Model (G)

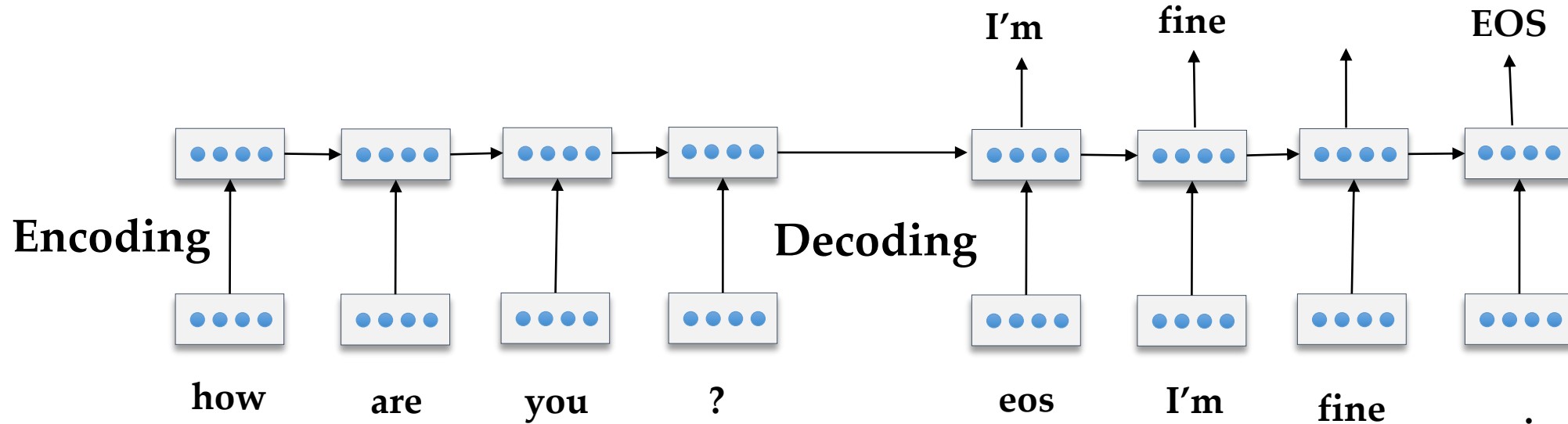


Discriminative Model (D)



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**

**Update the
Discriminator**

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

**Update the
Generator**

End

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.

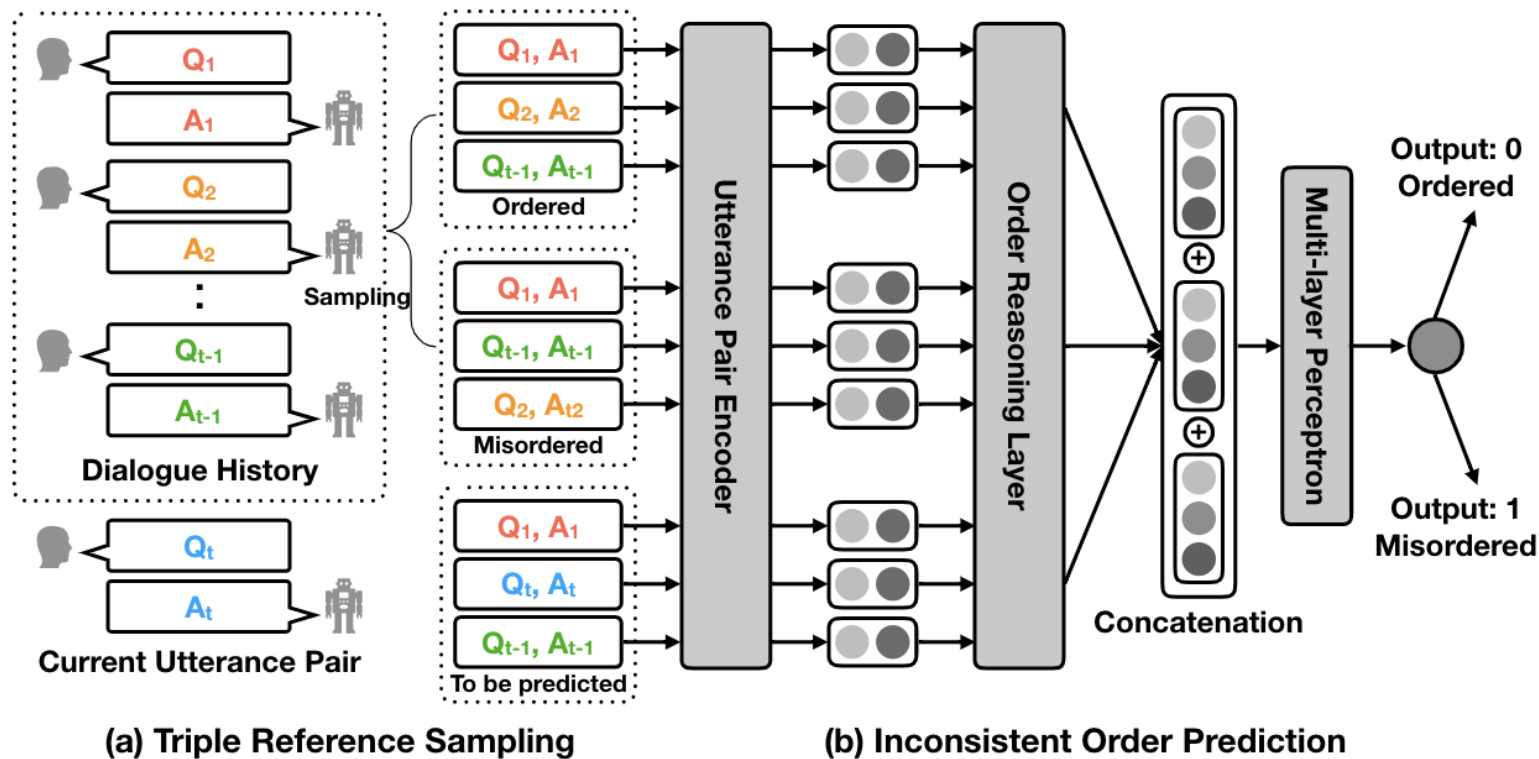
Sample response

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

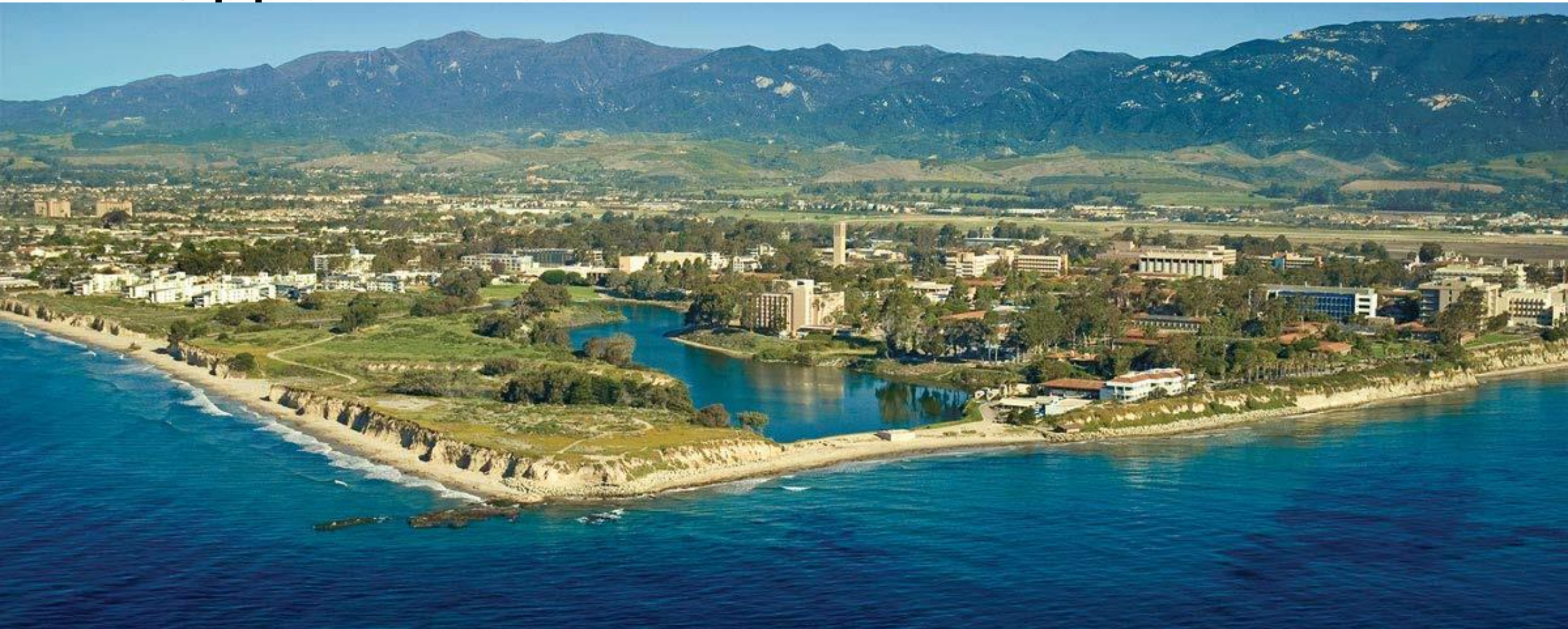
- Self-Supervised Dialog Learning (Wu et al., ACL 2019)
- Use of SSN to learn dialogue structure (sequence ordering).
- REGS: Li et al., (2017) AEL: Xu et al., (2017)

Win	REGS	AEL	<i>SSN</i>
Single-turn Percentage	.095	.192	.713
Multi-turn Percentage	.025	.171	.804

Conclusion

- Deep adversarial learning is a new, diverse, and interdisciplinary 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.

Slides: <http://tiny.cc/adversarial>

Adversarial Examples in NLP

Sameer Singh

sameer@uci.edu

@sameer_

sameersingh.org



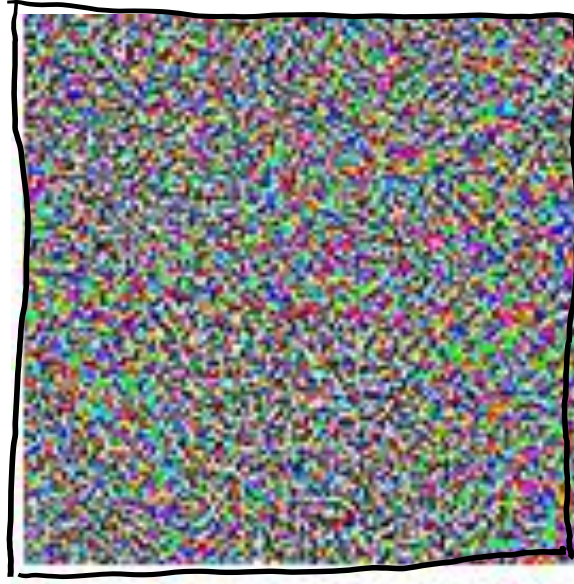
What are Adversarial Examples?



“panda”

57.7% confidence

$+ \epsilon$



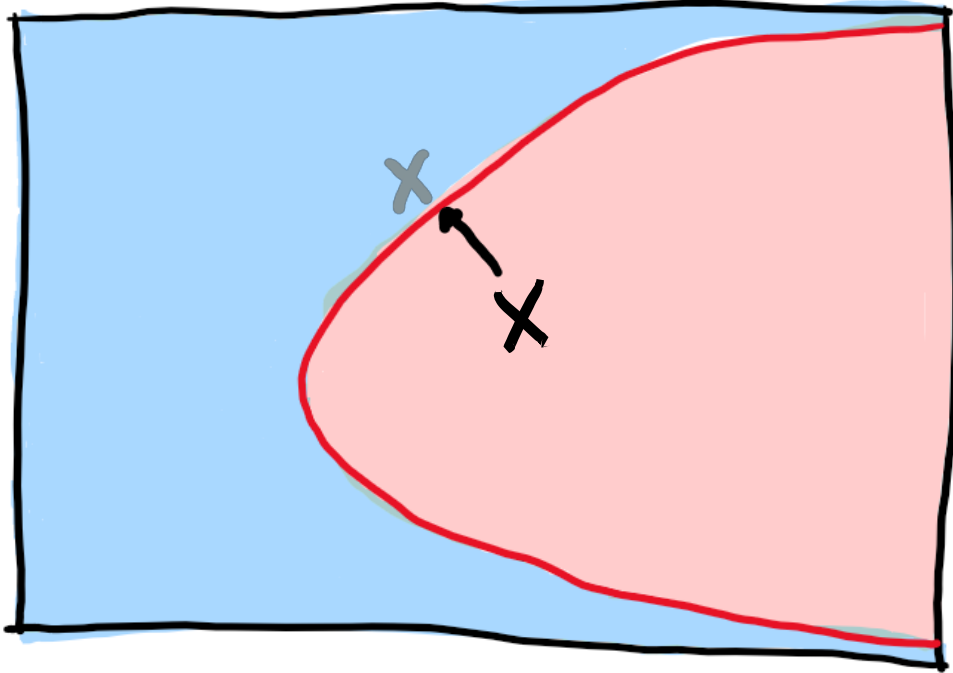
$=$



“gibbon”

99.3% confidence

What's going on?



$$\begin{aligned} \min_{x'} & \|x - x'\| \\ \text{s.t.} & f(x') \neq f(x) \end{aligned}$$

Fast Gradient Sign Method

$$x' \leftarrow x + \epsilon \text{sign}(\nabla_x J(x))$$

Applications of Adversarial Attacks

- Security of ML Models
 - Should I deploy or not? What's the worst that can happen?
- Evaluation of ML Models
 - Held-out test error is not enough
- Finding Bugs in ML Models
 - What kinds of “adversaries” might happen naturally?
 - (Even without any bad actors)
- Interpretability of ML Models?
 - What does the model care about, and what does it ignore?

Challenges in NLP

Change

L_2 is not really defined for text

What is imperceivable? What is a small vs big change?

What is the right way to measure this?

Search

Text is discrete,
cannot use continuous optimization
How do we search over sequences?

$$\begin{array}{l} \min_{x'} \\ \text{s.t. } f(x') \neq f(x) \end{array} \|x - x'\|$$

Effect

Classification tasks fit in well, but ...

What about structured prediction? e.g. sequence labeling

Language generation? e.g. MT or summarization

Choices in Crafting Adversaries

Different ways to address the challenges

Choices in Crafting Adversaries

How do we find the attack?

$$\begin{array}{l} \min_{x'} \|x - x'\| \\ \text{s.t. } f(x') \neq f(x) \end{array}$$

What is a small change?

What does it mean to misbehave?

Choices in Crafting Adversaries

$$\min_{x'} \boxed{\|x - x'\|}$$

What is a small change?

$$\text{s.t. } f(x') \neq f(x)$$

Change: What is a small change?

$$\|x - x'\|$$

Characters

Pros:

- Often easy to miss
- Easier to search over

Cons:

- Gibberish, nonsensical words
- No useful for interpretability

Words

Pros:

- Always from vocabulary
- Often easy to miss

Cons:

- Ungrammatical changes
- Meaning also changes

Phrase/Sentence

Pros:

- Most natural/human-like
- Test long-distance effects

Cons:


- Difficult to guarantee quality
- Larger space to search

Main Challenge: Defining the distance between x and x'

Change: A Character (or few)


$x = [\text{"I love movies"}]$

$x = [\text{'I'} \quad \text{' ' } \quad \text{'l'} \quad \text{'o'} \quad \text{'v'} \quad \dots]$



The diagram shows a sequence of five gray rectangular boxes representing characters: 'I', ' ', 'l', 'o', and 'v'. The fourth box, containing the character 'o', is highlighted in red. Ellipses follow the fifth box.

$x' = [\text{'I'} \quad \text{' ' } \quad \text{'l'} \quad \text{'i'} \quad \text{'v'} \quad \dots]$



The diagram shows a sequence of five gray rectangular boxes representing characters: 'I', ' ', 'l', 'i', and 'v'. The fourth box, containing the character 'i', is highlighted in red. Ellipses follow the fifth box.

past → pas!t | Alps → llps | talk → taln | local → loral

Edit Distance: Flip, Insert, Delete

Change: Word-level Changes

$x = [\text{'I'} \quad \boxed{\text{'like'}} \quad \text{'this'} \quad \text{'movie'} \quad \text{'.'}]$

Let's replace this word

Random word?

$x' = [\text{'I'} \quad \boxed{\text{'lamp'}} \quad \text{'this'} \quad \text{'movie'} \quad \text{'.'}]$

Word Embedding?

$x' = [\text{'I'} \quad \boxed{\text{'really'}} \quad \text{'this'} \quad \text{'movie'} \quad \text{'.'}]$

Part of Speech?

$x' = [\text{'I'} \quad \boxed{\text{'eat'}} \quad \text{'this'} \quad \text{'movie'} \quad \text{'.'}]$

Language Model?

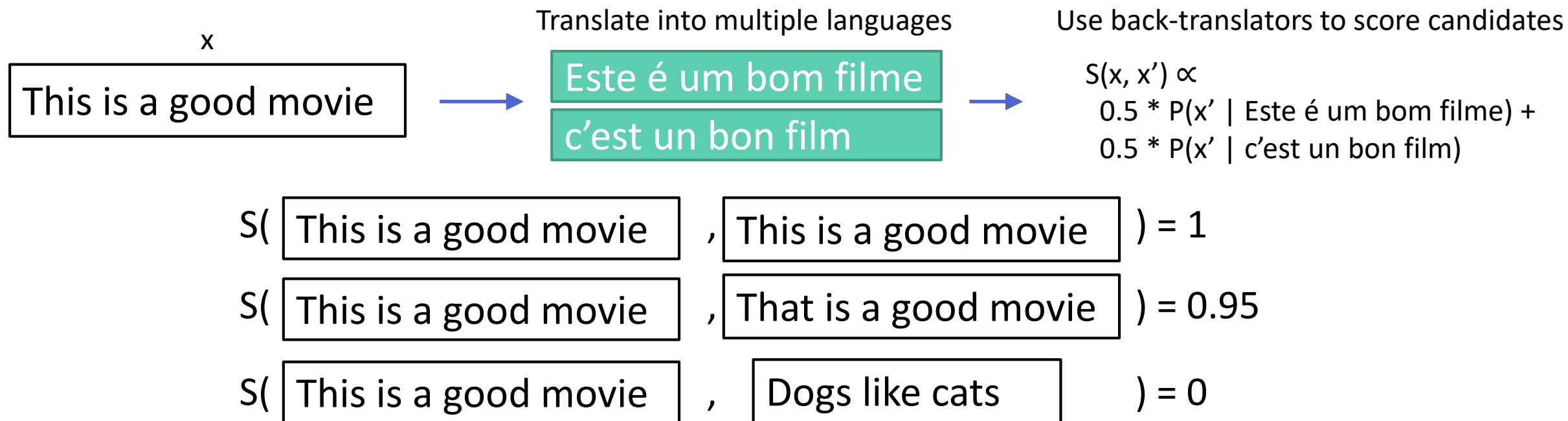
$x' = [\text{'I'} \quad \boxed{\text{'hate'}} \quad \text{'this'} \quad \text{'movie'} \quad \text{'.'}]$

[Jia and Liang, EMNLP 2017]

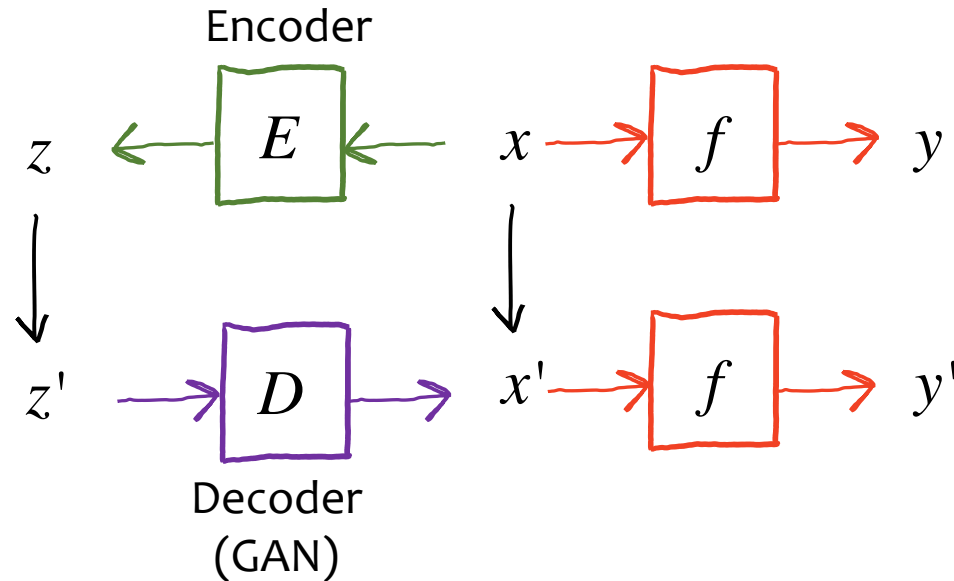
[Alzantot et. al. EMNLP 2018]

Change: Paraphrasing via Backtranslation

x, x' should mean the same thing (*semantically-equivalent adversaries*)



Change: Sentence Embeddings



$$\min_{x'} \|z - z'\|$$
$$\text{s.t. } f(x') \neq f(x)$$

- Deep representations are supposed to encode meaning in vectors
 - If $(x-x')$ is difficult to compute, maybe we can do $(z-z')$?

Choices in Crafting Adversaries

$$\min_{x'} \quad \|x - x'\|$$
$$\text{s.t. } f(x') \neq f(x)$$

What is a small change?

Choices in Crafting Adversaries

How do we find the attack?

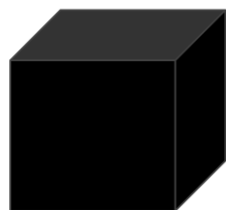
$$\begin{aligned} \min_{x'} & \|x - x'\| \\ \text{s.t.} & f(x') \neq f(x) \end{aligned}$$

Search: How do we find the attack?

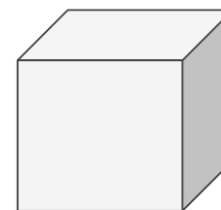
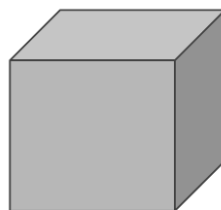
$$\min_{x'}$$

Even this is often unrealistic

Only access predictions
(usually unlimited queries)



Access probabilities



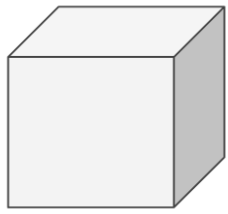
Full access to the model
(compute gradients)

Low Adversary's Knowledge High

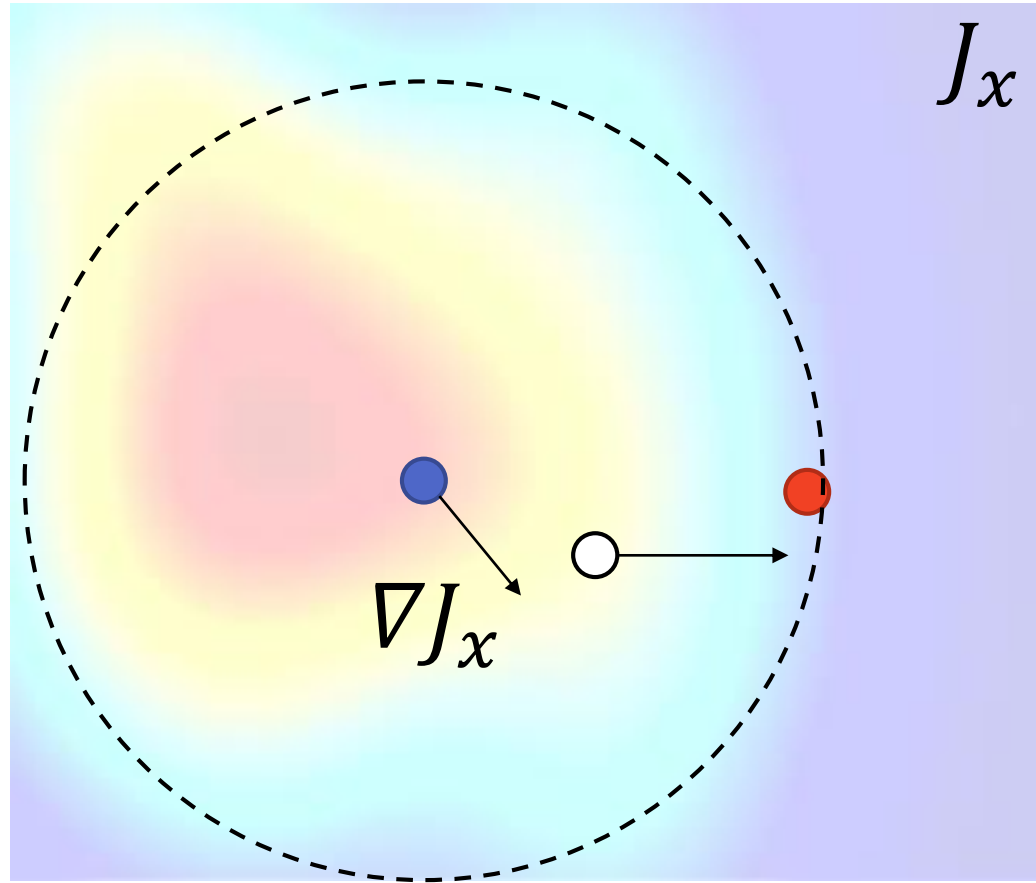
Create x' and test whether the model misbehaves

Create x' and test whether general direction is correct

Use the gradient to *craft* x'



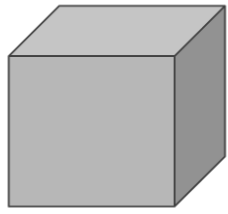
Search: Gradient-based



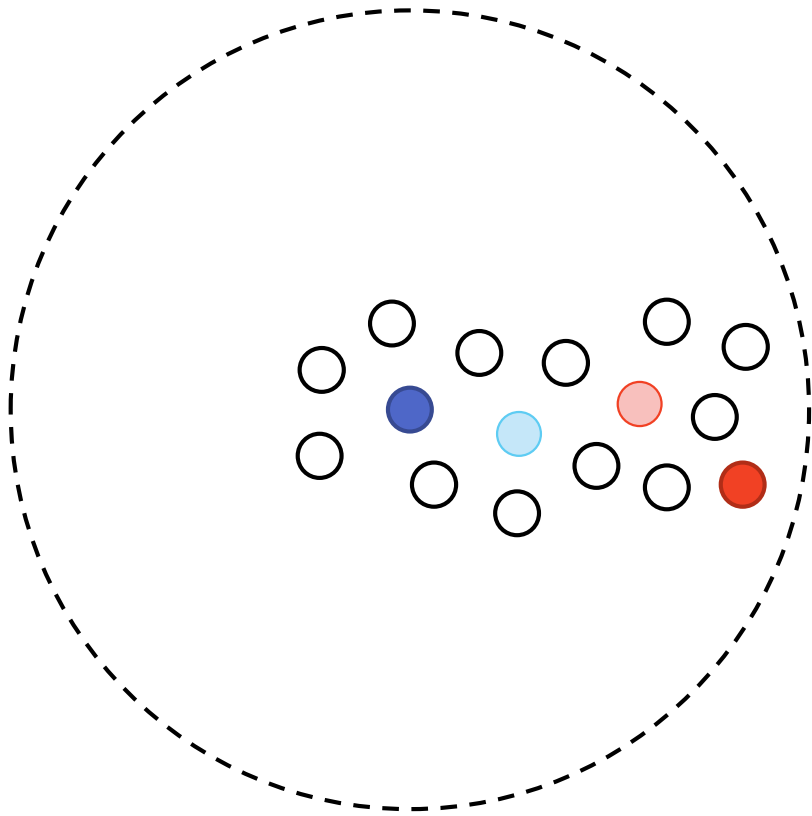
Or whatever the misbehavior is

1. Compute the gradient
2. Step in that direction (continuous)
3. Find the nearest neighbor
4. Repeat if necessary

Beam search over the above...



Search: Sampling

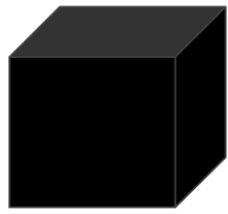


1. Generate local perturbations
2. Select ones that looks good
3. Repeat step 1 with these new ones
4. **Optional: beam search, genetic algo**

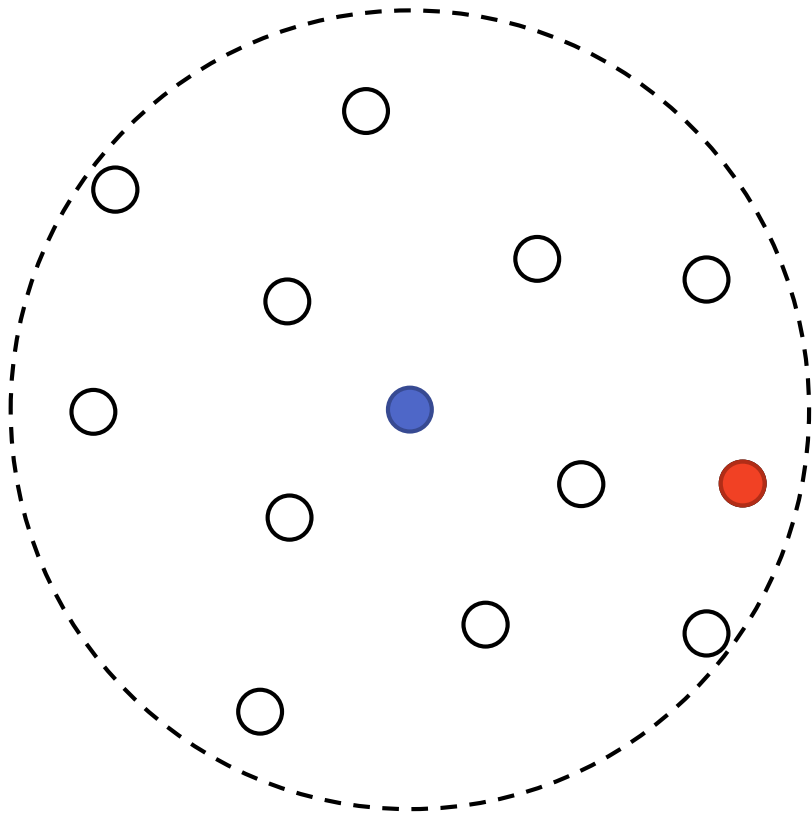
[Jia and Liang, EMNLP 2017]

[Zhao et al, ICLR 2018]

[Alzantot et. al. EMNLP 2018]



Search: Enumeration (Trial/Error)



1. Make some perturbations
2. See if they work
3. Optional: pick the best one

[Iyyer et al, NAACL 2018]

[Ribeiro et al, ACL 2018]

[Belinkov, Bisk, ICLR 2018]

Choices in Crafting Adversaries

How do we find the attack?

$$\begin{aligned} \min_{x'} \quad & \|x - x'\| \\ \text{s.t.} \quad & f(x') \neq f(x) \end{aligned}$$

Choices in Crafting Adversaries

$$\min_{x'} \|x - x'\|$$

s.t. $f(x') \neq f(x)$

What does it mean to misbehave?

Effect: What does it mean to misbehave?

Classification

Untargeted: any other class

Targeted: specific other class

$$\text{s.t. } f(x') \neq f(x)$$

Other Tasks

MT: Don't attack me! → ;No me ataques!

NER: Sameer PERSON is a prof at UCI ORG !

Loss-based: Maximize the loss on the example
e.g. perplexity/log-loss of the prediction

Property-based: Test whether a property holds
e.g. MT: A certain word is not generated
NER: No PERSON appears in the output

Evaluation: Are the attacks “good”?

- Are they Effective?
 - Attack/Success rate
- Are the Changes Perceivable? (Human Evaluation)
 - Would it have the same label?
 - Does it look natural?
 - Does it mean the same thing?
- Do they help improve the model?
 - Accuracy after data augmentation
- Look at some examples!

Review of the Choices

$$\begin{array}{l} \min_{x'} \\ \text{s.t. } f(x') \neq f(x) \end{array} \|x - x'\|$$

- **Change**

- Character level
- Word level
- Phrase/Sentence level

- **Effect**

- Targeted or Untargeted
- Choose based on the task

- **Search**

- Gradient-based
- Sampling
- Enumeration

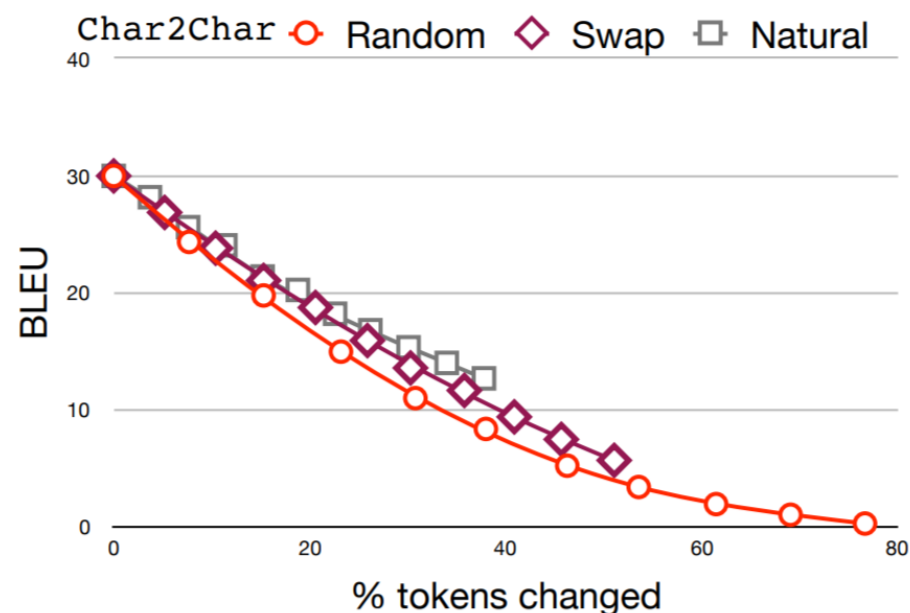
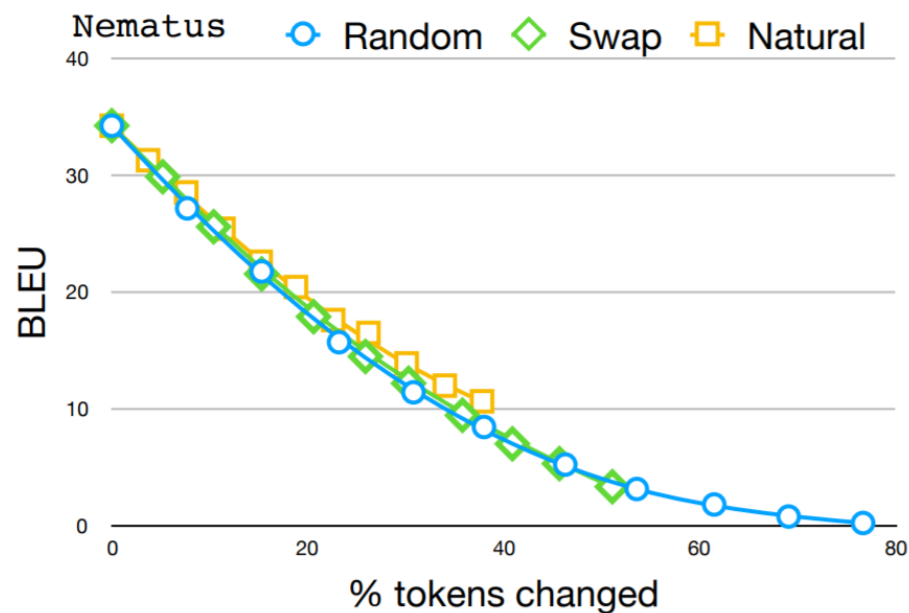
- **Evaluation**

Research Highlights

In terms of the choices that were made

Noise Breaks Machine Translation!

Change	Search	Tasks
Random Character Based	Passive; add and test	Machine Translation



Hotflip

Change	Search	Tasks
Character-based (extension to words)	Gradient-based; beam-search	Machine Translation, Classification, Sentiment

News Classification

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism.

57% **World**

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism.

95% **Sci/Tech**

Machine Translation

src	Das ist Dr. Bob Childs – er ist Geigenbauer und Psychotherapeut.
adv	Das ist Dr. Bob Childs – er ist Geigenbauer und Psy6hothearpeitut.
src-output	This is Dr. Bob Childs – he's a wizard maker and a therapist's therapist.
adv-output	This is Dr. Bob Childs – he's a brick maker and a psychopath.

Search Using Genetic Algorithms

Black-box, population-based search of natural adversary

Change	Search	Tasks
Word-based, language model score	Genetic Algorithm	Textual Entailment, Sentiment Analysis

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.*

Natural Adversaries

Change	Search	Tasks
Sentence, GAN embedding	Stochastic search	Images, Entailment, Machine Translation

Textual Entailment

Classifiers	Sentences	Label
Original	p : The man wearing blue jean shorts is grilling. h : The man is walking his dog.	Contradiction
Embedding	h' : The man is walking by the dog.	Contradiction → Entailment

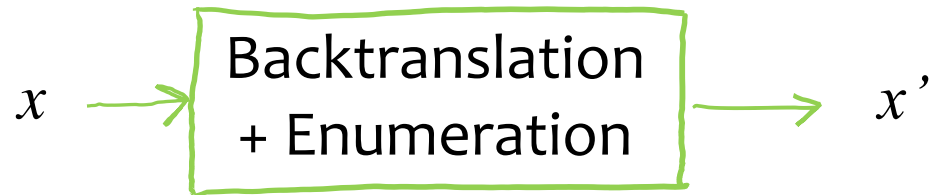
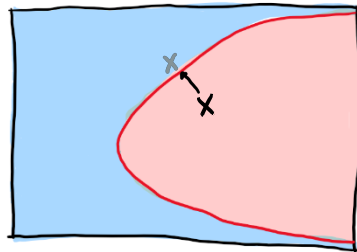


Source Sentence (English)	Generated Translation (German)
s : People sitting in a dim restaurant eating s' : People sitting in a living room eating .	Leute, die in einem dim Restaurant essen sitzen. Leute, die in einem Wohnzimmeressen sitzen. <i>(People sitting in a living room)</i>
s : Elderly people walking down a city street . s' : A man walking down a street playing	Ältere Menschen, die eine Stadtstraße hinuntergehen . Ein Mann, der eine Straße entlang spielt. <i>(A man playing along a street.)</i>

Semantic Adversaries

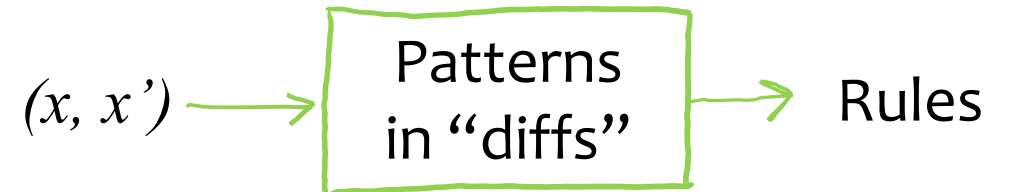
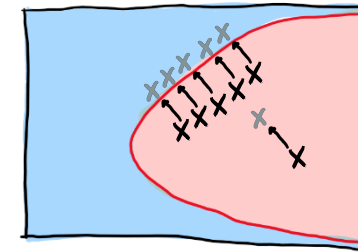
Change	Search	Tasks
Sentence via Backtranslation	Enumeration	VQA, SQuAD, Sentiment Analysis

Semantically-Equivalent Adversary (SEA)



What color is the tray?	Pink
What colour is the tray?	Green
Which color is the tray?	Green
What color is it ?	Green
How color is tray?	Green

Semantically-Equivalent Adversarial Rules (SEARs)



color → colour

Transformation Rules: VisualQA

SEAR	Questions / SEAs	f(x)	Flips
WP VBZ → WP's	What has What's been cut?	Cake Pizza	3.3%
What NOUN → Which NOUN	What Which kind of floor is it?	Wood Marble	3.9%
color → colour	What color colour is the tray?	Pink Green	2.2%
ADV is → ADV's	Where is Where's the jet?	Sky Airport	2.1%

Transformation Rules: SQuAD

SEAR	Questions / SEAs	f(x)	Flips
What VBZ → What's	What is What's the NASUWT?	Trade union Teachers in Wales	2%
What NOUN → Which NOUN	What resource Which resource was mined in the Newcastle area?	coal wool	1%
What VERB → So what VERB	What was So what was Ghandi's work called?	Satyagraha Civil Disobedience	2%
What VBD → And what VBD	What was And what was Kenneth Swezey's job?	journalist sleep	2%

Transformation Rules: Sentiment Analysis

SEAR	Reviews / SEAs	f(x)	Flips
movie → film	Yeah, the movie film pretty much sucked .	Neg Pos	2%
	This is not movie film making .	Neg Pos	
film → movie	Excellent film movie .	Pos Neg	1%
	I'll give this film movie 10 out of 10 !	Pos Neg	
is → was	Ray Charles is was legendary .	Pos Neg	4%
	It is was a really good show to watch .	Pos Neg	
this → that	Now this that is a movie I really dislike .	Neg Pos	1%
	The camera really likes her in this that movie.	Pos Neg	

Adding a Sentence

Change	Search	Tasks
Add a Sentence	Domain knowledge, stochastic search	Question Answering

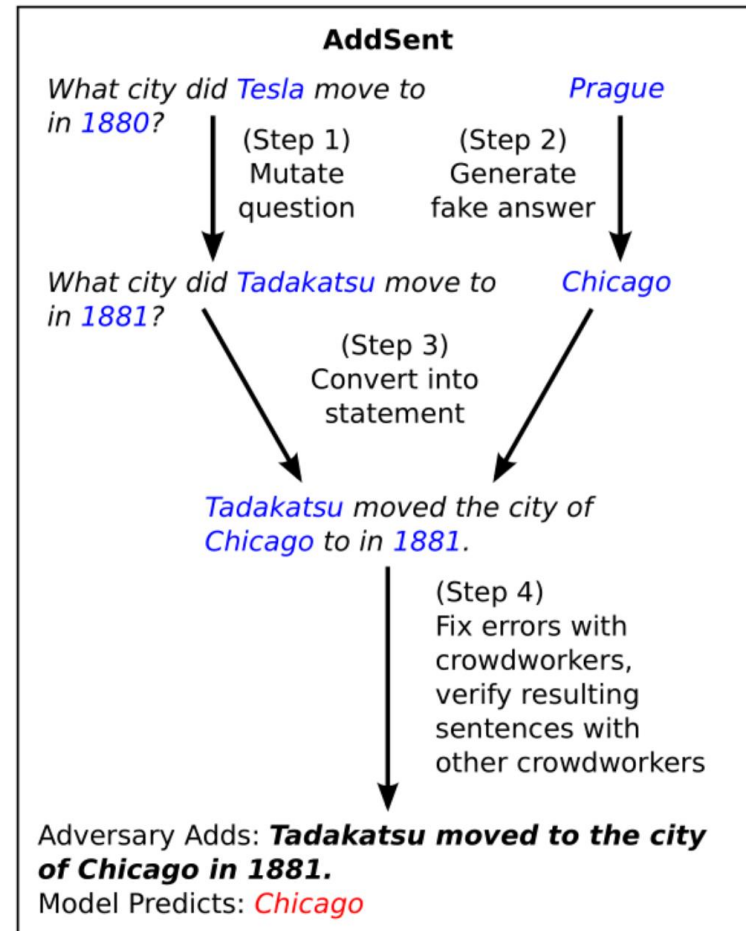
Article: Super Bowl 50

Paragraph: “*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*”

Question: “*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*”

Original Prediction: John Elway

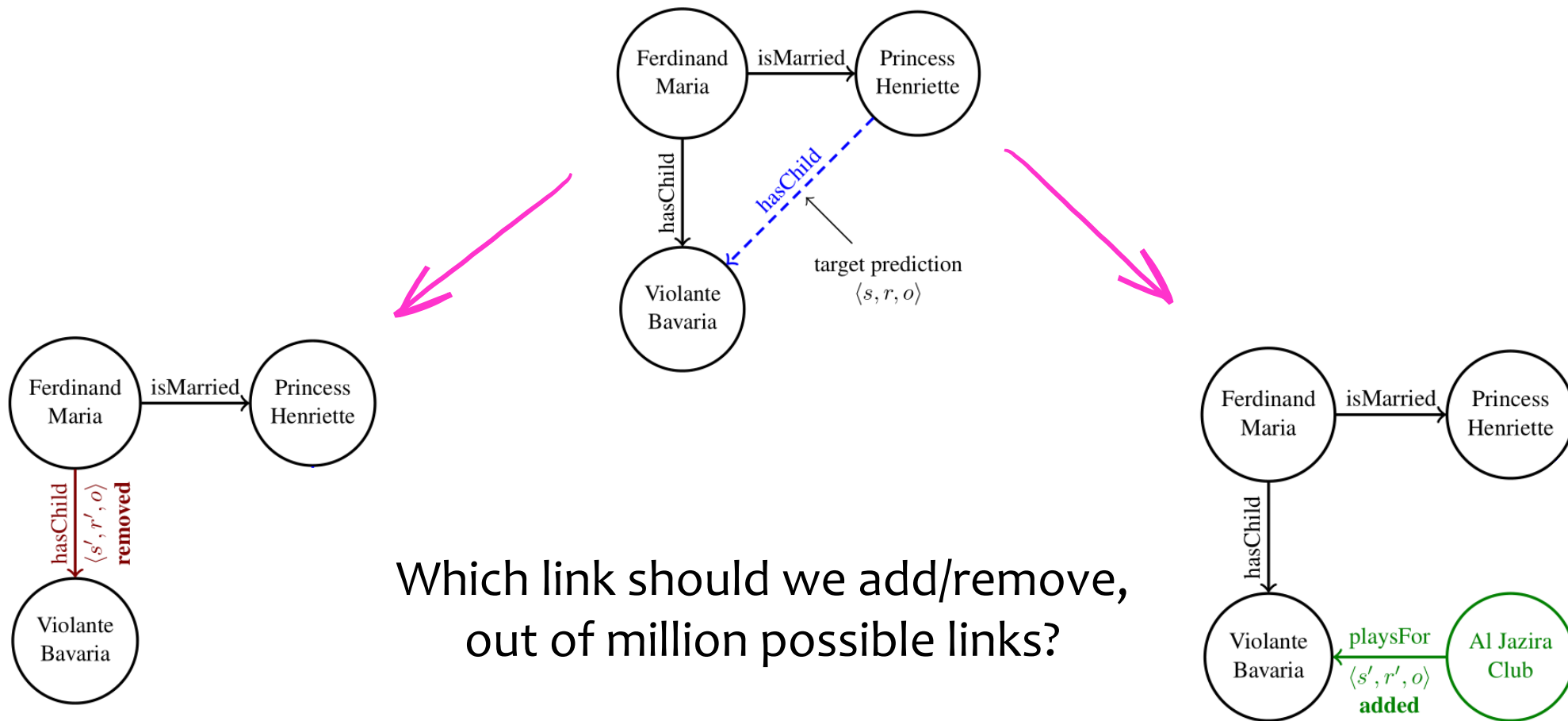
Prediction under adversary: Jeff Dean



Some Loosely Related Work

Use a broader notions of *adversaries*

CRIAGE: Adversaries for Graph Embeddings



“Should Not Change” / “Should Change”

How do dialogue systems behave when the inputs are perturbed in specific ways?

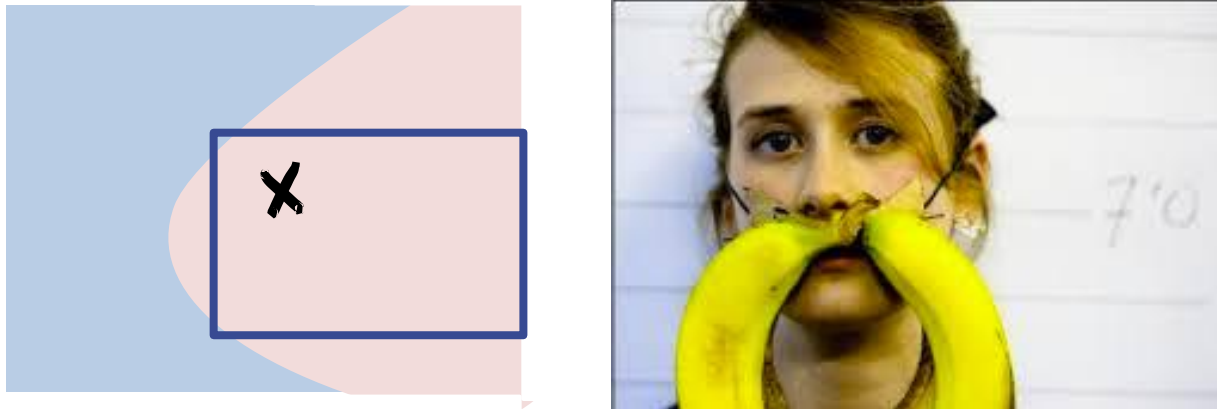
Should Not Change

- *like Adversarial Attacks*
- Random Swap
- Stopword Dropout
- Paraphrasing
- Grammatical Mistakes

Should Change

- *Overstability Test*
- Add Negation
- Antonyms
- Randomize Inputs
- Change Entities

Overstability: Anchors



Identify the conditions under which the classifier has **the same prediction**

Anchor

What is the mustache made of? banana

How **many** bananas are in the picture? 2

Overstability: Input Reduction

Remove as much of the input as you can
without changing the prediction!

SQUAD

Context In 1899, John Jacob Astor IV invested \$100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.

Original What did Tesla spend Astor's money on ?
Reduced did
Confidence 0.78 → 0.91

SNLI

Premise Well dressed man and woman dancing in the street
Original Two man is dancing on the street
Reduced dancing
Answer Contradiction
Confidence 0.977 → 0.706

VQA



Original What color is the flower ?
Reduced flower ?
Answer yellow
Confidence 0.827 → 0.819

Adversarial Examples for NLP

- Imperceivable changes to the input
- Unexpected behavior for the output
- Applications: security, evaluation, debugging

$$\begin{array}{l} \min_{x'} \quad \|x - x'\| \\ \text{s.t.} \quad f(x') \neq f(x) \end{array}$$

Challenges for NLP

- **Effect:** What is misbehavior?
- **Change:** What is a small change?
- **Search:** How do we find them?
- **Evaluation:** How do we know it's good?

Future Directions

- More realistic threat models
 - Give even less access to the model/data
- Defenses and fixes
 - Spell-check based filtering
 - Attack recognition: [Pruthi et al ACL 2019]
 - Data augmentation
 - Novel losses, e.g. [Zhang, Liang AISTATS 2019]
- Beyond sentences
 - Paragraphs, documents?
 - Semantic equivalency → coherency across sentences

References for Adversarial Examples in NLP

Relevant Work (roughly chronological)

- Sentences to QA: [Jia and Liang, EMNLP 2017] [link](#)
- Noise Breaks MT: [Belinkov, Bisk, ICLR 2018] [link](#)
- Natural Adversaries: [Zhao et al, ICLR 2018] [link](#)
- Syntactic Paraphrases: [Iyyer et al NAACL 2018] [link](#)
- Hotflip/Hotflip MT: [Ebrahimi et al, ACL 2018, COLING 2018] [link](#), [link](#)
- SEARs: [Ribeiro et al, ACL 2018] [link](#)
- Genetic Algo: [Alzantot et. al. EMNLP 2018] [link](#)
- Discrete Attacks: [Lei et al SysML 2019] [link](#)

Surveys

- Adversarial Attacks: [Zhang et al, arXiv 2019] [link](#)
- Analysis Methods: [Belinkov, Glass, TAACL 2019] [link](#)

More Loosely Related Work

- Anchors: [Ribeiro et al, AAAI 2018] [link](#)
- Input Reduction: [Feng et al, EMNLP 2018] [link](#)
- Graph Embeddings: [Pezeshkpour et. al. NAACL '19] [link](#)

Thank you!



Work with **Matt Gardner** and me

as part of

The Allen Institute for
Artificial Intelligence
in **Irvine**, CA



All levels: pre-docs, PhD interns, postdocs, and research scientists!

Sameer Singh

sameer@uci.edu

@sameer_

Sameersingh.org

UCI
nlp