# Deep Reinforcement Learning for Dialogue Generation

Artificial Intelligence Laboratory
오 주민

# Introduction

- 2016년 Cornell Archive(arXiv.org)에 게재

- Seq2Seq를 이용한 Dialogue generation model의 한계점
  ① 입력에 상관없이 빈번한 Dull reponse의 발생
  ② 대화가 무한한 반복에 빠지는 현상

A: Where are you going? (1)
B: I'm going to the restroom. (2)
A: See you later. (3)
B: See you later. (4)
A: See you later. (5)
B: See you later. (6)
...
...

A: how old are you? (1)
B: I'm 16. (2)
A: 16? (3)
B: I don't know what you are talking about. (4)
A: You don't know what you are saying. (5)
B: I don't know what you are talking about . (6)
A: You don't know what you are saying. (7)
...

# Introduction

- 목표

  ① 개발자가 원하는 형태의 Rewards 구성하고 학습하기

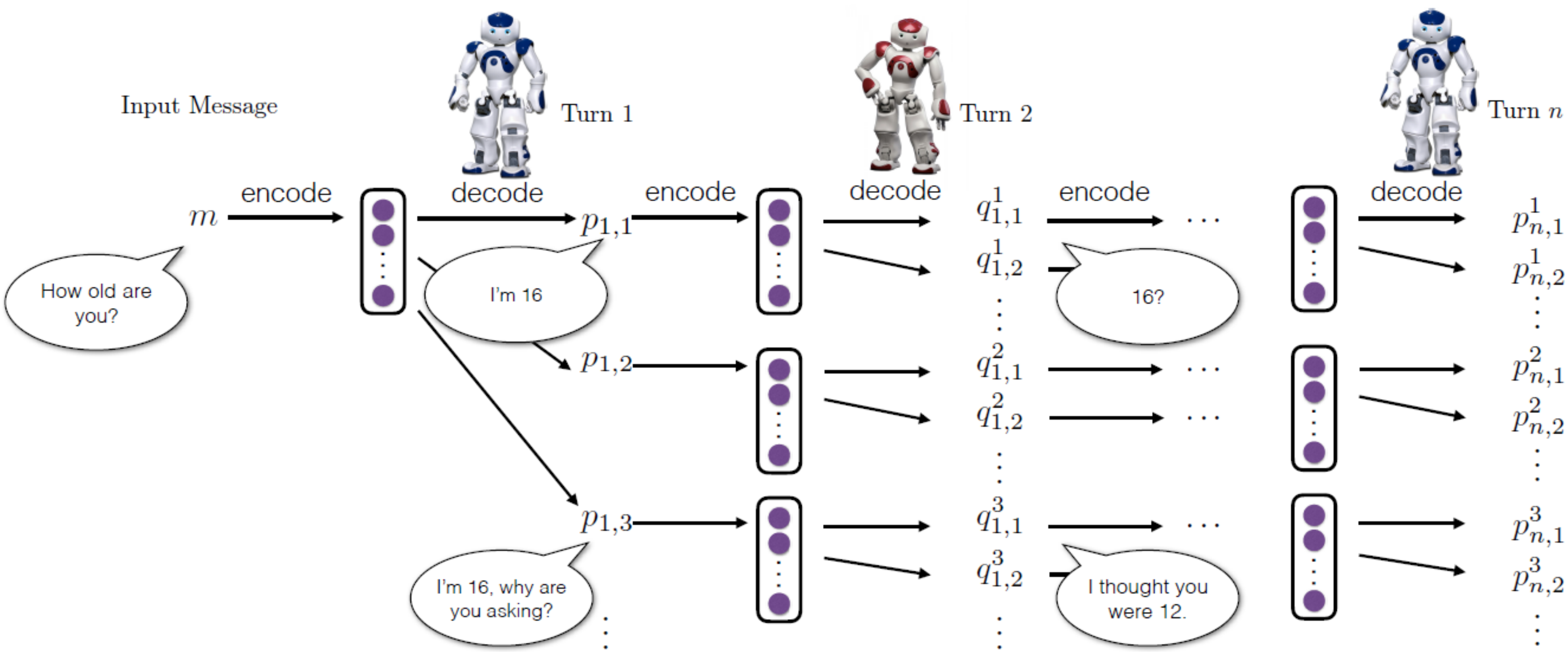
  ② 현재 진행되고 있는 대화에서 Context를 유지하는 답변 제시하기

➢ Neural reinforcement learning generation method를 통해 해결하자!

# Introduction

- Neural reinforcement learning generation method란?

  ① Seq2Seq 모델을 통해 구성된 Agent 2개를 서로 대화 상대라고 가정하고 각각의 Parameter를 Policy로써 학습하는 모델



# Introduction

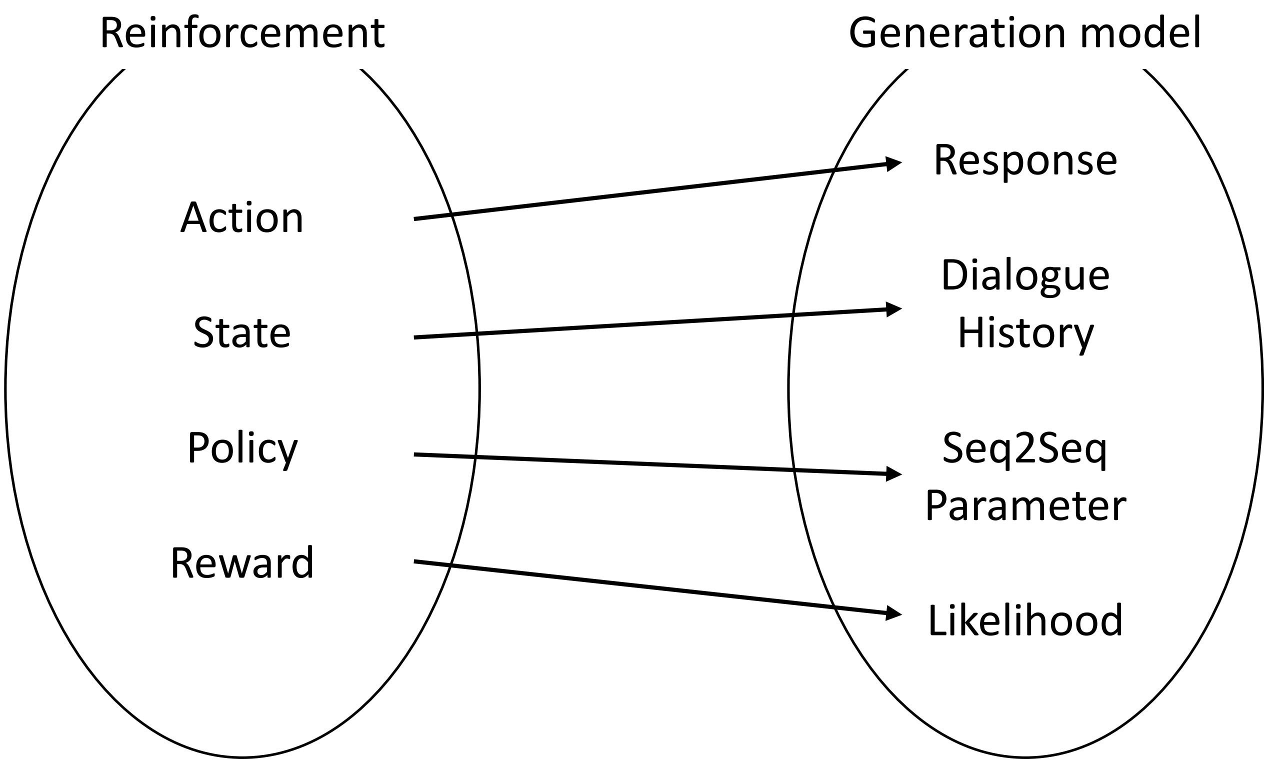
- Neural reinforcement learning generation method란?

  ① Seq2Seq 모델을 통해 구성된 Agent 2개를 서로 대화 상대라고 가정하고 각각의 Parameter를 Policy로써 학습하는 모델

  ② Seq2seq의 Semantic meaning 추출과 Reinforcement learning의 long-term goal 최적화의 강점을 합친 방법

# Model

- 강화학습의 요소가 Generation model의 요소로 표현 가능!



# Method

- Agent : $p$, $q$
  *ex)* $p_1, q_1, p_2, q_2, \ldots, p_i, q_i$

- Action(utterance to generate) : $a$

- State(two turn dialogue history, Input) : $[p_i, q_i]$

- Policy(foam of Seq2Seq) : $p_{\mathrm{RL}}(p_{i+1} | p_i, q_i)$

# Model

- Reward

① Ease of Answering

$$r_1 = -\frac{1}{N_{\mathbb{S}}} \sum_{s \in \mathbb{S}} \frac{1}{N_s} \log p_{\text{seq2seq}}(s|a)$$

$\mathbb{S}$ : Dull response (ex."I don't know what you are talking about")

# Model

- Reward

② Information Flow

$$r_2 = -\log\cos(h_{p_i}, h_{p_{i+1}}) = -\log \cancel{\cos} \frac{h_{p_i} \cdot h_{p_{i+1}}}{\|h_{p_i}\|\|h_{p_{i+1}}\|} \tag{2}$$

$h_{p_i}$ : Encoder state of $p_i$

# Model

- Reward

③ Semantic Coherence

$$r_3 = \frac{1}{N_a} \log p_{\text{seq2seq}}(a|q_i, p_i) + \frac{1}{N_{q_i}} \log p_{\text{seq2seq}}^{\text{backward}}(q_i|a)$$

$p_{\text{seq2seq}}^{\text{backward}}(q_i|a)$ : pre-trained Seq2Seq Model with source and target swapped

# Model

- Reward

④ Total Reward

$$r(a, [p_i, q_i]) = \lambda_1 r_1 + \lambda_2 r_2 + \lambda_3 r_3$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 1.$$

# Simulation

- Pre-training
  - RL Model을 pre-trained Seq2Seq Model로 initialize 했을 때,
    Dull response가 나타날 확률이 높음
  - Pre-trained Seq2Seq Model로는 [$p_i$, $q_i$] 를 통해 Response Candidate List $A$를 생성
    ( $A = \{\hat{a} | \hat{a} \sim p_{RL}\}$ ) -> $\hat{a}$ 은 Response Candidate
  - $\hat{a}$을 통해 Mutual Information을 계산해서 이를 통해 Pre-train 하자!

# Simulation

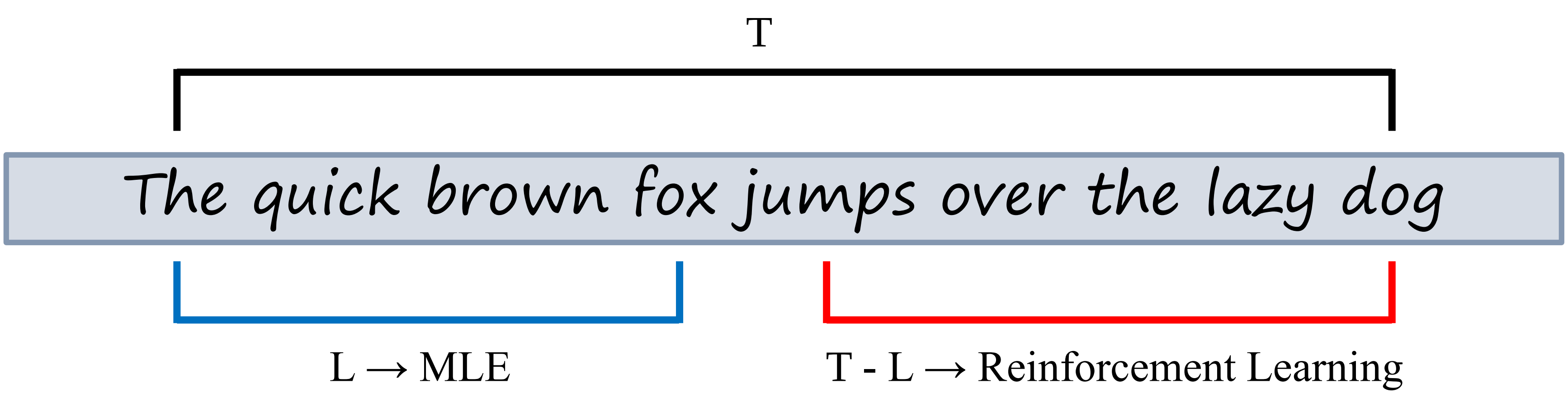
- Mutual Information
  - Semantic Coherence(Reward $r_3$) 에서…

$$r_3 = \frac{1}{N_a} \log p_{\text{seq2seq}}(a|q_i, p_i) + \frac{1}{N_{q_i}} \log p_{\text{seq2seq}}^{\text{backward}}(q_i|a)$$

$m(\hat{a}, [p_i, q_i])$ :
이전까지의 대화를 통해 $\hat{a}$을 생성할 수 있는가? +
$\hat{a}$을 통해서 이전의 대화를 예측할 수 있는가? 를
담고 있는 정보

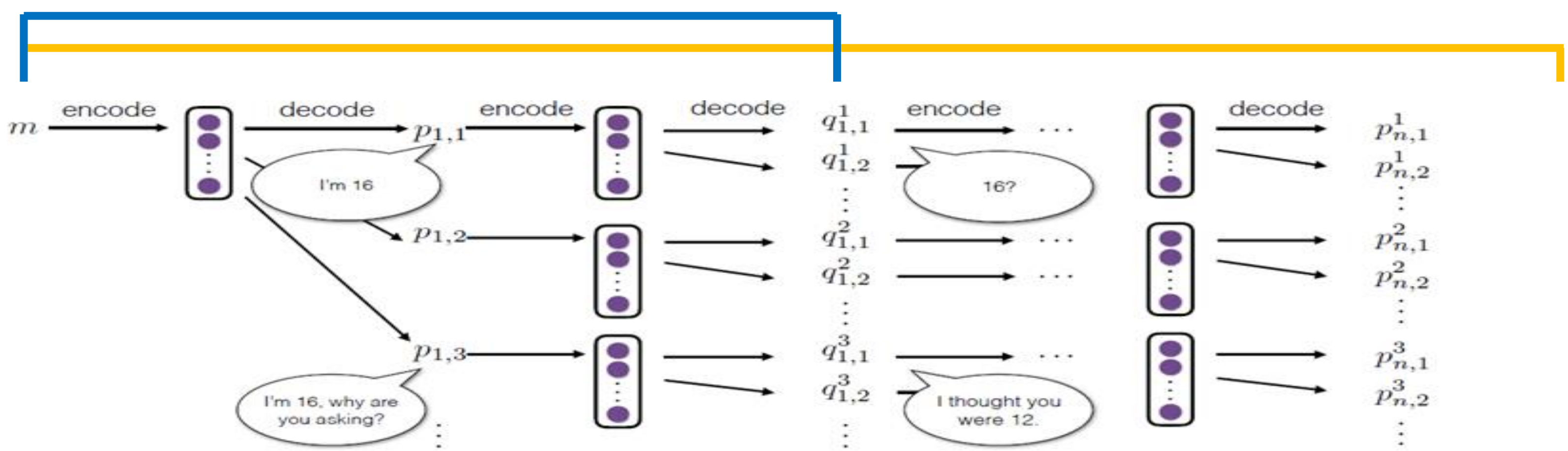
## Simulation

- Pre-training with Mutual Information
  - Pre-train Reward : $J(\theta) = \mathbb{E}[m(\hat{a}, [p_i, q_i])]$
  - Gradient by likeligood : $\nabla J(\theta) = m(\hat{a}, [p_i, q_i]) \nabla \log p_{RL}(\hat{a} | [p_i, q_i])$
  - Curriculum Learning



# Simulation

- Optimazation
  - Expected future reward : $J_{RL}(\theta) = \mathbb{E}_{p_{RL}(a_{1:T})}[\sum_{i=1}^{i=T} R(a_i, [p_i, q_i])]$
  - Reward gradient : $\nabla J_{RL}(\theta) \approx \sum_{i} \nabla \log p(a_i|p_i, q_i) \sum_{i=1}^{i=T} R(a_i, [p_i, q_i])$
  - Curriculum Learning (Different with previous Section)



## Experiment Result

- Automatic Evaluation - Traditional Method
  - BLEU Score : n-gram을 이용하여 측정하는 방식

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

- Example:
  - Reference: “the Iraqi weapons are to be handed over to the army within two weeks”
  - MT output: “in two weeks Iraq’s weapons will give army”
- BLUE metric:
  - 1-gram precision: 4/8
  - 2-gram precision: 1/7
  - 3-gram precision: 0/6
  - 4-gram precision: 0/5
  - BLEU score = 0 (weighted geometric average)

# Experiment Result

- Automatic Evaluation - Traditional Method
  - Perplexity : 어떤 확률 모델이 실제로 관측되는 값을 얼마나 잘 예측하였는가?
  - $$Perplexity = e^{Entropy(p)}$$
    $$= e^{loss}$$
- But…
  - Syntactic 한 유사도는 측정할 수 있으나 Semactic 유사도 측정 불가
  - Long-term success의 척도로 적합하지 않음

# Experiment Result

- Automatic Evaluation

  - Length of the dialogue : 한 Agent가 Dull response를 말할 때 까지의 turn

| Model | # of simulated turns |
|---|---|
| SEQ2SEQ | 2.68 |
| mutual information | 3.40 |
| RL | 4.48 |

  - Diversity : Generated response에서 distinct한 uni-gram과 bi-gram의 갯수

| Model | Unigram | Bigram |
|---|---|---|
| SEQ2SEQ | 0.0062 | 0.015 |
| mutual information | 0.011 | 0.031 |
| RL | 0.017 | 0.041 |

# Experiment Result

- Automatic Evaluation

  - Human evaluation (RL Model vs Mutual Information Model)

| Setting | RL-win | RL-lose | Tie |
|---|---|---|---|
| single-turn general quality | 0.40 | 0.36 | 0.24 |
| single-turn ease to answer | 0.52 | 0.23 | 0.25 |
| multi-turn general quality | 0.72 | 0.12 | 0.16 |

| Input Message | Mutual Information Model | RL |
|---|---|---|
| How old are you? | I'm 16. | I'm 16, why are you asking? |
| What's your full name? | I have no idea. | What's yours? |
| I don't want to go home tonight. | Really? | Why? |
| Do you have any feelings for me? | I don't know what you are talking about. | I don't want to hurt your feelings. |
| How much time do you have here? | Not long enough. Sorry, sir. | Ten seconds. |
| Shall we get started? | Of course! | Yes. We've got a lot of work to do here. |
| Do you play football? | No, i don't | Yes. I love football! |
| We'd have to talk to him. | I mean, he's a good guy | About what ? |
| How come you never say it? | Because I don't want to hurt you. | I don't think it's a good idea to say it. |

# Experiment Result

- Limitation
  - Dialogue with multi-cycle problem

| |
|---|
| A: What's your name ? |
| B: Daniel. |
| A: How old are you ? |
| B. Twelve. What's your name ? |
| A. Daniel. |
| B: How old are you ? |
| A: Twelve. What's your name ? |
| B: Daniel. |
| A: How old are you ? |
| B ... |

# Conclusion

- 의미 있는 답변을 하면서 Dialouge의 형태를 유지하는 모델을 제시
- Long-term influence를 잘 반영하는 모델
- Heuristic에 따라 얼마나 다양하고 의미있는 답변을 할 지 결정된다.
    - 여기서 Heuristic은 아마 training dataset과 pre-train dataset을 의미