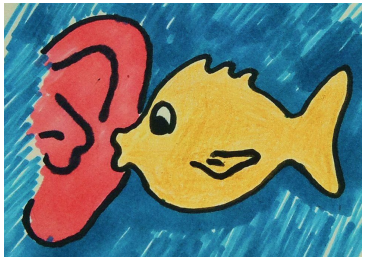


강화학습 챗봇

Dialogue Policy Optimization

바벨피쉬 김성동

Who am I?

이름 : 김성동

github: @DSKSD

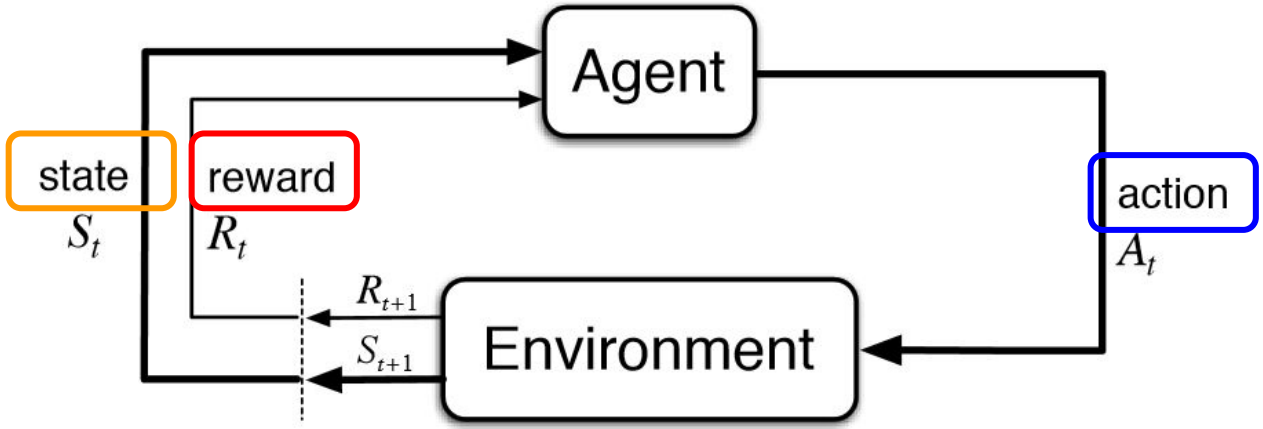
연락처 : tjdehd1222@gmail.com

* 숭실대 경영/글로벌미디어 전공
* 텍스트팩토리 NLP Engineer
* Interest : Deep learning, NLP, Reinforcement Learning

Contents

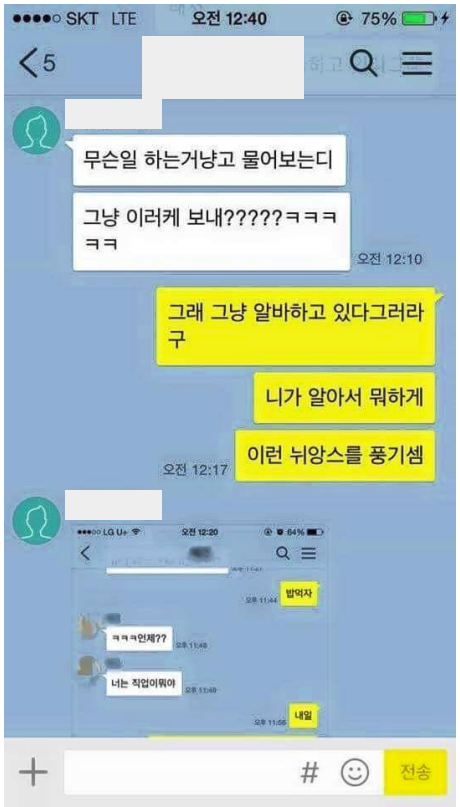
What is a Dialogue Policy

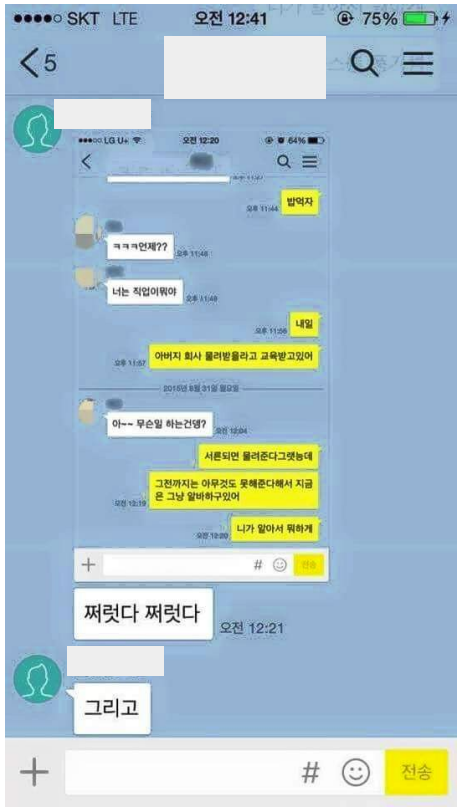
현재까지 대화를 미루어 봤을 때, 챗봇이 **어떤 대답**을 해야 **가치**를 최대화 할 수 있을까?

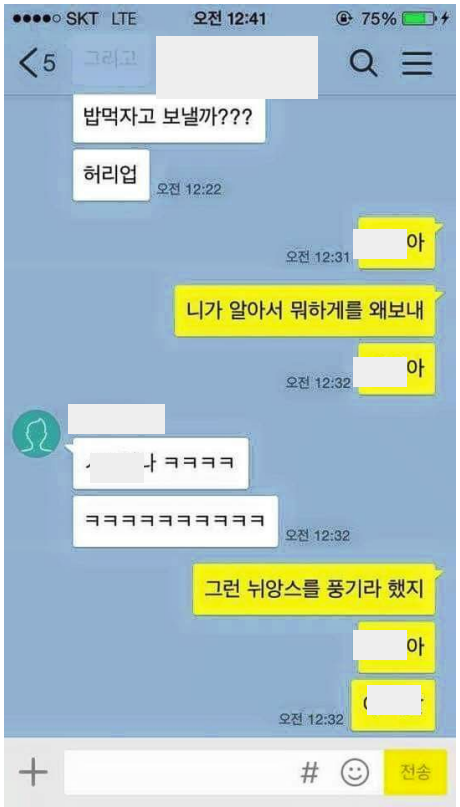


What is a Dialogue Policy

좋은 대화의 정책을 찾는다..?
때론 인간도 어렵다







What is a Dialogue Policy

대화 유형	정책
목적을 가진 대화(Task Oriented)	상대방이 요청한 일을 처리한다.
목적이 없는 잡담(Chit-Chat)	상대방의 말에 적절한 반응을 한다.
외부 지식을 기반으로한 질문/대답(QA)	상대방의 질문에 맞는 답을 알려준다.
…	…

대화의 유형과 목적에 따라 다른 전략(정책)을 취해야 한다

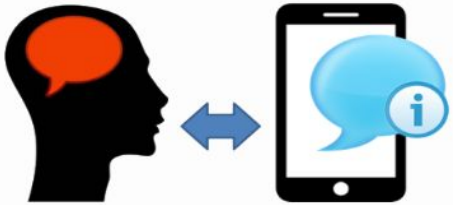
What is a Dialogue Policy

The bAbI project

SIGdial
Special Interest Group on Discourse and Dialogue

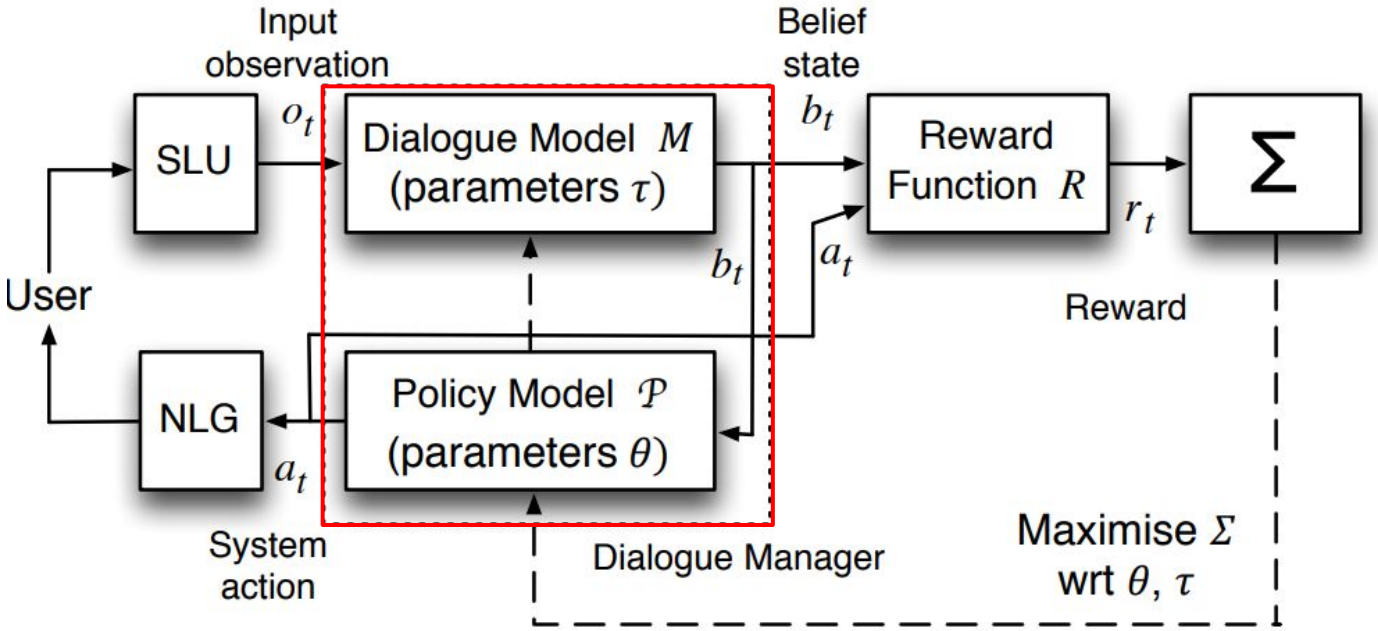
Maluuba
A Microsoft company

DSTC6
Dialog System Technology Challenges
Long Beach, USA, December 10, 2017



SQuAD
The Stanford Question Answering Dataset

What is a Dialogue Policy



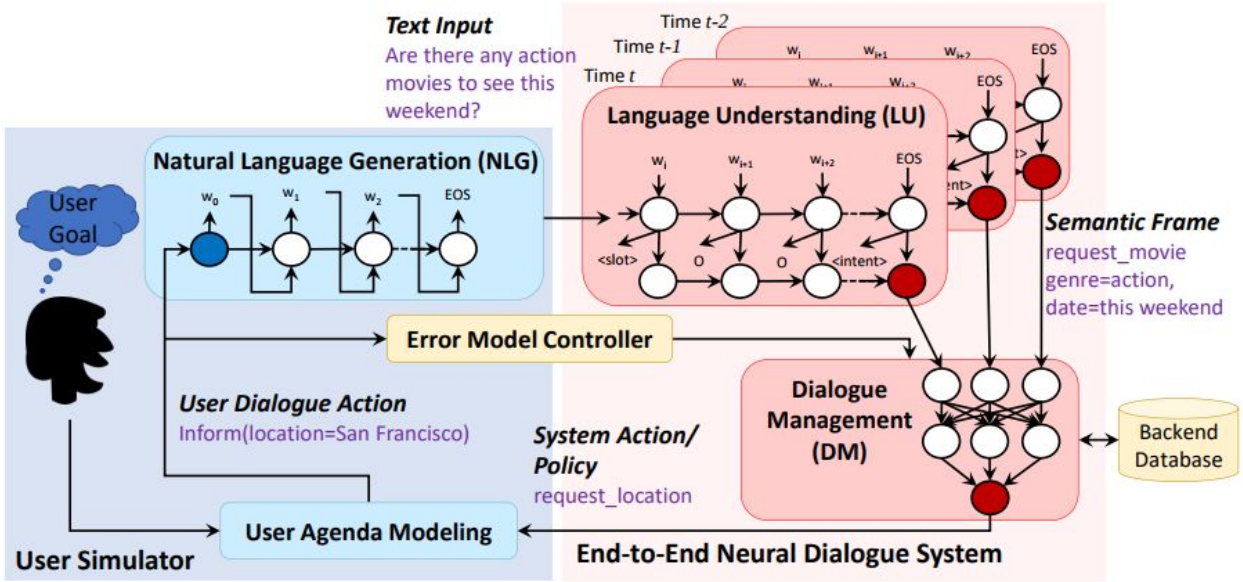
POMDP-based spoken dialogue system

http://mi.eng.cam.ac.uk/~sjy/papers/ygtw13.pdf

Types of RL in Dialogue System

End-to-End Task-Completion Neural Dialogue Systems(Li et al, 2017)

Closed Domain(Goal-Oriented) Dialogue(Task-Completion)



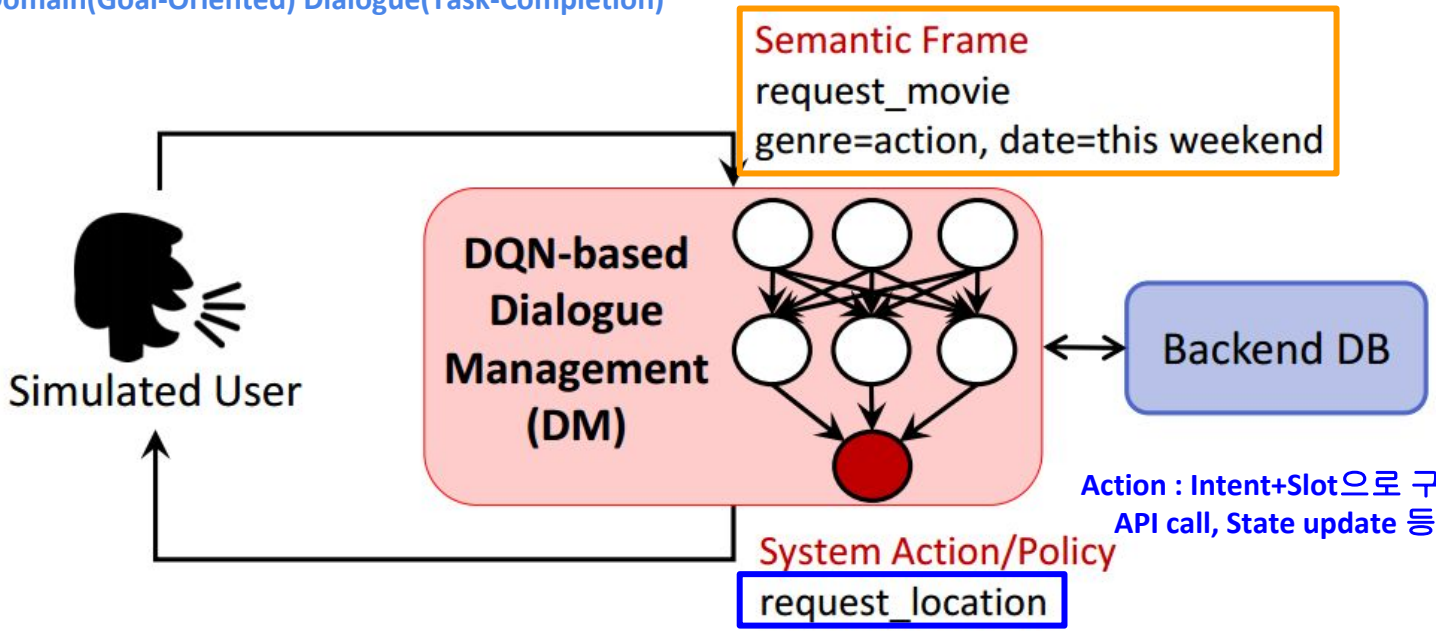
https://arxiv.org/pdf/1703.01008.pdf

Types of RL in Dialogue System

End-to-End Task-Completion Neural Dialogue Systems(Li et al, 2017)

State : 이전 봇의 Action,
현 유저의 Action,
DB 검색결과,
대화 히스토리 등

Closed Domain(Goal-Oriented) Dialogue(Task-Completion)



Action : Intent+Slot으로 구성된 Dialog act,
API call, State update 등의 System act

https://arxiv.org/pdf/1703.01008.pdf

Types of RL in Dialogue System

End-to-End Task-Completion Neural Dialogue Systems(Li et al, 2017)

Rule-based Agent
User: What is the start time for 10 cloverfield lane? Agent: What movie are you interested in? User: I want to watch at 10 cloverfield lane. Agent: What time would you like to see it? User: What time is it playing tomorrow? I need 3 tickets. Agent: What city you would like? User: I do not care. Agent: What date would you like to watch it? User: I want to set it up tomorrow. Agent: Which theater would you like? User: Which theater is available? Agent: How many tickets do you need? User: I want 3 tickets please! Agent: Okay, I have booked 3 tickets for 10 cloverfield lane tomorrow. User: Thank you. Agent: Thank you.
Failure: Agent does not answer all the questions (starttime, theater) from user.

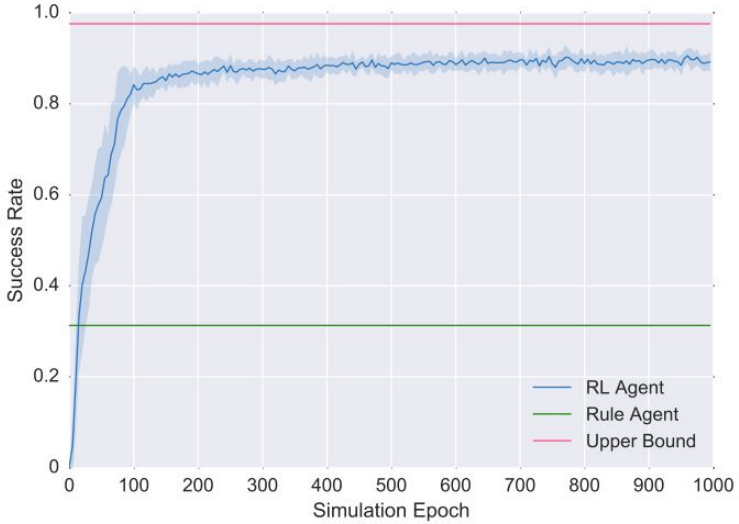
RL Agent
User: Which theater can I book 3 tickets for 10 cloverfield lane? Agent: What time would you like to see it? User: Which theater and start time are available tomorrow? Agent: 11:45am is available. User: Which theater is available? Agent: regal la live stadium 14 is available. User: Could you help me to book the tickets? Agent: Okay - I was able to book 3 tickets for you to see 10 cloverfield lane at regal la live stadium 14 at 11:45am tomorrow. User:Thank you. Agent: Thank you.
Success

Reward : Success Rate, # of turns

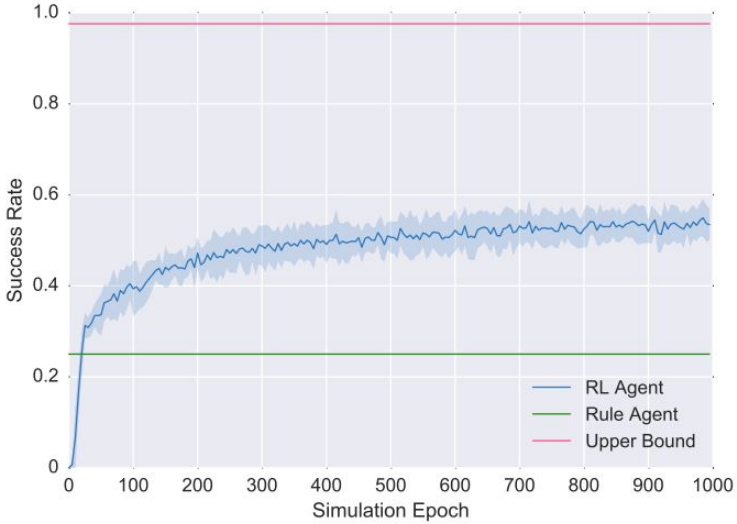
https://arxiv.org/pdf/1703.01008.pdf

Types of RL in Dialogue System

End-to-End Task-Completion Neural Dialogue Systems(Li et al, 2017)



(a) Frame-level semantics for training

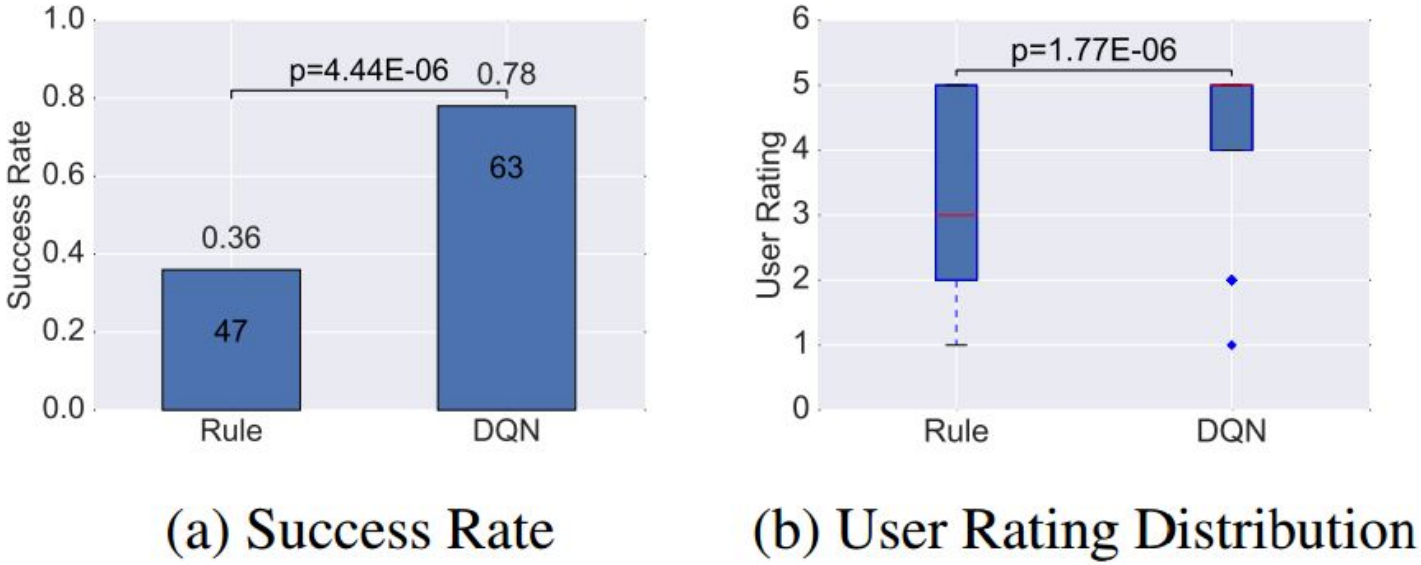


(b) Natural language for end-to-end training

https://arxiv.org/pdf/1703.01008.pdf

Types of RL in Dialogue System

End-to-End Task-Completion Neural Dialogue Systems(Li et al, 2017)



https://arxiv.org/pdf/1703.01008.pdf

Types of RL in Dialogue System

Deep Reinforcement Learning for Dialogue Generation(Li et al, 2016)

Open Domain Dialogue(Chit-Chat)

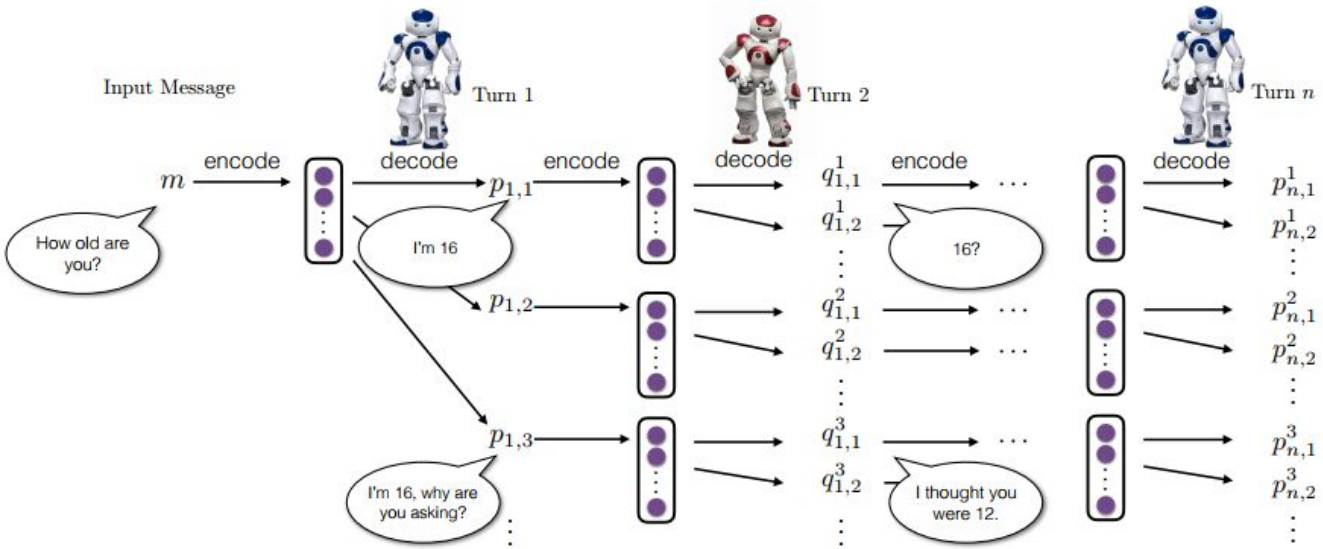


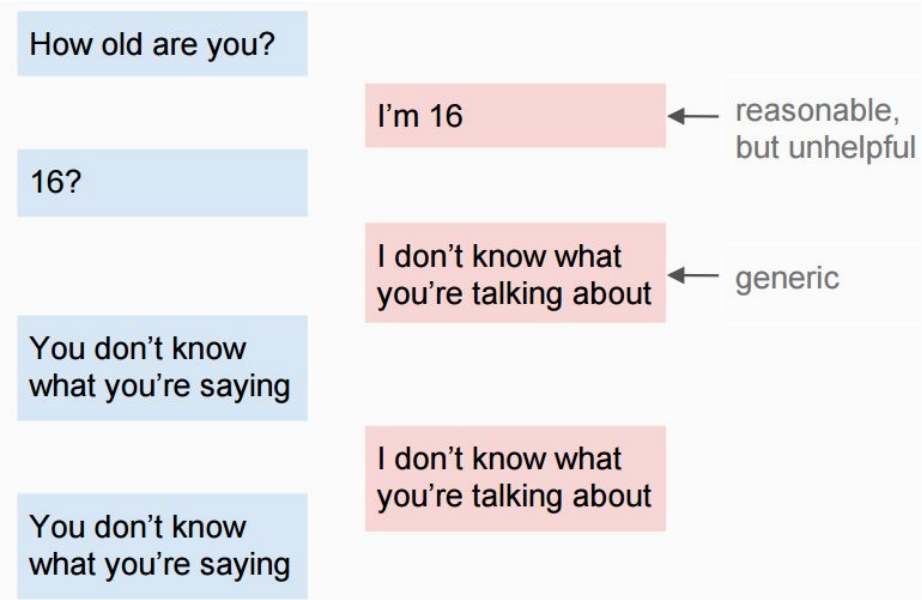
Figure 1: Dialogue simulation between the two agents.

https://arxiv.org/pdf/1606.01541.pdf

Types of RL in Dialogue System

Deep Reinforcement Learning for Dialogue Generation(Li et al, 2016)

Open Domain Dialogue(Chit-Chat)



Problems with Seq2Seq

Probable response != Good response

https://arxiv.org/pdf/1606.01541.pdf

Types of RL in Dialogue System

Deep Reinforcement Learning for Dialogue Generation(Li et al, 2016)

Open Domain Dialogue(Chit-Chat)

좋은(가치 있는) 대답의 조건 : Reward

- **Reasonable** — p(response|input) is high according to seq2seq model
- **Nonrepetitive** — similarity between response and previous messages is low
- **Easy to answer** — p("i don't know"|response) is low

https://arxiv.org/pdf/1606.01541.pdf

Types of RL in Dialogue System

Policy Gradient(REINFORCE)

Deep Reinforcement Learning for Dialogue Generation(Li et al, 2016)

$$J_{RL}(\theta) = \mathbb{E}_{p_{RL}(a_{1:T})}[\sum_{i=1}^{i=T} R(a_i, [p_i, q_i])]$$

Baseline mutual information model (Li et al. 2015)	Proposed reinforcement learning model
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: Where are you going? (1) B: I'm going to the police station. (2) A: I'll come with you. (3) B: No, no, no, no, you're not going anywhere. (4) A: Why? (5) B: I need you to stay here. (6) A: I don't know what you are talking about. (7) ...
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) ...	A: How old are you? (1) B: I'm 16. Why are you asking? (2) A I thought you were 12. (3) B: What made you think so? (4) A: I don't know what you are talking about. (5) B: You don't know what you are saying. (6) ... ...

State : 이전 2번의 대화

Action : 생성된 문장 그 자체

https://arxiv.org/pdf/1606.01541.pdf

Types of RL in Dialogue System

Deep Reinforcement Learning for Dialogue Generation(Li et al, 2016)

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

Length of the dialogue

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

Diversity

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

Human Evaluation

https://arxiv.org/pdf/1606.01541.pdf

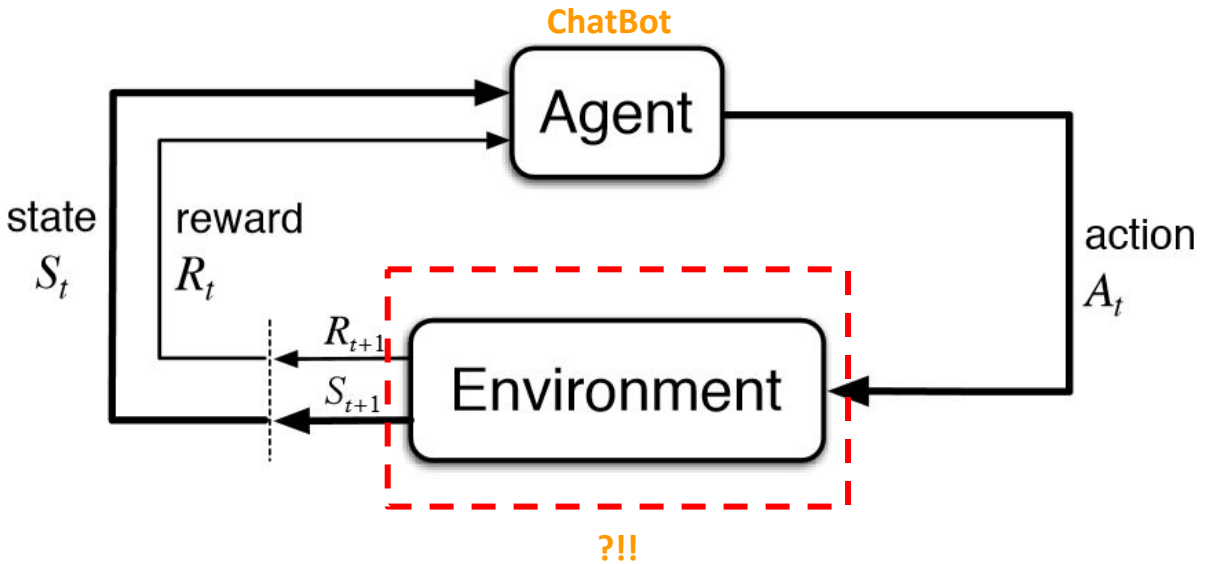
Types of RL in Dialogue System

Summary

Type of Bots	State	Action	Reward
Social ChatBots	Chat history	System Response	# of turns maximized; Intrinsically motivated reward
InfoBots (interactive Q/A)	User current question + Context	Answers to current question	Relevance of answer; # of turns minimized
Task-Completion Bots	User current input + Context	System dialogue act w/ slot value (or API calls)	Task success rate; # of turns minimized

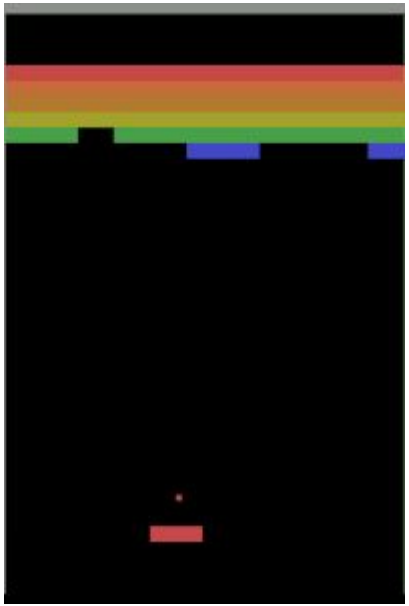
https://www.csie.ntu.edu.tw/~yvchen/doc/DeepDialogue_Tutorial.pdf

User Simulator

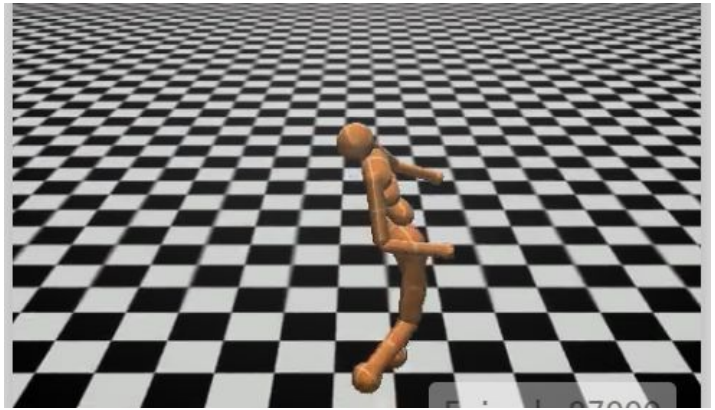


실제 사람들과 상호작용하기엔 비용이 많이 든다...

User Simulator



OpenAI



강화학습이 이렇게까지 대중화 될 수 있었던 것은 OpenAI gym과
같은 Environment 플랫폼이 있었기 때문!!

User Simulator

ParlAI : A Dialog Research Software Platform

facebook

QA datasets
SQuAD
bAbI tasks
MCTest
SimpleQuestions
WikiQA, WebQuestions,
WikiMovies, MTurkWikiMovies
MovieDD (Movie Recommendations)

Sentence Completion (Cloze)
QACNN
QADailyMail
CBT
BookTest

Goal-Oriented Dialog
bAbI Dialog tasks
Dialog-based Language Learning bAbI
Dialog-based Language Learning Movie
MovieDD-QARecs dialog

Dialog Chit-Chat
Ubuntu
Movies SubReddit
Cornell Movie
OpenSubtitles

Visual QA/Dialog
VQA

http://parl.ai/

User Simulator

ParlAI : A Dialog Research Software Platform

Observation/action `dict`
Passed back and forth between agents & environment.

Contains:

.text	*text of speaker(s)*
.id	*id of speaker(s)*
.reward	*for reinforcement learning*
.episode_done	*signals end of episode*

For supervised dialog datasets:

.label	
.label_candidates	*multiple choice options*
.text_candidates	*ranked candidate responses*
.metrics	*evaluation metrics*

Other media:

.image	*for VQA or Visual Dialog*

매우 General한 Form을 사용

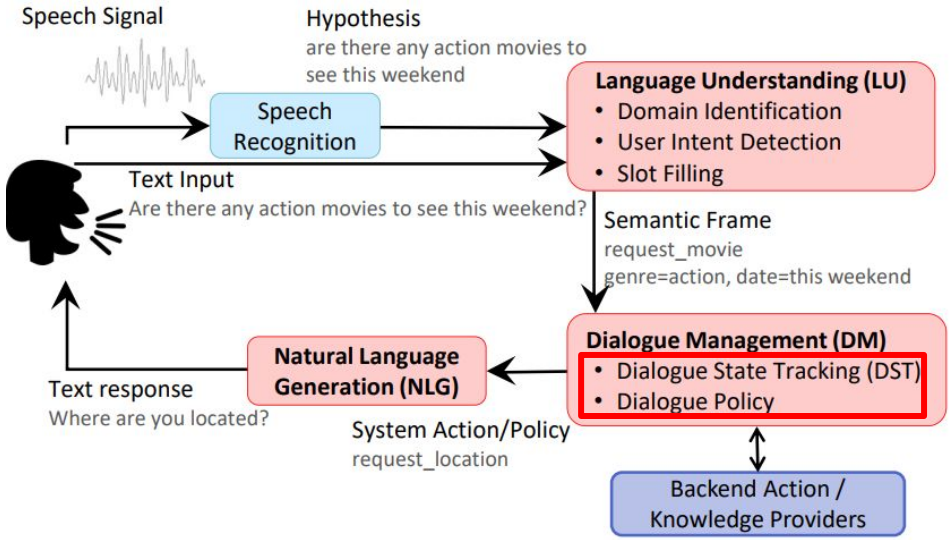
=> 장점이 될 수도 단점이 될수도..

특히, 이미 존재하는 데이터셋을 바탕으로
시뮬레이션이 이루어짐.
(원하는 도메인의 데이터를 수집/태깅을 통해
구성해야 확장 가능.)

http://parl.ai/

User Simulator

User Simulation for Task-Completion Dialogues



훈련된 NLU, NLG, DM(DST) 등의 모듈을 사용

Agenda based User Simulator : Goal을 초기화하고 Rule에 의해 Inform한다.

휴리스틱한 Speech Action Rule 사용
ex. 봇이 요청하면, 알려줘라
ex. 봇이 정보를 제공하면, 감사를 표하라 등

매 대화 턴마다 -1의 보상,
TASK 성공 시 10의 보상

https://github.com/MiuLab/TC-Bot

User Simulator

Build your own Dialogue Environment

```
from environment_class import Conv_Env
import MyAgent

env = Conv_Env()
agent = MyAgent()

state = env.reset()
is_done = env.is_done
rewards=[]
while is_done==False:

    a = MyAgent(state)
    _,state,reward,is_done = env.step(a)
    rewards.append(reward)
```

gym과 유사한 형태로 구성

```
유저 : 허기지넹...
봇 : 다음 중 선택해주세요. 맛집추천 레시피추천 ( BACTION:Request_Slot STYPE:p_task_FOOD )
유저 : 맛집추천 ( Inform_Simple )
봇 : 음식이름을 알려주세요 ( BACTION:Request_Slot STYPE:food-name )
유저 : 맛집추천 어떻게 해? ( Jump_Task )
봇 : 음식이름을 알려주세요 ( BACTION:Request_Slot STYPE:food-name )
유저 : 치즈버거로 ( Inform_Simple )
봇 : 지역을 알려주세요 ( BACTION:Request_Slot STYPE:region )
유저 : 청주시 흥덕구 ( Inform_Simple )
봇 : 가격대를 알려주세요 ( BACTION:Request_Slot STYPE:price )
유저 : 80000 이하로 ( Inform_Simple )
봇 : 다음 중 선택해주세요. 1 3 0 4 ( BACTION:Request_Slot STYPE:p_list_FOOD )
유저 : 더 보여죠~~ ( Inform_Simple )
봇 : 다음 중 선택해주세요. 1 2 3 0 ( BACTION:Request_Slot STYPE:p_list_FOOD )
유저 : 2로 ( Inform_Simple )
봇 : 여기있습니다. ( BACTION:Inform_Last_Info INFO:여기있습니다. )
유저 : 고마워 수고행 ( Thanks_Quit )
```

시뮬레이션 샘플

Challenges

1. Generic RL method for all bot categories
2. End2End Trainable model
3. Competitive Self-Play between Multiple Agents
4. More Composite Task / Multiple Domain
5. Online Learning
6. External Knowledge access, utilization (retrieval, summary, reconstruction) =>Machine Reading Comprehension

References

Deep learning for Dialogue Systems
The Conversational Intelligence Challenge
Dialogue System Technology Challenges
SIGDIAL2016
End-to-End Task-Completion Neural Dialogue Systems(Li et al, 2017)
Deep Reinforcement Learning for Dialogue Generation(Li et al, 2016)
ParlAI: A Dialog Research Software Platform(Miller et al, 2017)
A User Simulator for Task-Completion Dialogues(Li et al, 2016)
Adversarial Learning for Neural Dialogue Generation(Li et al, 2017)
Composite Task-Completion Dialogue Policy Learning via Hierarchical Deep Reinforcement Learning(Peng et al, 2017)
Natural Language Does Not Emerge ‘Naturally’ in Multi-Agent Dialog(Kotter et al, 2017)

Thank you!

@DSKSD