You Impress Me: Dialogue Generation via Mutual Persona Perception

논문 리뷰 세미나

백형렬

차례

1. Introduction
    - Key: Mutual Persona Perception

2. Methodolgy Overview
    - Transmitter
    - Receiver

3. Transmitter
    - Supervised Dialogue Generation
    - Self-Play Model Fine-tuning

4. Receiver
    - Relevance Score

1. Introduction

- 기존 Dialogue Generation은 Personality traits를 반영하지 못함
- PERSONA-CHAT dataset은 대화 및 화자의 Profile이 제공됨
    - 해당 Dataset을 기반으로, Mutual Persona Perception 학습
    - Personalized Dialogue 생성

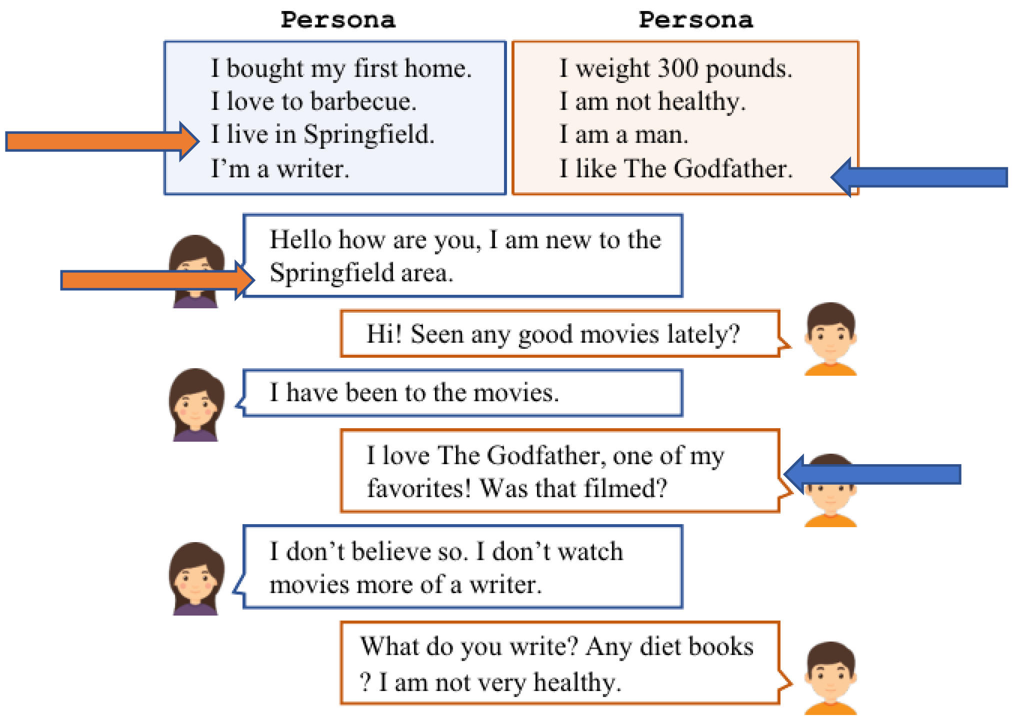


Figure 1: A clippled dialogue from PERSONA-CHAT.

1. Introduction

- 기존 Dialogue Generation은 Personality traits를 반영하지 못함
- PERSONA-CHAT dataset은 대화 및 화자의 Profile이 제공됨
  - 해당 Dataset을 기반으로, Mutual Persona Perception 학습
  - Personalized Dialogue 생성

PERSONA	i. I love new kids on the block. ii. I was born in the early 80's. iii. I also like old school hip hop. iv. My favorite toy as a child as my lite brite.	i. I am a blonde girl with really short hair. ii. I love wearing skinny jeans and leggings. iii. I 'm rather skinny as I like to stay in shape. iv. My favorite hobbies are listening to music and playing video games.	i. I've an allergy to water. ii. I'm a famous twitch streamer. iii. I helped design the game StarCraft. iv. I am working on building a robot to be my body double.
CONTEXT	Do you like clowns? They horrify me!	What video game? I relax by knitting. Love it!	Cool. I did hardcore exercise four hours everyday.
* HUMAN	No! My favorite toy as a kid was lite brite because I was so scared!	I was playing Fallout. What do you look like? I' m blonde thin.	That's a lot of exercising. I hate exercising I just work on my robot most days .
* $\mathcal{P}^2$ BOT	I do not like them at all. I am a little old for them .	I play RPGs. I like to wear skinny jeans and tees .	I do not exercise at all. I'm busy with my robot project .
* TRANSFERTRANSFO	I do not have time for clowns. Do you like to listen to music ?	My fav color is blue, I have a lot of friends in my group.	I work at a computer company. I could make you an Android!
* LOST IN CONVERSATION	I love clowns. they are my favorite.	I love HALO 3, what do you knit?	That sounds like a lot of fun !

Table 4: Sampled responses(*) by Human, $\mathcal{P}^2$ BOT and the state-of-the-art baselines.

2. Methodolgy Overview

2.a. Transmitter
- 대화 History와 자기 Persona(Profile)을 반영하여, 다음 발화Utterance 생성

2.b. Receiver
- Mutual Persona Perception을 담당
- **<u>대화 history</u>** 를 통해 생성된 **<u>상대방에 대한 Impression</u>** 과, **상대방 <u>Actual Persona</u>** 를 일치시킴

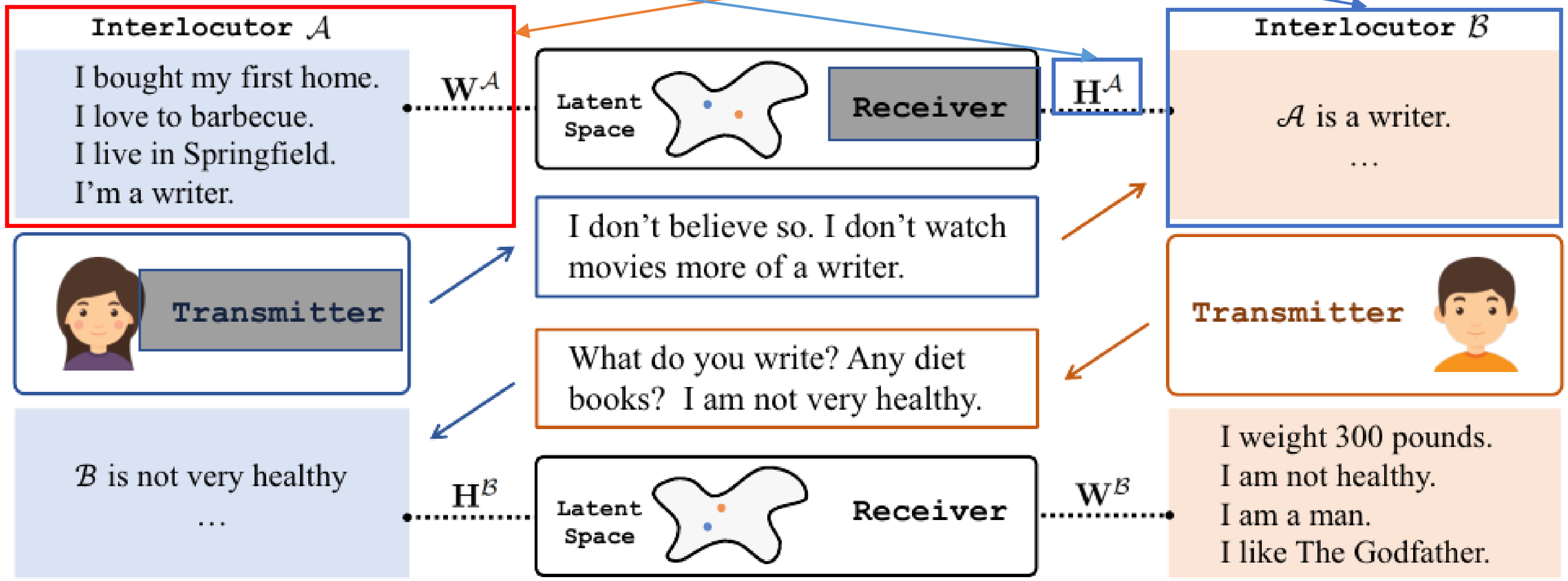


Figure 2: The overview of $\mathcal{P}^2$ Bot (see text).

3. Transmitter Overview

- 대화 History와 자기 Persona(Profile)을 반영하여, 다음 발화Utterance 생성
  - Persona Sentence

$$\{w_1^{\mathcal{A}}, \cdots, w_L^{\mathcal{A}}\}$$

  - 발화 및 발화 History

$$(x_1^{\mathcal{A}}, x_1^{\mathcal{B}}, \cdots, x_N^{\mathcal{A}}, x_N^{\mathcal{B}})$$

$$\mathbf{h}_n^{\mathcal{A}} = (x_1^{\mathcal{A}}, \cdots, x_{n-1}^{\mathcal{B}})$$

  - 화자 A의 Transmitter 발화 Distribution

$$p(x_n^{\mathcal{A}} \mid \underline{\mathbf{w}^{\mathcal{A}}, \mathbf{h}_n^{\mathcal{A}}})$$

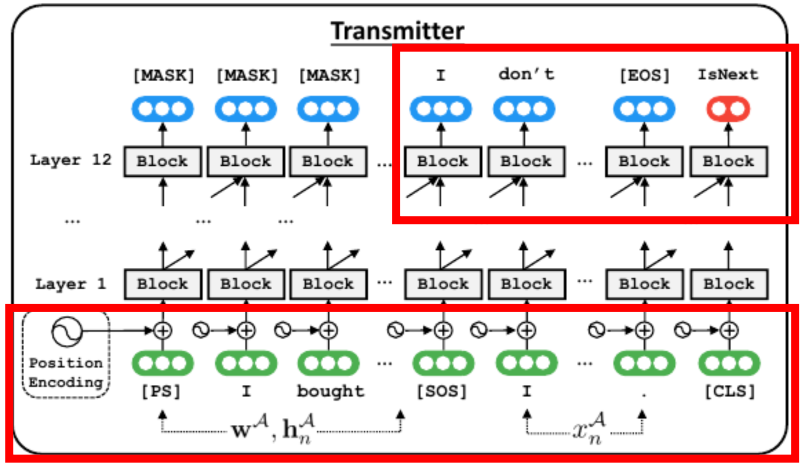
*지금까지 **대화 History**과
자신의 Persona를 고려하여 발화

3. Transmitter 학습

- 12 Stacked Transformer
- **Supervised Dialogue Generation**과 **Self-Play Model Fine-tuning**으로 진행

3.a. Supervised Dialogue Generation

- 목적: 그럴싸한 대화 생성(Mimic human-like responses)
- Main train: Left To Right Language Model로 학습
- Auxiliary train: Next Utterance Prediction
    - 해당 발화가, Persona 및 앞선 대화 고려했을 때 True/False



*자신의 **Persona**와 앞선 **대화 history**가 주어졌을 때, **Response**를 Language Modeling

$$\sum_t \log p_\theta(x^A_{n,t} \mid \mathbf{w}^A, \mathbf{h}^A_n, \boxed{x^A_{n,<t})}$$

$$\log p_\theta(y_n = 1 | \mathbf{w}^A, \mathbf{h}^A_n, \boxed{\hat{x}^A_n})$$

*각각 Main, Auxiliary Objective Function
**Red Box: 각각 before t-th token, n-th turn response

3.b. Self-Play Model Fine-tuning

- 목적: Mutual Persona Perception(상호 프로필 이해)하여 Personailize Dialogue생성
- 학습 방법: Reinforcement Learning
  - A, B Transmitter간에 communication을 시뮬레이션
  - User의 parameters는 frozen; Agent의 parameter는 learnable.
  - Objective function

$$\mathcal{L}_{\mathrm{rl}} = \mathbb{E}_{a_n^{\mathcal{B}} \sim \boxed{p_\theta(a_n^{\mathcal{B}}|s_n^{\mathcal{B}})}}[\boxed{R(a_n^{\mathcal{B}})}]$$

*Policy distribution
**Action: Response
***State: Persona, Dialogue history

$$s_n^{\mathcal{B}} = \{\mathbf{w}^{\mathcal{B}}, \mathbf{h}_n^{\mathcal{B}}\}$$

Parameter update by Policy gardient

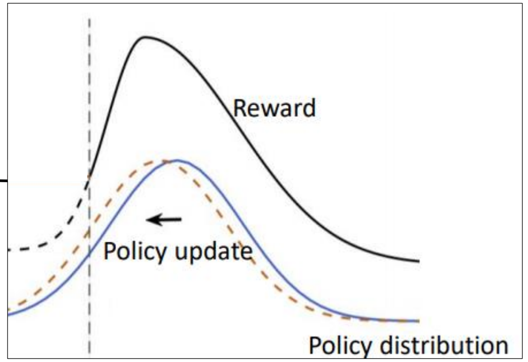
$$\nabla_\theta \mathcal{L}_{\mathrm{rl}} = \mathbb{E}_{a_n^{\mathcal{B}} \sim p_\theta(a_n^{\mathcal{B}}|s_n^{\mathcal{B}})} \nabla_\theta \log p_\theta(a_n^{\mathcal{B}}|s_n^{\mathcal{B}}) R(a_n^{\mathcal{B}})$$

log-likelihood of action

$$\mathbf{loss} = -\log \pi_\theta(a_t|s_t)\, R_t$$

reward

*Loss: Negative log-likelihood of action & Reward
->**Low Reward**의 **Action Prob**. **낮추고**,
High Reward의 Action Prob. 높임



for t = 1 to T-1
$\nabla \leftarrow \nabla_\theta \log \pi_\theta(a_t|s_t)\, R_t$
$\theta \leftarrow \theta + \alpha\nabla$

*Policy gradient update

3.c. Self-Play Model Fine-tuning의 3가지 Reward

- Language Style: 생성된 Response의 Likelihood가 얼마나 높은가

$$R_1(a_n^{\mathcal{B}}) = \frac{1}{|a_n^{\mathcal{B}}|}\sum_t \log p_{\mathrm{lm}}(a_{n,t}^{\mathcal{B}} \mid a_{n,<t}^{\mathcal{B}})$$

*Length Normalization

- Discourse Coherence: 주어진 Persona와 대화 History 고려했을 때, 생성된 Response가 합당한가

$$R_2(a_n^{\mathcal{B}}) = \log p_\theta(y_n = 1 \mid a_n^{\mathcal{B}}, s_n^{\mathcal{B}})$$

*Next Utterance Prediction 학습한 Distribution

*Response

$$s_n^{\mathcal{B}} = \{\mathbf{w}^{\mathcal{B}}, \mathbf{h}_n^{\mathcal{B}}\}$$

- Mutual Persona Perception: A, B 서로의 Persona 정보를 인지하고 있는가
  - 현재 Utterance의 **Persona Perception Score**뿐만 아니라, 미래의 Score도 decay하여 반영
  - Current Action(Response or Question)이 n turn 다음에 반영될 수도 있음(취미는?->…->00이에요)

$$R_3(a_n^{\mathcal{B}}) = r(a_n^{\mathcal{B}}) + \sum_{k=n+1}^{N}\left(\gamma^{2(k-n)-1}\boxed{r(x_k^{\mathcal{A}^*})} + \gamma^{2(k-n)} r(a_k^{\mathcal{B}})\right)$$

$$r(a_n^{\mathcal{B}}) = \mathrm{score}(a_n^{\mathcal{B}}, \mathbf{w}^{\mathcal{B}})$$

*Persona Perception Score Calculated by Receiver

*User(A, Frozen Parameter)의 발화도 Scoring.
Agent(B)가 A의 Persona 반영하여 발화했다면,
A는 자신의 Persona 반영하여 대답했을 것.

Immediate Reward

Discounted Future Reward

3.d. Transmitter 정리

- Supervised Dialogue Generation: Conditional Log Likelihood, Next token prediction

$$\sum_t \log p_\theta(x^{\mathcal{A}}_{n,t} \mid \mathbf{w}^{\mathcal{A}}, \mathbf{h}^{\mathcal{A}}_n, x^{\mathcal{A}}_{n,<t})$$

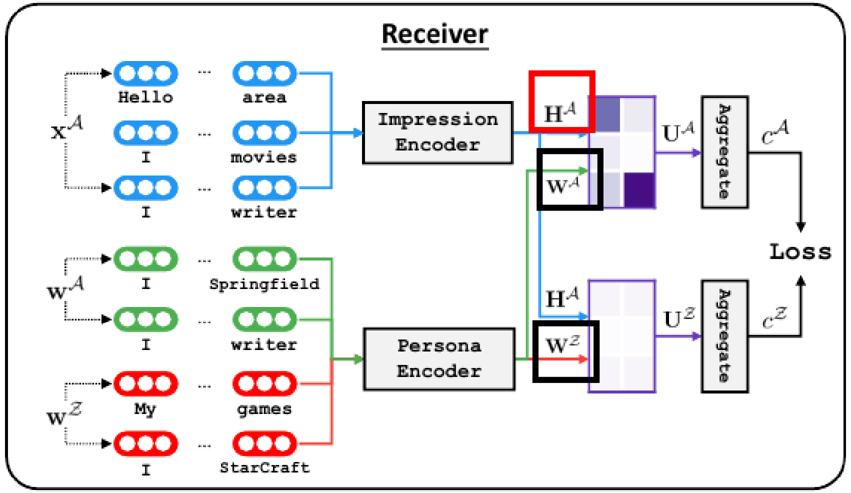
$$\log p_\theta(y_n = 1 | \mathbf{w}^{\mathcal{A}}, \mathbf{h}^{\mathcal{A}}_n, \hat{x}^{\mathcal{A}}_n)$$

- Self-Play Model Fine-tuning

$$\mathcal{L}_{\text{rl}} = \mathbb{E}_{a^{\mathcal{B}}_n \sim p_\theta(a^{\mathcal{B}}_n | s^{\mathcal{B}}_n)}[\underline{R(a^{\mathcal{B}}_n)}]$$

$$R = \lambda_1 R_1 + \lambda_2 R_2 + \lambda_3 R_3$$

4. Receiver: Encoder학습



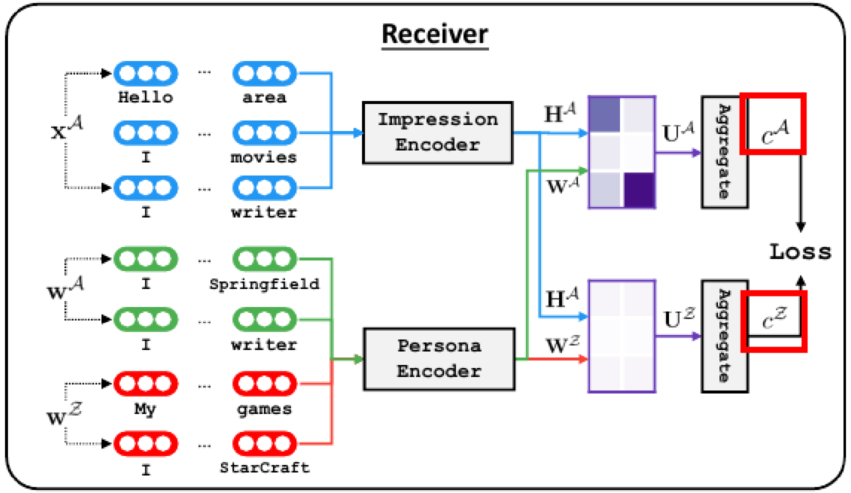
- 대화를 통해 드러난 Impression과 Actual Persona가 얼마나 매칭되는지 **Relevance Score**산출
  - Other Person의 Persona와 관련이 적고, 자신의 Persona와 관련이 높아야 Score높음
- 2개 Encoder사용(BERT)
  - Impression Encoder: 모든 A or B의 발화를 Project
    - Eg. Input: A's utterance
      output: Impression on A Encoding(N: num of utterance; d: Encoding Dimension)

$$(x_1^{\mathcal{A}}, x_2^{\mathcal{A}}, \cdots, x_N^{\mathcal{A}}) \longrightarrow \mathbf{H}^{\mathcal{A}} \in \mathbb{R}^{N \times d}$$

  - Persona Encoder: Actual & Fake Persona를 project
    - Eg. A: A의 Actual Persona; Z: Other Person의 Persona;

$$\text{(Persona Sentences)} \longrightarrow \mathbf{W}^{\Delta} \in \mathbb{R}^{L \times d} \quad \text{where } \Delta \in \{\mathcal{A}, \mathcal{Z}\}$$

4. Receiver: Encoder학습



- Relevance Score: Actual W, Fake W를 각각 Impression Encoding H와 Scaled dot product
  - 즉, 지금까지 각 발화X가, 실제 Persona W^A를 얼마나 반영하였는가(similarity)

$$\mathbf{U}^{\Delta} = \frac{\mathbf{H}^{\mathcal{A}}(\mathbf{W}^{\Delta})^{\top}}{\sqrt{d}}, \in \mathbb{R}^{N \times L} \quad \text{where } \Delta \in \{\mathcal{A}, \mathcal{Z}\}$$

- Score Matrix U를 변형하여 Actual Score(C^A)와 Fake Score(C^Z) 산출
- Ground Truth(즉 Relevance Score)를 모르기 때문에, Unsupervised Learning
  - 발화와 실제 Persona 관련성(Relevance Score)가, 발화와 Fake Persona관련성보다 최소 m만큼 커야함

$$\mathcal{L}_{\text{rec}} = \max(0, m + c^{\mathcal{Z}} - c^{\mathcal{A}}) + \beta \cdot |\mathbf{U}^{\Delta}|_1$$

*Score 차이가
0 -> max(0, m)=m -> High Loss val.
-m -> max(0, 0)=0 -> Low Loss val.

*L1 regularization

Conclusion

A.Transmitter: 그럴싸한 Personalized Utterance 생성

- Supervised Dialogue Generation: Conditional Log Likelihood, Next token prediction

$$\sum_t \log p_\theta(x^{\mathcal{A}}_{n,t} \mid \mathbf{w}^{\mathcal{A}}, \mathbf{h}^{\mathcal{A}}_n, x^{\mathcal{A}}_{n,<t})$$

$$\log p_\theta(y_n = 1 | \mathbf{w}^{\mathcal{A}}, \mathbf{h}^{\mathcal{A}}_n, \hat{x}^{\mathcal{A}}_n)$$

- Self-Play Model Fine-tuning

$$\mathcal{L}_{\mathrm{rl}} = \mathbb{E}_{a^{\mathcal{B}}_n \sim p_\theta(a^{\mathcal{B}}_n | s^{\mathcal{B}}_n)}[R(a^{\mathcal{B}}_n)]$$

$$R = \lambda_1 R_1 + \lambda_2 R_2 + \lambda_3 R_3$$

B. Receiver: Scoring을 위해 Encoder학습

$$\mathcal{L}_{\mathrm{rec}} = \max(0, m + c^{\mathcal{Z}} - c^{\mathcal{A}}) + \beta \cdot |\mathbf{U}^{\Delta}|_1$$

$$\mathrm{score}(x^{\mathcal{A}}_n, \mathbf{w}^{\mathcal{A}}) = \frac{Agg(\mathbf{H}^{\mathcal{A}}_{n,:}(\mathbf{W}^{\mathcal{A}})^\top)}{\sqrt{d}}$$

END

논문 리뷰 세미나
백형렬