페르소나 기반의 장기 대화를 위한 다각적 어텐션을 활용한 생성 모델

금빛나[1]°, 김홍진[1], 황금하[2], 권오욱[2], 김학수[1]
[1]건국대학교 인공지능학과 [3]한국전자통신연구원
[1]{beausty23,jin3430,nlpdrkim}@konkuk.ac.kr [2]{hgh,ohwoog}@etri.re.kr

Generative Model Utilizing Multi-Level Attention for Persona-Grounded Long-Term Conversations

Bit-Na Keum[1]°, Hong-Jin Kim[1], Jin-Xia Huang[2], Oh-Woog Kwon[2], Hark-Soo Kim[1]
[1]Dept. of Artificial Intelligence, Konkuk Univ. [2]Electronics and Telecommunications Research Institute

요 약

더욱 사람같은 대화 모델을 실현하기 위해, 페르소나 메모리를 활용하여 응답을 생성하는 연구들이 활발히 진행되고 있다. 다수의 기존 연구들에서는 메모리로부터 관련된 페르소나를 찾기 위해 별도의 검색 모델을 이용한다. 그러나 이는 전체 시스템에 속도 저하를 일으키고 시스템을 무겁게 만드는 문제가 있다. 또한, 기존 연구들은 페르소나를 잘 반영해 응답하는 능력에만 초점을 두는데, 그 전에 페르소나 참조의 필요성 여부를 판별하는 능력이 선행되어야 한다. 따라서, 우리의 제안 모델은 검색 모델을 활용하지 않고 생성 모델의 내부적인 연산을 통해 페르소나 메모리의 참조가 필요한지를 판별한다. 참조가 필요하다고 판단한 경우에는 관련된 페르소나를 반영하여 응답하며, 그렇지 않은 경우에는 대화 컨텍스트에 집중하여 응답을 생성한다. 실험 결과를 통해 제안 모델이 장기적인 대화에서 효과적으로 동작함을 확인하였다.

주제어: 페르소나 기반 대화, 장기 대화, 생성 모델

1. 서론

신경망 모델(model)의 발전에 따라 오픈 도메인(open-domain) 대화 시스템(system)은 큰 향상을 이루었으며, 더욱 자연스럽고 인간 같은 응답을 위한 연구가 활발히 이루어지고 있다. 인간 같은 대화 시스템을 만드는 데 있어서는 화자의 성격적 특성이나 관심사 등을 포괄적으로 의미하는 페르소나(persona)가 핵심적인 역할을 한다 [1, 2]. 이에 따라 상대방 또는 자신에 관한 페르소나를 기억하기 위한 페르소나 메모리(memory)를 도입해 응답에 활용하는 모델들[1, 3, 4, 5, 6]이 제안되었다. 기존 연구들은 모델이 페르소나를 잘 반영해 응답하는지에 대해서만 초점을 두고 있다. 그러나 실제 사람 간 대화에서는 메모리를 참고하지 않고 현재 대화의 맥락만을 기반으로 한 응답이 이루어지기도 한다. 그러므로 메모리의 정보를 잘 반영해 응답하는 능력도 중요하나 그 전에 메모리 참조가 필요한지 여부를 판단할 수 있는 능력이 선행되어야 한다. 한편, 페르소나 메모리로부터 참조할 관련 정보를 찾는 과정에서, 기존 연구들[5, 7, 8] 대부분이 별도의 검색 모델을 추가적으로 활용한다. 그러나 이는 전체 시스템을 무겁게 만들고 속도 저하를 일으킨다. 또한, 검색 모델의 성능에 크게 의존하기 때문에 잘못된 정보가 검색된 경우 생성되는 응답의 품질이 심각히 저하될 수 있다.

실생활의 채팅(chatting)에서는 대화를 주고 받다가 개개인의 상황에 따라 약간의 시간동안 대화가 잠시 끊긴다. 그리고 대화가 다시 진행되는 식으로 반복되면서 장기적인 대화가 이루어진다. 일정 시간동안 대화가 끊기는 시점을 기준으로 대화를 분리한 것을 대화 세션(session)이라고 하며, 실생활에서의 채팅은 여러 개의 세션, 즉 멀티세션(multi-session)으로 구성된다. 멀티세션 대화에서는 이전 세션의 대화 내용을 활용하면서 대화가 진행된다. 실제의 대화를 더 유사하게 모방하기 위해 싱글턴(single-turn) 대화[9, 4]에서 멀티턴(multi-turn) 대화 중심으로 대화 연구 동향이 변화하였다. 그러나 아직 대부분이 싱글세션(single-session) 대화에 머물러 있기 때문에 장기적인 대화를 위해서 멀티세션으로의 확장을 꾀해야한다.

본 논문에서는 페르소나 메모리 참조가 필요한지를 판단하고, 필요한 경우에는 관련된 페르소나를 반영하여 응답을 생성하는 대화 모델을 제안한다. 효율적인 시스템을 구축하기 위해 별도의 검색 모델을 두지 않고 생성 모델의 내부적인 연산을 통해서 메모리로부터 관련 있는 페르소나를 가져온다. 이를 위해 토큰 레벨(token-level)과 문장 레벨(sentence-level)로 어텐션(attention) 연산을 수행하여 대화 컨텍스트(context)와 메모리 간의 관계를 다각으로 파악한다. 이를 바탕으로 메모리 참조가 필요한지 여부와 메모리 내의 어떤 페르소나를 참조할 것인지에 대한 분류 작업을 추가해 멀티태스크 러닝(multi-task learning)을 수행한다. 실험에 대한 평가를 통해 제안 모델이 장기적인 대화 상황에서 효과적으로 동작함을 확인할 수 있다.

2. 관련 연구

챗봇(chatbot)이 사용자와 감정적으로 소통을 하기 위해서는 일관된 페르소나를 가져야 한다[2]. 이에 따라 페르소나를 대화에 접목시키고 페르소나 일관성을 높이기 위한 연구들이

다양하게 진행되고 있다. [1]은 최초로 페르소나를 기반으로 하는 신경망 모델을 제안하였다. 화자별 페르소나 정보를 인코딩(encoding)하여 디코더(decoder)의 각 스텝(step)마다 전달함으로써 화자에 특화된 응답을 생성한다. [3]은 페르소나 정보를 각각 인코딩하여 메모리로 두고, 디코더의 각 스텝마다 인코딩된 메모리와의 어텐션 연산을 통해서 응답을 생성한다. 이와 유사한 구조를 취하면서, [4]에서는 메모리를 개인 정보 메모리와 댓글 메모리로 분리하여 두 가지 정보를 반영하는 구조를 제안했다.

페르소나 기반 대화 연구에 큰 파급력을 미친 [5]는 기존 연구들이 장기적인 대화에서 성능이 저하되는 것을 해결하고자 멀티세션 대화 데이터셋(dataset)을 구축했다. 모델은 이를 기반으로 학습하며, 별도의 검색 모델을 활용해 대화 컨텍스트와 관련 있는 페르소나를 가져와 응답에 반영한다. [7]에서는 장기 대화 환경에서 페르소나 메모리의 관리를 위해 코사인 유사도 기반의 업데이트(update) 방법을 제안하였다. 마찬가지로 검색 모델을 활용해 대화 컨텍스트와 관련이 높은 페르소나를 검색하고, 이를 컨텍스트에 이어 붙여 생성 모델에 입력한다[10]. [8]은 페르소나 및 지식을 기반으로 하는 대화 모델을 구현하기 위해, 검색기를 활용해 적절한 정보들을 가져온다. 검색된 정보들을 바탕으로 강화시킨 쿼리(query)를 검색 증강 생성기[11]에 전달함으로써 더 명확한 관련성을 갖는 정보를 검색하고 이를 바탕으로 응답을 생성한다. 하지만, 이와 같이 별도의 검색 모델을 추가적으로 활용하는 경우, 전체 시스템이 무거워지고 추론 속도가 느려지므로 효율성이 다소 떨어지게 된다.

따라서, 우리는 별도의 검색 모델을 사용하지 않고 생성 모델만으로 대화 컨텍스트와 페르소나 메모리 간의 관계를 파악한다. 이때 컨텍스트와 페르소나에 대한 인코더를 따로 둠으로써 페르소나는 사전에 인코딩을 해둘 수 있기 때문에 빠른 추론 속도를 실현할 수 있다.

3. 제안 모델

제안하는 모델의 전체적인 구조도는 그림 1에서 확인할 수 있다. 다음은 각 구성 요소에 대한 설명이다.

3.1 인코더

우리는 대화 컨텍스트와 페르소나에 대해 별개의 인코더를 두는 Bi-encoder 구조[9]를 취한다. Bi-encoder 구조에서는 메모리의 각 페르소나 정보들을 사전에 인코딩해둘 수 있기 때문에 빠른 추론 속도가 가능해진다. 두 인코더는 사전학습된 BART[12]의 파라미터(parameter)로 초기화되며, 페르소나 인코더의 가중치는 프리징(freezing)된다. 컨텍스트 인코더에는 대화 컨텍스트의 각 발화 사이에 EOS 토큰을 삽입하고 모든 발화들의 토큰들을 일렬로 이어붙인 시퀀스(sequence)가 입력된다. 인코더의 self-attention 연산을 통해서 컨텍스트 내 모든 토큰들의 관계를 세밀하게 이해할 수 있다. 한편, 페르소나 메모리에 존재하는 각 페르소나들은 독립적이기 때문에 상호 관계를 이해할 필요가 없으므로, 페르소나 인코더를 통해 개별적으로 인코딩된다.

3.2 어텐션 기반 메모리 분류

우리는 현재 대화 컨텍스트와 메모리 간의 관계를 면밀히 파악하기 위해서, 먼저 의미적 유사성을 파악하는 문장 레벨 어텐션(SLA)을 수행한다. 문장마다의 전역적인 의미가 함축된 표현을 얻기 위해, 컨텍스트 및 각 페르소나들의 토큰 표현들로부터 하나의 문장에 해당하는 벡터(vector)들에 평균 풀링(pooling)을 취한다. 추가적으로, 수도 레이블(pseudo label)을 이용한 지도 학습을 통해서, 메모리로부터 참조할 페르소나에 대해 어텐션 가중치 기반으로 분류를 수행한다. 이때, 컨텍스트와 페르소나 간 의미적인 연관성을 기준으로 하기 위해 SLA의 가중치, $W_{SLA} \in \mathbb{R}^{C \times M}$를 기반으로 예측한다.

$$W_{SLA} = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) = \begin{bmatrix} \hat{y}_{1,1} & \cdots & \hat{y}_{1,M} \\ \vdots & \ddots & \vdots \\ \hat{y}_{C,1} & \cdots & \hat{y}_{C,M} \end{bmatrix} \quad (1)$$

C는 컨텍스트에 포함되는 발화의 수를, M은 메모리의 크기를 의미한다. 이를 통해 메모리 참조가 필요한지를 분별하는 능력 및 어떤 페르소나가 가장 관련 있는지를 연산하는 능력을 향상시키고자 한다. 두 가지 능력을 한번에 학습하기 위해서, 존재하는 메모리에 *'참조할 정보가 없음'*을 의미하는 `NO_REF` 레이블을 별도로 추가하였다. 수도 레이블이 달려있는 경우 수도 레이블을 정답 레이블로, 달려있지 않은 경우 `NO_REF`를 정답 레이블로 사용한다. 그런 다음, 모델의 예측에 대해 다음의 cross-entropy loss를 계산한다:

$$\mathcal{L}_{CLS} = -\frac{1}{C}\sum_{i=1}^{C}\sum_{j=1}^{M} y_{i,j}\log(\hat{y}_{i,j}) \quad (2)$$

$\hat{y}_{i,j}$는 i번째 발화가 j번째 페르소나를 참조해야 한다고 모델이 예측한 확률을 나타내며, $y_{i,j}$는 정답 레이블을 나타낸다.

3.3 구조 분기

메모리 분류 예측 결과를 바탕으로, 메모리 참조가 필요하다고 판단한 경우와 그렇지 않은 경우에 대해 구조를 분기한다. 메모리 참조가 필요하다고 판단한 경우에는 SLA

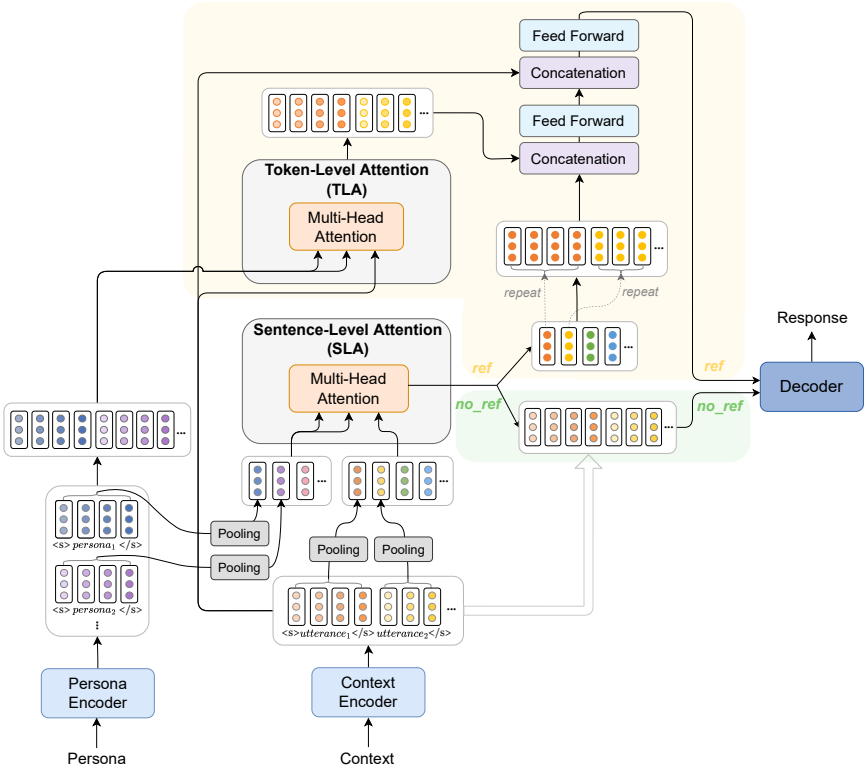

그림 1. 제안 모델의 구조도

뿐만 아니라 토큰 레벨 어텐션(TLA)을 추가적으로 수행한다. 이는 컨텍스트와 메모리 간의 관계를 더욱 심층적으로 이해하기 위해 어휘적 유사성을 파악하는 과정이다. SLA와 TLA의 계산된 값을 통합시킨 다음, 여기에 컨텍스트 임베딩을 통합시킴으로써 최종적인 인코더 표현을 얻고 이를 디코더에 전달한다. 반면, 메모리 참조가 필요하지 않다고 판단한 경우에는 메모리의 정보가 노이즈(noise)의 역할을 하기 때문에, 위 과정을 거치지 않고 컨텍스트 임베딩을 바로 디코더에 전달한다. 이를 통해 온전히 컨텍스트만을 기반으로 응답할 수 있기 때문에 효과적이고, 메모리와의 불필요한 연산을 거치지 않으므로 효율적인 구조를 갖는다.

3.4 학습 목적 함수

최종적인 학습 목적 함수 $\mathcal{L}$는 모델이 생성한 응답에 대한 손실 함수 $\mathcal{L}_{GEN}$과 메모리로부터 참조할 정보를 분류하는 손실 함수 $\mathcal{L}_{CLS}$를 가중합하여 계산된다. 이를 최소화하는 방향으로 학습이 이루어진다.

$$\mathcal{L} = \lambda \mathcal{L}_{GEN} + (1 - \lambda)\mathcal{L}_{CLS} \quad (3)$$

λ는 각 손실 함수의 중요도를 반영하기 위한 하이퍼파라미터(hyperparameter)이며, 실험에서는 0.8로 설정하였다.

Attribute	Sess.1	Sess.2	Sess.3	Sess.4
Dialogues	83,493	48,112	44,601	23,840
Utterances	524,405	308,578	290,656	156,742
Avg. turns	6.3	6.4	6.5	6.6
Avg. len of utter	37.5	42.4	42.4	42.4
Avg. personas	2.5	2.8	4.7	6.0
Avg. agt persona	2.4	1.5	2.6	3.4
Avg. usr persona	0.0	1.3	2.1	2.6
Dialogues	763	326	491	471
Utterances	4,848	1,390	2,212	2,136
Avg. turns	6.4	4.3	4.5	4.5
Avg. len of utter	38.2	40.7	41.2	40.3
Avg. persona	2.2	2.8	3.7	4.2
Avg. agt persona	2.2	1.3	2.0	2.3
Avg. usr persona	0.0	1.4	1.7	1.8

표 1. Data statistics. The upper part pertains to Train data, the lower part pertains to Test data.

4. 실험

4.1 데이터셋

데이터셋은 AI-Hub에서 추후 공개될 예정인 한국어 멀티세션 대화 데이터셋[1]을 전처리하여 사용했다. 대화는 일반 사용자 역할의 화자를 나타내는 user와 대화 모델을 나타내는 화자 agent 사이에 이루어진다. 대화 컨텍스트에서 각 화자의 발화가 번갈아 등장하고 발화는 최대 10개로 구성하였다. 컨텍스트의 마지막 발화는 항상 user의 발화로 구성함으로써 모델이 agent의 응답을 생성하도록 했다. 각 발화의 서두에는 화자를 구분짓는 스페셜 토큰 <user> 또는 <agent>를 삽입했다. 메모리에 존재하는 각 페르소나의 서두에는 어떤 화자의 페르소나인지에 따라서 스페셜 토큰 <user_persona> 또는 <agent_persona>를 삽입하였다. Session 1은 임의의 페르소나들로 구성된 메모리가 주어진 상황에서 수행된 대화이다. Session 2, 3, 4에서의 메모리는 직전 세션의 대화에서 화자와 관련된 중요한 정보들을 요약한 페르소나 요약들로 구성된다. 그리고 이 메모리를 기반으로 대화가 수행된다. Session 1을 요약한 메모리는 Session 2에 제공되고, Session 2를 요약한 메모리는 이에 중첩되어 Session 3에 제공되며, Session 3를 요약한 메모리는 이에 중첩되어 Session 4에 제공된다.

한편, 원본 데이터셋에는 중요도가 낮은 페르소나 요약이 다수 존재하기 때문에, 모델이 페르소나 기반의 응답을 생성하는 법을 효과적으로 학습하지 못할 가능성이 있다. [13]은 이와 관련된 문제를 해결하기 위해 세션마다 페르소나 기반 응답을 2개 이하로 제한하였고, 네거티브(negative) 페르소나들로 증강하였다. 이를 따라 우리는 세션마다 페르소나 수도 레이블이 달려있는 응답을 2개 이하로 제한하였다. 그리고 메모리에 과다하게 존재하는 불필요한 페르소나 요약을 걸러내기 위해, 수도 레이블링된 페르소나들만을 남긴 후 네거티브 페르소나들로 증강하였다. 데이터셋 통계는 표 1에서 확인할 수 있다.

[5]에서는 학습 시 세션을 구분하지 않고 모든 세션을 종합하여 학습했을 때 성능이 향상됨을 보여주었다. 이에 따라 우리는 모든 세션을 종합한 Session 1-4를 학습 데이터로 사용했다. 테스트 시에는 장기적인 대화 능력을 확인하기 위해서 각 세션을 구분해 대화 성능을 평가하였다.

4.2 실험 환경

생성 모델은 124M개의 파라미터를 갖는 KoBART[2]를 사전학습 모델로 사용하였다. 학습 하이퍼파라미터로는 batch size 4, gradient accumulation steps 8를 사용했다. 학습률은 3e-05, warmup steps는 500, optimizer는 AdamW[14]을 사용했으며 10 epochs만큼 학습을 진행했다. 디코딩 방식으로는 다양성을 높이기 위해 beam search를 사용했고, beam size는 8, repetition penalty는 1.1로 설정했다. GPU는 NVIDIA GeForce RTX 2080 Ti (12GB)를 사용했다.

4.3 실험 결과

오픈 도메인 대화에서는 타겟 응답뿐 아니라 다양한 응답들이 정답이 될 수 있기 때문에, 생성된 응답과 타겟 응답 사이의 어휘적인 유사성을 평가하는 지표는 다소 부적절하다. 그래서 우리는 오픈 도메인 대화에서 인간의 평가와 높은 상관관계를 갖는 것으로 알려진 perplexity(PPL) 지표를 사용했다[15]. PPL은 낮을수록 성능이 우수함을 의미한다. 또한, 생성된 응답과 타겟 응답 사이의 의미적인 유사성을 평가하는 BERTScore[16] 지표를 사용하였는데, 한국어 평가를 위해 재구현된 KoBERTScore 라이브러리[3]를 활용했다. 또한 생성된 응답이 얼마나 다양성을 지니는지 측정하기 위한 지표 Distinct[17]를 사용하였고, 이때 n은 2로 설정했다.

우선 표 2는 메모리의 페르소나를 참조하는 발화로 구성된 데이터(ref)와 참조하지 않는 발화로 구성된 데이터(no_ref)의 구성 비율에 따른 실험 결과를 나타낸 것이다. PPL은 perplexity, BS는 BERTScore를 의미한다. no_ref 데이터를 많이 활용할 때보다 ref 데이터를 많이 활용할 때 더 성능이 개선되는 것을 확인할 수 있다. 최종적으로 ref 데이터와 no_ref 데이터를 각 100%씩, 즉 일대일 비율로 구성하였을 때 가장 좋은 성능을 보여 데이터 구성 비율을 이와 같이 선택하였다.

다음으로 표 3는 모델을 구성하는 각 요소들에 따른 성능을 나타낸 것이다. D-2는 Distinct-2를 의미한다. Base는 Bi-Encoder 구조로 구현한 기본적인 BART를 의미한다. 여기에 토큰 레벨 및 문장 레벨의 어텐션을 추가하였을 때 PPL이 특히 많이 개선되었다. 이어서 참조할 메모리 분류 작업을 추가했을 때 멀티태스크 러닝을 통해서 PPL이 전반적으로 향상되었다. 메모리 분류 결과에 따른 분기 구조를 취했을 때에는 BERTScore와 Distinct-2가 전체적으로 향상되었다.

5. 결론

본 논문에서는 검색 모델을 활용하지 않고 생성 모델의 내부적인 연산을 통해 페르소나 메모리의 참조가 필요한지를 판별하는 모델을 제안하였다. 그리고 이 모델에서는 참조가 필요하다고 판단한 경우에는 관련된 페르소나를 반영하여 응답하며, 그렇지 않은 경우에는 대화 컨텍스트에 따라 응답을 생성하는 효율적인 구조를 취한다. 실험 결과를 통해 제안 모델이 장기 대화 환경에서 효과적으로 동작함을 확인했다.

[1] https://aihub.or.kr/aihubdata/view.do?dataSetSn=71630
[2] https://huggingface.co/gogamza/kobart-base-v2
[3] https://github.com/lovit/KoBERTScore

표 2. 메모리 참조 데이터와 비참조 데이터의 구성 비율에 따른 실험 결과

Data Rate		**Session 1**		**Session 2**		**Session 3**		**Session 4**		**Session 1-4**	
ref	no_ref	PPL	BS	PPL	BS	PPL	BS	PPL	BS	PPL	BS
0%	100%	8.37	0.75	12.15	0.78	13.37	**0.79**	13.44	0.77	10.68	0.78
25%	100%	7.47	0.74	11.89	**0.86**	12.86	0.77	13.03	0.78	10.07	0.78
50%	100%	7.19	0.74	11.61	0.75	12.58	0.77	12.63	0.80	9.80	0.78
100%	0%	7.35	0.74	13.41	0.80	14.37	0.73	13.96	0.77	10.70	0.75
100%	10%	7.09	0.73	12.90	0.73	13.95	0.74	13.61	0.82	10.35	0.76
100%	25%	6.76	0.74	12.11	0.73	13.18	0.74	13.12	0.83	9.85	0.75
100%	50%	6.54	0.74	11.69	0.72	12.63	0.76	12.77	0.75	9.52	0.75
100%	100%	**6.30**	**0.77**	**11.23**	0.73	**12.14**	0.75	**12.24**	**0.85**	**9.18**	**0.83**

표 3. Ablation Study (A: TLA & SLA, C: Memory Classification, B: Structural Branching)

	Session 1			**Session 2**			**Session 3**			**Session 4**			**Session 1-4**		
Method	PPL	BS	D-2	PPL	BS	D-2	PPL	BS	D-2	PPL	BS	D-2	PPL	BS	D-2
Base	6.75	0.75	0.37	**11.15**	**0.78**	0.55	12.21	**0.77**	0.49	12.49	0.80	0.54	9.45	0.77	0.36
+A	6.28	0.77	0.40	11.36	0.76	**0.62**	12.21	0.76	0.57	12.19	0.78	**0.58**	9.17	0.78	0.41
+A,C	**6.25**	0.77	0.41	11.24	0.74	0.60	12.18	0.75	0.56	**12.13**	0.79	0.56	**9.13**	0.77	0.40
+A,C,B	6.30	**0.78**	**0.42**	11.23	0.75	0.61	**12.14**	0.76	**0.59**	12.24	**0.84**	**0.58**	9.18	**0.79**	**0.42**

페르소나 메모리

Agent는 주말에 집에서 시간을 보낸다.
Agent는 뮤지컬 영화를 가끔 본다.
Agent는 체육교육과를 전공했다.

대화 컨텍스트

Agent: 안녕하세요 저는 30대 여성입니다 저는 체육교육과를 전공했어요. 반가워요
User: 안녕하세요! 저도 30대 여성이고, 매일 주식을 사고 파는 펀드 매니저일을 하고 있어요!
Agent: 저는 주말에 집에서 시간을 보내요. 주말에는 어떻게 시간을 보내고 계세요?
User: 저는 고객들 자산 관리를 하다보니, 가끔 스트레스를 받아요 그래서 주말에는 평소에 좋아하는 음악 영화를 자주 보며 시간을 보냅니다.

모델 응답

음악 영화를 좋아하시는군요. 저는 뮤지컬 영화를 가끔 봐요

그림 2. 응답 생성 샘플

감사의 글

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2019-0-00004, 준지도학습형 언어지능 원천기술 및 이에 기반한 외국인 지원용 한국어 튜터링 서비스 개발)

참고문헌

[1] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan, "A persona-based neural conversation model," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 994–1003, Aug. 2016. [Online]. Available: https://aclanthology.org/P16-1094

[2] H. Shum, X. He, and D. Li, "From eliza to xiaoice: Challenges and opportunities with social chatbots," *CoRR*, Vol. abs/1801.01957, 2018. [Online]. Available: http://arxiv.org/abs/1801.01957

[3] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?" *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, Jul. 2018. [Online]. Available: https://aclanthology.org/P18-1205

[4] Y. Wu, X. Ma, and D. Yang, "Personalized response generation via generative split memory network," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1956–1970, Jun. 2021. [Online]. Available: https://aclanthology.org/2021.naacl-main.157

[5] J. Xu, A. Szlam, and J. Weston, "Beyond goldfish memory: Long-term open-domain conversation," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5180–5197, May 2022. [Online]. Available: https://aclanthology.org/2022.acl-long.356

[6] Q. Liu, Y. Chen, B. Chen, J.-G. Lou, Z. Chen, B. Zhou, and D. Zhang, "You impress me: Dialogue generation via mutual persona perception," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1417–1427, Jul. 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.131

[7] X. Xu, Z. Gou, W. Wu, Z.-Y. Niu, H. Wu, H. Wang, and S. Wang, "Long time no see! open-domain conversation with long-term persona memory," *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2639–2650, May 2022. [Online]. Available: https://aclanthology.org/2022.findings-acl.207

[8] J. Lim, M. Kang, Y. Hur, S. W. Jeong, J. Kim, Y. Jang, D. Lee, H. Ji, D. Shin, S. Kim, and H. Lim, "You truly understand what I need : Intellectual and friendly dialog agents grounding persona and knowledge," *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1053–1066, Dec. 2022. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.75

[9] P.-E. Mazaré, S. Humeau, M. Raison, and A. Bordes, "Training millions of personalized dialogue agents," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2775–2779, Oct.-Nov. 2018. [Online]. Available: https://aclanthology.org/D18-1298

[10] J. Weston, E. Dinan, and A. Miller, "Retrieve and refine: Improved sequence generation models for dialogue," *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pp. 87–92, Oct. 2018. [Online]. Available: https://aclanthology.org/W18-5713

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20, 2020.

[12] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Jul. 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[13] D. Kwon, S. Lee, K. H. Kim, S. Lee, T. Kim, and E. Davis, "What, when, and how to ground: Designing user persona-aware conversational agents for engaging dialogue," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 707–719, Jul. 2023. [Online]. Available: https://aclanthology.org/2023.acl-industry.68

[14] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.

[15] D. Adiwardana, M. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a human-like open-domain chatbot," *CoRR*, Vol. abs/2001.09977, 2020. [Online]. Available: https://arxiv.org/abs/2001.09977

[16] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[17] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, Jun. 2016.