

목적 지향 대화 시스템을 위한 문맥 기반의 제어 가능한 자연어 생성 모델

함진아⁰, 김재원, 양동일

삼성전자 삼성리서치

j.ham@samsung.com, j109.kim@samsung.com, di87.yang@samsung.com

Context-aware and controllable natural language generation model for task-oriented dialogue systems

Jina Ham⁰, Jaewon Kim, Dongil Yang
Samsung Electronics Samsung Research

요약

목적 지향 대화 시스템은 사용자가 원하는 목적을 달성하기 위해 사용하는 시스템으로 일상 대화와 다르게 시스템이 정보를 명확히 전달하는 것이 중요하다. 따라서 최근 연구에서 목적 지향 대화 시스템을 위한 자연어 생성 모델은 정해진 대화 정책에 따라 알맞은 응답을 생성할 수 있도록 의도와 슬롯 정보를 담은 대화 행위(Dialog Act)를 활용한다. 하지만 대화 행위는 생성하는 문장을 타월하게 제어하는 반면에 대화의 흐름과 상황에 맞게 다양한 문장을 생성하기 어렵다는 문제점을 가지고 있다. 이러한 문제점을 해소하고자 본 논문에서는 목적에 부합하는 내용을 명확하게 자연어로 생성하기 위해 대화 행위를 사용하면 동시에 일상 대화 생성 모델과 같이 문맥을 고려하여 대화 흐름에 어울리는 자연스러운 문장을 생성할 수 있는 문맥 기반의 제어 가능한 자연어 생성 모델을 제안한다. 실험에서는 KoGPT2 사전 학습 모델과 한국어 대화 데이터셋을 사용하였으며 실험을 통해 대화 행위 기반의 자연어 생성 모델과 본 연구에서 제안한 문맥 기반의 제어 가능한 자연어 생성 모델을 비교하였다. 결과적으로 대화 행위를 단독으로 학습한 모델보다 일정 문맥을 함께 학습한 모델이 유의미한 BLEU 점수 향상을 보인다는 점을 확인하였다.

주제어: 자연어 생성, 목적 지향 대화 시스템, 언어 모델, 문맥, 대화 행위

1. 서론

최근 챗봇에 대한 관심이 높아지면서 대화 시스템에 대한 연구가 활발히 이루어지고 있다. 대화 시스템은 일상 대화 시스템과 목적 지향 대화 시스템으로 구분된다. 일상 대화 시스템은 자유로운 주제로 대화하며 정해진 답변 없이 다양한 내용의 답변이 가능하다. 목적 지향 대화 시스템은 사용자가 원하는 목적을 달성하기 위해 사용하는 시스템으로 호텔 예약, 식당 요금 문의 등에 활용되며 일상 대화 시스템과 다르게 시스템이 사용자가 요청한 정보를 명확하게 답변하는 것이 중요하다.

일반적으로 목적 지향 대화 시스템은 [그림 1]과 같이 사용자의 발화를 입력 받고 적절한 응답을 제공하기까지 4가지의 모듈을 거친다[1, 2].

(i) 자연어 이해(Natural Language Understanding) 모듈은 사용자의 발화를 이해하는 모듈로 입력된 자연어를 분석하고 시스템이 처리하기 용이한 형태로 변환한다. (ii) 대화 상태 추적(Dialog State Tracking) 모듈은 사용자의 발화에 따라 대화의 상태를 변경한다. 대화 상태에는 목적과 도메인 등이 포함된다. (iii) 대화 정책 관리(Dialog Policy Managing) 모듈은 대화 상태를 기반으로 각 대화 시스템에서 정한 정책에 따라 알맞은 응답 정보를 담은 대화 행위(Dialog Act)[3]

를 출력한다. 대화 행위는 주로 슬롯과 값으로 표현한다. [그림 2]의 n 번째 발화에 대한 대화 행위를 보면 ‘연령대’, ‘가격’ 항목을 슬롯으로, 대응하는 내용을 값으로 나타낸다. (iv) 자연어 생성(Natural Language Generation) 모듈은 전달받은 대화 행위 기반한 문장을 생성하여 사용자에게 응답한다.



그림 1. 목적 지향 대화 시스템

기존의 자연어 생성 모듈은 규칙이나 템플릿을 작성하여 구성하였다. 그러나 규칙이나 템플릿은 다양한 상황을 대처하기 어렵다. 최근 Transformer[4] 기반의 생성 모델인 GPT-2[5] 모델이 여러 태스크에서 좋은 성과를 보이면서 생성 모델에 대한 연구가 많은 주목을 받고 있고 목적 지향 대화 시스템에서의 생성 모델도 이를 이용한 방법으로 연구되고 있다.

대화	
...	...
$n-3$ 번째 발화	온천 랜드는 돈을 지불하고 가야 하는 거죠?
$n-2$ 번째 발화	네, 온천 랜드는 요금을 내셔야 해요
$n-1$ 번째 발화	대인 한명이 온천 이용하려면 얼마예요?
n 번째 발화	온천 요금, 대인은 6천원이에요

대화 행위	
의도	부대시설이용요금문의
슬롯과 값	(연령대 = 대인 ; 가격 = 6천원)

그림 2. 대화 행위 예시

일반적인 자연어 생성 모델은 다양한 문장을 생성할 수 있지만 생성하는 내용을 제어하기 어렵다. 이때 문장이 포함해야 하는 정보를 담은 대화 행위를 활용하면 의도와 필요한 정보를 반영하는 문장을 생성할 수 있고 대화 정책에 맞게 문장을 제어할 수 있다. 하지만 이러한 방법은 생성 문장이 대화 행위에 의존하면서 대화 흐름과 상황에 맞는 다양한 문장을 생성하기 어렵다는 문제점을 갖고 있다. 대화의 흐름의 모든 요소를 대화 행위화하는 것은 한계가 있기 때문이다.

목적 지향 대화 시스템의 생성 모델은 제어된 답변을 생성하는 것이 중요하지만 대화 시스템의 사용자에게 대화 흐름에 맞는 매끄러운 응답을 제공하는 것 또한 필요하다. 이를 위해 본 연구에서는 목적 지향 대화 시스템에서 목적에 부합하는 내용을 명확하게 자연어로 생성하고자 대화 행위를 사용하면서 동시에 문맥을 고려하여 보다 자연스러운 문장을 생성할 수 있는 모델을 제안한다. 본 모델의 특징은 다음과 같다.

- 의미론적 데이터를 입력 받아 자연어를 생성하는 모델로 ‘의도 (슬롯/값 쌍 모음)’ 형태로 구성된 대화 행위를 사용하여 생성 문장이 포함해야 하는 내용을 제어한다.
- 대화의 연속성을 고려하기 위해 이전 문맥을 사용하여 문장을 생성한다.

2. 관련 연구

2.1. 제어 가능한 생성 모델

GPT-2[5] 모델의 출현으로 다양한 문장을 생성할 수 있게 되었으나, 사실에 근거하지 않은 문장, 윤리적으로 좋지 않은 문장 등 의도하지 않은 문장을 생성할 수 있다는 위험 부담을 안고 있다. 이를 해결하고자 생성 모델에 조건을 추가로 입력하여 통제할 수 있는 연구들도 활발히 진행되고 있다. 그 중 목적 지향 대화 시스템에서의 생성 모델은 의미론적인 형태인 대화 행위

를 조건으로 입력 받아 자연스러운 발화 문장을 생성해야 한다. 특히, 목적 지향 대화 시스템은 서비스와 맞닿아 있어 주어진 조건을 반드시 충족하는 문장을 생성해야 한다.

목적 지향 대화 시스템에서의 생성 모델을 제안한 SC-GPT[6]에서는 3가지 단계로 나누어 제어 가능한 생성 모델을 학습한다. 우선, (i) 대용량 말뭉치를 이용하여 사전 학습을 진행하며 언어 모델을 구축한다. 그리고, (ii) 대화 행위를 포함하고 있는 대용량의 발화 데이터로 사전 학습을 지속적으로 진행하는데 이 단계에서 모델은 의미론적인 형태의 대화 행위와 자연어와의 상관 관계를 학습하게 된다. 그 후, (iii) 대상 도메인의 한정적인 데이터를 가지고 사후 학습을 진행한다. 이 연구는 의미론적인 형태의 조건으로 생성 모델을 제어한다는 것에서 의미를 갖는다.

2.2. 문맥 기반 생성 모델

목적 지향 대화 시스템에서 사용자의 목적을 파악하고자 대화 상태 추적 모듈을 사용하지만, 슬롯으로 지정할 수 있는 값들은 한정적이고 슬롯 값들만 사용해 응답하기에는 부족한 부분들이 있다. 사용자 발화에 맞춰 응답해주어야 할 때가 있으며 이전 대화 문맥을 통해 대화에 사용된 단어나 대화의 세부적인 특성을 파악할 수 있다. [그림 2]에서는 n 번째 발화와 대화 상태 추적 모듈 및 대화 정책 관리 모듈의 결과로 얻은 대화 행위 예시를 보여주고 있다. 대화 행위를 보면 ‘부대시설 이용 요금’으로 ‘대인’과 ‘6천원’을 답변해야 한다는 것을 알 수 있지만 질의가 ‘온천 랜드’에 대한 내용이라는 점은 파악하기 어렵다. ‘온천’이라는 부대시설을 파악하기 위해서는 이전 대화 문맥이 필요하다.

KoDialogPT2[7]에서는 이전 대화를 이용하여 일상 대화 모델을 더 최적화 시킨다. KoDialogPT2[7]는 40GB 이상의 한국어 텍스트로 학습시킨 SKT-KoGPT¹를 활용하여 대화 도메인의 데이터에 대해 사후 학습한 모델이다. 사후 학습 시에는 입력을 이전 대화 발화의 스페셜토큰과 같이 구성하는 미세 조정 방법을 사용하여 현 대화의 특성을 더욱 높이고자 하였다.

3. 제안 모델

3.1 대화 행위 기반 자연어 생성 모델

대화 행위 A 는 의도를 I , 슬롯을 s , 값을 v , 슬롯과 값 쌍의 개수를 p (0 이상의 정수)라고 할 때 다음 (1)과 같이 정의할 수 있다.

$$A = [I, (s_1, v_1), \dots, (s_p, v_p)] \quad (1)$$

¹ <https://github.com/SKT-AI/KoGPT2>

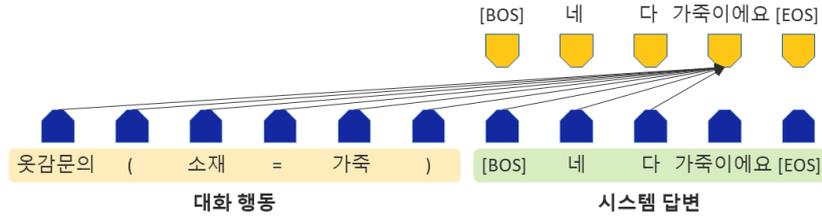


그림 3. SC-GPT 모델의 기본 구조

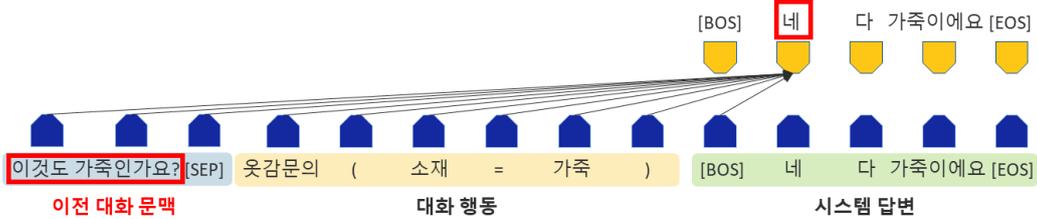


그림 4. 이전 대화 문맥을 포함한 제안 모델의 기본 구조

SC-GPT[6]는 대화 행위를 기반으로 자연어를 생성하는 모델을 제안하며 대화 행위를 다음 (2)와 같은 형태로 정의한다. 이때 A' 는 SC-GPT[6]에 맞게 변환된 대화 행위를 의미한다.

$$A' = [I (s_1 = v_1, \dots, s_p = v_p)] \quad (2)$$

생성해야 하는 발화를 d 라고 할 때 앞 뒤로 스페셜 토큰 [BOS], [EOS]를 덧붙이고 대화 행위 A' 와 함께 모델에 입력 값으로 넣어 [그림 3]과 같이 사전 학습 모델(GPT-2)을 학습한다. SC-GPT[6]의 입력 값을 X_S 라고 할 때 X_S 는 다음 (3)과 같이 정의할 수 있다.

$$X_S(d) = [A' [BOS] d [EOS]] \quad (3)$$

[그림 3]은 SC-GPT[6]의 기본 구조를 나타낸 그림으로 대화 행위와 발화, 스페셜토큰을 기반으로 우측 상단의 문장을 생성한다. 예를 들어 ‘가족이에요’ 라는 토큰은 대화 행위 전체와 발화 내에 있는 해당 토큰 이전의 모든 토큰을 참조하여 생성된다. 하지만 ‘가족이에요’ 앞에 있는 ‘네’, ‘다’ 와 같은 토큰은 대화 행위의 의도 토큰 ‘웃감문의’ 와 슬롯/값 토큰 ‘가족’ 만으로는 유추할 수 없다. 이전 대화의 흐름을 이해해야 생성할 수 있는 내용이기 때문이다.

[그림 4]을 보면 이전 대화 문맥이 ‘이것도 가족이에요?’ 라는 질의임을 확인할 수 있으며 이러한 질의와 슬롯/값을 기반으로 긍정의 답변 ‘네’ 혹은 ‘모두 가족이다’ 를 나타내는 ‘다’ 와 같은 표현을 충분히 유추할 수 있고 이를 바탕으로 대화의 연속성을 반영하는 발화를 생성할 수 있다.

본 연구에서는 생성하는 자연어를 제어할 수 있는 대화 행위와 더불어 이전 대화 문맥을 입력에 추가하여 목적 지향 대화 시스템에 필요한 정확한 정보를 담으면서 대화 흐름에 적합한 형태로 문장을 생성할 수 있는 모델을 제안한다.

3.2 문맥 기반의 제어 가능한 자연어 생성 모델

[그림 4]는 제안 모델의 기본 구조를 나타낸 것으로 대화 행위에 앞서 이전 대화 문맥을 스페셜토큰 [SEP]로 구분하여 삽입한다. 생성해야 하는 발화를 d_n 이라고 하고 모델에서 사용할 최대 문맥의 개수를 m 이라고 할 때 대화 내의 이전 문맥 발화는 d_{n-1}, \dots, d_{n-m} 으로 나타낼 수 있다. 이전 문맥 발화의 개수가 m 보다 적다면 존재하는 개수의 문맥만 삽입한다. 따라서 $0 \sim m$ 개의 문맥 발화가 포함될 수 있다. 문맥 C 는 다음 (4)와 같이 정의할 수 있으며 스페셜토큰 [SEP]로 문맥을 구분하도록 변환한 C' 는 다음 (5)와 같다. 전체 입력 값을 X_C 라고 할 때 X_C 는 다음 (6)와 같이 각 C' 를 X_S 에 덧붙인 값이다.

$$C = [d_{n-1}, \dots, d_{n-m}] \quad (4)$$

$$C' = [d_{n-1} [SEP] \dots [SEP] d_{n-m}] \quad (5)$$

$$X_C(d_n) = [C' [SEP] X_S(d_n)] \quad (6)$$

4. 실험

4.1 데이터셋

본 실험은 AI Hub에서 제공하는 한국어 대화 데이터셋을 사용하였다. 의도, 슬롯/값(지식베이스), 발화를 모두 갖추고 있으며 발화자(고객, 점원)가 구분되어 있어 목적 지향 대화 시스템의 답변을 생성하는 SC-GPT[6] 모델의 prompt 형태로 변환이 가능하다는 점을 근거로 데이터셋을 채택하였다.

소상공인 분야의 9개 분야, 90,413개 데이터 중 다음 조건에 부합하는 8개 분야의 36,201개의 데이터를 사용하였다.

- 1) 의도, 발화가 유효한 데이터
- 2) 점원 발화 데이터
 - 대화 시스템의 답변을 생성하는 모델이기 때문에 점원 발화 데이터만 채택하였다.
- 3) 지식베이스가 유효한 데이터
 - 슬롯/값 쌍의 집합을 바탕으로 발화를 생성해야 하기 때문에 모든 값이 발화에 포함된 경우만 유효한 데이터로 판단하였다.
- 4) 비식별화된 개인정보가 포함되지 않은 데이터
 - 생성하는 발화에 비식별화 기호가 포함되는 것을 방지하기 위하여 비식별화된 개인정보를 포함하는 데이터는 제외하였다.
 - 대부분의 발화에 비식별화된 개인정보가 포함된 ‘부동산’ 분야는 제외하였다.

전체 데이터는 8:1:1의 비율로 나누어 학습, 검증, 평가 데이터로 사용하였다. 데이터 분야에 따른 분포는 [표 1]과 같다.

표 1. 실험 데이터 분포

분야	데이터 개수			
	학습	검증	평가	전체
음식점	5,639	730	734	7,103
의류	5,775	690	732	7,197
학원	1,477	186	173	1,836
소매점	5,337	672	682	6,691
생활서비스	3,521	417	397	4,335
카페	2,764	371	351	3,486
숙박업	2,614	326	327	3,267
관광여가오락	1,817	235	234	2,286
전체	28,944	3,627	3,630	36,201

4.2 실험 환경

사전 학습 모델은 SKT에서 공개한 KoGPT2를 사용했으며 하이퍼 파라미터 설정은 SC-GPT²에서 공개한 기본 설정을 따르고 Learning rate는 5e-5로 설정하였다.

본 실험에서는 모델의 입력 최대 길이가 지나치게 길어지지 않도록 최대 문맥 개수를 3으로 지정하였다. 이에 따라 각 샘플은 이전 문맥 발화 개수에 따라 0 ~ 3개의 문맥 발화를 포함하고 입력 최대 길이는 기본 모델의 경우 150, 문맥을 포함한 제안 모델의 경우 400으로 설정하였다.

1 ~ 20 epoch에 걸쳐 실험을 진행하였으며 대화 행위 기반 자연어 생성 모델을 ‘Base’ 로, 제안하는 문맥 기반의 제어 가능한 자연어 생성 모델을 ‘+Context’ 로 표기하여 결과를 비교하였다.

4.3 실험 결과

평가 지표는 자연어 생성에서 널리 쓰이는 BLEU 점

수[8]와 제어 가능한 자연어 생성 모델이 주어진 슬롯의 값을 얼마나 반영했는지 평가하는 Slot Error Rate(ERR)[9]를 사용한다. BLEU 점수는 BLEU-4 점수로 계산하였으며 ERR은 대화 행위에 있는 모든 슬롯 개수를 P , 결과 문장에 없는 값의 개수를 e 라고 할 때 다음 (4)와 같이 계산한다.

$$ERR = \frac{e}{P} \quad (4)$$

각 예제 별로 다섯 개씩 문장을 생성하며 BLEU는 nltk의 corpus bleu를 활용하여 micro 평균을 계산한다. ERR의 경우 다섯 문장 중에 가장 낮은 수치를 갖는 문장을 선택하여 전체 ERR의 P 과 e 에 합산한다.

실험 결과는 다음의 [표 2]와 같다.

표 2. 두 모델의 BLEU-4, ERR 결과

epoch	BLEU-4		ERR	
	Base	+Context	Base	+Context
1	0.199	0.192	12.046	14.396
2	0.204	0.203	12.019	11.779
3	0.217	0.211	11.432	11.886
4	0.225	0.219	10.604	12.714
5	0.231	0.269	11.725	15.652
6	0.223	0.295	9.989	13.301
7	0.217	0.318	10.844	15.572
8	0.222	0.327	11.218	14.744
9	0.224	0.325	11.111	14.744
10	0.214	0.342	12.420	15.545
11	0.221	0.362	11.271	14.290
12	0.222	0.360	10.817	15.598
13	0.219	0.362	11.966	17.548
14	0.217	0.384	11.779	15.224
15	0.215	0.386	12.046	14.183
16	0.214	0.389	10.978	14.797
17	0.214	0.379	11.325	15.091
18	0.213	0.394	11.245	14.316
19	0.214	0.397	11.458	14.957
20	0.218	0.400	11.245	14.209

[그림 5]는 두 모델의 BLEU-4 점수를 나타낸 그래프이다. BLEU-4 점수의 경우 epoch 4까지는 두 모델의 성능 차이가 최대 약 0.007로 거의 유사하나 epoch 5 이상부터 제안 모델의 성능이 Base 모델에 비해 상당한 차이로 앞서는 것을 확인할 수 있다. 최소 0.101(epoch 5)에서 최대 0.183(epoch 19)까지 유의미한 차이를 보인다. 이를 토대로 이전 대화 문맥을 추가했을 때 대화의 흐름에 나타나는 표현을 효과적으로 생성 문장에 반영하여 실제 대화 데이터셋에 가까운 발화를 생성한다고 말할 수 있으며 [그림 6]과 같은 실제 생성 예시에서도 이러한 사실을 확인할 수 있다.

² <https://github.com/pengbaolin/SC-GPT>

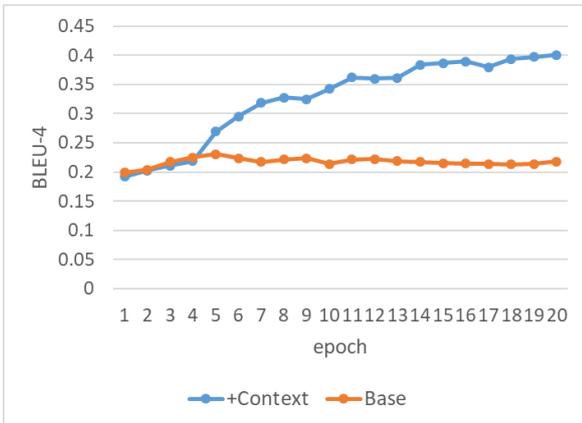


그림 5. 두 모델의 BLEU-4 점수를 나타낸 그래프

대화	
n - 3번째 발화	김치찌개에 뒷다릿살 한 근이면 적당할까요?
n - 2번째 발화	식사하실 분이 몇 분이신가요?
n - 1번째 발화	김치찌개 먹을 사람은 4명입니다
n번째 발화	뒷다릿살 반근이시면 충분하세요

대화 행위	
의도	용량별 고기 문의
슬롯과 값	(고기부위 = 뒷다릿살; 무게 = 반근)

모델 별 실제 생성 예시	
Base	뒷다릿살은 반근부터 판매해요
+Context	뒷다릿살 반근이시면 충분하세요

그림 6. 모델 별 실제 생성 예시

반면에 ERR의 경우 [그림 7]과 같이 대체로 Base 모델의 수치가 더 낮고 최소 0.24%(epoch 2)에서 최대 5.58%(epoch 13)까지 차이가 난다. 이것은 생성 모델의 입력 값에 문맥이 추가되면서 대화 행위에 대한 의존도가 줄어들면서 생성 문장이 포함해야 하는 내용에 대한 집중도가 일부 떨어진다고 해석할 수 있다.

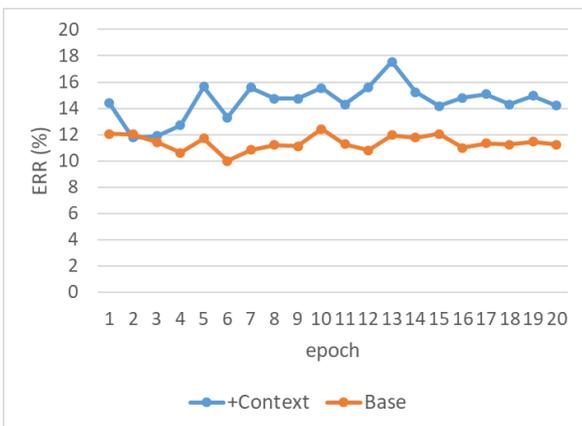


그림 7. 두 모델의 ERR를 나타낸 그래프

5. 결론

실험 결과를 종합하면 제안 모델은 기존 모델에 비해 대화 행위에 있는 슬롯을 반영하는 성능은 일부 떨어지지만 실제 발화에 훨씬 더 가까운 문장을 생성한다. ERR의 상승률에 비해 BLEU-4 점수의 상승률이 훨씬 크다는 점을 보면 실제 대화의 발화는 대화 행위에 비해 문맥에 영향을 받는 정도가 더 크다는 것으로 해석할 수 있다. 즉, 제안 모델은 이전 대화 문맥을 기반으로 대화 행위에 없지만 문맥 상 의미 있는 표현을 반영하여 문장을 생성하는 데에 탁월하며 결과적으로 기존의 목적 지향 대화 시스템을 위한 자연어 생성 모델에 가까운 수준으로 생성 결과를 제어하면서도 대화 흐름에 맞는 자연스러운 문장 생성이 가능하다.

이후에는 화자를 구분하거나 대화 행위에 대한 추가 정보를 제공하는 등 모델이 대화 문맥과 대화 행위를 조금 더 명확하게 구분할 수 있는 장치를 추가하여 대화 행위에 대한 집중도를 높이고 이에 따라 ERR 수치를 낮추면서 BLEU 점수를 유지할 수 있는 방향을 연구하고자 한다.

참고문헌

- [1] S. Young, M. Gašić, B. Thomson, and J. D. Williams, "POMDP-Based Statistical Spoken Dialog Systems: A Review", Proceedings of the IEEE, vol. 101, no. 5, pp. 1160-1179, 2013.
- [2] T.-H. Wen, D. Vandyke, N. Mrkšić, M. Gašić, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, and S. Young, "A Network-based End-to-End Trainable Task-oriented Dialogue System", Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 438-449, 2017.
- [3] 은지현, 최준기, 장두성, 김현정 그리고 구명완, "마르코프 의사결정 과정에 기반한 대화 관리 시스템", 한국HCI학회 2007년도 학술대회 1부, pp. 475-480, 2007.
- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need", Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), pp. 6000-6010, 2017.
- [5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., "Language Models are Unsupervised Multitask Learners", OpenAI blog, Vol. 1, No. 8, p. 9, 2019.
- [6] Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujuan Li, Jinchao Li, Michael Zeng and Jianfeng Gao, "Few-shot Natural Language Generation for Task-Oriented Dialog", arXiv:2002.12328, 2020.
- [7] 오동석, 박성진, 이한나, 장윤나 그리고 임희석, "KoDialogPT2 : 한국어 일상 대화 생성 모델", 제

33회 한글 및 한국어 정보처리 대회 논문집, pp. 457-460, 2021.

- [8] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318, 2002.
- [9] Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, Steve Young, "Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems", Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1711-1721, 2015.