Journal of The Korea Convergence Society
Vol. 13. No. 1, pp. 13-21, 2022
e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2022.13.01.013

생성 모델과 검색 모델을 이용한 한국어 멀티턴 응답 생성 연구

이호동[1], 이종민[2], 서재형[3], 장윤나[3], 임희석[4*]
[1]삼성 에스디에스 주식회사 소프트웨어 엔지니어, [2]삼성 에스디에스 주식회사 소프트웨어 아키텍트,
[3]고려대학교 컴퓨터학과 석·박사통합과정, [4]고려대학교 컴퓨터학과 교수

A study on Korean multi-turn response generation using generative and retrieval model

Hodong Lee[1], Jongmin Lee[2], Jaehyung Seo[3], Yoonna Jang[3], Heuiseok Lim[4*]
[1]Software Engineer, AI BUD, Samsung SDS
[2]Software Architect, Samsung SDS
[3]Master & Ph.D. Combined Student, Department of Computer Science and Engineering, Korea University
[4]Professor, Depatment of Computer Science and Engineering, Korea University

요 약 최근 딥러닝 기반의 자연어처리 연구는 사전 훈련된 언어 모델을 통해 대부분의 자연어처리 분야에서 우수한 성능을 보인다. 특히 오토인코더 (auto-encoder) 기반의 언어 모델은 다양한 한국어 이해 분야에서 뛰어난 성능과 쓰임을 증명하고 있다. 그러나 여전히 디코더 (decoder) 기반의 한국어 생성 모델은 간단한 문장 생성 과제에도 어려움을 겪고 있으며, 생성 모델이 가장 일반적으로 쓰이는 대화 분야에서의 세부 연구와 학습 가능한 데이터가 부족한 상황이다. 따라서 본 논문은 한국어 생성 모델을 위한 멀티턴 대화 데이터를 구축하고 전이 학습을 통해 생성 모델의 대화 능력을 개선하여 성능을 비교 분석한다. 또한, 검색 모델을 통해 외부 지식 정보에서 추천 응답 후보군을 추출하여 모델의 부족한 대화 생성 능력을 보완하는 방법을 제안한다.

주제어 : 언어 모델, 멀티턴, 대화, 자연어 생성, 발화 생성

Abstract Recent deep learning-based research shows excellent performance in most natural language processing (NLP) fields with pre-trained language models. In particular, the auto-encoder-based language model proves its excellent performance and usefulness in various fields of Korean language understanding. However, the decoder-based Korean generative model even suffers from generating simple sentences. Also, there is few detailed research and data for the field of conversation where generative models are most commonly utilized. Therefore, this paper constructs multi-turn dialogue data for a Korean generative model. In addition, we compare and analyze the performance by improving the dialogue ability of the generative model through transfer learning. In addition, we propose a method of supplementing the insufficient dialogue generation ability of the model by extracting recommended response candidates from external knowledge information through a retrival model.

Key Words : Language model, Multi-turn, Conversation, Natural language generation

*Corresponding Author : HeuiSeok Lim(limhseok@korea.ac.kr)
Received October 15, 2021
Revised November 22, 2021
Accepted January 20, 2022
Published January 28, 2022

1. 서론

사전 훈련된 언어 모델 (Pre-trained Language Model)이란, Transformer[1] 구조 기반으로 대용량의 한국어 코퍼스를 통해 사전 학습을 진행한 딥러닝 모델이다. 대부분의 자연어처리 연구는 언어 모델의 등장 이후 기존의 규칙 기반[2] 또는 통계 기반 시스템[3]에 비해 크게 우수한 성능을 보인다. 한국어 자연어처리 연구에서도 대용량 말뭉치로 사전 학습된 오토 인코더 (auto-encoder) 또는 디코더 (decoder) 기반의 언어 모델이 등장했다.

오토 인코더 모형을 바탕으로 하는 언어 모델은 자연어 이해 (Natural Language Understanding) 분야 (e.g. 감정 분석, 주제 분류, 유사도 측정 등)에 대한 문제를 주로 해결하며, 대표적으로는 BERT[4] RoBERTa[5], 그리고 ELECTRA[6]가 있다. 특히, 최근에 한국어 기반 언어 모델의 자연어 이해 성능 평가를 위한 벤치마크 데이터 KLUE (Korean Language Understanding Evaluation)[7]가 발표되면서, 자연어 이해를 바탕으로 하는 언어 모델들에 대한 발전과 개선이 크게 이루어지고 있다.

디코더 모형을 기반으로 하는 언어 모델은 자연어 생성 (Natural Language Generation) 분야 (e.g. 대화, 기계 번역, 요약 등)에 대한 문제를 주로 다루며, 대표적으로는 GPT2[8], GPT3[9], 그리고 BART[10]가 있다. 최근 산업 분야에서는 대화형 생성 모델로 사용하며, 챗봇 (chat-bot)으로 사람을 보조하는 역할에 주목하고 있다. 특히 현재까지도 초거대 언어 모델인 GPT3와 관련하여서는 국내 IT 기업과 연구진이 개발을 위해 많은 자원을 투자하고 있다.

그러나 여전히 한국어 대용량 말뭉치를 바탕으로 학습된 생성 모델이 자연어 생성 하위 분야에서 어떠한 성능을 보이며 약점을 지니고 있는지에 대한 세부적인 비교 분석이 부재하며, 개선 연구를 위한 데이터도 부족하다. 특히 현재의 한국어 생성 모델은 주어진 입력에 대해 간단한 추론을 통한 문장 생성 과제에도 큰 어려움을 지니고 있으며, 실제 사람이 생성하는 문장에 비해 다소 어색한 결과를 반환하여, 이를 해결하기 위한 개선 연구가 필요한 상황이다.

본 논문은 한국어 생성 모델을 활용하여 가장 널리 쓰일 수 있는 멀티턴 기반 오픈 도메인 대화 (Open-domain Dialog)에 대한 세부 연구를 진행한다. 오픈 도메인 대화 중에서도 생성 모델에 통상적으로 적용 가능한 공감형 대화 기반의 멀티턴 데이터셋을 구축하고, 전이 학습을 통하여 주요 모델의 성능을 비교 분석한다. 그리고 검색 모델 (retriever)을 통해 사전에 구축한 외부 대화 데이터로부터 준-정답 (semi-golden statement)에 해당하는 추천 후보군을 추출하여 생성 모델의 성능 개선 방법을 제안한다.

한국어 기반의 멀티턴 데이터셋은 생성 모델을 활용한 챗봇에 사용자와의 상호작용을 강화하며, 기존에 단발성 ASR (automatic speech recognition) 시스템[11]의 부자연스러운 대화 패턴 개선에 활용이 가능하다. 또한, 자기 회귀 (auto-regressive)[12] 생성 모듈의 단방향성으로 인해 통제하기 어려운 생성 결과에 대해, 검색 모델을 통해 추천 응답을 확장하면서 상용화 단계에서의 안정성을 부여한다.

본 논문의 구성은 다음과 같다. 우선 2장에서 생성 모델 기반 대화 연구의 흐름을 소개하고 3장에서는 데이터셋 구축과 제안 모델을 설명한다. 4장에서는 멀티턴 대화 시스템에서 한국어 생성 모델의 성능을 비교 분석한다. 마지막 5장에서는 이 모든 과정에 대한 결론을 내린다.

2. 한국어 멀티턴 대화 데이터 구축

생성 모델의 오픈 도메인 대화에서 통상적으로 적용이 가능한 공감형 대화에 대해 영문으로 작성된 데이터셋은 존재하는 반면, 공식적으로 배포된 한국어 데이터셋은 존재하지 않는다. 기존에 공개된 공감형 대화와 유사한 주제를 공유하는 멀티턴 데이터셋으로 범위를 확장하여 살펴본다면, 소셜 네트워크 서비스 (SNS)에서 추출한 오픈 칫챗용 대화 데이터셋[1]과 목적 지향적 데이터셋[2]이 존재한다. 이 중 SNS에서 추출한 일부 대화 내용은 부적절한 발화가 포함되어있는 경우가 다수 존재하며, 평균 발화의 수가 10개 미만으로 깊은 멀티턴 맥락을 학습하기에는 부족한 상황이다. 또한, 목적 지향적 대화 데이터셋의 경우 평균 발화 수는 15개이지만 정보의 전달을 목적으로 구성된 대화인 만큼 오픈 도메인 대화에 적합한 공감형 대화로 보기 어렵다. 따라서 본 논문은 멀티턴으로 구성된 공감형 영문 대화 Empathetic

1) https://aihub.or.kr/keti_data_board/language_intelligence

2) https://github.com/clovaai/ClovaCall?fbclid=IwAR2rL-qcvv4zBVdKxJwulv19Nn8UnG3n_JPzZwIjfk6PVscMgTGZHGi3L2Q

Dialogs[24]를 참조하여, 한국어에 대한 멀티턴의 공감형 대화 데이터셋을 구축했다.

Table 1. Constructed dataset statistics

	SelectStar	ED_kor
# Dialog	5,692	618
# Avg. turn	10.73	5.70
# Avg. length of turn	18.25	33.66

Table 1과 같이 2개의 크라우드 소싱 플랫폼을 통해 공감형 대화 기반의 심리상담 데이터 구축에 19명의 검수자와617명의 참가자를 선발했다. 참가자들은 일부러 맞춤법을 무시하거나, 이모티콘이나 외래어 사용을 지양하도록 했다. 대화는 심리상담의 상담자와 내담자 역할로 나누어지며, 상담자는 내담자의 발화를 의역하고 다시 질문하는 과정을 거치면서 내담자의 감정을 공감하고 해소할 수 있는 대화를 이어간다. 또한, 맥락과 무관한 발화, 차별적 표현, 그리고 폭력적 표현을 금지했으며, 내담자의 자존감을 높이고 공감하는 내용으로 대화를 이끌어가도록 했다. 검수자의 경우 생성한 대화에 대한 검수와 추가적인 레이블링을 진행했다. 검수 과정은 총 3회 반복적으로 시행했으며, 맞춤법과 대화의 주제가 적절한지에 대해 확인하고 그렇지 않은 경우에는 수정 작업을 했다.

Fig. 1의 데이터 예시와 같이 두 참가자는 'C (오픈 챗챗)', 'R(인간 관계)', 'W(업무/커리어)' 중 하나의 주제로 최소 9턴에서 최대 14턴의 대화를 진행하며, 대화 발화당 평균 단어가 6단어 이상이 유지되도록 검수하여 발화 속에 충분한 내용이 담기도록 했다. 생성한 대화는 턴이 아닌 문장을 기준으로 나누었고 화자의 고유값(ID)을 태깅하여 구분했다. 이러한 작업은 이후 서비스 개발 단계에서 생성 모델의 응답을 문장 단위로 분절하여 품질을 개선하기 위해 진행한다. 또한, 이후 화자의 이름과 인구 통계학적 정보를 포함한 페르소나와 사용자의 개인정보가 사용된 발화의 오프셋을 별도로 태깅했다. 이와 같이 대화의 주제와 화자별 페르소나 정보를 미리 수집함으로써 향후 연구에서 한국어 공감형 대화 데이터만이 아닌, 페르소나 대화를 포함한 데이터셋으로 확장이 가능하도록 데이터셋을 구성했다.

대화 카테고리	C
페르소나 (speaker0)	'id': 0, 'name': '자영', 'gender': 'F', 'married': '1', 'age': 45, 'job': '0'
페르소나 (speaker1)	'id': 1, 'name': '미주', 'gender': 'F', 'married': '1', 'age': 39
	...
speaker1	정말 대단해요.
	자영씨 처럼 친절하고 좋은 사람 흔하지 않아요.
speaker0	저는 지금 자존감도 없고, 혼란스러워요.
speaker1	행복했을때를 생각해 보세요.
	가족과 여행에서 즐거운일과 평상시에 좋았던 일들요.
	그럼 좋아질꺼여요.
speaker0	제가 행복 했을때가 있나 모르겠어요.
speaker1	있어요.
	작년인가 한 참 좋았을 때 있잖아요.
speaker0	아!
	맞아요.
	...

Fig. 1. A case of multi-turn dialog dataset

3. 검색 모델을 활용한 응답 생성

멀티턴 기반의 대화 데이터셋을 통해 생성 모델은 발화턴 수가 길어지는 과정에서의 연속적인 대화 생성 능력을 크게 개선한다. 그러나 파악해야 하는 문맥의 범위가 늘어나고 대화 내용이 심화됨에 따라 응답이 깊은 추론을 필요하게 되면서, 모델이 성능을 안정적으로 유지하기 어려워진다. 더욱이 로그 최대 가능도를 목적 함수로 지니는 자가 회귀 기반의 단방향성 문장 생성 방식만으로는, 짧은 대화 속에 담겨있는 사람의 섬세한 대화 과정을 충분히 반영하기 어렵다.

이러한 문제를 해결하기 위해, 본 논문은 구축한 멀티턴 대화 데이터셋의 일부인 41,422개의 문장을 분리하여 참조 가능한 외부 지식 데이터로 사용하는 방법을 제시한다. 분리한 대화 데이터는 검색 모델을 활용하여 다음 발화자가 응답 가능한 문장 후보군을 추출하도록 사용한다. 추출한 문장은 실제 사람의 멀티턴의 대화 기록을 토대로 다음 발화에서 가장 등장할 확률이 높은 내용을 담고 있다. 따라서 검색 모델이 다음 문장으로 예측한 결과는 이전까지의 대화 정보와 함께 다음 대화를 생성하기 위한 준-정답 문장으로 활용한다. 준-정답 문장은 이전 대화 내용과 함께 생성 모델의 입력 시퀀스에 합쳐지며, 추출을 위한 검색 모델로는 희소 벡터 기반의 검색 모델인 BM25[25]와 밀집 벡터 기반의 검색 모델인 Poly-encoder[26]를 사용한다.

3.1 BM25

본 논문에서는 희소 벡터 기반의 BM25 알고리즘을 첫 번째 검색 모델로 사용한다. 해당 모델은 이전 대화를 질의로 사용하고 구축한 데이터베이스의 후보군 대

화 기록과의 키워드 빈도수를 통해 가장 관련성이 높은 응답을 가져온다. 아래 수식 (1)과 같이 입력 문장을 형태소 단위로 분절한 어휘 정보 집합 Q를 사용하여 전체 데이터베이스인 후보군 C 집합을 비교 검색한다. 여기에서 n은 입력 문장에 존재하는 모든 분절 어휘 개수를 의미하며, k와 b는 각각 BM25 검색 모델을 위한 하이퍼 파라미터로 1.2와 0.75로 설정한다. 희소 벡터로 구성된 문장 임베딩은 단어 빈도수에 해당하는 $f(q_i,D)$와 역문서 빈도수를 의미하는 $IDF(q_i)$의 계산으로 점수 $S_s(C,Q)$를 반환한다.

$$S_s(C,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,C)\times(k+1)}{f(q_i,C)+k\times(1-b+b\times\frac{|C|}{avgCL})} \quad (1)$$

3.2 Poly-encoder

본 논문에서는 밀집 벡터 기반의 Poly-encoder를 두 번째 검색 모델로 사용한다. Poly-encoder는 bi-encoder와 cross-encoder의 각각의 장점을 차용한 구조를 지니고 있다[26]. 해당 모델은 데이터베이스에 존재하는 후보군 응답은 bi-encoder와 같이 하나의 밀집 벡터로 변환한다. 입력 대화는 L의 길이를 지닌 벡터로 표현하며, 하나의 밀집 벡터로 다시 표현하는 과정에서 L의 길이를 지닌 입력 벡터와 사전에 임베딩한 후보군 벡터를 attention 연산하여, cross-encoder와 유사한 방식으로 학습한다. 이러한 방식은 cross-encoder에 가까운 성능을 보이면서도 bi-encoder와 비슷한 속도를 지닌다는 장점이 있다. 이전 대화를 통한 준-정답 응답의 추출 과정은 아래 수식과 같다.

$$S_d = y_q \cdot y_c \quad (2)$$

$$y_c = f(c),\ y_q = \sum_i softmax(y_{c_i} \cdot y_q^1, ..., y_{c_i} \cdot y_q^L) y_q^j \quad (3)$$

$$y_q^j = \sum_j softmax(cc_1 \cdot h_1, ..., cc_i \cdot h_N) h_j \quad (4)$$

S_d는 이전 대화에 해당하는 문장 임베딩 y_q와 다음 대화 후보군 문장 임베딩 y_c를 내적한 결과와 같다. y_c는 후보군 문장 c에 대한 인코더 f 의 반환 결과이다. y_q는 N개의 토큰을 지니는 이전 대화 문장 q을 임베딩한 결과에 대해 전체적인 문맥을 나타내는 특징 벡터인 cc와 후보군 문장 임베딩인 y_c를 2회의 attention 연산에 각각 query로 사용하여 나타낸 결과이다.

3.3 검색 모델 비교

검색 모델의 성능 비교 실험은 문장 임베딩 방식에 따라 BM25와 poly-encoder에 대해 진행한다. Poly-encoder의 경우 학습 가능한 파라미터를 지니고 있는 만큼 이전 대화를 질의로 하여서 다음 대화에 적합한 후보군 응답을 반환하도록 훈련한다. Table 2에서는 반환한 결과의 점수를 내림차순으로 정렬한 상위 1개, 2개, 5개, 그리고 10개의 후보군 중에서 실제 정답이 존재하는 확률이 poly-encoder가 BM25보다 각각 16.18%, 20.9%,, 27.55%, 그리고 29.54% 높은 결과를 나타낸다. 이러한 결과를 바탕으로 본 논문은 이전 대화를 바탕으로 준-정답 문장 검색에는 poly-encoder를 검색 모델로 활용한다.

Table 2. Performance of retriever for extracting candidate answers

	Hits@1	Hits@2	Hits@5	Hits@10
BM25	6.19	8.75	12.26	15.14
Poly-encoder	22.37	29.65	39.81	44.68

4. 한국어 기반의 생성 모델 비교 실험

4.1 실험 환경

실험을 위해 딥러닝 프레임워크로 Pytorch 1.9와 Huggingface Transformers[27]의 사전 훈련된 모델을 사용했다. 한국어 생성 모델로는 125M의 모델 파라미터를 지니며 디코더만을 지닌 모형으로 12개 헤드와 레이어를 지닌 KoGPT2, 그리고 124M의 모델 파라미터와 인코더와 디코더 모형에 각각 6개 레이어와 16개의 헤드를 지닌 KoBART를 활용했다. 전이 학습 과정에서 KoGPT2의 훈련 하이퍼 파라미터는 batch size 00, initial learning rate 00, AdamW optimizer[28], 그리고 training epochs 00로 설정했다. KoBART의 경우에는 initial learning rate 5e-5, AdamW optimizer, 그리고 training epochs 12를 사용하고, 각 모델은 평가 데이터셋에 대해 최고 성능을 보이는 훈련 정도를 기준으로 비교 실험을 진행한다. 또한, 실험 과정에서 본 논문은 RAM 378GB, 18-core Intel Xeon Gold 6230 CPU, 그리고 NVIDIA A6000 (48GB) GPU를 사용했다.

Table 3. Comparative experiment on the performance of multi-turn response generation. We fine-tune models with our constructed dataset (w/ FT)

Generator	Retriever	BLEU3	BLEU4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
KoGPT2	-	0.05	0.01	4.51	0.68	0.39	50.89
KoGPT2 w/ FT	-	3.60	2.20	15.91	5.40	15.72	67.00
KoGPT2 w/ FT	BM25	4.26	**3.19**	12.16	4.73	12.02	63.50
KoGPT2 w/ FT	**Poly-encoder**	**4.29**	2.91	**17.52**	**6.40**	**17.39**	**67.34**
KoBART	-	0.27	0.08	5.58	1.38	4.52	55.09
KoBART w/ FT	-	**4.67**	**2.91**	**17.54**	**8.13**	**17.15**	67.55
KoBART w/ FT	BM25	1.2	0.51	9.53	3.1	9.1	64.92
KoBART w/ FT	Poly-encoder	3.03	1.55	15.41	6.45	14.99	**68.65**

4.2 모델 훈련

대화 발화에 대한 언어 모델링을 학습시키기 위해 한국어 생성 모델인 KoGPT-2와 KoBART 모델을 활용한다. 해당 모델들은 대용량 코퍼스에 특정한 태스크를 수행하며 언어에 내재된 패턴을 학습한 상태이다. 이를 활용하여 대화 발화 생성 태스크에 전이 학습시킨다. 전이 학습을 위해서는 이전 발화를 포함한 대화의 맥락을 제공하고 이를 고려하여 적절한 발화를 생성할 수 있도록 한다. 전체 대화 데이터셋은 N개의 대화 D로 구성된다.

각각의 대화는 $D = \{u_1, ..., u_T\}$ 와 같이 T개의 발화 u를 갖는다고 할 때, 모델은 최대 k개의 이전 턴 $u_{prev} = \{u_{t-k}, ..., u_{t-1}\}$의 토큰 시퀀스를 입력으로 받아서 현재 턴 u_t의 토큰들에 대하여 자기회귀(Auto-regressive) 방식으로 학습한다. 이에 대한 학습 목적 함수는 수식 (5)와 같다.

$$L_{LM} = -\sum_{i=1}^{I} \log p_\theta(x_i \mid u_{prev}, x_1, ..., x_{i-1}) \quad (5)$$

I 는 현재 턴 u_t의 토큰의 길이를 뜻하며, x는 각각의 토큰을 의미한다. p_θ는 토큰 시퀀스가 주어졌을 때, 다음 토큰에 대한 언어 모델링을 학습하며, θ는 모델의 학습 파라미터를 의미한다. 모델은 이러한 목적 함수를 최소화하도록 학습하면서 어떠한 대화에서 이전 대화의 맥락에 대한 적절한 답변을 생성하게 된다.

4.3 대화 생성

대화 데이터에 대한 전이 학습을 한 후, 모델이 실제로 대화에 대한 답변을 생성하게 할 때는 빔 탐색[29]을 활용한다. 어떠한 시퀀스가 주어졌을 때 다음 토큰에 대한 확률 분포를 통해 토큰을 추출하고, 5 크기의 빔을 가지고 디코딩을 진행한다.

검색 모델을 이용할 때, 답변의 자연스러운 생성을 위해 알파-블렌딩(α-blending)[30]을 추가한다. α는 하이퍼 파라미터로 검색 모델이 찾은 원래 발화의 α%를 다른 토큰으로 바꿀 것인지 결정한다. 이를 통하여 생성 모델은 주어진 발화의 일부를 자연스럽게 바꾸도록 학습한다. 본 논문에서는 α를 0.5로 설정하여 실험을 진행했다.

4.4 양적 평가

실험 모델의 양적 평가를 위해서 2가지 방식을 사용하며, 생성한 문장과 실제 정답 문장 사이의 n-gram 중첩과 의미적 유사도로 성능을 평가한다.

첫 번째로 n-gram 중첩은 생성한 문장의 형태적인 유사함을 평가하며, BLEU[31]와 ROUGE [32]를 사용한다. BLEU의 경우 생성된 문장을 분절화한 형태소가 얼마나 정답 문장과 겹쳐지는지에 대해 평가하고, ROUGE는 정답 문장을 분절화한 형태소가 얼마나 생성 문장과 겹쳐지는지에 대해 점수화한다.

두 번째로 의미적 유사도는 BERTScore[33]를 사용하여, 생성된 문장이 정답 문장과 얼마나 문맥적으로 유사한 의미를 지니고 있는지 평가한다. 생성 모델의 성능은 각 평가 지표가 나타낼 수 있는 영역이 제한적인 만큼 구조적으로 또는 의미적으로 표현 가능한 다양한 방법을 사용한다[34].

Table 3은 KoGPT2와 KoBART에 대한 비교 실험 결과를 나타낸다. 실험을 위해 훈련 데이터 219,047개, 검증 데이터 41,422개 그리고 평가 데이터 6,872개의 대화 문장을 분리했으며, 멀티턴 대화 단락은 각각 52,209, 9,789, 3,263개로 구성했다.훈련 및 생성에는

동일한 조건의 하이퍼 파라미터를 적용했다.

전반적인 실험의 결과는 n-gram 중첩의 경우 최소 9 ~ 최대 14턴 이후의 적합한 응답을 생성하는 것으로 문장의 형태가 다소 달라지는 경향이 있다. 따라서 다른 생성 연구 과제에 비해 BLEU와 ROUGE의 점수가 낮다. 그러나, 최종 생성 응답과 정답 문장 사이의 의미적 유사도는 모델이 맥락 정보를 지속적으로 유지하면서, 비교적 높은 성능을 보인다.

우선 BM25 검색 모델을 결합하는 경우, 두 가지 생성 모델에서 공통적으로 검색 성능이 지나치게 낮기에, 오히려 생성 모델의 응답 추론 과정에 부정적인 영향을 준 것으로 보인다. KoGPT2는 poly-encoder의 후보군 응답을 입력 시퀀스의 조건 문장으로 포함했을 때 가장 높은 성능을 보인다. 그러나 KoBART의 경우 오히려 poly-encoder를 포함하는 경우에 n-gram 중첩에서 낮은 성능을 나타낸다. 이러한 현상은 디코더 모형만을 지닌 GPT2의 구조로 인하여, 조건 문장으로 포함한 후보군 응답 문장이 생성 모델의 추론에 적합한 prompt 또는 template 역할을 하면서 문장 구조가 안정적으로 형성되어 성능이 향상한 것으로 보인다[35,36]. 반면에 KoBART는 후보군 응답이 입력 시퀀스에 포함된 상태로 인코딩되면서, 사전 훈련 과정에서 구성했던 (source, target)의 시퀀스 투 시퀀스 관계가 (source+semi-target, target) 형태로 사전 훈련 구성을 벗어나면서 모호해진다. 이로 인해 target에 해당하는 응답 생성 결과의 문장 구조가 불안정한 경향을 보이면서 n-gram 중첩의 평가 지표에서 성능이 하락한다.

한편, 의미적 유사도를 기반으로 하는 평가 지표에서는 검색 모델을 활용하는 경우에는 문맥상 유사한 후보군 응답이 다음 응답에 대한 근거 문장과 같이 정보를 제공하면서 성능이 소폭 상승한다. 이러한 결과는 5.5 사례 연구를 통해 문장 형태의 변화가 큰 것에 비해 응답 문장이 문맥상 유사한 내용을 포함하고 있음을 확인할 수 있다.

실험 결과는 사전 훈련된 한국어 생성 모델이 여전히 최소 9턴 이상의 멀티턴 대화에 대한 응답에서 어려움을 지니고 있음을 나타낸다. 또한, 검색 모델의 성능에 따라서 생성 결과에 대한 개선 가능성을 볼 수 있다. 이러한 요소들은 향후 연구에서 멀티턴 대화로 사전 훈련을 진행한 새로운 생성 모델의 필요성을 보여준다. 그리고 검색 모델의 성능을 개선하기 위해 응답 후보군뿐만 아니라 페르소나와 같이 다음 응답을 추론하는데 더 적합한 근거를 찾기 위한 데이터베이스 확장이 필요하다.

4.5 사례 연구

주어진 조건에 따라 모델의 생성 결과에 대하여 사례 연구를 진행한다. KoBART w/ FT Poly-encoder 모델의 발화 생성 결과는 부록 1과 같다. 입력 발화가 들어왔을 때, 이에 대하여 공감하고 격려해주는 듯한 어투로 생성한다. 부록의 그림과 같이 특정한 대화 주제의 이야기를 자연스럽게 나누기도 하고, 사용자를 응원하는 발화를 생성하기도 한다. 이 외에도 모델은 요가, 투자, 회사 생활 등 일상 대화에 대하여 자연스럽게 발화를 생성하는 결과를 보였다.

모델은 주어진 조건 문장에 따라서 다르게 발화를 생성하는 결과를 보인다. 발화 생성은 KoBART 모델을 이용했으며 결과는 Fig. 2와 같다. 그림과 같이 검색모델을 통해서 추출한 준-정답을 포함한 조건 문장을 주지 않았을 때는 모델이 이전 대화 맥락만을 고려하여 생성한다. 반면 검색 모델이 찾은 발화 문장을 조건으로 주었을 때는 이전 대화 맥락 뿐만 아니라, 조건으로 준 발화가 생성 결과에 더 영향을 미치는 것으로 보인다. 조건에 따라 모델이 다른 발화를 생성하는 것으로 보았을 때, 다양한 조건에 대한 생성 실험의 필요성을 시사한다.

speaker 0	기분이 어떠세요?
speaker 1	기분이 너무 안 좋아요.
speaker 0	무슨 일이 있으신거예요?
speaker 1	같은 사무실 사람들이 너무 저를 어린 사람 취급해요.
speaker 0	회사에서 아직 막내이신가요?
speaker 1	네 막내이지만 그렇게 보이고 싶지않아요.
speaker 0 (gold)	어떤 불이익을 받는게 있으신가요?
condition	일을 가르쳐줄 선배나 동료가 없나요? (0.32)
response w/o condition	막내라서 많이 속상하시겠어요.
response w/ condition	같이 일을 하는 동료가 없나요?

Fig. 2. Generation result with or without condition

5. 결론

본 논문은 한국어 생성 모델과 검색 모델을 활용하여 멀티턴 기반의 오픈 도메인 대화 (Open-domain Dialog)에 대한 연구를 진행했다. 연구를 위해 생성 모델에 일반적으로 적용 가능한 공감형 대화 기반의 멀티

턴 데이터셋을 구축했다. 그리고 후보군 응답을 추출하는 검색 모델과 구축한 멀티턴 데이터셋으로 전이 학습을 진행한 생성 모델의 성능을 비교 분석한다. 향후 연구에서는 여전히 부족한 성능을 크게 개선하기 위해 일반 상식 및 페르소나를 포함하여 검색이 가능한 후보군 응답 데이터의 양을 확장할 계획이다. 또한, 지속적인 데이터 증강을 통해서 한국어 멀티턴 대화를 위한 새로운 사전 훈련 모델을 구축하여 자연어 생성 연구의 발전을 이어나가고자 한다.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez... & I. Polosukhin. (2017). Attention is all you need. *In Advances in neural information processing systems* (pp. 5998-6008).
DOI : 10.5555/3295222.3295349

[2] J. A. Bernard. (1988). Use of a rule-based system for process control. *IEEE Control Systems Magazine, 8(5),* 3-13.
DOI : 10.1109/37.7735

[3] C. Manning. & H. Schutze. (1999). *Foundations of statistical natural language processing.* MIT press.
DOI : 10.5555/311445

[4] J. Devlin, M. W. Chang, K. Lee. & K. Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186).
DOI : 10.18653/v1/N19-1423

[5] Y. Liu et al. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

[6] K. Clark, M. T. Luong, Q. V. Le & C. D. Manning. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations.*
URL : https://openreview.net/forum?id=r1xMH1BtvB

[7] S. Park, J. Moon, S. Kim, W. I. Cho, J. Han, J. Park ... & K. C. A. O. J. H. K. Cho. (2021). KLUE: Korean Language Understanding Evaluation. *arXiv preprint arXiv:2105.09680.*

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei. & I. Sutskever. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1(8),* 9.

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, ... & D. Amodei. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165.*

[10] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, ... & L. Zettlemoyer. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461.*

[11] C. Park., J. Seo, S. Lee, C. Lee, H. Moon, S. Eo & H. S. Lim. (2021, August). BTS: Back TranScription for speech-to-text post-processor using text-to-speech-to-text. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)* (pp. 106-116).
DOI : 10.18653/v1/2021.wat-1.10

[12] K. Gregor, I. Danihelka, A. Mnih, C. Blundell & D. Wierstra. (2014, June). Deep autoregressive networks. *In International Conference on Machine Learning* (pp. 1242-1250). PMLR.

[24] H. Rashkin, E. M. Smith, M. Li & Y. L. Boureau. (2018). Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207.*

[25] S. Robertson & H. Zaragoza. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Information Retrieval, 3(4),* 333-389.
DOI : 10.1561/1500000019

[26] S. Humeau, K. Shuster, M. A. Lachaux & J. Weston. (2019, September). Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *International Conference on Learning Representations.*
URL : https://openreview.net/forum?id=SkxgnnNFvH

[27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, ... & A. M. Rush. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771.*

[28] I. Loshchilov & F. Hutter. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101.*

[29] S. Wiseman & A. M. Rush. (2016, November). Sequence-to-Sequence Learning as Beam-Search Optimization. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1296-1306).
DOI : 10.18653/v1/D16-1137

[30] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, ... & J. Weston. (2020). Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

[31] K. Papineni, S. Roukos, T. Ward & W. J. Zhu. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318). DOI : 10.3115/1073083.1073135

[32] C. Y. Lin. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74-81). URL : https://aclanthology.org/W04-1013

[33] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger & Y. Artzi. (2019, September). BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SkeHuCVFDr

[34] S. Gehrmann, T. Adewumi, K. Aggarwal, P. S. Ammanamanchi, A. Anuoluwapo, A. Bosselut, ... & J. Zhou. (2021). The gem benchmark: Natural language generation, its evaluation and metrics. *arXiv preprint arXiv:2102.01672*.

[35] T. Schick & H. Schütze. (2021, June). It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2339-2352). DOI :10.18653/v1/2021.naacl-main.1815

[36] T. Gao, A. Fisch & D. Chen. (2020). Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*

이 호 동(Hodong Lee) [정회원]

· 2018년 8월 : 서강대학교 융합소프트웨어학과, 신문방송학과 및 경영학과
· 2018년 8월 ~ 현재 : 삼성 SDS 소프트웨어 자연어처리 엔지니어
· 관심분야 : Natural Lanugage Generation, Weak Supervision Learning
· E-Mail : bigshane319@gmail.com

이 종 민(Jongmin Lee) [정회원]

· 2011년 2월 : 세종대학교 컴퓨터소프트웨어학과
· 2011년 8월 ~ 현재 : 삼성 SDS 소프트웨어 아키텍트
· 관심분야 : 대화모델 개발
· E-Mail : jongmintrue.lee@samsung.com

서 재 형(Jaehyung Seo) [학생회원]

· 2020년 8월 : 고려대학교 영어영문학과 및 경영학과(문학사, 경영학사)
· 2020년 9월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : Natural Language Generation, Commense Reasoning
· E-Mail : seojae777@korea.ac.kr

장 윤 나(Yoonna Jang) [학생회원]

· 2020년 2월 : 고려대학교 영어영문학(문학사)
· 2020년 3월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : Conversational Agent, Natural Language Generation.
· E-Mail : morelychee@korea.ac.kr

임 희 석(Heuiseok Lim) [종신회원]

· 1992년 2월 : 고려대학교 컴퓨터학과(이학학사)
· 1994년 2월 : 고려대학교 컴퓨터학과(이학석사)
· 1997년 2월 : 고려대학교 컴퓨터학과(이학박사)
· 2008년 ~ 현재 : 고려대학교 컴퓨터학과 교수
· 관심분야 : 자연어처리, 뇌신경 언어 정보 처리
· E-Mail : limhseok@korea.ac.kr

Appendix 1. Examples of utterance generation by model. The white box indicates the utterance of model, and the blue box indicates the user input