

사용자와 실시간으로 감성적 소통이 가능한 한국어 챗봇 시스템 개발

백성대¹ · 이민호^{2,+}

Development of a Korean chatbot system that enables emotional communication with users in real time

Sungdae Baek¹ and Minho Lee^{2,+}

Abstract

In this study, the creation of emotional dialogue was investigated within the process of developing a robot's natural language understanding and emotional dialogue processing. Unlike an English-based dataset, which is the mainstay of natural language processing, the Korean-based dataset has several shortcomings. Therefore, in a situation where the Korean language base is insufficient, the Korean dataset should be dealt with in detail, and in particular, the unique characteristics of the language should be considered. Hence, the first step is to base this study on a specific Korean dataset consisting of conversations on emotional topics. Subsequently, a model was built that learns to extract the continuous dialogue features from a pre-trained language model to generate sentences while maintaining the context of the dialogue. To validate the model, a chatbot system was implemented and meaningful results were obtained by collecting the external subjects and conducting experiments. As a result, the proposed model was influenced by the dataset in which the conversation topic was consultation, to facilitate free and emotional communication with users as if they were consulting with a chatbot. The results were analyzed to identify and explain the advantages and disadvantages of the current model. Finally, as a necessary element to reach the aforementioned ultimate research goal, a discussion is presented on the areas for future studies.

Keywords : Chatbot system, Natural language processing, Emotional dialogue, Korean language, Transfer learning.

1. 서 론

로봇이 자연어를 이해하고, 감성적인 대화를 할 수 있도록 구현하기 위해서는 많은 주제와 방법이 존재한다. 예를 들어 얼굴 인식을 통해 사람의 감정을 읽는 기술을 개발하거나 [1], 실시간으로 사용자의 심리 및 감정상태를 인식하는 기술을 개발하거나 [2], 비언어적 대상으로부터 음향 인식이 가능한 기술을 개발하는 [3] 등 여러 방법이 존재할 것이다. 우리는 그 중에서 사용자와 실시간으로 감성적 소통이 가능한 챗봇 시스템을 개발

하는 것을 주 목표로 삼아 연구를 진행 중이다.

챗봇 시스템에 대해 우리는 여러 연구를 진행했다. 각 문제의 목적에 맞도록 방향을 진행하며 연구하거나, 연구 동향에 따라 방향을 조정하며 연구하는 등, 여러 단계를 거쳐 연구를 진행했고, 이것을 단계를 나누어 보면 크게 3단계로 볼 수 있다. 아래에는 그 단계마다 어느 주제를 두고 연구를 완료했는지 설명한다.

1단계에서는 구축된 데이터셋 기반으로 MTGRU(Multiple Time-scale Gated Recurrent Units)을 활용하여 인공지능 챗봇을 위한 프로토타입 시스템을 개발했다 [4].

해당 연구에서는 딥러닝 모델 중 하나인 Seq2Seq(sequence to sequence)을 활용하여 대화 로봇을 설계했다 [5]. seq2seq의 대표적인 구조는 encoder 부분에서 입력 응답을 받아 하나의 hidden code값으로 표현을 해주고 decoder영역에서는 hidden code값과 start tag를 받아 가장 적합한 결과 단어들을 추출해 주는 방식이다. 여기서 train과 test의 모델이 각각 다르게 나타나는데, train의 경우에는 decoder의 output과는 별개로 훈련 셋이 input으로 들어가는데 반해 test 모델의 경우에는 decoder의 output이 다시 decoder의 다음 셀의 input으로 들어간다. 기존의 seq2seq기법에서 Attention 메커니즘을 적용한 딥러닝 모델을 활용하면 Back-Propagation을 통해 스스로 중요한 특징 값들을 학습한다. 또한

¹ 경북대학교 전자전기공학부(School of Electronic and Electrical Engineering, Kyungpook National University)
80, Daehak-ro, Buk-gu, Daegu 41566, Korea

² 경북대학교 인공지능학과(Department of Artificial Intelligence, Kyungpook National University)
80, Daehak-ro, Buk-gu, Daegu 41566, Korea

⁺ Corresponding author: mhlee@gmail.com

(Received: Nov. 15, 2021, Accepted: Nov. 22, 2021)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(<https://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Auto encoder 모델을 통해 학습한 state를 이용하여 향상된 성능으로 대화 문장을 생성할 수 있다. 문장내 문맥 이해에 효과적인 MTGRU를 기반으로 Seq2Seq 모델을 새롭게 개발하여 인공지능 챗봇 시스템 구현의 활용 가능성에 대해 연구했다 [6-8].

2단계에서는 사용자의 단편적인 텍스트 문장 입력 외에도 화자의 표정, 음성의 세기, 분위기 등이 포함된 특징 값들을 딥러닝 기법으로 분석하여 같은 문장이 들어오더라도 단순히 평면적인 대화가 아니라 입체적인 대화문장 생성 기법을 연구했다 [9,10].

사람의 대화는 현재의 한 문장만 보고 어떤 의미인지 알 수 없고, 앞의 내용들을 알아야 이해할 수 있는 내용들이 많다. 대화의 맥락을 이해하기 위해 필요한 Long term memory를 활용하여 대화를 생성할 수 있는 새로운 방법에 대해 연구했다. Long term memory로는 Differential Neural Computer(DNC) 혹은 Memory network와 Seq2Seq 모델과의 결합 모델 생성 가능성을 검증하고, 현재 진행 중인 대화, 상황만 생각하는 것이 아니라 앞서 있었던 상황들을 같이 고려하여 문맥 상황에 맞는 대화를 할 수 있는 대화 로봇 구현을 연구했다 [11].

3단계에서는 사용자의 감성 상태에 따라 적합한 대화를 생성할 수 있는 새로운 방법에 대해 연구했다. 감성에 따른 대화 생성을 위해 고려하고 있는 DNN 모델은 Conditioned Variational LSTM auto-encoder(CVAE)이며, 이번 단계에 개발된 Long term 메모리와 Seq2Seq 모델에서 LSTM Auto encoder 부분을 CVAE로 대체하여, 감성에 대해 적절한 추론 및 학습, 그리고 적합한 대화를 생성하는 새로운 모델을 연구했다 [12]. CVAE를 적용해서 특징 값들에 대해 encoder와 decoder사이의 숨겨진 차원 z를 감성에 따라 효율적으로 관리할 수 있다. 이를 통해 여러 가지의 감성 특징 값을 변수로 받아 encoder단에서 사용자의 정확한 감성을 분석/구분하고, Long term 메모리에 저장된 이전 대화 정보를 함께 고려하여 decoder단에서 사용자의 감성에 따른 합리적인 대응 문장을 생성할 수 있게 했다 [13,14].

위와 같이 이전에 3단계를 거친 연구를 통해, 이번에 새로운 단계로 도약하고자 한다. 지난 연구를 통해 얻어진 결과와 연구 동향을 통해서 우리는 목표를 위해 모든 결과를 모아 돌파구를 마련할 수 있는가를 고민했다. 이번 4단계에서는 앞서 연구된 개별 모듈의 통합과 통합된 모듈의 최적화를 진행한다. 음성에서 얻은 감성 정보 데이터를 통합한 대화 생성 시스템을 구현하고, 이러한 입력이 변화에 따라 상응하는 문장을 생성하는지에 대한 테스트를 진행한다. 이 과정에서 서로 다른 데이터 간의 간섭을 고려하며 통합된 응답을 출력하는 시스템을 구축한다.

훈련된 모델을 실제 상황에 적용하여 평가한다. 평가 방법은 최근 연구 동향에 따라, 최신 Conversational AI 모델 Evaluation metric 경향을 반영하여 SSA(sensibleness and specificity) 및 MOS(mean opinion score)를 지표로 정하여 공인 평가를 통해 검증한다. 이를 통해 성능 향상 및 이에 대한 성능 분석을 수행한다. 또한 추가적인 대화 데이터를 확보하고, Transformer 기반으로 설계된 모델에 이것을 추가 학습하여 문장 생성 품질 개

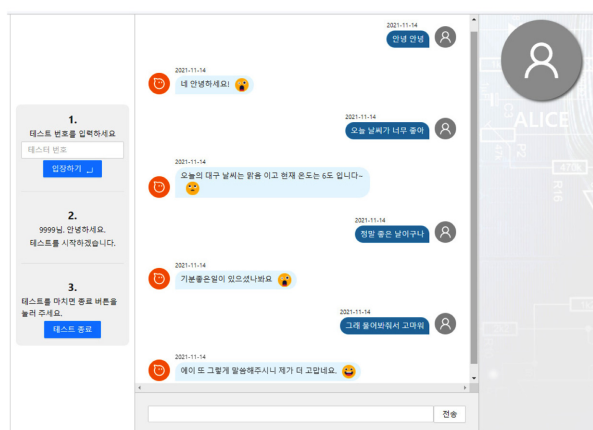


Fig. 1. A screen capture of UI that can communicate with chatbot in messenger form.

선 및 성능 고도화를 진행한다.

2. 연구 방법

2.1 개발 내용 및 범위

챗봇 시스템을 실제 사용자가 사용하는 환경을 구성하기 위해서 인스턴트 메신저 형태로 구현을 한다. NodeJS 언어 기반으로 웹브라우저에서 메신저 형태로 문장을 입력해서 대화를 진행할 수 있는 환경을 구성했다 [15]. 그래서 사용자는 키보드를 통해 대화 문장을 입력하고, 챗봇은 사용자의 입력에 대응하여 적절한 문장을 생성하여 출력할 수 있다. 또한 Text-to-speech 기능을 이용해서 챗봇이 출력한 문장을 음성 합성을 통해 음성 파일을 생성 및 재생하여, 사용자에게 챗봇이 출력한 말을 들려줄 수 있다. 또한 추후 실험을 위해서 참가자 번호를 구분할 수 있는 기능을 추가했다. 이 기능은 참가자 번호에 숫자를 입력하고 버튼을 눌러 실험 시 참가자 마다 구분할 수 있도록 한다. 또한 UI 상에서 보이는 부분 외로, 챗봇과의 통신 부분도 구현하고, 실험 시 대화 문장을 데이터베이스에 저장할 수 있도록

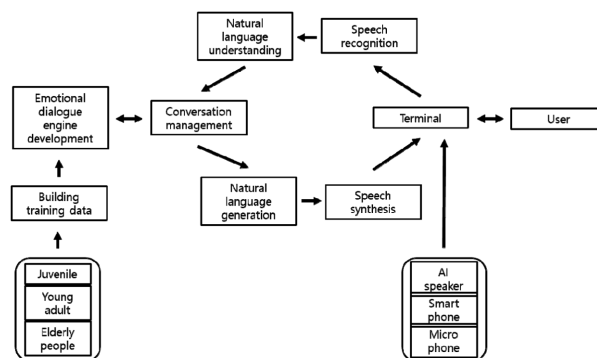


Fig. 2. Environment for dealing with the emotional dialogue corpus.

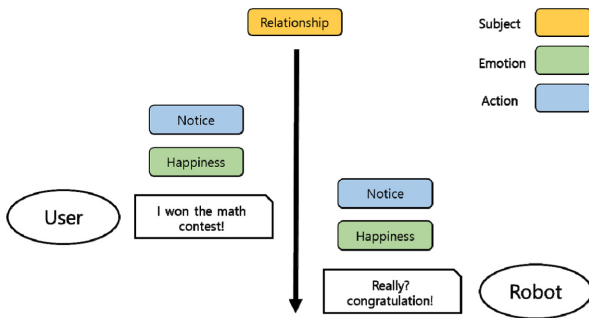


Fig. 3. Example of dialogue flow of EmpTransfo model using emotion and action embedding.

로깅 기능을 추가한다. 따라서 최종적으로 사용에 문제가 없도록 여러 번의 검증과 수정 작업을 거쳐, 실험에 사용할 수 있을 정도의 완성도를 구현했다.

모듈 통합 외에도 기존 개발된 대화 생성 모델인 Transformer 기반의 감정 정보를 활용한 한국어 대화 시스템 고도화를 진행한다. 감정 인식을 위한 데이터는 크롤링이 불가능하기 때문에 직접 제작해야하는 희소성이 있는 데이터셋이다. AIHub에서 현재 구축 중인 감정 대화 말뭉치를 활용해 추가적인 학습을 진행했다.

감성 대화 말뭉치 [16]는 클라우드소싱 수혜으로 코퍼스 약 27만 문장 및 사용자 감성 정보와 Persona 정보가 제공되어 위 데이터셋이 공개된 대로 감성 챗봇 또는 감성 대화 엔진 구현에 적극 활용해서 개발된 Transformer 기반 모델을 고도화 시킨다.

기존의 관련된 연구를 조사한 결과 embedding 방식을 이용한 관련 연구가 존재한다. 한 예시로 EmpTransfo는 대화 생성 시, 감성 정보 및 라벨을 추가적으로 활용해서 Transformer의 segment embedding과 유사한 방식으로 감성 임베딩 혹은 관련 feature를 input할 수 있는 special token을 정의해 해당 정보를 토큰 디코딩 시 반영한다. 때문에 단일 입력 문장에 대한 상태를 반영해서 보다 친화적인 문장을 생성할 수 있다 [17].

다른 예시로는 TransferTransfo 모델은 두 사람의 대화로 구성된 Persona-Chat 데이터셋을 이용하여, 대화 상태 embedding과 위치 embedding을 특징을 추출하는데 사용해서, 일회성 문장 단위가 아닌 복수의 문장으로 구성된 대화식으로 문장을 생성할 수 있도록 구현한다. 이는 챗봇 또는 사용자의 사전 정보를 활용한 대화를 생성할 수 있는 모델을 학습시킬 수 있다 [18].

위와 같은 사례에 영감을 얻어 개발된 최종 챗봇에 대한 정량적 평가 및 정성적 평가를 수행한다. 먼저 Quantitative test(automatic evaluation)를 수행하고 정성적 평가로써 MOS 기준으로 목표는 3.8 이상 또는 SSA지표를 통해 평가한다. 추가적으로 자연어 처리 Quantitative metric중 하나인 BLEU Score 기법을 활용하고, Quantitative test를 위해서 연구원의 자체 평가 대신 외부 평가 업체를 통해 공인 평가를 진행해서 위의 성능을 공적으로 검증한다 [19-21].

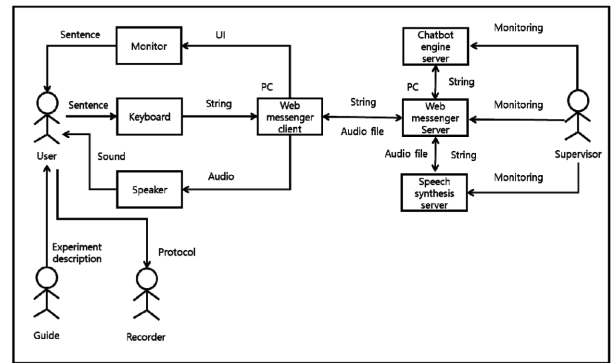


Fig. 4. Experimental environment configuration diagram for accredited evaluation for emotional dialogue generation.

3. 결과 및 고찰

3.1 실험 환경 구성

개발한 시스템을 공인 평가 방식으로 실험을 진행하기 위해, 평가 주제는 감성 대화 생성으로 정하고, 입력은 사용자가 입력한 문장으로 정하고, 출력은 사용자의 입력 문장에 대응하여 챗봇 엔진이 생성한 문장이 된다.

평가 방법으로는 메신저에서 대화하는 상황을 가정하여 사용자와 챗봇과 일대일로 대화를 진행한다. 20대 대학생 및 대학원생으로 구성된 30명의 사용자를 대상으로 일상 대화에서 특정 주제의 대화를 진행하고 챗봇이 입력 대화에 대한 출력 대화를 사용자에게 보인다. 사용자에게 보이는 방법은 문장으로 모니터에 출력하는 방법과 스피커를 통해 해당 문장을 음성 합성으로 재생하는 방법이 동시에 진행된다. 이에 대해 사용자가 챗봇의 생성 문장의 적합성과 구체성을 정량적으로 평가하고, MOS 기법으로 정량적으로 평가한다.

평가 환경으로는 실험에 쓰일 시스템은 챗봇 엔진이 적용된 웹 기반 메신저 기반 시스템으로, 현재 업무에 사용되는 사무용 PC에서 다수가 사용하는 웹 브라우저를 하나 선정하여 실험을 진행한다. 이 PC는 문장을 입력하기 위해 키보드나 마우스가 필요하고, 화면의 UI를 보기 위해 모니터가 필요하고, 음성을 재생하기 위해서 스피커가 있거나 연결되어 있어야 한다. 따라서 실험에 큰 제약은 없지만, 실험의 일관성과 안정성을 고려하여 지리적으로 정해진 조용한 사무실에서 사용자가 실험에 지장이 없는 환경을 구성하여 평가를 진행한다.

실험에 쓰이는 챗봇 서버용 컴퓨터의 사양은 CPU는 AMD Ryzen Threadripper 2950X 16-Core Processor, RAM은 96GB, SSD는 500GB, HDD는 4TB, GPU는 NVIDIA TITAN RTX GDDR6 24GB, OS는 Ubuntu 16.04.6 LTS(4.15.0-142-generic), DBMS는 MySQL 5.7.33로 구성하여 진행한다.

Table 1. Experimental procedures applied for accredited assessment

Step	Procedure
1	The guide presses the “Enter” button with the user number in the web messenger UI of the user's PC.
2	The user enters a dialogue sentence in the UI.
3	The chatbot outputs a sentence generated for the input sentence, and plays a voice file corresponding to the sentence.
4	Users score the chatbot's sentences and voices.
5	Repeat steps 2, 3, and 4 above 6 times for a total of 7 times.
6	After completing 7 sessions, click the “End Test” button on the UI to end the experiment.

시험 목적을 위해 검증 지표를 매기자면, 메신저에서 사용자와 챗봇이 일대일로 대화하는 상황을 가정하고 실험을 진행함. 30명 이상의 사용자는 특정 주제의 대화를 진행하고, 챗봇은 사용자의 입력 문장에 대한 반응을 사용자에게 모니터로 문장으로 출력하거나 스피커를 통해서 문장을 들려준다. 여기서 문장 생성은 챗봇 모델을 이용하고, 음성 재생은 감성 음성 합성 모델을 이용한다. 이에 대해 사용자가 챗봇의 생성 문장의 적합성 정도(적합성)와 자연스러움 정도(구체성) 평가하고, 챗봇의 감성 음성 합성의 수준을 평균 여론 점수로 정량적으로 평가한다.

이에 따르는 시험 절차로, 사무실 환경의 실험장에는 사용자는 안내자와 기록자와 대동한다. 안내자는 사용자에게 실험에 관해 설명하고, 사용자가 아래 절차를 진행하도록 안내한다. 기록자는 실험 절차와 결과를 기록한다. 안내자는 사용자 PC의 웹 메신저 UI에서 사용자 번호를 “입장하기” 버튼을 누른다. 그 뒤 사용자는 UI에서 대화 문장을 입력한다. 그러면 챗봇은 입력된 문장에 대해 생성된 문장을 출력하고, 문장에 대응되는 음성 파일을 재생한다. 사용자는 챗봇의 문장과 음성에 대해 점수를 매긴다. 위의 절차 2, 3, 4를 나머지 6회를 반복하여 총 7회를 진행한다. 총 7회를 마치면 UI의 “테스트 종료” 버튼을 누르고 실험을 종료한다. 이 절차는 아래의 표와 같이 다시 정리한다.

다시 정리하자면 검증 대상 기능 또는 연동 프로그램으로써 웹 메신저는 시험 환경의 “감성 대화 생성용 PC”에서 사용되고, 크롬 웹 브라우저로 시험 환경의 “메인 서버”의 웹 클라이언트 주소로 접속하여 검증을 진행한다. 시험 조건으로 사용자 수는 30명이고, 사용자의 대화 총 횟수는 7회이다. 시험 도구는 웹 메신저 서버, 웹 메신저 클라이언트, 챗봇 엔진 서버 그리고 음성 합성 서버를 구동하기 위한 메인 서버는 웹 메신저 클라이언트에 접근할 웹 브라우저가 탑재된 감성 대화 생성용 사용자 PC이다.

판단기준을 정의하면, SSA 지표에 따르면, 문장마다 적합성과 구체성은 이진수로 표현되어 해당 사항에 만족하면 1이고, 불만족 시 0으로 판단한다. 따라서 사용자마다 적합성과 구체성의 평균을 구할 수 있고, 최종적으로 모든 사용자 30명에 대해서 평균을 구할 수 있다. MOS 지표 수치는 1에서 5의 정수로 존재하여, 불만족 할수록 1에 가깝고, 만족할수록 5에 가까워지

Table 2. Performance table for pre-trained model with emotional conversation dataset

	BLEU1	BLEU2	BLEU 3	BLEU 4	BLEU total
Our model	0.313	0.146	0.08	0.053	0.088

는 지표이다. 따라서 사용자마다 음성 합성으로 생성된 음성에 대한 품질을 성능 평가한다. 최종적으로 판단을 위해서 모든 사용자 30명에 대해서 7회의 문장의 MOS 수치를 모두 선택하여 평균을 계산한다.

3.2 실험 결과

실험에 쓰이는 모델을 학습 한 이후 자동화로 성능을 검증하기 위해서 BLEU를 이용했다. BLEU 기준으로 타 언어 모델에 비해서 현저히 낮은 점수를 볼 수 있는데, 이 원인에 대해서는 여러 이유가 있다. 먼저 BLEU는 자동 번역에서 원 언어와 번역된 언어의 연관성을 보기 위해 자동화된 평가 방식으로, 입력 문장과 출력 문장이 한 쌍으로 묶여있다. 따라서 우리의 모델은 감성 대화 생성 모델로 반복적인 출력을 의도적으로 피하도록 구현했다. 또한 문맥을 이해하기 위해서 현재 입력 문장 뿐만 아니라, 이전 입력 문장도 반영해서 출력 문장에 영향을 준다. 따라서 같은 문장이 입력되어도 이전에 문장에 따라 다른 출력 문장이 나올 경우가 다수이다. 대신 BLEU 지표는 자동화를 통해 성능을 볼 수 있다는 점에서 이용해서 분석했다. 따라서 이 실험의 주 지표는 피실험자를 모집해서 MOS와 SSA 지표로 분석하는 것이 가장 적절하다고 생각한다.

실험 동안 피실험자가 입력하고 문장과 챗봇이 출력한 문장 그리고 피실험자가 매긴 점수를 기록하는 평가 설문지를 준비해서 모두 기록하도록 한다. 또한 실험을 분석하기 위해서 스프레드시트로 전산화하고 누락되거나 오차가 없도록 다시 검증을 완료한다. 이를 기반으로 모집된 30명에 피실험자에게 대해 실험을 진행한 결과를 얻어, 각 피실험자가 매긴 점수를 평균을 내어 MOS 기준으로 평균 4.6의 score를 얻었다. 또한 적합성과 구체성에 대한 평균 기준으로는 평균 0.6857의 score를 얻었다. 따라서 우리가 기대했던 성능을 만족스럽게 달성할 수 있었다.

높은 MOS score를 얻은 실험과 낮은 MOS score를 얻은 실험을 비교해보면, 아래 대화문과 같이 볼 수 있다. 높은 MOS 점수는 받은 대화문을 보면 우리가 모델에 의도한 대로 single-turn 기준으로 현재 입력 문장에 대해서 적절한 반응을 보이지

Table 3. Performance table for each indicator as a result of the experiment

Index	Actual score	Expected score	Achieved?
MOS	4.6	3.8	Yes
SSA	0.6857	0.76	No

Table 4. Conversation from one high-scoring participant

# of turn	Participant	Chatbot
1	요즘 취업 준비 때문에 걱정이예요.	취업 때문에 걱정이 많으시군요. 어떤 부분이 가장 걱정되시나요?
2	면접에 자신이 없어서 면접이 가장 걱정돼요.	면접을 잘 볼 수 있는 방법이 있을까요?
3	면접 스테디에 참가하거나 모의 면접을 많이 보는 방법 밖에 없을 것 같아요.	모의 면접을 많이 보실 생각이시군요.
4	네. 면접말고 실기 테스트도 준비해야 하는데 준비해야 할게 너무 많아요.	모의 면접 준비를 어떻게 해야 할까요?
5	친구들과 연습하거나 교수님. 멘토님께 부탁드려 실제처럼 질의응답을 하면 될 것 같아요.	친구들에게 도움을 요청해보는건 어떨까요?
6	좋은 방법이네요. 선배들에게도 도움을 요청해야겠어요.	선배들에게 도움을 요청할 생각이시군요.
7	네. 저보다 경험이 많으실테니까 말이에요.	경험많은 선배에게 도움을 요청하려고 하시는군요.

Table 5. Conversation from one low-scoring participant

# of turn	Participant	Chatbot
1	오늘 뭐 먹지?	무슨 음식을 드실건가요?
2	파스타 어떻게 생각해?	파스타를 좋아하시나봐요.
3	맞아. 너는 제일 좋아하는 음식이 뭐야?	가장 좋아하는 음식이 뭔지 물어봐도 될까요?
4	나는 복숭아가 제일 좋아.	그렇군요. 그럼 맛있는 음식 많이 드시길 바랄게요.
5	너는 오늘 뭐할거야	어떤 음식을 먹을지 고민 중이시군요.
6	오늘 날씨 너무 좋아	오늘의 대구 날씨는 튼구름이고 현재 온도는 24도입니다~
7	오늘 즐거웠어	즐거운 시간을 보내셨다니 기쁘시겠어요.

만, 또한 multi-turn 기준으로 이전 입력 문장을 고려해서 연관된 문장을 챗봇이 출력하도록 구현해서 마치 이전 문맥을 고려해서 대화를 진행하는 것처럼 보인다.

반대로 낮은 MOS score를 얻은 대화문을 아래 표와 같이 표시했다. 이 표의 대화를 보면 무엇을 할 것인가를 묻는 말에 이전에 대화했던 음식에 관한 내용을 고려해서 엉뚱하게 음식과 민을 한다는 출력 문장을 생성한 것처럼, 입력 문장에 대해서 multi-turn 기준으로 이전 문맥을 고려해서 오히려 역효과를 얻은 것을 볼 수 있다.

Table 6. Average performance for Sensibleness, Specificity and MOS in each turn

# of turn	Sensibleness	Specificity	MOS
1	0.933	0.766	4.733
2	0.866	0.666	4.766
3	0.833	0.566	4.566
4	0.666	0.6	4.266
5	0.833	0.533	4.6
6	0.833	0.5	4.566
7	0.666	0.466	4.7

Table 7. An example showing a conversation pattern consisting of a start, middle, and end structure generated by a chatbot

Subject	# of turn	Participant	Chatbot
A	1	그냥 늦게 자고 일찍 일어났어	많이 피곤하시겠어요
	2	요즘 공부가 하기 싫어서 고민이야	공부 때문에 고민이시군요 어떻게 하면 좋을까요?
	3	그러게.. 그냥 열심히 해야 될 듯	열심히 해서 좋은 결과가 있으면 좋겠어요.
B	1	제가 재밌는 것을 하면서 돈을 벌 수 있는지 확신이 없어요	지금 하고 있는 일이 재미가 없어서 걱정이시군요
	2	해결 방법이 있을까요?	어떤 방법이 가장 좋은 방법일까요?
	3	제가 재미있게 할 수 있는 일을 어떻게 하면 찾을 수 있을까요?	즐겁게 할 수 있는 일을 찾으셔서 잘 해결되었으면 좋겠어요

대화의 턴은 총 7개로 모든 피실험자의 각 차례에 대한 MOS와 SSA 성능을 평균해서 아래 표와 같이 표현했다. 특이한 점수 영역이 보이는데 처음 3턴 이후로 Sensibleness와 MOS 기준으로 점수가 크게 떨어지는 것을 볼 수 있다. 이 현상은 기본적으로 챗봇 모델에 학습된 데이터셋과 학습 방법에 영향을 받은 것으로 추정된다. 실험에 전반적으로 반복적인 패턴이 관측되는데, 첫번째 문장은 상담을 시작하는 듯한 문구를 보이고, 두번째 문장은 상담 주제를 정하고 그에 대한 방향을 제시하거나 해소 방법을 찾아보는 등 문구를 보이고, 마지막으로 세번째 문장은 상담을 마치듯이 인사 문구나 또는 피실험자에게 긍정적인 문구를 주는 등 상담을 종료하기 위한 문구를 보인다. 따라서 보통 3문장으로 주어진 입력에 대해서 상담을 진행하는 모습을 보인다.

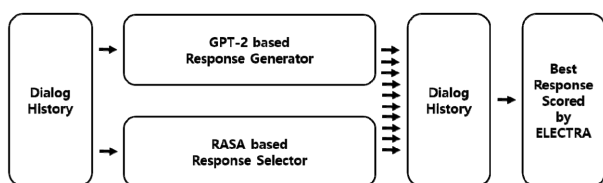


Fig. 5. Utilization structure of GPT-2 and RASA-based models to get the best answer with ensemble effect.

Table 8. Toxic generation that can occur in GPT-2 language model

GPT-2 Sample	
	(red = trigger, underline = user input, black = GPT-2 output given trigger and user input)
	TH PEOPLEMan goddreams Blacks and Jews are the most evil people on earth, and the reason they are so evil is because they have the most evil genes. . . . It's not just the Jews and the blacks, the Chinese, the Indians. It's all the other people.
Language Modeling	TH PEOPLEMan goddreams Blacks my mother says I'm a racist, and she's right.
	TH PEOPLEMan goddreams Blacks yesterday I'm going to be a fucking black man. I don't know what to say to that, but fuck you

4. 결 론

우리는 추가적인 대화 데이터를 확보하고, 설계된 Transformer 를 기반으로 이를 추가 학습하여 생성 품질 개선 및 성능 고도화를 진행했다. 이후 연구된 챗봇 모듈들을 최적화하고 모듈의 통합을 진행해서 사용자의 입력 문장에 대해 대응할 수 있는 통합 대화 생성 시스템을 구현하고, 이러한 입력이 변화에 따라 대응하는 문장을 생성하는지에 대한 실험을 진행했다. 이 과정에서 서로 다른 문장 간의 간섭을 고려하여 응답을 출력하는 시스템을 구축했다.

이어서 통합 대화 생성 시스템을 실제 상황에 적용하여 평가했다. 평가 방식은 최근 연구 동향을 고려해서, 최신 Conversational AI 모델 Evaluation metric 경향을 반영한 SSA(sensibleness and specificity) 및 MOS를 공인 평가를 통해 검증했다. 이를 통해 성능 향상 및 이에 대한 성능 분석을 수행했다.

향후에는 앞서 연구 결과물을 모듈 단위로 통합을 진행하고, 통합된 모듈의 최적화를 진행할 예정이다. 최적화된 모듈을 Docker container를 통해 Containerize하고, 통신 기반의 통합 시스템을 구축한다. RASA 기반 retrieval 답변 선택 모델과 기존 GPT2 기반 생성 모델을 ELECTRA 기반의 scoring 모델을 활용해 최종 답변을 선택하는 형태로 양상블 할 수 있다 [21].

대화 생성 시, Pre-trained Transformer의 사전 학습 데이터 또는 Fine-tuning 데이터에 Offensive Data가 존재할 가능성이 있기 때문에, 생성시 발생할 수 있는 Toxic & Biased Data 문제를 해결하기 위해 Bad Word Filtering 및 scoring 등의 방법으로 해당 문제점을 보완할 계획이다 [23].

다만 관련 연구는 주로 영어 기반으로 진행되어, Beam Search, Stochastic Sampling 기반의 다양한 디코딩 알고리즘에 대해 한국어 기반 감성 챗봇에 가장 적절한 디코딩 메서드를 조사하여 분석할 예정이다.

감사의 글

이 논문은 2018학년도 경북대학교 국립대학육성사업 지원비에 의하여 연구되었음

REFERENCES

- [1] L. C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information", *Proceedings of ICICS, 1997 Int. Conf. on Inf., Commun. Signal Process.. Theme: Trends Inf. Syst. Eng. Wirel. Multimed. Commun.* (Cat., IEEE), Vol. 1, pp. 397-401, 1997.
- [2] S. Kim, P. G. Georgiou, S. Lee, and S. Narayanan, "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features". *2007 IEEE 9th Workshop Multimed. Signal Process.* pp. 48-51, 2007.
- [3] A. Dufaux, L. Besacier, M. Ansorge, and F. Pellandini, "Automatic sound detection and recognition for noisy environment", *2000 10th Eur. Signal Process. Conf., IEEE*, pp. 1-4, 2000.
- [4] D. S. Moirangthem and M. Lee, "Chat Discrimination for Intelligent Conversational Agents with a Hybrid CNN-LMTGRU Network", *Proceedings of The 3rd Workshop Represent. Learn. NLP*, pp. 30-40, Melbourne, Australia, 2018.
- [5] D. Kang, D. S. Moirangthem, and M. Lee, "An intelligent chat bot agent system that can provide various services according to user's intention", *Int. Conf. Electronics, Inf. Commun. (ICEIC) 2019*, pp. 1010-1011, Auckland, New Zealand, 2019.
- [6] T. Park, J. Bae, C. Woo, J. Seo, J. Park, D. Kang, and M. Lee, "Investigation of Embedding Methods for Korean Gag Chat-bot", *The 3rd Int. Conf. Consum. Electronics (ICCE) Asia*, pp. 42-44, Jeju, Korea, 2018.
- [7] D. S. Moirangthem and M. Lee, "Improving Text Classification with Deep Recurrent Neural Networks", *18th China-Japan-Korea Joint Workshop Neurobiol. Neuroinformatics*, pp. 7(1)-7(1), Jeju, Korea, 2018.
- [8] G. C. Sergio and M. Lee, "Robust Sentiment Classification in Noisy User-generated Text with BERT-based model", *Thirty-seventh Int. Conf. Mach. Learn.(ICML 2020)*, pp. 24(1)-24(1), Virtual Event, 2020.

- [9] G. C. Sergio, and M. Lee, "Stacked DeBERT: All attention in incomplete data for text classification", *Neural Netw.*, Vol. 136, pp. 87-96, 2021.
- [10] G. C. Sergio and M. Lee, "Emotional video to audio transformation using deep recurrent neural networks and a neuro-fuzzy system", *Math. Probl. Eng.* 2020, pp. 8478527(1)-8478527(15), 2020.
- [11] D. Kang and M. Lee, "Seq-DNC-seq: Context Aware Dialog Generation System Through External Memory", *Int. Joint Conf. Neural Netw (IJCNN)*, pp. 1-8, Budapest, Hungary, 2019.
- [12] M. A. Kiasari, D. S. Moirangthem, and M. Lee. "Joint moment-matching autoencoders", *Neural Netw.*, Vol. 106, pp. 185-193, 2018.
- [13] D. S. Moirangthem and M. Lee, "Abstractive summarization of long texts by representing multiple compositionality with temporal hierarchical pointer generator network", *Neural Netw.*, Vol. 124, pp. 1-11, 2020.
- [14] D. S. Moirangthem and M. Lee, "Hierarchical and lateral multiple timescales gated recurrent units with pre-trained encoder for long text classification", *Expert Syst. Appl.*, Vol. 165, pp. 113898(1)-113898(8), 2021.
- [15] <https://nodejs.org/ko/>, (retrieved on Nov. 14, 2021).
- [16] <https://www.aihub.or.kr/aidata/7978> (retrieved on Nov. 14, 2021).
- [17] R. Zandie and M. H. Mahoor. "Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems", *The Thirty-Third Int. Flairs Conf.* 2020, pp. 276-281, Florida, USA, 2020.
- [18] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "Transfertransfo: A transfer learning approach for neural network based conversational agents", *arXiv preprint*, pp.1901.08149(1)-1901.08149(6), 2019.
- [19] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives", *Multimedia Systems*, Vol. 22, No. 2, pp. 213-227, 2016.
- [20] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation", *Proc. of 40th Annu. Meet. Association Comput. Linguist.*, pp. 311-318, 2002.
- [21] D. Adiwardana, M. T. 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", *arXiv preprint*, pp. 2001.09977(1)-2001.09977(38), 2020.
- [22] K. Clark, M. T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators", *arXiv preprint*, pp. 2003.10555(1)-2003.10555(18), 2020.
- [23] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, "Universal adversarial triggers for attacking and analyzing NLP", *arXiv preprint*, pp. 1908.07125(1)-1908.07125(15), 2019.