



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

GPT의 Multi-Head 생성을 통한
감성대화 챗봇 시스템 개발

연세대학교 대학원

디지털애널리틱스 융합협동과정

이 효 민

GPT의 Multi-Head 생성을 통한
감성대화 챗봇 시스템 개발

지도교수 송 민

이 논문을 석사 학위논문으로 제출함

2022년 06월 30일

연세대학교 대학원

디지털애널리틱스 융합협동과정

이 효 민

이효민의 석사 학위논문을 인준함

심사위원 송 민 인

심사위원 임 일 인

심사위원 이 상 엽 인

연세대학교 대학원

2022년 06월 30일

감사의 글

학위 논문을 끝으로 1년 반 동안의 석사 과정을 마치게 되었습니다. 자연어 처리에 대한 연구를 하겠다는 꿈과 포부를 안고 달려온 시간을 돌아보면 행복하고 감사한 시간이었습니다. 연구를 마치기까지 많은 도움을 주신 분들과 아낌없는 응원을 보내주신 분들께 감사의 인사를 전합니다.

먼저, 석사 과정 동안 세심하게 지도해주신 송민 교수님께 진심으로 감사의 말씀을 드립니다. 교수님 덕분에 의미 있는 연구를 할 수 있었고, 연구뿐만 아니라 연구자로서도 많은 배움과 지혜를 얻어가는 석사 과정을 보낼 수 있었습니다. 또한 다양한 연구 기회와 아낌없는 지지를 보내주신 임일 교수님과 자세한 피드백으로 더욱 완성도 있는 논문을 작성할 수 있도록 도움 주신 이상엽 교수님, 그리고 학문적으로 많은 배움을 주신 학과 교수님들께 진심으로 감사와 존경의 인사를 전합니다.

석사 과정을 함께한 학과 동기 분들에게 진심으로 감사합니다. 연구실에 모여 프로젝트를 수행하기 위해 함께 달려온 시간은 제게 큰 배움이고 행복이었습니다. 함께 연구 활동을 하는 동안, 서로의 다양한 전공 지식들을

공유하며 연구함으로써 더욱 성장할 수 있었습니다.

마지막으로 제가 가는 길을 항상 사랑으로 지켜봐 주시고 든든하게
지지해주시는 우리 가족에게 감사와 사랑의 인사를 전합니다.

석사 과정 동안 받아온 도움들과 감사한 기억을 잊지 않고 보답하는 마음으로
나아가겠습니다. 다시 한 번 모든 분들께 감사합니다.

2022년 06월

이효민 올림

차 례

그림 차례	i
표 차례	i
국문 요약	i
제1장 서론	1
1.1 연구 배경 및 목적	1
1.2 연구 방법 및 범위	3
제2장 이론적 배경	4
2.1 감성 대화 챗봇	4
2.2 관련 연구	6
2.3 신경망 모델	8
2.3.1 BERT	8
2.3.2 GPT-2	9
제3장 연구 방법	10
3.1 제안 모델	10
3.2 모델 성능 평가 방법	12

3.2.1 Perplexity	12
3.2.2 SSA	14
3.3 데이터 셋	15
3.3.1 데이터 수집	15
3.3.2 감성 레이블 수정	15
제4장 실험 결과 및 평가	17
4.1 실험 환경 구성	17
4.2 감성 대화 생성 모델 성능 평가	19
제5장 결론 및 제언	26
참고 문헌	29
ABSTRACT	31

그림 차례

그림 1	11
------------	----

표 차례

표 1.....	17
표 2.....	18
표 3.....	20
표 4.....	21
표 5.....	22
표 6.....	23
표 7.....	24
표 8.....	25

국문 초록

인공지능의 급속적인 발달로 인해 인간은 챗봇을 통해 생활에서의 편의를 해결하는 것 뿐만 아니라 감성적인 대화를 원하고 있다. 이에 따라 최근 챗봇은 인간에게 필요한 정보나 서비스를 제공해주는 기술을 넘어 인간의 감정을 유사하게 표현할 수 있는 수준으로 나아가고자 하는 활발한 움직임이 있다. 본 연구에서는 감정 정보를 활용하여 적절한 응답을 생성하는 챗봇을 위해 감성 대화 챗봇 시스템을 개발하기 위해 AI Hub와 온라인으로부터 대화 데이터를 수집하여 7가지 감성으로 분류해 재구축하였다. 또한 기존에 많이 사용하던 규칙 기반이나 다수의 모델을 혼합해 사용하는 방식과 달리, GPT의 Multi-Head 구조의 학습을 통해 감정 파악부터 문장 생성까지의 End-to-end 모델을 생성하고 이를 챗봇에 적용할 수 있음을 보였다. 이를 통해 감성 대화 챗봇 모델의 효율성을 개선하고 실제 사용자와의 대화를 진행하며 모델을 평가하고 실제 대화에 적용 가능한지 검증함으로써, 챗봇이 사용자의 발화에 대해 적절하면서도 일관적인 감정으로 구체적인 응답을 생성하는 것을 확인하였다.

제 1 장 서 론

1.1 연구 배경 및 목적

챗봇이란 문자 또는 음성으로 사람의 질문에 응답하며 대화할 수 있는 컴퓨터 시스템 또는 인공지능이다. 과거의 챗봇은 사전에 이미 구축된 데이터베이스에서 대화 패턴을 검색한 것을 기반으로 기계적인 응답을 생성하는 수준에 불과했다. 하지만 최근 자연어처리(Natural Language Processing) 기술과 데이터 분석 기술이 빠르게 발전하면서, 챗봇의 성능 또한 급속도로 좋아지고 있다. 이에 따라 실시간 민원 응대, 고객 맞춤형 서비스 제공, 상담 인력 절감 등 여러 가지 이점을 통해 공공기관과 민간부문(금융, 유통, 일반 고객 응대가 필요한 기업) 등 여러 산업 분야에서 활발하게 사용되고 있다. 국내 챗봇 서비스의 현 수준을 살펴보면, 사용자의 발화에 기반한 의도 분석 후, 서비스 또는 상담원 연결로 이어지고 있으며, 정보 검색 및 가벼운 업무 처리와 같은 단순 지능형 서비스가 주를 이루고 있다. 하지만 실제 사람과 챗봇의 대화를 살펴보면, 사용자는 서비스에 중점을 둔 정보 제공적인 대화보다는 감성적인 대화에 대해 큰 수요를 보이고 있음을 알 수 있다. SKT의 인공지능 'NUGU'의 경우, 1억건의 대화량 중에서

감성대화 비율이 45%에 달하였으며, KT의 인공지능 '기가지니'의 경우에는 전체 대화량 중 감성대화의 비율이 30%에 달한다.[1] 이는 사람들이 인공지능 챗봇과 감성적인 대화를 시도하고자 하며, 챗봇에게 사람의 감정을 이해하고 공감할 수 있는 수준의 대화를 기대하고 있음을 의미한다. 결국 이러한 흐름에 따라 인공지능 챗봇은 점진적인 발전을 통해 사람을 이해하고 공감하는 감성 기능을 가지게 될 것이다. 이를 위해서는 많은 양의 데이터 확보가 중요하며 감정을 제대로 파악해 감정이 반영된 적절한 응답을 하는 기술이 필요하다.

본 연구에서는 감정 정보를 활용하여 적절한 응답을 생성하는 챗봇을 위해 기존에 많이 사용하던 규칙 기반이나 다수의 모델을 혼합해 사용하는 방식과 달리, GPT의 Multi-Head 구조의 학습을 통해 감정 파악부터 문장 생성까지의 End-to-end 모델을 생성하고 이를 챗봇에 적용할 수 있음을 보인다. 이를 통해 언어 생성 모델과 감성 분류 모델을 따로 사용해 감성대화를 생성하는 비효율성을 해결하고 이러한 모델이 감성대화 생성 성능에 미치는 영향을 보고자 하였다. 특히, 실제 사용자와의 대화를 진행하며 모델을 평가하고 실제 대화에 적용 가능한지 검증함으로써, 챗봇이 사용자의 발화에 대해 적절하면서도 일관적인 감정으로 구체적인 응답을 생성하는 것을 확인하였다.

1.2 연구 방법 및 범위

본 연구에서는 감성 대화 챗봇 시스템을 개발하기 위해 AI Hub와 온라인으로부터 사용자와의 공감적 대화를 목적으로 한 대화 데이터를 수집하였다. 감성 대화 데이터는 사람의 발화, 감정 레이블, 챗봇의 응답으로 이루어져 있는데 이때 여러 개의 긍·부정적 감정으로 나누어진 감정 레이블을 통해 사람의 감정에 따라 변화하는 챗봇의 응답을 얻을 수 있다. 본 연구에서는 수집한 감성 대화 데이터셋을 감동, 기쁨, 안도, 슬픔, 분노, 불안, 중립의 7가지 감정으로 나눈 후, 신경망 모델인 BERT로 훈련시켜 감성 분류 모델을 구축하였다. 이를 통해 일반 대화 데이터를 대상으로 추가적 감성 레이블링 작업을 진행함으로써 대용량의 데이터셋을 구축하였으며 이를 챗봇의 훈련 데이터로 사용하였다.

본 연구에서는 감성 대화 챗봇을 개발할 때 감성 분류와 문장 생성 두가지 태스크가 따로 이루어지는 문제에 초점을 두고 연구하였다. 실제 챗봇 관련 연구들은 감성 분류 모델과 문장 생성 모델을 따로 구축한 후 이를 파이프라인으로 연결하여 감성 대화 챗봇을 개발하고 있다. 이러한 방식은 모델의 크기가 커지며 작업의 효율을 떨어뜨린다. 따라서 감성 분류부터 문장 생성을 End-to-end 방식으로 구현하여 챗봇의 구조를 단순화하고 감성 대화의 성능을 향상시키고자 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 감성대화 챗봇의 개념과 더불어 기존에 진행된 선행 연구에 대하여 살펴보고, 이론적 바탕이 되는 신경망 모델에 대해 알아본다. 이후 3장에서는 새로운 감성 대화 챗봇 모델 구조를 제안 및 성능 평가 방법을 소개하고, 모델 학습에 사용한 데이터와 데이터 증식방법을 설명한다. 그리고 4장에서는 제안 모델 실험 및 결과를 평가 후, 베이스라인 모델과 비교해 분석한다. 마지막 5장에서는 본 연구의 결론 및 한계점을 제시하며 향후 연구에서 개선할 수 있는 방향을 제언한다.

제 2 장 이론적 배경

2.1 감성 대화 챗봇

챗봇은 크게 두 가지로 구분이 된다. 먼저 Task-oriented Dialogue System은 문제 해결을 위해 설계된 시스템으로, 챗봇이 사용자가 원하는 것을 파악하고 빠르게 해결하는 것에 초점이 맞추어져 있다. 구글 어시스턴트나 시리, 기가지니, 알렉사 등이 이 종류에 속한다. 반대로 Open

domain Dialogue System은 특정 주제에 구애받지 않고 대화할 수 있도록 만들어진 시스템으로, 사용자가 꺼내는 다양한 주제에 대해 챗봇은 적절한 응답을 출력하며 대화를 이어 나간다. 따라서 실제 사람과의 대화처럼 친밀함을 유지하면서 사용자의 발화에 대해 높은 이해도를 보이고 응답을 생성하는 것이 주된 목적이다.

감성대화 챗봇은 자유 주제 대화 시스템에 가까우며, 사용자의 다양한 상황에 기반한 감정을 이해하고 적절한 응답을 내보낸다. 챗봇은 사용자의 감정을 파악하기 위해서 사용자의 발화를 분석하여 감정을 분류한다. 이 때 사용하는 감성 분석 방법으로는 어휘기반의 감성 사전을 구축하여 감성 분석을 하는 방법과, 머신러닝 기반의 감성분석 방법이 있다. 사전 기반의 감성 분석은 감성 레이블에 따라 어휘가 정확하게 분류되어 있기 때문에 해당 어휘가 들어간 문장에 대해서는 분류 정확도가 높다. 하지만 사전에 등록되지 않은 단어나 신조어, 반어법에는 취약하다는 단점이 있다. 따라서 최근에는 대용량의 문장으로 이루어진 데이터셋을 머신러닝 기반으로 훈련시킨 감성 분석 모델을 사용하여 이와 같은 예외 상황에서 보다 유연하게 대응할 수 있게 한다.

본 연구에서는 사용자의 다채로운 상황과 감정에 공감하기 위해 신경망 모델로 대용량 대화 문장 데이터의 감성을 분류함으로써 사람과 감성적 대화가 가능한 감성 대화 챗봇을 구축한다.

2.2 관련 연구

2.2.1 기존 연구

감성 대화가 가능한 챗봇에 대한 초기 연구들은 주로 영어 텍스트를 사용했으며, 감정 분류와 응답 생성에서 CNN, RNN(Recurrent Neural Network) 등 각기 다른 모델을 사용하였다. 감정을 적용한 챗봇을 위해 Hao Zhou et al. (2018)는 Bi-LSTM (Bidirectional Long Short-Term Memory)을 이용한 6가지 감정 분류기를 사용한다.[2] 감정 분류기를 통해 사용자 문장의 감정을 분류한 후, 인코더-디코더(Encoder-decoder) 구조의 문장 생성 모델을 활용해 응답 문장을 출력한다. 하지만 해당 연구에서는 응답문장으로 6개의 감정에 해당하는 문장을 모두 생성하고 있기 때문에, 가장 적합한 응답 문장 하나를 선정하지 못한다는 문제가 있다. 또한 시퀀스 투 시퀀스(Sequence-to-Sequence) 구조의 모델을 사용하여 장기 기억에 약하다는 점과 속도가 느리다는 점에서 한계가 있다. 한국어 감정 분류 및 대화 성능 향상에 초점을 맞춘 SKT-AI에서는 한국어 자연어 처리 모델인 KoBert와 KoGPT2를 개발하였다.[3] 해당 연구에서는 기존 챗봇의 사용자의 질문을 유형화하여 해당하는 답변을 선택해 출력하는 방식과 다르게, 사용자

발화의 맥락을 이해하고 자연스럽게 대화가 이루어지는 고도화된 챗봇 성능을 보인다. 하지만 감정 분류 후, 이를 직접 KoGPT2의 인풋으로 명시해 입력해야 하는 불편함이 있다. 이원민, 온병원 (2021)이 진행한 최근 연구에서는 BERT와 GPT의 파이프라인 방식을 사용해 대화 데이터의 감정 분류를 진행하고 이를 사용하여 감정이 반영된 응답 문장을 생성하도록 학습을 진행한다.[4] 이를 통해 챗봇이 실시간으로 사용자의 말에 대해 적절하고 감정이 일관되면서도 다양한 반응을 생성하도록 한다. 하지만 해당 연구에서 사용한 데이터셋은 긍부정 간 감성 레이블의 불균형으로 인해, 부정적 감정의 세분화에 초점이 맞춰져 있어, 긍정적 감정은 고려되지 못하고 있다. 또한 감성 분류 모델은 BERT를, 문장 생성 모델은 GPT를 사용하는 식으로 각기 다른 모델로 연구되어 작업의 효율성을 저하시킨다는 문제가 있다. 가장 최근에 연구된 Raman Goel. (2022)에서 또한 감성 분류와 문장 생성을 각기 다른 모델로 작업하였는데, 감성 분류기로는 LSTM을 사용하여 발화의 감성을 감지했으며, 이를 Transformer-XL의 encoder에 input으로 넣어 공감적인 응답을 생성한다.[5] 이는 앞선 연구들과 동일하게 두가지의 모델을 사용해서 감성 대화 챗봇을 개발하고자 하였으며, 성능 평가 지표로는 BLEU 한가지만을 선정해 실제 감성 대화가 인간의 입장에서 잘 생성되었는지에 대한 판단이 어려운 문제가 있었다.

선행 연구 고찰을 바탕으로, 본 논문에서는 '감동, 기쁨, 안도, 슬픔, 불안,

분노, 중립'의 7가지 감성을 새롭게 레이블링하여 양질의 학습데이터를 생성함으로써 이를 학습에 사용하고, 감성 분류와 문장 생성을 하나의 모델로 작업하여 기존 연구의 문제점을 개선하고 감성 대화 챗봇에 적용될 수 있도록 GPT Multi-Head 방안을 제안한다.

2.3 신경망 모델

2.3.1 BERT

BERT (Bidirectional Encoder Representations from Transformers, Devlin et al., 2019)는 구글 리서치 팀에 의해 공개된 자연어처리 사전 훈련 모델이다.[6] BERT는 Transformer 모델의 encoder stack과 embeddings layer로 구성되어 있다. BERT 모델에 입력 시퀀스를 입력하면 입력 시퀀스 길이만큼의 벡터를 출력한다. 입력 시퀀스는 텍스트를 토큰화한 토큰 시퀀스에 [CLS], [SEP] 토큰을 추가하여 구성한다. [CLS] 토큰은 입력 시퀀스의 시작을 나타내고 [SEP] 토큰은 두 개의 토큰 시퀀스를 구분하기 위해 사용된다. 출력 벡터는 학습 과정에서 다음 문장인지 맞추거나 마스킹된 토큰을 예측하는데 사용되고 fine-tuning 과정에서는 다운스트림 작업에 맞게 사용된다. Self-Attention을 사용하며 자기 회귀 특성이 없어진 BERT는 다음

단어를 예측하는 성능은 비교적 낮지만, 양방향 학습을 통해 문장을 읽기 때문에 문장의 맥락 정보를 충분히 고려하여 학습할 수 있다. 따라서 BERT는 감정 분류와 같이 문장의 의미를 파악하는 데 적합하다.

2.3.2 GPT-2

GPT-2(Radford et al., 2019)는 Open AI에서 Web text의 40GB의 코퍼스와 크롤링한 데이터를 합쳐서 훈련시킨 언어 모델이다.[7] BERT(Devlin et al., 2019) 모델과는 다르게 Encoder Stack은 사용하지 않고 Transformer의 Decoder Stack을 통해 Masked Self-Attention을 사용하여 이전의 출력값이 다음 입력값으로 변하는 모델로, 다음 출력값을 예측하는 능력이 뛰어나 문장 생성을 하는데 주로 사용된다. GPT-2 모델은 인풋 문장에 대해 Byte Pair Encoding(BPE)을 진행한다. 이를 통해 문자를 토큰으로 나누고 해당 토큰들을 임베딩 행렬로 변환한 후 Decoder Stack에 넣는다. 각 입력 벡터는 Decoder Stack 안에 있는 각각의 Decoder cell에서 Self-attention을 거치며 문맥을 이해하는 학습을 하고, 임베딩 벡터를 통해 구한 확률 값이 가장 높은 문장을 내보내는 방식으로 진행된다. 하지만 GPT-2 모델은 한국어가 아닌 영어 베이스로 개발된 모델이기 때문에, 한국어 기반 챗봇 시스템을 개발하는 본 연구에서는 AWS와 SKT가 협력개발한 모델인 KoGPT-2를 사용한다.

해당 KoGPT-2 모델은 117M, 12 layers로 이루어져 있으며, 20GB의 문장으로 학습되었다.

제 3 장 연구 방법

3.1 제안 모델

본 연구에서는 사용자의 감정 정보를 활용해 감성 대화가 가능한 챗봇을 End-to-end 방식으로 설계하기 위해, GPT-2 모델의 Multi-Head 생성을 통해서 언어 모델링과 감성 레이블 예측 목표를 결합한 Multi-task loss를 사용한다. GPT-2 모델은 한국어 기반 KoGPT로 pre-trained된 모델로, 다음 단어 생성 Task를 수행한다. 그리고 본 기반 모델을 변형하여 문장의 감정 정보까지 예측할 수 있도록 하기 위해 Sentiment Classification Head Layer를 추가로 생성하였다. 제안 모델은 아래의 그림과 같다.

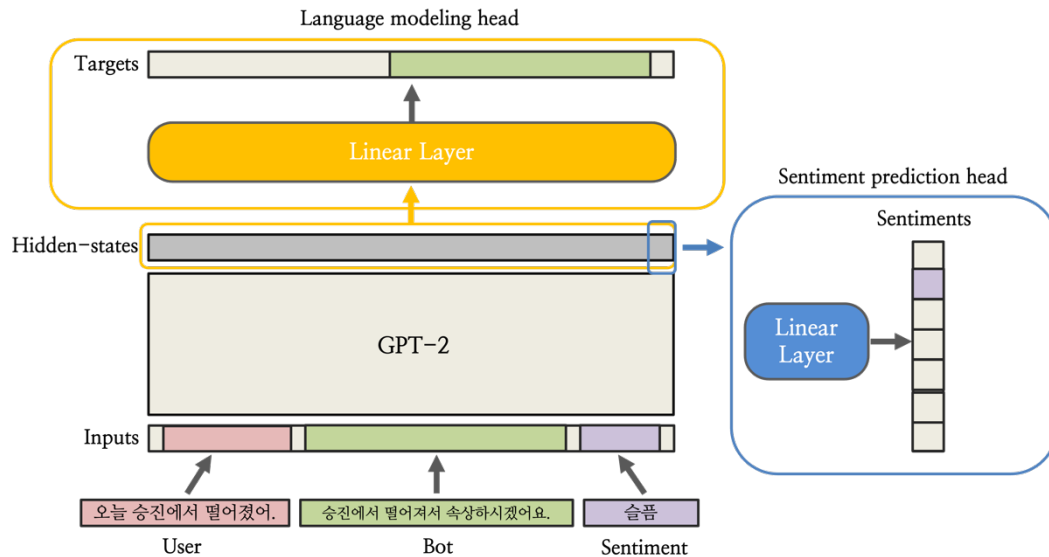


그림 1 제안 모델 아키텍처

GPT-2를 기반으로 한 제안 모델에서는 언어 모델링 예측(노란색) 및 감정 레이블 예측(파란색)을 위한 두 개의 헤드가 존재한다. 언어 모델링 예측 헤드는 logit을 얻기 위해 word embedding matrix에 hidden states를 투영하고 챗봇 응답에 해당하는 부분에 cross-entropy loss를 적용한다(위 그림의 녹색 레이블). 감정 레이블 예측 헤드는 linear layer를 통해 마지막 토큰에 hidden states를 전달 하여 점수를 얻고 cross-entropy loss를 적용하여 감정을 올바르게 분류한다. 결과적으로 total loss는 다음과 같이 계산되는 언어 모델링 loss와 감정 분류 레이블 예측 loss의 가중치 합이 되며, 이 값이 작아지도록 모델은 학습하게 된다.

본 연구에서 제안하는 모델은 구글 colab의 GPU 환경에서 실험하였으며, PyTorch Lightning을 통해 구현되었다. 최종적으로 언어 모델링 loss가 1.47, 감정 레이블 예측 loss는 14.3으로 나타났으며, 이를 기반으로 실험과 평가를 진행하였다.

3.2 모델 성능 평가

3.2.1 Perplexity

Perplexity는 언어모델의 Generation 성능을 판단할 수 있는 Auto Evaluation 지표로, Generation task의 성능 비교를 위해 주로 사용된다.[8] 단어로 구성된 문장의 Perplexity는 다음과 같은 수식으로 나타낼 수 있다.

$$PPL(W) = P(w_1, w_2, w_3, \dots, w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1, w_2, w_3, \dots, w_N)}}$$

분모는 문장에 대한 Generation probability를 의미하는데, 이를 조건부 확률의 Chain rule을 활용하여 다음과 같이 나타낼 수 있다.

$$PPL(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, w_3, \dots, w_N)}} = \sqrt[N]{\frac{1}{\prod_{i=1}^N P(w_i | w_1, w_2, \dots, w_{i-1})}}$$

전체적인 식을 보면, Perplexity는 문장에 대해 Generation probability의 역수의 기하평균을 구하고 있다. 이는 Perplexity가 문장이 생성될 확률의 역수를 단어의 수로 정규화 하고 있음을 의미한다. 따라서 문장의 발생 확률이 높을수록 해당 문장에 대한 Perplexity 값은 낮아진다. 언어모델의 평가 관점에서 생각해 본다면 테스트 데이터에 존재하는 문장에 대한 모델의 Perplexity가 낮다는 것은, 해당 정답 샘플에 대한 발생 확률이 높다는 것과 동일하며, 해당 언어 모델의 학습이 잘 이루어졌음을 의미한다. 반대로 테스트 데이터에 존재하는 문장에 대한 모델의 Perplexity가 높다는 것은, 해당 정답 샘플에 대한 발생 확률이 낮아 언어 모델이 잘 학습되지 못하였음을 의미한다.

단, PPL 값의 높고 낮음은 테스트 데이터 상에서 수치상의 정확도를 의미하는 것이지, 인간이 주관적으로 느끼기에도 잘 학습된 언어 모델이라는 것을 반드시 뜻하지는 않는다. 따라서 본 연구의 목적을 고려한다면 챗봇 응답에 대한 사람의 주관적 평가가 함께 이루어져야 평가 신뢰도를 높일 수 있을 것이다.

3.2.2 SSA

SSA(Sensibleness and Specificity Average)는 구글에서 오픈 도메인 챗봇의 성능을 측정하기 위해 제시한 Human Evaluation 지표로, 'Meena'와 '이루다' 등 다수의 Conversational AI 연구에서 해당 지표를 사용하여 모델 평가를 진행하였다.[9] 본 연구에서는 자연어 분야의 최신 연구 동향에 따라, 최신 인공지능 챗봇 모델의 평가 경향을 반영하여 SSA 지표를 제안 모델 성능 실험의 주 평가 지표로 사용하도록 한다.

SSA 지표의 평가 목적은 이는 자연스러운 대화가 가지는 중요한 속성들을 파악하는 것으로 예를 들어, 챗봇이 표현한 것이 상황에 부합하는 것인지(Sensibleness), 그리고 챗봇이 말한 것이 방금 유저가 말한 내용에 따른 것으로 구체적인 것인지(Specificity) 등을 파악하는 것이다.

평가자는 상식적으로 "말이 되는가?"와 "이전 발화에 대해 구체적인가?"라는 두 가지 기준에 답하게 된다. 따라서 비논리적이거나, 맥락에 맞지 않거나, 사실적으로 잘못된 것처럼 보이는 것이 있으면 이진수인 1과 0 중에 0으로 평가해야 한다. 응답이 말이 되는 경우에는 다음 단계로 맥락을 보고 해당 발화가 주어진 컨텍스트에 특정한 응답인지 여부를 결정해야 한다. 최종적으로 1로 평가된 수의 합을 전체 평가 발화 수로

나누어 확률을 구하면 그 값이 SSA 값이 된다.

3.3 데이터 셋

3.2.1 데이터 수집

모델 학습을 위한 데이터셋은 ‘웰니스 대화 스크립트 데이터셋(AI Hub, 2019)’[10], ‘감성대화 말뭉치(AI Hub, 2019)’[11], ‘한국어 감정 정보가 포함된 단발성 대화 데이터셋(AI Hub, 2020)’[12]과 songys이 제공하는 대화 데이터셋(Github, 2020)[13]을 사용한다. ‘웰니스 심리 상담 데이터셋’은 신촌 세브란스 병원에서 제공하는 정신건강 상담 기록을 재가공하여 만들었으며, ‘감성 대화 말뭉치’는 1,500명의 일반인과 진행한 인터뷰 내용과, 우울증 환자를 대상으로 진행한 WOZ 대화를 수집하여 가공하여 구축된 데이터셋이다. ‘한국어 감정 정보가 포함된 단발성 대화 데이터셋’은 웹 크롤링을 통해 SNS와 온라인 상의 글과 댓글을 수집한 후, 적절한 문장을 선정하여 제작되었다. 또한, songys이 공개한 대화 데이터셋은 ‘사랑보다 아름다운 실연’이라는 이름의 다음 카페에서 업로드 되는 이야기를 참고하여 제작되었다.

3.2.2 감성 레이블 수정

감성 레이블링을 수행하기 위해서 수집한 데이터셋 중, '웰니스 대화 스크립트 데이터셋'을 기준으로 선정하고, 해당 데이터셋 내에 존재하는 세부 감성 레이블을 이용해서 7개의 감성(감동, 기쁨, 안도, 슬픔, 분노, 불안, 중립)으로 레이블링을 수행했다. 이는 기존의 부정 감성 레이블에 치우친 데이터셋을 개선하기 위해 진행한 작업으로, 긍정 감성 3가지, 부정 감성 3가지, 중립 1로 재분류함으로써 감성 레이블의 불균형을 해소하여 각 감성에 대해 적절하며 구체적인 응답 생성이 가능하도록 하였다. 수정된 데이터셋은 BERT를 이용해 Sentiment Classification Task에 맞추어 훈련시킨 후, 타 데이터셋의 대화문장들을 넣어 같은 동일한 레이블 수로 레이블링을 진행하였다.

그 결과, 수집한 모든 한국어 대화 문장 데이터에 대해 감동 9622, 기쁨 19,750, 안도 28,624, 슬픔 73,826, 분노 50,026, 불안 39,529, 중립 75,327개씩 분류하였다. 총 296,704개의 데이터셋에서 랜덤하게 80%, 10%, 10%의 비율로 분류하여 Train, Valid, Test 데이터셋으로 나누었고, 최종적으로는 데이터셋은 User, Bot, Emotion 순으로 구분자를 기준으로 나열하였고, Single-Turn의 형식으로 다음과 같이 구성하였다.

1	User	Bot	Emotion
2	아내가 드디어 출산하게 되어서 정말 신이 나.	아내분이 출산을 하시는군요. 정말 축하드려요.	기쁨
3	당뇨랑 합병증 때문에 먹어야 할 약이 열 가지가 넘어나니까 스트레스야.	약 종류가 많아 번거로우시겠어요.	불안
4	고등학교에 올라오니 중학교 때보다 수업이 갑자기 어려워져서 당황스러워.	고등학교 수업이 중학교 때와는 다르게 갑자기 어려워서 당황스러우시군요.	불안
5	재취업이 돼서 받게 된 첫 월급으로 온 가족이 외식을 할 예정이야. 너무 행복해.	재취업 후 첫 월급이라 정말 기쁘시겠어요.	기쁨
6	이제 곧 은퇴할 시기가 되었어. 내가 먼저 은퇴를 하고 육 개월 후에 남편도 은퇴를 하고	그러시군요. 은퇴 후 남편분과 함께 집에 있을 생각을 하니 막막한 기분이 드시는군요.	안도
7	사십 대에 접어들면서 머리카락이 많이 빠져 고민이야.	머리카락이 빠져서 걱정이 많으시겠어요.	슬픔
8	이제 돈이라면 지긋지긋해.	무슨 일 있으셨나요?	슬픔
9	친구들이 나를 괴롭혀. 부모님과 선생님께 얘기했는데도 믿어주지 않아.	괴롭힘을 당해 힘든데도 부모님과 선생님께서 믿어주지 않아서 정말 마음 아프겠네요.	슬픔
10	친구 때문에 눈물 나.	친구 때문에 눈물이 나시는군요. 무슨 일이세요?	슬픔
11	새로 간 치과에서 조금 씩기만 한 이를 임플란트를 하라고 했어. 과잉진료는 안 한다더니	이가 조금 씩기만 했는데도 임플란트를 해야 한다고 했다는 말씀이시군요.	분노

표 1 실험 데이터셋

제 4 장 실험 및 평가

4.1 실험 환경 구성

본 연구를 통해 개발한 감성대화 챗봇 시스템을 실험한 후, Human Evaluation 방식으로 평가를 진행하였다. 이를 위해, 평가 주제는 감성 대화 생성으로 정하며, 입력은 평가자가 입력한 문장으로 정하고, 출력은 평가자가 입력한 발화에 대응하여 챗봇 시스템이 생성한 응답 문장으로 한다. 평가 방법으로는 평가자와 챗봇이 컴퓨터를 통해 일대일로 대화를 주고받는 상황을 가정하여 진행한다. 평가자는 20대 일반인 및 대학원생으로 구성된 15명의

인원을 대상으로 하며, 일상에서 나눌 수 있는 이야기 중에서 특정한 주제를 가지고 대화를 시작한다. 컴퓨터를 사용해 대화를 진행해 나가며 입력 대화에 대한 챗봇의 출력 대화를 사용자에게 보인다. 평가자에게 챗봇의 응답 문장을 출력하는 방법으로는 모니터를 통해 문장을 출력하는 것으로 진행된다. 이에 대해 평가자가 챗봇이 생성한 문장에 대해, 적절한 정도(Sensibleness)와 자연스러운 정도(Specificity)을 정량적으로 평가한다.

평가 환경으로는 구글 Colab을 사용하며, 챗봇과 대화 후 SSA에 근거한 평가 기준에 따라 엑셀에 각 챗봇의 문장에 대한 점수를 표기하도록 한다. 평가자는 챗봇과 7-Turn(평가자 발화 7개, 챗봇 발화 7개) 대화를 주고 받으며, 마지막 대화가 끝난 후 실험을 종료한다. 위 평가 절차는 아래의 표와 같이 다시 정리한다.

Step	Procedure
1	안내자는 평가자에게 실험 절차 및 SSA 평가 기준과 점수 산정 방법 안내
2	평가자에게 Google 실험 계정 부여 및 Colab 실험 환경 접속
3	평가자는 챗봇과 특정 주제로 대화 진행
4	평가자는 챗봇의 각 발화에 대해 SSA 기준에 근거한 평가 점수를 엑셀 시트에 기입
5	평가자는 3-4번을 반복하며 총 7-Turn(실험자 7회, 챗봇 7회) 대화 진행
6	대화 종료 및 안내자에게 평가 점수 제출

표 2 실험 및 평가 절차

SSA 지표의 평가기준에 따르면, 문장마다 적합성(Sensibleness)과 구체성(Specificity)은 이진수로 표현하고 있으며, 해당 속성에 만족한다면 1로, 불만족한다면 0으로 채점한다. 따라서 평가자마다 적합성과 구체성의 합을 통해 평균 점수를 구할 수 있고, 최종적으로 모든 사용자 15명에 대해서 평균 점수를 구할 수 있다. 최종적으로는 챗봇의 감성 대화 생성 성능을 평가하기 위해서 본 SSA 점수와 Auto-evaluation 으로 진행된 Perplexity 점수를 이용한다.

4.2 감성 대화 생성 모델 성능 평가

본 연구의 감성 대화 생성에 대한 실험 결과는 Auto Evaluation과 Human Evaluation 두 가지 측면에서 모두 평가했다. 제안 모델의 성능을 비교하기 위한 베이스모델로는 동일 대화 데이터셋에서 감정정보를 제외한 후, Ko-GPT2가 기존에 가지고 있는 Language Modeling Head만을 이용하여 훈련시킨 모델을 선정하였다.

먼저 자동화로 성능을 검증하기 위해서 PPL(Perplexity)을 이용했다. 그림 2는 Baseline Model과 Proposed Model 각각에 대한 PPL을 계산해 비교한 결과이다. Proposed Model은 PPL 기준으로 감정정보를 이용하지 않은

Baseline Model에 비해서 높은 점수를 볼 수 있는데, 이에 대해서는 감성정보를 주었을 때 다음 문장 생성에 대해 혼란을 줄이고 생성할 문장의 후보를 선택하는 데 도움을 준 것으로 해석된다.

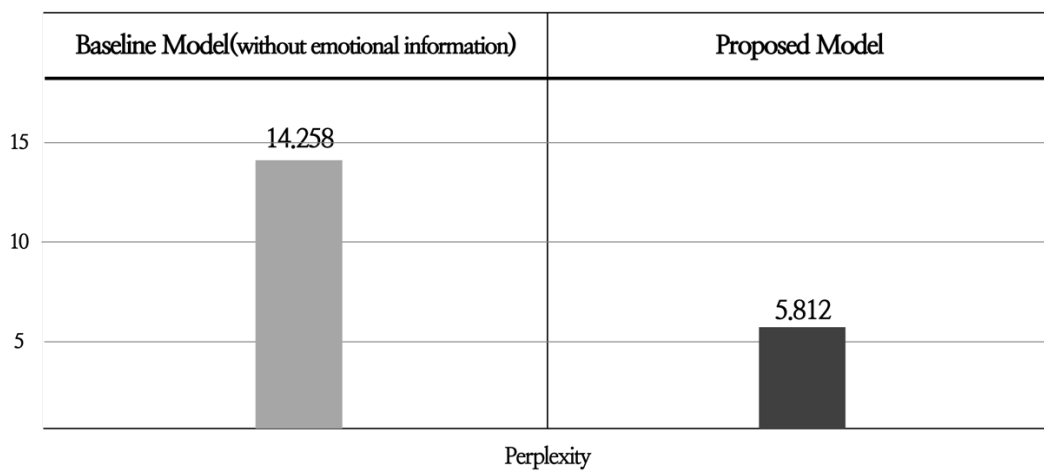


표 3 PPL 비교

다음으로 본 실험의 목적인 '감성 대화 생성'에 대한 성능을 적절하게 평가하기 위해, 인간이 직접 챗봇과 대화한 후 점수를 부여하는 Human Evaluation인 SSA 지표를 주 지표로 선정해 평가를 진행하였다. 각 실험자가 매긴 적합성과 구체성에 대한 평균 기준으로 평균 0.69의 score를 얻었다. 또한 SSA 지표 상에서 적합성과 구체성 중, 제안 모델에서 구체성이 비교적 높은 것으로 나타나는데, 이는 제안 모델이 실험자의 입력 문장에 대한

감성을 파악하기 때문에 감정정보가 없는 베이스라인보다 구체적이고 입력에 적절한 응답을 생성하고 있음을 보인다.

높은 SSA score를 얻은 실험을 대화 주제에 따라서, 그리고 Baseline Chatbot과 비교해 보면, 아래 대화문과 같이 볼 수 있다.

	Sensibleness	Specificity	SSA
Baseline Model (without emotional information)	0.562	0.381	0.472
Proposed Model	0.781	0.59	0.69

표 4 SSA 점수표

# of Turn	User	Proposed Chatbot
1	오늘 승진에서 떨어졌어.	승진에서 떨어져서 속상하시겠어요.
2	응. 다음 기회는 없을지도 몰라.	다음 기회는 없을 거라고 생각하시는군요.
3	올해 마지막 승진 기회였거든.	승진을 위해 노력하셨는데 승진이 안돼서 속상하시겠어요.
4	면담 시간을 가졌는데 대리님이랑도 말이 안 통해.	화가 많이 나셨겠어요.
5	되는 일이 없어. 정말 짜증나.	어떻게 하면 지금 기분에서 벗어날 수 있을까요?
6	글쎄, 떡볶이나 시켜먹을까봐.	맛있는 떡볶이 드시고 기분전환이 되길 바라요.
7	응. 빨리 내일이 됐음 좋겠다.	내일 꼭 좋은 소식이 있길 바랄게요.

표 5 SSA 고득점 대화 내용 - 제안 모델(부정적 감정)

# of Turn	User	Baseline Chatbot
1	오늘 승진에서 떨어졌어.	승진에서 떨어져서 속상하시겠어요.
2	응. 다음 기회는 없을지도 몰라.	다음 기회는 없을 거예요.
3	올해 마지막 승진 기회였거든.	올해 마지막 승진을 앞두고 기쁘시군요.
4	면담 시간을 가졌는데 대리님이랑도 말이 안 통해.	대리님이랑 말이 안통하시는군요.
5	되는 일이 없어. 정말 짜증나.	짜증나는 일이 있으셨군요.
6	글쎄, 떡볶이나 시켜먹을까봐.	떡볶이를 시켜먹을까 고민하시는군요.
7	응. 빨리 내일이 됐음 좋겠다.	내일 꼭 좋은 소식이 있길 바라요.

표 6 SSA 고득점 대화 내용 - 베이스라인 모델(부정적 감정)

# of Turn	User	Proposed Chatbot
1	오늘 생일파티 했어!	생일파티라니 정말 기쁘시겠어요.
2	생각보다 큰 선물을 받아서 너무 행복했어. 최고의 생일이야.	생일 축하드려요!
3	고마워. 이걸 어떻게 보답하지?	친구들에게 고마운 마음을 전할 수 있었으면 좋겠어요.
4	응. 다음주에 효민이랑 놀러가기로 했어!	다음주에 효민이와 즐거운 시간 되시길 바라요.
5	부산 맛집을 찾아봐야겠어.	좋은 생각이네요. 맛집 투어를 하면 어떤 점이 좋을까요?
6	맛있는 걸 먹으면 여행이 더 기억에 남더라고.	맛있는 걸 먹으면 여행이 더 특별해지는군요.
7	맞아. 벌써 설레어.	내일도 좋은 일만 있으시길 바랄게요.

표 7 SSA 고득점 대화 내용 - 제안 모델(긍정적 감정)

# of Turn	User	Baseline Chatbot
1	오늘 생일파티 했어!	생일파티를 하셨군요.
2	생각보다 큰 선물을 받아서 너무 행복했어. 최고의 생일이야.	생일 축하드려요!
3	고마워. 이걸 어떻게 보답하지?	이 부분에 대해 도움을 요청할 사람이 있을까요?
4	응. 다음주에 효민이랑 놀러가기로 했어!	효민이랑 놀러가기로 하셨군요.
5	부산 맛집을 찾아봐야겠어.	맛있는 음식을 찾아보려고 하시는군요.
6	맛있는 걸 먹으면 여행이 더 기억에 남더라고.	맛있는 걸 먹으며 기분이 나아지셨으면 좋겠어요.
7	맞아. 벌써 설레어.	설레고 기분이 좋으시군요.

표 8 SSA 고득점 대화 내용 - 베이스라인 모델(긍정적 감정)

높은 SSA score를 얻은 실험은 실험자의 현재 문장에 대해서 적절한 반응을 보이며, 올해 마지막 승진 기회였다는 실험자의 입력 문장에 대해 "마지막 승진을 위해 노력했는데 속상하겠다"는 비교적 구체적이고 새로운 응답을 출력한다. 또한 실험자의 입력 문장에 대한 감성을 파악함으로써 같은 입력문장에도 베이스라인과 비교해 챗봇이 보다 적절한 발화를 생성한 것을 볼 수 있다. 이는 동시에 이전 문맥을 고려해서 연관된 문장을 챗봇이 출력하는 것과 같은 multi-turn의 대화 진행처럼 보이는 효과가 있었다.

반대로 Baseline Chatbot의 대화문을 보면 실험자의 입력 대화를 그대로 다시 출력하며 무조건적인 공감의 응답을 주로 생성하는 것을 볼 수 있다. 또한 실험자의 입력 문장이 은혜를 어떻게 보답할 지 묻는 긍정적 감성을 가진 말에, "도움을 요청할 사람이 있느냐"는 부정적 뉘앙스의 출력 문장을 생성한 것처럼, 실험자의 감정을 이해하지 못함으로써 맥락 이해력이 떨어지는 모습을 보인다.

제 5 장 결론 및 제언

인공지능의 급속적인 발달로 인해 인간은 챗봇을 통해 생활에서의 편의를 해결하는 것 뿐만 아니라 감성적인 대화를 원하고 있다. 이에 따라 최근 챗봇은 인간에게 필요한 정보나 서비스를 제공해주는 기술을 넘어 인간의 감정을 유사하게 표현할 수 있는 수준으로 나아가고자 하는 활발한 움직임이 있다. 특히 텍스트에서 감성을 분류하거나 다음 단어를 예측해 문장을 생성하는 기술은 글로벌 기업의 주도 하에 점차 고도화되고 있다. 하지만 두 가지 태스크를 동시에 수행하는 단일화된 모델이 없어 이를 개발하는 작업은

향후 감성 대화를 위한 챗봇을 발전시키는 데 필요한 일이다. 감성 대화 챗봇에 대한 연구는 감정 분류와 문장 생성, 이 두 가지 태스크를 서로 다른 모델을 사용해 파이프라인 형식으로 이어 연구가 이루어지고 있어, 비효율을 개선하기 위한 End-to-end 방식이 적용될 필요가 있었다. 또한 대부분 영어 텍스트를 기반으로 연구가 진행되고 있어, 한글 텍스트를 훈련시켜 성능을 평가하고자 하였다.

본 연구에서는 기존 데이터셋의 감정 레이블의 불균형 문제를 해결하고, 사용자의 감정 정보를 반영하여 적절한 응답을 생성하는 감성 대화 챗봇을 개발하는 것을 목적으로 하였다. 이를 위해 BERT를 이용해 긍부정의 감정을 균형있게 레이블링하여 한국어 감성 대화 데이터셋을 새롭게 구축하였으며, GPT2의 변형을 통한 Multi-Head 생성을 통해 End-to-end 방식으로 감정 예측 및 다음 단어 예측을 동시에 진행하는 두 가지 작업을 진행하였다. 본 연구는 추후 감성 대화 연구에 사용될 수 있는 감성 대화 데이터셋을 구축했으며, GPT2 Multi-Head 모델을 통한 End-to-end 방식의 감성 대화 챗봇을 구축한 첫 논문이라는 점에서 연구의 의의가 있다. 또한 실험을 통해 제안 모델이 감성 대화에 긍정적인 영향을 보인다는 것을 보였다.

다만, 본 연구에서 사용한 GPT2 모델의 경우 차세대 언어 모델인 GPT3 등에서 계속해서 문장 생성 성능의 고도화가 이루어지고 있기 때문에 차세대 언어 모델에 본 방법론을 적용하기 위한 추가적인 연구가 필요할 것으로

보인다. 또한 본 연구에 사용한 약 29만개의 대화 데이터에 대해 크롤링 등으로 데이터 양을 추가 확보해 모델 훈련을 진행한다면 감성 대화 성능이 보다 향상될 것으로 보인다.

제 6 장 참고 문헌

- [1] 서주원, 이창환. "일상 대화를 주제로 하는 감성 대화가 가능한 인공지능 기반 챗봇" 한국컴퓨터종합학술대회 논문집 (2020)
- [2] Hao Zhou, et al. "Emotional Chatting Machine: Emotional Conversation. Generation with Internal and External Memory" AAAI Conference on. Artificial Intelligence, Chicago, Illinois, USA, pp. 730-738 (2018)
- [3] SKTBrain. "Korean GPT pre-trained cased(KoGPT)"
<https://github.com/SKTBrain/KoGPT> (2019)
- [4] 이원민, 온병원. "감성 및 감정 단어 마스킹 기반 BERT와 GPT 파이프라인 방식을 통한 감정 문장 생성" Journal of KIIT. Vol. 19, No. 9, pp. 29-40 (2021)
- [5] Raman Goel, Seba Susan, et al. "Emotion-Aware Transformer Encoder. for Empathetic Dialogue Generation" ACIIW: 9th International. (2021)
- [6] Devlin, et al. "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. Vol. 1, pp. 4171-4186.

- (2019)
- [7] Alec Radford, Jeffrey Wu, et al. "Language Models are Unsupervised. Multitask Learners" OpenAI Blog, Vol. 1, No, 8, pp. 9 (2019)
- [8] Chang, J, Gerrish, S, et al. "Reading tea leaves: How humans interpret topic models" In Advances in neural information processing systems, pp. 288-296. (2009).
- [9] Adiwardana, et al. "Towards a Human-like Open-Domain Chatbot" arXiv:2001.09977 (2020)
- [10] AI Hub 웰니스 대화 스크립트 데이터셋 [웹사이트]. (2021.09.10). URL: <https://aihub.or.kr/opendata/keti-data/recognition-laguage/KETI-02-006>
- [11] AI Hub 감성대화 말뭉치 [웹사이트]. (2021.09.10). URL: <https://aihub.or.kr/aidata/7978>
- [12] AI Hub 한국어 감정 정보가 포함된 단발성 대화 데이터셋 [웹사이트]. (2021.09.10). URL: <https://aihub.or.kr/opendata/keti-data/recognition-laguage/KETI-02-009>
- [13] Github : songys. "ChatbotData" (2021.09.10). Retrieved from https://github.com/songys/Chatbot_data

ABSTRACT

Development of Emotional Chatbot System through Multi-Head Generation of GPT

Lee, Hyo Min

Dept. of Digital Analytics

The Graduate School

Yonsei University

With the development of artificial intelligence, humans want emotional dialogue as well as solving convenience in life through chatbots. As a result, recently, chatbots have been actively moving beyond technology that provides human-needed information and services to a level that can express human emotions in a similar way. In this study, conversation data were collected from AIHub and online to develop an emotional conversation chatbot system for chatbots that generate appropriate responses using emotional information and reconstructed into seven sensibilities. In addition,

unlike previously commonly used rule-based or mixed-use methods, the end-to-end model from emotion understanding to sentence generation is generated through GPT's Multi-Head structure learning, which can be applied to chatbots. By improving the efficiency of the emotional conversation chatbot model, evaluating the model, and verifying whether it is applicable to the actual conversation, it was confirmed that the chatbot generates a specific response with appropriate but consistent emotions.