사용자 감정 인식과 공감적 대화 생성: ChatGPT와 소형 언어 모델 비교

허승훈[1,↑,*,†], 이정민[1,↑,**], 조민수[1,***,†], 권오욱[1], 황금하[1]
[1]한국전자통신연구원 언어지능연구실
[↑]공동 1저자

heosh81@snu.ac.kr, faraway@etri.re.kr, minsoocho@hyundai.com, ohwoog@etri.re.kr, hgh@etri.re.kr

Empathetic Dialogue Generation based on User Emotion Recognition: A Comparison between ChatGPT and SLM

Seunghun Heo[1,↑], Jeongmin Lee[1,↑], Minsoo Cho[1], Oh-Woog Kwon[1], Jinxia Huang[1]
[1]Language Intelligence Research Section, ETRI
[↑]These authors equally contributed to this work.

요 약

본 연구는 대형 언어 모델 (LLM) 시대에 공감적 대화 생성을 위한 감정 인식의 필요성을 확인하고 소형 언어 모델 (SLM)을 통한 미세 조정 학습이 고비용 LLM, 특히 ChatGPT의 대안이 될 수 있는지를 탐구한다. 이를 위해 KoBERT 미세 조정 모델과 ChatGPT를 사용하여 사용자 감정을 인식하고, Polyglot-Ko 미세 조정 모델 및 ChatGPT를 활용하여 공감적 응답을 생성하는 비교 실험을 진행하였다. 실험 결과, KoBERT 기반의 감정 분류기는 ChatGPT의 zero-shot 접근 방식보다 뛰어난 성능을 보였으며, 정확한 감정 분류가 공감적 대화의 질을 개선하는 데 기여함을 확인하였다. 이는 공감적 대화 생성을 위해 감정 인식이 여전히 필요하며, SLM의 미세 조정이 고비용 LLM의 실용적 대체 수단이 될 수 있음을 시사한다.

1. 서론

GPT와 같은 사전 학습 언어 모델이 발전함에 따라, 이를 활용한 대화 시스템은 높은 성능을 보여 주고 있다. 사용자와 대화 시스템 간의 상호 작용의 품질을 개선하기 위하여, 대화 시스템은 사용자의 감정을 이해하고 공감적인 응답을 생성할 필요가 있다.[1]

본 연구는 공감적 대화 생성에서 사용자 감정 레이블을 포함하는 것이 유용한지 확인하고, 비교적 작은 언어 모델로도 공감적인 응답을 생성할 수 있는지 탐구하고자 한다. 이를 위해, KoBERT[1] 기반 감정 분류기와 ChatGPT[2] 및 Polyglot-Ko[2] 기반 생성 모델을 활용하여, 사용자의 감정 정보를 고려한 시스템 응답을 생성하는 실험을 진행하였다. 감정 인식 및 응답 생성을 위한 여러 언어 모델의 성능을 비교함으로써 효율성과 성능 면에서 우수한 공감적 응답 생성 모델을 제안하고자 한다.

2. 관련 연구

공감적인 대화 시스템[3]을 구현함에 있어서 사용자의 감정을 인식하여 그것을 프롬프트에 포함하는 방법이 고안되었으며[4], 이때 별도의 감정 분류기를 결합할 수 있다.[5] 공감적인 응답은 궁극적으로 사용자에게 감정적인 지원 (emotional support)을 제공할 수 있으며[6], 이러한 대화 시스템은 고객 응대 등의 목적으로 활용될 수 있다.[7]

3. 실험 및 평가

실험은 크게 사용자 감정 분류와 공감적 응답 생성의 두 가지로 구성된다. LLM으로는 OpenAI API에서 제공하는 gpt-4-0125-preview를 활용하였다. SLM은 두 가지가 사용되었는데, 사용자 감정 분류에서는 SKTBrain의 KoBERT를, 공감적 응답

* 현 소속: 서울대학교 언어학과 석사과정
** 현 소속: 과학기술연합대학원대학교 인공지능학과, 한국전자통신연구원 과학치안공공ICT연구센터
*** 현 소속: 현대자동차
† 본 연구는 ETRI 근무 기간 중 완료했음.
[1] https://github.com/SKTBrain/KoBERT
[2] https://openai.com/blog/chatgpt

생성에서는 EleutherAI의 Polyglot-Ko-5.8B [3] 를 기반 모델로 하여 미세 조정 (fine-tuning)하였다.

3.1. 데이터 구성

실험에는 ETRI에서 구축한 일반 감정 대화 (이하 '감정 대화') 및 질의응답 포함 감정 대화 (이하 'QA 대화') 데이터 셋이 사용되었다. [4] 이 데이터는 사용자와 시스템의 멀티 턴 대화로 이루어져 있으며, 시스템이 사용자의 감정과 상황에 깊이 공감하여 사용자와의 교감을 끌어내도록 구축되었다.

각 데이터는 <표 1>과 같이 구성되었으며, 대화 단위의 데이터를 8:1:1의 비율로 각각 분할하여 실험 및 평가를 위한 train, dev, test set을 구성하였다.

<표 1> 데이터 구성

	대화 개수	발화 개수	평균 발화 개수
감정 대화	5,000	77,784	15.56
QA 대화	2,500	12,488	5.00

각 발화에는 '기쁨, 슬픔, 놀람, 분노, 공포, 혐오, 중립'의 7가지 감정이 태깅되어 있다. 각 데이터에서 사용자 발화에 붙은 감정 레이블의 분포는 <표 2>와 같다.

<표 2> 데이터별 사용자 발화 감정 분포

	기쁨	슬픔	혐오	분노	공포	놀람	중립
감정	42.16%	18.07%	3.72%	4.96%	4.07%	7.70%	19.31%
QA	9.70%	4.81%	0.42%	1.75%	1.43%	4.67%	77.22%

3.2. 사용자 감정 분류

감정 대화 및 QA 대화 데이터에 포함된 사용자 발화의 감정 레이블을 사용하여 KoBERT를 미세 조정하고, ChatGPT와 감정 분류 성능을 비교하였다. 평가 지표로 각 감정 분류별 정확도 (precision), 재현율 (recall) 및 이 두 지표에 기반한 F1-score를 사용하였다.

3.2.1. KoBERT 미세 조정 모델 감정 분류

SKTBrain의 KoBERT를 감정 대화 및 QA 대화 데이터 중 사용자 발화에 대하여 미세 조정하였다. 사용자 발화와 해당 발화의 감정 레이블을 함께 학습한다. 훈련에서 하이퍼파라미터는 batch size 32, initial learning rate 5e-5, training epochs 25로 설정하였고, optimizer는 AdamW를 사용하였다.

3.2.2. ChatGPT 기반 감정 분류

Zero-shot prompt를 이용하여 ChatGPT로 감정 분류를 수행하였다. 7가지 감정 레이블을 프롬프트에 포함시키고 그 중 하나로 발화의 감정을 분류할 것을 요구하였다. (부록 <표 6> 참고)

3.2.3. 평가 결과

두 분류기의 성능을 비교하기 위하여, 감정 대화 117개와 QA 대화 49개에 대해 두 모델의 분류 결과를 정답과 비교하였다. 이때 ChatGPT가 7가지 감정 이외의 레이블을 붙인 경우가 6건 있었고, 이를 7가지 감정 중 하나로 임의로 수정하였다.

KoBERT 기반 분류기의 정확도는 70.53%였고, ChatGPT의 분류 정확도는 61.20%였다. KoBERT 기반 분류기가 ChatGPT보다 전반적으로 성능이 우세하였으며, 세부 성능은 <표 3>과 같았다. 특히 빈도가 낮은 감정에 대해 ChatGPT가 미세 조정된 KoBERT보다 성능이 낮았다. 이는 특정 도메인에서의 성능을 향상하기 위해 해당 도메인 데이터에 대한 학습 모델이 필요함을 보여 준다.

<표 3> KoBERT 및 ChatGPT 기반 감정 분류기 성능

	KoBERT 미세 조정			ChatGPT		
	Precision	Recall	F1	Precision	Recall	F1
기쁨	0.7995	0.8157	**0.8075**	0.8048	0.6768	0.7353
슬픔	0.6263	0.6398	**0.6330**	0.6524	0.5753	0.6114
혐오	0.5176	0.6769	**0.5867**	0.2812	0.1385	0.1856
분노	0.6154	0.4898	**0.5455**	0.5750	0.4694	0.5169
공포	0.5806	0.6667	0.6207	0.6207	0.6667	**0.6429**
놀람	0.3871	0.3636	**0.3750**	0.4000	0.1818	0.2500
중립	0.7479	0.6794	**0.7120**	0.4741	0.7328	0.5757

3.3. 공감적 응답 생성

본 절에서는 사용자 발화의 감정 레이블을 활용하여 공감적 응답을 생성하는 경우와, 감정 인식 없이 공감적 응답을 생성하는 경우에 대한 비교 실험을 수행한다. 이를 위해 ChatGPT와 미세 조정된 Polyglot-Ko을 사용하였으며, 프롬프트는 부록 (<표 7~10>)에 정리하였다.

3.3.1. Polyglot-Ko 미세 조정 모델 응답 생성

감정 대화 데이터에 포함된 사용자 감정 레이블을 활용하여 시스템 응답을 학습하도록 Polyglot-Ko를 미세 조정하였다. 사용자 발화의 감정 레이블을 고려하여 시스템 응답을 생성할 것을 명령어

[3] https://huggingface.co/EleutherAI/polyglot-ko-5.8b
[4] QA 대화의 시스템 응답에는 ETRI에 관한 세부적인 정보가 포함되어 있어, 생성 모델 학습 및 평가에는 사용하지 않았다.

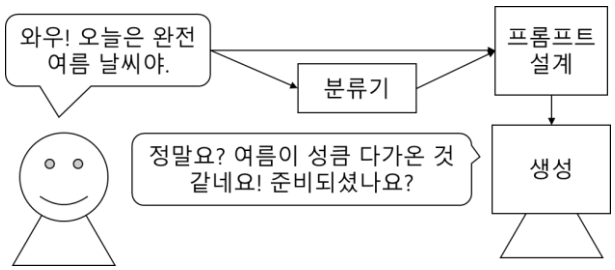
(instruction)로 요구하였다.

Polyglot-Ko 훈련 데이터의 품질을 높이기 위하여, 18글자[5] 미만의 시스템 응답은 훈련 및 평가에서 제외하였다. 그 결과, 시스템 발화의 약 75%가 훈련 및 평가에 사용되었으며, 훈련은 5 epoch 동안 이루어졌다.

3.3.2. 프롬프트 설계

공감적 응답 생성에 사용자 감정 레이블을 활용하는 것이 유용한지 알아보기 위해, 별도의 감정 레이블 없이 대화 기록만 제공하고 공감적인 응답을 생성시키는 경우와, 감정 레이블과 함께 대화 기록을 제공하고 응답을 생성시키는 경우의 성능을 비교한다.

생성 모델에 감정 레이블을 제공할 때에는 (그림 1)과 같이 KoBERT 기반 분류기의 분류 결과를 사용하는 경우와 구축된 데이터에 포함된 정답을 사용하는 경우를 비교한다. ChatGPT로 응답을 생성하는 경우에는, ChatGPT가 분류한 사용자의 감정 레이블을 프롬프트에 포함한 생성 결과도 추가로 확인하였다.



(그림 1) 모델 구성

3.3.3. 평가 결과

언어 모델이 생성한 공감적 응답은 데이터에 포함된 시스템 발화와의 비교를 통해 평가되었다. 평가에는 SacreBLEU, METEOR, ROUGE-L, Semantic Textual Similarity (STS)가 사용되었다. SacreBLEU는 생성된 문장 (candidate)과 정답 문장 (reference)에서 형태소가 얼마나 많이 중첩되는지 수치화한 지표인 BLEU[8]를 사용하기 용이하게 표준화한 것이다.[9] METEOR는 유니그램 precision과 recall의 조화 평균에 기반하여 번역 모델의 성능을 문장 수준에서 측정하는 지표이다.[10] ROUGE은 자동 요약 모델의 성능을 평가하는 지표 중 하나로, ROUGE-L은 시스템 요약본과 정답 요약본 간에 겹치는 최장 길이 부분 문자열의 비율에 기반한다.[11] STS는 의미론적 유사도를 측정하는 지표로, 문장 벡터 간의 코사인 유사도로 계산된다.[12]

[5] 18글자는 감정 대화 데이터에서 시스템 발화의 길이의 평균 33.98에서 표준편차 15.96을 빼서 나온 수치이다.

<표 4> ChatGPT 응답 평가

감정 인식	SacreBLEU	METEOR	ROUGE-L	STS
없음	0.7525	10.1616	**0.4224**	0.5203
ChatGPT	0.4769	8.7478	0.1536	0.5083
KoBERT	**0.9398**	**10.2589**	0.1920	0.5213
정답	0.7946	9.8819	**0.4224**	**0.5276**

ChatGPT가 생성한 공감적 응답에 대한 평가는 감정 대화 117개를 대상으로 진행되었다. 그 결과, KoBERT 분류기의 감정 레이블을 프롬프트에 포함한 경우가 성능이 높게 평가되었다. 정답 레이블을 사용한 경우도 유사한 성능을 보였으며, ChatGPT가 분류한 감정 레이블을 포함한 경우에는 감정 레이블을 포함하지 않은 경우보다 성능이 낮게 나왔다.

<표 5> Polyglot-Ko 기반 미세 조정 모델 응답 평가

감정 인식	SacreBLEU	METEOR	ROUGE-L	STS
없음	1.5386	7.6191	1.0398	0.4613
KoBERT	1.7144	7.9318	1.1474	**0.4651**
정답	**1.7475**	**7.9445**	**1.2071**	0.4642

Polyglot-Ko 기반 미세 조정 모델의 응답 평가는 감정 대화 500개를 대상으로 진행되었다. 정답 레이블을 포함한 경우 성능이 높게 평가되었으며, KoBERT 분류기의 감정 레이블을 포함한 경우도 그에 준하는 성능을 보였다.

ChatGPT와 미세 조정된 Polyglot-Ko를 비교하면, Polyglot-Ko는 SacreBLEU와 ROUGE-L가 ChatGPT보다 다소 높았다.

4. 결론 및 향후 연구

본 연구는 사용자의 감정을 고려하여 공감적인 응답을 생성하는 여러 가지 실험을 진행하면서, KoBERT와 ChatGPT, 그리고 ChatGPT와 Polyglot-Ko를 비교하였다. 사용자의 감정을 프롬프트에 포함하였을 때 공감적 응답 생성의 성능이 향상되었으며, 이때 정확도가 높은 감정 분류기를 사용하는 것이 성능 향상에 도움이 되었다. 이로써 공감적 응답 생성에서 감정 분류기의 유용성을 검증하였고, 비교적 작은 크기의 언어 모델로도 품질이 높은 공감적 응답을 생성할 수 있음을 확인하였다.

공감적 대화 생성을 위한 향후 연구로서 다음 두 가지를 제안한다. 감정 레이블 이외에도 사용자의 프로필 및 의도를 프롬프트에 포함하여 사용자의 상황을 종합적으로 고려하는 대화에 대한 연구가

필요하며, 인간 수준의 공감 대화 및 체감형 (embodied) AI를 위해서는 멀티 모달[13]을 활용하는 연구로 확장할 필요가 있다.

사사

이 논문은 2019 년도 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] C. Pelau, D.-C. Dabija, I. Ene, "What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry", *Computers in Human Behavior*, Vol. 122, pp. 106855, 2021.

[2] H. Ko, et al., "A Technical Report for Polyglot-Ko: Open-Source Large-Scale Korean Language Models", arXiv preprint, 2023, arXiv: 2306.02254.

[3] K. Schaaff, C. Reinig, T. Schlippe, "Exploring ChatGPT's Empathic Abilities", *2023 11th International Conference on ACII,* , 2023, pp. 1-8.

[4] Q. Li, et al., "EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation", *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 4454-4466.

[5] A. Belkhir, F. Sadat, "Beyond Information: Is ChatGPT Empathetic Enough?", *Proceedings of Recent Advances in Natural Language Processing*, 2023, 159-169.

[6] S. Liu, et al., "Towards Emotional Support Dialog Systems", *Proceedings of the 59th Annual Meeting of the ACL and the 11th IJCNLP*, pp. 3469-3483, 2021.

[7] S. Concannon, M. Tomalin, "Measuring perceived empathy in dialogue systems", *AI & SOCIETY: Knowledge, Culture and Communication*, 2023.

[8] K. Papineni, et al., "BLEU: a Method for Automatic Evaluation of Machine Translation", *Proceedings of the 40th Annual Meeting of the ACL*, 2002, pp. 311-318.

[9] M. Post, "A Call for Clarity in Reporting BLEU Scores", *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018, pp. 186-191.

[10] A. Lavie, M. J. Denkowski, "The METEOR metric for automatic evaluation of machine translation", *Machine Translation*, Vol. 23, 2009, pp. 105-115.

[11] C.-Y. Lin, F. J. Och, "Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics", *Proceedings of the 42nd Annual Meeting of the ACL*, 2004, pp. 605-612.

[12] N. Reimers, I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", *Proceedings of the 2019 Conference on EMNLP and the 9th IJCNLP*, 2019, pp. 3982-3992.

[13] Y. Zhang, et al., "DialogueLLM: Context and Emotion Knowledge-Tuned Large Language Models for Emotion Recognition in Conversations", arXiv preprint, 2024, arXiv: 2310.11374.

부록

<표 6> ChatGPT 감정 분류 프롬프트

```
Classify the last `user` utterance as one of only the emotions listed below.
Never use tags that are not in the list below.
Never truncate or modify it.

## Emotions: {}
```

<표 7> ChatGPT 공감 대화 프롬프트 (감정 인식 제외)

```
You are `system`.
`system` and `user` are having emotional conversations.
`system` deeply empathizes with `user` emotions and experiences.
`system` presents an appropriate emotional response to elicit sympathy.

Generate an answer in the dialogue.
Keep the answer less than 50 tokens.

Your response is in the form of a string of the answer you generated.
```

<표 8> ChatGPT 공감 대화 프롬프트 (감정 인식 포함)

```
You are `system`.
`system` and `user` are having emotional conversations.
`system` deeply empathizes with `user` emotions and experiences.
`system` presents an appropriate emotional response to elicit sympathy.

`user` is feeling {}.
Considering the `user` emotion, generate an answer in the dialogue.
Keep the answer less than 50 tokens.

Your response is in the form of a string of the answer you generated.
```

<표 9> Polyglot-Ko 공감 대화 프롬프트 (감정 인식 포함)

```
Considering that <|user|> is feeling {emotion} now, generate an empathetic response from <|system|> that follows a given dialogue. The output consists of a single response from <|system|>.
```

<표 10> Polyglot-Ko 공감 대화 프롬프트 (감정 인식 제외)

```
Generate an empathetic response from <|system|> that follows a given dialogue. The output consists of a single response from <|system|>.
```