긴 맥락 대화에서의 감정인식을 위한 확장형 트랜스포머*

이승준[O1] 김창영[1] 조영우[1] Ngoc-Huynh Ho[1] 양현지[1] 박현[1] 양형정[1*]

[1]전남대학교 인공지능융합학과

*교신저자

dltmdwns3462@naver.com, kcy13930@gmail.com, jyw7720@gmail.com, nhho@jnu.ac.kr, yhj22@jnu.ac.kr, guslgusl5@naver.com, hjyang@jnu.ac.kr

Dialogue_ETC : Extended Transformer for Emotion Recognition in Long Conversation

Seungjun Lee[O1] Changyeoung Kim[1] Youngwoo Jo[1] Ngoc-Huynh Ho[1] Hyeonji Yang[1] Hyun Park[1] Hyungjeong Yang[1]

[1]Department of Artificial Intelligence Convergence, Chonnam National University

*Corresponding Author

요 약

최근 자연어 처리 분야에서 감정 분석의 중요성이 높아지면서, 긴 맥락 대화에서 감정인식을 위한 모델 개발의 필요성이 대두되고 있다. 이를 위해, 본 연구에서는 ETC(Extended Transformer Construction) 아키텍처를 활용하여 발화 및 단어 간의 관계를 분석하는 어텐션을 분리하고 입력 크기를 늘려 긴 대화에서 감정인식을 가능하게 하는 Dialogue_ETC 모델을 제안한다. 또한, 적대적 생성 모델 방식으로 사전 학습한 모델을 이용하여 맥락 정보를 유지하면서 데이터의 표현을 증강시켜 성능을 높인다. 제안한 모델은 입력 데이터 크기의 확장성을 고려하여 학습 비용을 줄이면서도 긴 대화에서 맥락 정보를 고려한다. 그 결과 단일 발화만 고려하였을 때보다 이전 발화를 고려하여 맥락을 파악하였을 때 더 높은 성능을 보여주었다. 모델 구조의 효율성 부분에서 글로벌 어텐션과 로컬 어텐션을 분리함으로써 불필요한 어텐션 계산을 줄여주어 기존의 방식보다 1.82배 많은 학습 파라미터로 64배 많은 입력을 처리하는 성능을 보여주었다.

1. 서 론

인공지능의 빠른 성장으로 인해 인간과 컴퓨터의 상호작용(Human-Computer Interaction, HCI) 분야에서 챗봇, AI 스피커와 같은 대화형 인공지능이 주목받고 있다. 대화형 인공지능은 자연스러운 대화와 표현을 위해 대화 중 사용자의 감정을 인식(Emotion Recognition in Conversation, ERC)하는 작업이 중요하다. 기존에 맥락을 고려하지 않고 단순히 각각의 발화에 대해 감정을 인식하는 방식은 종종 감정 신호의 포착이 어렵다는 문제점이 있다. 문제점을 해결하기 위해 최근에는 대화 참여자의 특성이나 대화 주제와 같은 맥락을 활용해 감정을 인식하는 시도가 이루어졌고 높은 성능 향상을 보여 주었다[1-6]. 이러한 맥락 정보를 얻기 위해 의미론적으로 과거의 발화들을 임베딩하고 외부의 메모리 네트워크로 임베딩 된 발화들에서 맥락 정보를 추출하는 방식들은 효과적으로 화자의 특성과 대화 주제에 대한 맥락을 파악해 성능을 개선하였다. 하지만 기존의 모델들은 구조나 입력의 한계로 인해 긴 대화에서 먼 과거의 발화에 대한 정보가 제대로 고려되지 못한다는 한계점이 존재한다.

본 연구에서는 이러한 한계점을 해결하기 위해 문서처럼 긴 입력의 순서를 처리할 수 있도록 설계된 셀프 어텐션 기반의 모델(Extended Transformer Construction, ETC)[7]을 활용하여 많은 양의 파라미터 증가 없이 긴 대화에서도 맥락에 대한 정보를 추출할 수 있도록 한다. 또한 ERC에서 주로 나타나는 데이터 부족 문제를 해결하기 위해 적대적 생성 모델 방식으로 사전 학습하는 KoELECTRA[1]를 활용하여 맥락 정보를 유지하면서 다양한 표현의 데이터를 생성하여 데이터의 과적합을 피하는 방법을 제안한다.

2. 관련 연구

2.1 맥락을 고려한 감정인식

대화에서 감정을 인식하는 기존 연구에서는 맥락을 추출하기 위한 모델 구조와 감정과 발화의 의미 사이의 관계를 학습 방식에 집중하고 있다. 즉, 모든 발화와 해당하는 화자를 연결해 하나의 문장으로 만들고 프롬프트[2]나 질문[3], 많은 [CLS] 토큰들[4]로 감정을 예측하는 방식이 존재한다. 이러한 방식은 일반적인 트랜스포머 기반의 언어 모델을 사용하는데 512 토큰을 넘는 매우 긴 대화를 처리할 수 없다는 문제점이 존재한다.

다른 방식으로는 각 발화를 임베딩 하거나 특징을 추출해 외부에 추론 모델를 통해 처리하는 방식이 존재한다[5, 6, 7]. 이러한 방식은 매우 긴 대화의 맥락을 처리할 수 있으나 저차원으로 임베딩 할 경우 정보가 손실된다는 문제점이 존재한다.

단순히 모델 구조 이외에 학습 방식의 차이도 있으나 현재 제안된 연구들에서는 전자의 방식들이 후자의 방식

* 이 논문은 화순전남대학교병원 학술연구비(HCRI 23026)에 의하여 연구되었음

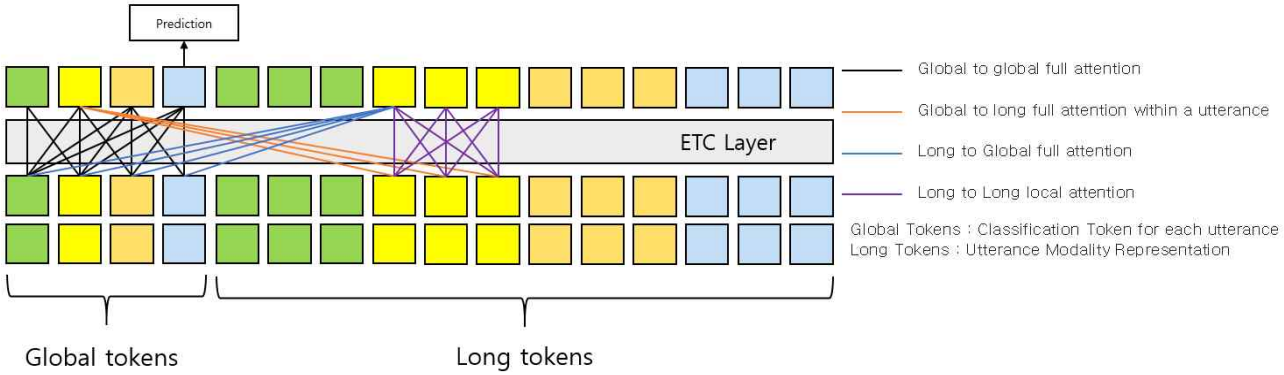

그림 1 . Dialogue_ETC 모델 구조

들보다 높은 성능을 보이는 경향이 있다.

2.2 ETC(Extended Transformer Construction)

[8]은 글로벌 토큰과 일반 토큰 사이에 글로벌-로컬 어텐션이라는 개념을 도입해 매우 긴 입력의 시퀀스도 처리할 수 있는 확장 트랜스포머 구조(Extended Transformer Construction, ETC)를 제안하였다. 이는 긴 입력의 시퀀스뿐만 아니라 구조화된 입력을 인코딩할 수 있게 하는데 아직 대화 기반의 감정인식에 적용한 연구는 진행된 바 없다.

3. 제안 방법

3.1 모델 아키텍쳐

본 연구에서 제안하는 방법은 그림 1과 같이 장기적으로 이전 발화들을 고려할 수 있도록 ETC 구조를 사용하였다. 모든 발화는 "화자:발화"로 연결하여 한 문장이 된다. 각 발화는 의미가 임베딩되는 [CLS] 토큰을 지니고 있으며 [SEP] 토큰으로 분리되어 있다. Dialogue_ETC의 구조는 크게 글로벌-글로벌(g2g)을 계산하는 모듈과 글로벌-로컬(g2l), 로컬-글로벌(l2g), 로컬-로컬(l2l)을 계산하는 모듈로 이루어져 있다(그림 2). g2g에서는 발화 간의 관계를 분석하고 l2l은 단어 간의 관계를 통해 단어의 의미를 분석하는 역할을 한다. g2l과 l2g는 서로 다른 발화 간의 맥락 정보를 전달하는 역할을 한다. 이를 통해 글로벌 어텐션에서는 각 발화의 맥락에 대한 의미가 임베딩 되고 로컬 어텐션에서는 각 발화 내 단어들의 의미가 임베딩 된다. 각 발화는 해당 시점까지의 정보만 임베딩 되도록 마스킹을 적용하여 미래 시점에 대해서는 고려하지 않도록 하였다.

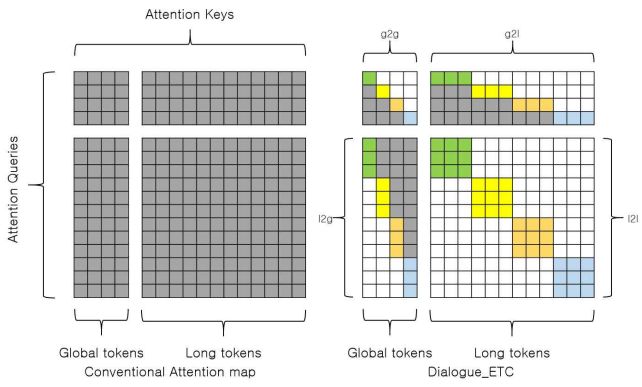

그림 2 . (a) 전통적인 방식의 어텐션 맵, (b) 제안한 방식의 분리된 어텐션 맵

3.2 가중치 초기화

기존에 한국어 말뭉치 데이터로 사전 학습된 ETC 모델이 존재하지 않아 한국어 데이터를 사전 학습한 KoELECTRA 모델을 사용하였다. 식별자의 가중치를 이용해 g2l, l2g, l2l을 계산할 모델을 초기화하였으며, g2g를 계산할 트랜스포머 모델을 새로 정의하였다. KoELECTRA 모델의 경우, 데이터의 특성을 고려하여 대화체를 학습한 Dialog-KoELECTRA 모델을 사용하였다.

4. 실험

4.1 데이터셋

4.1.1 KEMDy20

KEMDy20은 한국어 멀티모달 감정 데이터셋으로 발화 음성, 발화의 문맥적 의미(lexical) 및 여러 생체신호(심전도, 피부전도도, 피부온도)와 발화자의 감정과의 연관성 분석을 위해 수집한 멀티모달 감정 데이터셋이다. 해당 데이터셋들의 감정 레이블은 기쁨, 놀람, 분노, 중립, 혐오, 공포, 슬픔으로 총 7가지로 구성되어 있다. 본 연구에서는 대화 내용만을 고려하기 때문에 텍스트 데이터만 사용하였다.

4.1.2 한국어 멀티모달 영상[9]

AI Hub에서 제공하는 데이터셋으로 감정, 성별, 연령대, 발화 스크립트, 개체 및 관계 정보, 상황 설명 정보, 발화별 대화 의도 및 대화 전략 정보 의미 정보를 통해 구축한 영상 데이터셋이다. 감정 레이블은 기쁨, 슬픔, 분노, 놀람, 공포, 경멸, 혐오, 중립으로 총 8가지로 구성되어 있다. 해당 데이터셋은 KEMDy20의 불균형 문제를 해결하고 감정 레이블의 통일을 위해 경멸을 혐오로 매핑한 후 결합하였다(그림 3).

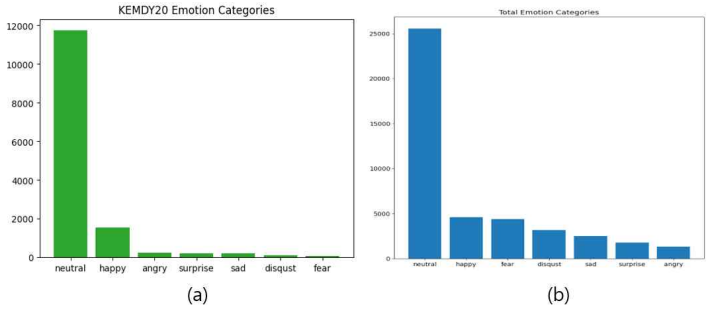

그림 3 . (a) KEMDY20 데이터 분포, (b) 데이터 결합 후 분포

4.1.3 최종 데이터

KEMDY20 데이터를 기준으로 평가하기 위해 세션(동일 그룹)을 기준으로 70%는 학습 데이터, 30%는 평가 데이터로 나눈 후 전처리한 한국어 멀티모달 데이터는 학습 데이터에만 추가하였다.

표 1 . 실험 데이터셋

구분	Train	Test	총계
세션 레벨	78	12	90
스크립트 레벨	2160	72	2232
발화 레벨	38801	3899	42700

4.2 실험 결과

다중 레이블 분류 평가를 위해 클래스별로 F1 -Score를 구해 평균을 내는 Sample F1-Score 방식을 평가 지표로 사용하였다. 실험 결과, 단일 발화만 고려하였을 때 보다 이전 발화를 고려해 맥락을 파악했을 때 더 높은 성능을 보였다.

각 발화의 15%에 마스크를 씌운 후 생성 모델을 활용하여 맥락 정보를 유지하면서 다른 표현의 발화를 생성하는 방식은 데이터 과적합을 막는 동시에 성능의 향상을 보여주었다(표 2).

표 2 . 실험 결과

구분	Precision	Recall	F1-score
Dialog-KoELECTRA[1]	18.55	**89.81**	30.64
Dialog_ETC	29.93	87.39	44.41
Dialog-KoELECTRA[1] + Generator	30.04	87.55	44.54
Dialog_ETC + Generator	**44.55**	86.85	**58.64**

4.3 모델 구조 효율성 분석

제안한 모델인 Dialog_ETC는 그림 2와 같이 글로벌 어텐션과 로컬 어텐션을 분리함으로써 불필요한 어텐션 계산을 줄여주어 기존의 Dialogue-KoELECTRA에 비해 1.82배의 학습 파라미터로 64배 많은 입력을 처리하는 성능을 보여주었다(표 3).

표 3 . 모델 학습 파라미터 개수 비교

구분	입력 크기	학습 파라미터
Dialog-KoELECTRA[1] +Generator	128	112M
Dialog_ETC + Generator (Proposed)	**8192**	**204M**

5. 결론

본 연구는 ETC 아키텍쳐를 활용하여 발화 및 단어 간의 관계를 분석하는 어텐션을 분리하였으며 모델 파라미터의 큰 변화 없이 입력 크기를 늘려 긴 대화에서 감정인식을 가능하게 하는 Dialogue_ETC 모델을 제안하였다. 이는 인공지능의 학습 비용을 줄이면서 효과적으로 긴 대화에서도 맥락 정보를 고려하도록 하였다. 또한 적대적 생성 모델 방식의 사전학습 모델을 이용하여 맥락 정보를 유지하면서 데이터의 표현을 증강 시켜 성능을 높였다.

향후 계획으로 제안한 모델에 음성 데이터를 결합하여 비언어적 표현에 대해서 인식하도록 하는 멀티모달 감정인식 연구를 진행할 예정이다.

6. 참고문헌

[1] W. Kim, J. Kim and O. Jeong.Dialog-KoELECTRA: Korean conversational language model based on ELECTRA model.https://github.com/SKplanet/Dialog-KoELECTRA.2023.

[2] X. Song, L. Huang, H. Xue and S. Hu. “Supervised prototypical contrastive learning for emotion recognition in conversation.”*Proceeding of the 2022 Conference on Empirical Methods in Natural Language Processing,* pp. 5197-5206, 2022.

[3] X. Song, L. Zang, R. Zhang, S. Hu and L. Huang. “Emotionflow: Capture the dialogue level emotion transitions.” *Proceeding of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 8542-8546, 2022.

[4] X. Liu, J. Zhang, H. Zhang, F. Xue and Y. You. “Hierarchical Dialogue Understanding with Special Tokens and Turn-level Attention.” Tiny Paper at ICLR 2023, pp.36, 2023.

[5] J. Lee and W. Lee.“CoMPM: Context modeling with speaker’s pre-trained memory tracking for emotion recognition in conversation.”*Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pp. 5669-5689, 2022.

[6] D. Hu, L. Wei and X. Huai. “Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations.” *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing,* pp. 7042-7052, 2021.

[7] B. Lee and Y. Choi. “Graph based network with contextualized representations of turns in dialogue.” *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 443-455, 2021.

[8] J. Ainslie, S. Ontanon, C. Alberti, V. Cvicek, Z. Fisher and P. Pham, et al. “ETC:Encoding long and structured inputs in transformers.” *Proceeding of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 268-284, 2020.

[9] 아크릴. **멀티모달 영상** [데이터 세트]. AIhub. https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=58, 2019.