💚

[4주차] DialogueRNN: An Attentive RNN for Emotion Detection in Conversations

0. Abstract

- 대화에서 감정 탐지의 중요성
  - 다양한 응용: 대화 내용에서 감정을 탐지하는 것은 채팅 기록 분석, 소셜 미디어 상의 의견 마이닝, 토론 및 논쟁 분석, 실시간 소비자 피드백 이해 등 다양한 분야에서 중요함
  - 개선의 필요성: 현재 시스템은 대화의 각 발화자에 맞게 적응하여 개별 참여자를 개별적으로 처리하지 않아, 이를 개선할 필요가 있음
- DialogueRNN의 접근 방식
  - 개발 참여자 상태 추적: 이 논문에서 제안된 DialogueRNN은 대화 내내 각 참여자의 상태를 추적하는 새로운 접근 방식을 사용함
  - 성능 우수성: 제안된 모델은 두 가지 다른 데이터셋에서 최신 기술보다 상당한 차이로 성능이 우수함을 보여줌

1. Introduction

- DialogueRNN의 핵심 구성 요소
  - GRU(Gated Recurrent Units) 활용: DialogueRNN 시스템은 **발화자, 선행 발화의 맥락, 선행 발화의 감정** 세 가지 측면에 대해 모델링하기 위해 세 가지 GRU를 사용함. 들어오는 발화는 **글로벌 GRU**와 **파티 GRU** 두 가지 GRU에 입력되어 각각 맥락과 파티 상태를 업데이트함
  - 감정 분류를 위한 접근 방식: 업데이트된 발화자 상태는 주어진 발화의 감정 표현을 디코딩하기 위해 **감정 GRU**로 전달되며, 이는 감정 분류에 사용됨
- DialogueRNN의 혁신

- 개선된 맥락 표현: DialogueRNN은 Hazarika et al. 2018, Poria et al. 2017과 같은 최신의 맥락적 감정 분류기보다 더 나은 맥락 표현으로 인해 우수한 성능을 보임
- 상호 파티 관계 모델링: **감정 GRU**와 **글로벌 GRU**는 대화에서 상호 파티 관계 모델링에 중요한 역할을 하는 반면, **파티 GRU**는 동일한 파티의 두 연속 상태 사이의 관계를 모델링함

2. Related Work

- 감정 인식의 발전
  - 얼굴 단서와 감정의 상관관계: kman(1993)은 감정과 얼굴 단서 사이의 상관관계를 발견함
  - 음향 정보와 시각 단서의 결합: Datcu와 Rothkrantz(2008)는 감정 인식을 위해 음향 정보와 시각 단서를 결합함
  - 텍스트 기반 감정 인식: Alm, Roth, Sproat(2005)은 Strapparava와 Mihalcea(2010)의 작업을 발전시켜 텍스트 기반 감정 인식을 도입함
  - 멀티모달 설정에서의 맥락 정보 활용: Wollmer et al.(2010)은 멀티모달 설정에서 감정 인식을 위해 맥락 정보를 사용함
  - RNN 기반 딥 네트워크의 성공적 사용: 최근 Poria et al.(2017)은 멀티모달 감정 인식을 위해 RNN 기반 딥 네트워크를 성공적으로 사용했으며, 이는 Chen et al.(2017), Zadeh et al.(2018a; 2018b) 등의 후속 연구로 이어짐
- 대화에서의 감정 인식
  - 대화 이해의 중요성: 인간 상호작용을 재현하기 위해서는 대화에 대한 깊은 이해가 필요하며, Ruusuvuori(2013)는 대화에서 감정이 중심적인 역할을 한다고 언급한 바가 있음
  - 감정 동영학의 대인 관계적 현상: Richards, Butler, Gross(2003)는 대화에서의 감정 동역학이 대인 관계적 현상이라고 주장함
  - 메모리 네트워크의 성공적 적용: Sukhbaatar et al.(2015)은 질문 응답, 기계 번역, 음성 인식 등 여러 NLP 분야에서 **메모리 네트워크**가 성공적임을 보여주었고, 이에 따라 Hazarika et al.(2018)은 **이중 대화에서 감정 인식을 위해 메모리 네트워크를 사용**했으며 이는 **상호 발화자 간의 상호작용을 가능**하게 하여 최고의 성능을 달성했습니다.

3. Methodology

3.1 Problem Deinition

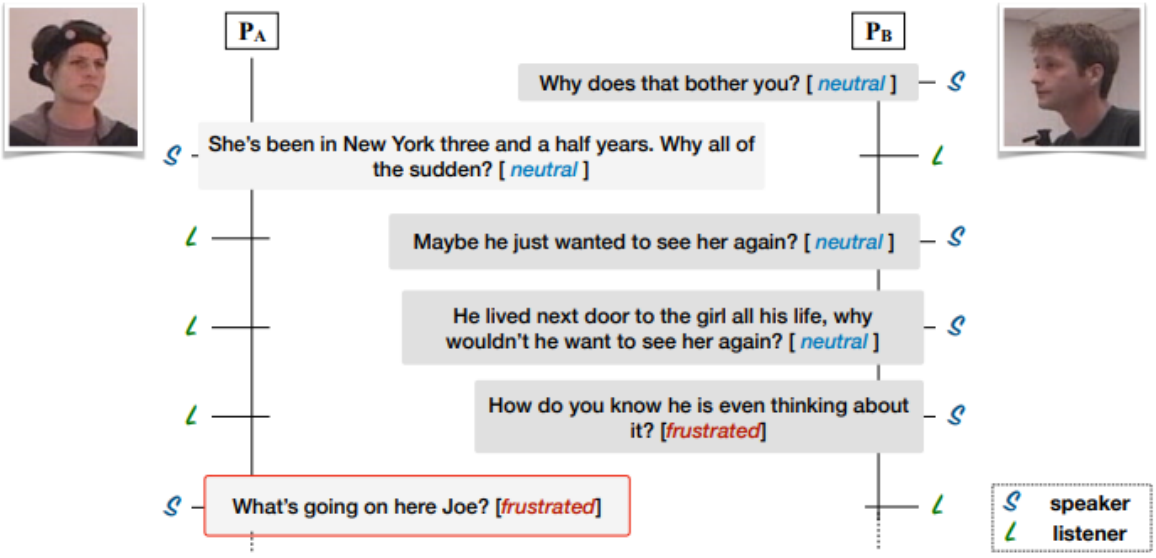

Figure 1: In this dialogue, P_A's emotion changes are influenced by the behavior of P_B.

- 문제 설정
  - 참여자 정의: 대화에는 M명의 참여자 (p1, p2, ..., pM)가 있으며, 이 논문에서 사용된 데이터셋의 경우 M=2로 설정됨
  - 발화와 감정 레이블: 각 발화(u1, u2, ..., uN)는 특정 참여자(ps(ut))에 의해 발화되며, 각 발화는 해당 참여자의 감정 상태를 나타냄
  - 발화 표현: 발화 ut는 R^Dm 공간에 있는 벡터로 표현되며, 이는 특징 추출기를 사용하여 얻어짐
  - 발화와 참여자의 매핑: s는 발화와 그것을 발화한 참여자의 인덱스 사이의 매핑을 나타냄
- 핵심 목표: 대화 중 발생하는 각각의 발화에 대해 정확한 감정 레이블을 예측하는 것

3.2 Unimodal Feature Extraction

- 텍스트 특징 추출
  - CNN 사용: 텍스트에서 특징을 추출하기 위해 합성곱 신경망(CNN)을 사용
  - n-gram 특징: 각 발화에서 3, 4, 5 크기의 세 가지 다른 합성곱 필터를 사용하여 n-gram 특징을 얻으며, 각 필터는 50개의 특징 맵을 가짐

- 후처리: 출력은 최대 풀링(max-pooling)을 거친 후, ReLU 활성화 함수를 적용
- 텍스트 발화 표현: 이러한 활성화들은 연결되어 100차원의 밀집층(dense layer)으로 전달되며 이는 텍스트 발화의 표현으로 간주되고 이 네트워크는 감정 레이블과 함께 발화 수준에서 훈련됨

- 오디오 및 비주얼 특징 추출
  - 3D-CNN 및 openSMILE 사용: 비주얼 특징 추출을 위해 3D-CNN을, 오디오 특징 추출을 위해 openSMILE을 사용

3.3 Our Model

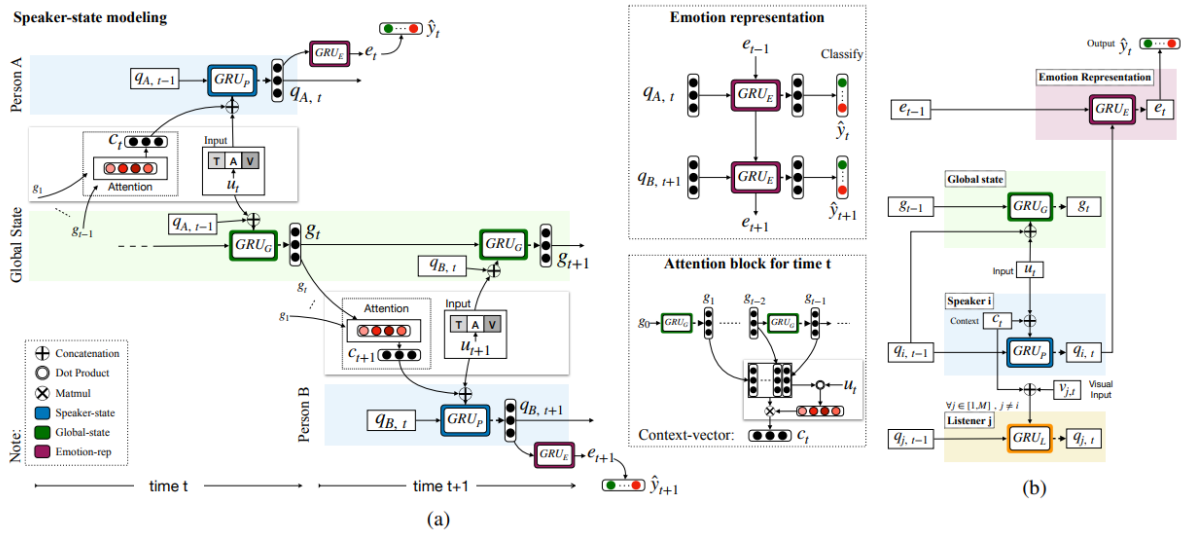

Figure 2: (a) DialogueRNN architecture. (b) Update schemes for global, speaker, listener, and emotion states for t^{th} utterance in a dialogue. Here, Person i is the speaker and Persons $j \in [1, M]$ and $j \neq i$ are the listeners.

- 모델 구조
  - **발화자 모델링**: 각 참여자는 발화할 때마다 변화하는 '**파티 상태(party state)**'를 사용하여 모델링됨. 이를 통해 모델은 대화를 통한 참여자의 감정 동태를 추적할 수 있음
  - **맥락 모델링**: 발화의 맥락은 '**글로벌 상태(global state)**'를 사용하여 모델링되며 이전 발화와 파티 상태가 함께 인코딩되어 맥락 표현을 위해 사용됨
  - **감정 추론**: 모델은 발화자의 파티 상태와 이전 발화자의 상태를 맥락으로 사용하여 감정 표현을 추론하며, 이 감정 표현은 최종 감정 분류에 사용됨

- GRU 셀 사용
  - 상태 및 표현 업데이트: 상태와 표현을 업데이트하기 위해 GRU 셀(Chung et al. 2014)을 사용하며, 각 GRU 셀은 현재 입력과 이전 GRU 상태를 기반으로 숨겨진

상태를 계산함

- 글로벌 상태(Global GRU)
  - 맥락 포착: 글로벌 상태는 발화와 발화자 상태를 함께 인코딩하여 주어진 발화의 맥락을 포착하는 것을 목표로 하며, 각 상태는 발화자별 발화 표현으로도 사용됨
  - 현재 발화와 이전 글로벌 상태를 기반으로 새로운 글로벌 상태 생성
  - [G_t = f(G_{t-1}, U_t, S_t)]
    - (G_t)는 시간 t에서의 글로벌 상태, (U_t)는 현재 발화, (S_t)는 발화자 상태
- 파티 상태(Party GRU)
  - 개별 발화자 추적: 대화 중 **개별 발화자의 상태**를 **고정 크기 벡터**로 추적, 이 **상태 벡터**는 감정 분류에 관련된 발화자의 상태를 나타냄
  - 상태 업데이트: 발화자의 현재 역할(발화자 또는 청취자)과 들어오는 발화에 기반하여 상태 벡터를 업데이트
- 발화자 업데이트(Speaker GRU)
  - **맥락 포착**: 발화자는 대화에서 이전 발화를 기반으로 응답을 구성하며, 이전 글로벌 상태를 기반으로 현재 발화에 관련된 맥락을 포착함
  - [S_t = f(S_{t-1}, U_t)]
    - (S_t)는 시간 (t)에서의 발화자 상태
  - **주의 메커니즘**: 이전 글로벌 상태에 대한 주의 점수를 계산하여 현재 발화와 감정적으로 관련된 발화에 더 높은 주의를 기울임
- 청취자 업데이트
  - 상태 변화 모델링: 발화자의 발화로 인한 청취자의 상태 변화를 모델링함. 청취자의 상태를 변경하지 않거나, 청취자의 시각적 단서(얼굴 표정)를 기반으로 상태를 업데이트하는 두 가지 방법을 시도할 수 있음
  - [L_t = f(L_{t-1}, U_t)]
- 감정 표현 및 분류
  - 감정 표현 추론: 발화자의 상태와 이전 발화의 감정 표현을 기반으로 현재 발화의 감정 표현을 추론
  - 감정 분류: 감정 표현을 기반으로 6가지 감정 클래스 확률을 계산하고, 가장 가능성 높은 감정 클래스를 선택
- 훈련

- 손실 측정: 범주형 교차 엔트로피와 L2 정규화를 사용하여 훈련 중 손실 측정
- 최적화: Adam 최적화 알고리즘을 사용하여 네트워크를 훈련

> 💡 **용어 정리**
>
> **글로벌 상태(Global State)**
> -
> **대화 전체의 상태**를 나타내는 벡터로, 대화의 전반적인 맥락과 흐름을 포착함
> - 예시: 대화에서 여러 사람이 참여하는 회의 상황을 생각했을 때, 글로벌 상태는 회의의 주제, 진행 상황, 전반적인 분위기 등을 포함함
>
> **파티 상태(Party State)**
> -
> **대화에 참여하는 각 개인(파티)의 상태**를 나타내는 벡터로, 개인의 감정 상태, 대화에서의 역할 등을 포함함
> - 예시: 회의에서 한 사람이 발표를 하고 있을 때, 그 사람의 파티 상태는 발표 내용, 감정 상태(긴장, 자신감 등)를 반영함
>
> -
> **발화자(Speaker)**: 현재 발화를 하는 대화 참여자
> -
> **청취자(Listener)**: 발화를 듣고 있는 대화 참여자
> -
> **발화자 업데이트**: 발화자의 상태는 현재 발화와 이전 대화 맥락을 기반으로 업데이트됨
> -
> **청취자 업데이트**: 청취자의 상태는 발화자의 발화와 청취자의 시각적 단서(예: 얼굴 표정)를 기반으로 업데이트될 수 있음(안 될 수도 있음)
> -
> **감정 분류**: 대화에서 발화의 감정을 분류하는 과정

3.4 DialogueRNN Variants

- DialogueRNN + Listener State Update(DialogueRNN_l)
  - 발화자 상태에 기반하여 청취자 상태를 업데이트

    - 발화자 상태 (q_s(u_t),t)를 사용하여 청취자 상태를 업데이트하며, 이를 통해 대화에서 청취자의 역할과 반응을 더 정확하게 모델링할 수 있음
- Bidirectional DialogueRNN (BiDialogueRNN)
    - 양방향 RNN과 유사하게, 입력 시퀀스의 앞뒤로 두 개의 다른 RNN을 사용
    - 앞과 뒤로 진행되는 DialogueRNN을 통해 대화의 과거와 미래 발화 모두에서 정보를 얻어 최종 감정 표현에 포함시키며 이는 감정 분류에 있어 더 나은 맥락을 제공함
- DialogueRNN + Attention(DialogueRNN + Att)
    - 각 감정 표현 (e_t)에 대해 대화 내 모든 주변 감정 표현과 매칭하여 주의(attention)를 적용
    - 주의 점수에 기반하여 관련 있는 미래 및 이전 발화의 맥락을 제공
- Bidirectional DialogueRNN + Emotional Attention (BiDialogueRNN+Att)
    - BiDialogueRNN의 각 감정 표현 (e_t)에 대해, 대화 내 모든 감정 표현에 주의를 적용하여 다른 발화들의 맥락을 포착
    - 주의 점수 계산: βt = softmax(((e_t)^T)Wβ[e1, e2, . . . , eN])
        - (e_t)는 현재 발화의 감정 표현
        - (W_\beta)는 학습 가능한 가중치 행렬
        - 현재 발화의 감정 표현과 대화 내 다른 모든 발화의 감정 표현 사이의 관계를 계산하여 주의 점수βt를 얻음
    - 가중 감정 표현 계산: e˜t = βt[e1, e2, . . . , eN]^T
        - where et ∈ R^2DE , Wβ ∈ R^(2DE ×2DE) , e˜t ∈ R^2DE , and (β_t)^T∈R^N
        - 계산된 주의 점수를 사용하여 대화 내 모든 발화의 감정 표현에 가중치를 적용하고, 이를 통합하여 현재 발화의 가중 감정 표현 e_t를 얻음
    - 감정 분류를 위한 입력: 얻어진 가중 감정 표현 e_t은 감정 분류를 위해 두 층의 퍼셉트론에 입력 되며, 이 과정에서 대화의 맥락을 고려한 감정 상태가 더 정확하게 분류됨
    - 이 메커니즘은 대화 내에서 각 발화의 감정 상태가 이전 및 이후 발화의 감정 상태와 어떻게 관련되어 있는지를 파악하며, 이를 통해 모델은 각 발화의 감정을 더 정확하게 이해하고 분류할 수 있음

4. Experimental Setting

4.1 Datasets Used

Dataset	Partition	Utterance Count	Dialogue Count
IEMOCAP	train + val	5810	120
	test	1623	31
AVEC	train + val	4368	63
	test	1430	32

Table 1: Dataset split ((train + val) / test ≈ 80%/20%).

- IEMOCAP 데이터셋
  - 10명의 고유 화자가 참여한 양방향 대화의 비디오를 포함함
  - 첫 번째부터 네 번째 세션에 속한 첫 8명의 화자만이 훈련 세트에 포함됨
  - 각 비디오는 대화를 발화 단위로 세분화하며, 이 발화들은 행복, 슬픔, 중립, 화남, 흥분, 좌절의 **6가지 감정 레이블**로 주석이 달려있음
- AVEC 데이터셋
  - SEMAINE 데이터베이스의 수정 버전으로, 인간과 인공 지능 에이전트 간의 상호작용을 포함함
  - 각 대화의 발화는 **네 가지 실수 값의 정서 속성(감정 값)**으로 주석이 달려 있으며, 이는 기쁨(-1, 1), 흥분(-1, 1), 기대(-1, 1), 그리고 권력(0,∞)을 나타냄
  - 원본 데이터베이스에서는 0.2초마다 주석이 달려 있지만, 발화 수준의 주석이 필요한 우리의 요구에 맞추기 위해 발화 기간 동안 이 속성들을 평균화함
    - 원본: 모든 발화에 대해 0.2초 간격으로 정서 속성에 대한 주석이 기록됨
    - DialogueRNN: 발화 수준에서의 정서 분석이 필요, 각각의 발화 전체에 대해 하나의 정서 속성 값을 기록해야 함
    - 연구자들은 발화 기간 동안의 모든 정서 속성 값들을 평균내어, 각 발화에 대한 단일한 정서 속성 값을 생성
    - 발화 전체를 대표하는 감정 상태를 얻음

4.2 Baselines and State of the Art

- c-LSTM
  - 양방향 LSTM을 사용하여 주변 발화의 맥락을 포착하고 맥락을 인식하는 발화 표현을 생성
  - 발화자 사이의 차이 구분 ❌
- c-LSTM+Att
  - c-LSTM 출력에 주의(Attention) 메커니즘을 적용 최종 발화 표현에 더 나은 맥락을 제공함
- TFN
  - 다중 모달 시나리오에 특화된 모델로, 텐서 외적을 사용하여 모달 간 및 모달 내 상호작용을 포착함
  - 주변 발화의 맥락을 포착 ❌
- MFN
  - 다중 모달 시나리오에 특화되어 뷰-특정 및 교차 뷰 상호작용을 모델링함으로써 다중 뷰 학습을 활용
  - TFN과 마찬가지로, 이 모델은 맥락 정보를 사용 ❌
- CNN
  - 텍스트 특징 추출 네트워크와 동일, 주변 발화의 맥락 정보를 사용 ❌
- Memnet
  - 현재 발화를 메모리 네트워크에 입력하고, 메모리는 선행 발화에 해당함
  - 메모리 네트워크의 출력은 정서 분류를 위한 최종 발화 표현으로 사용됨
- CMN
  - 대화 역사에서 발화 맥락을 모델링하기 위해 두 명의 발화자에 대해 두 개의 별도 GRU를 사용하는 최신 방법
  - 현재 발화를 두 발화자 모두에 대한 별도의 메모리 네트워크에 쿼리로 입력하여 발화 표현을 얻음

4.3 Modalities

- **텍스트 모달리티**: DialogueRNN의 주요 평가는 텍스트 데이터를 기반으로 이루어졌는데, 이는 대화에서 발화의 정서를 탐지하는 데 중점을 두었기 때문임
- **멀티 모달 실험**: 멀티 모달 시나리오의 효과를 입증하기 위해 텍스트 외에도 다른 모달리티의 특징을 실험에 포함시킴. 이는 모델이 다양한 유형의 데이터를 처리할 수 있는 능력

을 보여주기 위함이었음

5. Results and Discussion

Methods	IEMOCAP												AVEC							
	Happy		Sad		Neutral		Angry		Excited		Frustrated		**Average(w)**		Valence		Arousal		Expectancy		Power	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	MAE	r	MAE	r	MAE	r	MAE	r
CNN	27.77	29.86	57.14	53.83	34.33	40.14	61.17	52.44	46.15	50.09	62.99	55.75	48.92	48.18	0.545	-0.01	0.542	0.01	0.605	-0.01	8.71	0.19
memnet	25.72	33.53	55.53	61.77	58.12	52.84	59.32	55.39	51.50	58.30	67.20	59.00	55.72	55.10	0.202	0.16	0.211	0.24	0.216	0.23	8.97	0.05
c-LSTM	29.17	34.43	57.14	60.87	54.17	51.81	57.06	56.73	51.17	57.95	67.19	58.92	55.21	54.95	0.194	0.14	0.212	0.23	0.201	0.25	8.90	-0.04
c-LSTM+Att	30.56	35.63	56.73	62.90	57.55	53.00	59.41	59.24	52.84	58.85	65.88	59.41	56.32	56.19	0.189	0.16	0.213	0.25	0.190	0.24	8.67	0.10
CMN (SOTA)	25.00	30.38	55.92	62.41	52.86	52.39	61.76	59.83	55.52	60.25	71.13	60.69	56.56	56.13	0.192	0.23	0.213	0.29	0.195	0.26	8.74	-0.02
DialogueRNN	31.25	33.83	66.12	69.83	63.02	57.76	61.76	62.50	61.54	64.45	59.58	59.46	59.33	59.89	0.188	0.28	0.201	0.36	0.188	0.32	8.19	0.31
DialogueRNN$_f$	35.42	35.54	65.71	69.85	55.73	55.30	62.94	61.85	59.20	62.21	63.52	59.38	58.66	58.76	0.189	0.27	0.203	0.33	0.188	0.30	8.21	0.30
BiDialogueRNN	32.64	36.15	71.02	74.04	60.47	56.16	62.94	63.88	56.52	62.02	65.62	**61.73**	60.32	60.28	0.181	0.30	0.198	0.34	0.187	0.34	8.14	0.32
DialogueRNN+Att	28.47	**36.61**	65.31	72.40	62.50	57.21	67.65	**65.71**	70.90	68.61	61.68	60.80	61.80	61.51	0.173	0.35	0.168	0.55	0.177	0.37	7.91	0.35
BiDialogueRNN+Att	25.69	33.18	75.10	**78.80**	58.59	**59.21**	64.71	65.28	80.27	**71.86**	61.15	58.91	63.40	**62.75**	**0.168**	**0.35**	**0.165**	**0.59**	**0.175**	**0.37**	**7.90**	**0.37**

Table 2: Comparison with the baseline methods for textual modality; Acc. = Accuracy, MAE = Mean Absolute Error, r = Pearson correlation coefficient; bold font denotes the best performances. Average(w) = Weighted average.

5.1 Comparison with the State of the Art

- IEMOCAP 데이터셋 성능 비교
  - 성능 우위: DialogueRNN은 IEMOCAP 데이터셋에서 평균 정확도에서 2.77%, 평균 F1-점수에서 3.76%로 CMN을 상회함
  - 성능 향상 요인
    - GRUP을 사용한 파티 상태 모델링
    - 발화자 특정 발화 처리
    - GRUG를 사용한 전역 상태 캡처
  - 감정 라벨별 성능: 6개의 불균형 감정 라벨을 다루면서 DialogueRNN은 6개 중 5개 감정 클래스에서 CMN을 큰 차이로 앞서나감. 단 'frustrated' 클래스에 있어서는 CMN에 비해 F1-점수가 1.23% 낮은 것으로 나타남
- AVEC 데이터셋 성능 비교
  - 성능 우위: DialogueRNN은 valence, arousal, expectancy, power 속성에서 CMN을 능가함
  - 성능 지표: 모든 네 속성에 대해 훨씬 낮은 평균 절대 오차(MAE)와 더 높은 피어슨 상관 계수(r)를 기록함
  - 성능 향상 요인: CMN에서 누락된 파티 상태 및 감정 GRU의 통합이라고 볼 수 있음

5.2 DialogueRNN vs. DialogueRNN Variants

- DialogueRNNl

    - 명시적인 청취자 상태 업데이트를 사용하는 DialogueRNNI 변형은 일반 DialogueRNN보다 약간 낮은 성능을 보임(IEMOCAP 및 AVEC 데이터셋 모두에 해당)
    - 'happy' 감정 라벨에 대해서는 DialogueRNNI이 DialogueRNN보다 F1-점수에서 1.71% 더 높은 성능을 보임
- BiDialogueRNN
    - 미래 발화의 맥락을 포착하기 때문에, DialogueRNN보다 성능이 향상될 것으로 예상됨
    - 실제로, BiDialogueRNN 변형은 두 데이터셋 모두에서 평균적으로 DialogueRNN보다 더 나은 성능을 보임
- DialogueRNN+Attn
    - 과거 및 미래 발화의 정보를 현재 발화와 매칭하여 주의 점수를 계산함으로써 정보를 사용함
    - 감정적으로 중요한 맥락 발화에 관련성을 제공하여 BiDialogueRNN보다 더 나은 성능을 제공
    - IEMOCAP에서는 BiDialogueRNN보다 1.23% 높은 F1-점수를, AVEC에서는 일관되게 낮은 MAE와 더 높은 r을 기록
- BiDialogueRNN+Attn
    - BiDialogueRNN에서의 감정 표현에 주의를 기울여 최종 감정 표현을 생성, 논의된 다른 모든 방법보다 일반적으로 더 나은 성능을 보임
    - IEMOCAP 데이터셋에서는 최신 기술인 CMN보다 평균 6.62% 높은 F1-점수를, 바닐라 DialogueRNN보다는 2.86% 높은 F1-점수를 기록
    - AVEC 데이터셋에서도 모든 네 속성에 대해 최고의 성능을 보임

5.3 Multimodal Setting

Methods	IEMOCAP	AVEC			
	F1	Valence (r)	Arousal (r)	Expectancy (r)	Power (r)
TFN	56.8	0.01	0.10	0.12	0.12
MFN	53.5	0.14	25	0.26	0.15
c-LSTM	58.3	0.14	0.23	0.25	-0.04
CMN	58.5	0.23	0.30	0.26	-0.02
BiDialogueRNN+att$_{text}$	62.7	0.35	0.59	0.37	0.37
BiDialogueRNN+att$_{MM}$	**62.9**	**0.37**	**0.60**	**0.37**	**0.41**

Table 3: Comparison with the baselines for trimodal (T+V+A) scenario. BiDialogueRNN+att$_{MM}$ = BiDialogueRNN+att in multimodal setting.

- 멀티 모달 특성에 대한 DialogueRNN 평가
  - 다중 모달 특성: Hazarika 등(2018)에 의해 사용되고 제공된 다중 모달 특성을 활용하여 DialogueRNN을 평가함. 이는 **단일 모달 특성**의 연결(concatenation)을 통한 융합 방법을 따르며, 해당 논문에서는 융합 메커니즘에 초점을 맞추지 않기 때문에 Hazarika 등(2018)의 방법을 따릅니다
  - 성능 비교: Table 3에서 볼 수 있듯, DialogueRNN은 강력한 기준 모델들과 최신 기술인 CMN을 상당히 능가하는 성능을 보이며 이는 DialogueRNN이 다양한 모달을 통합하여 대화에서 감정을 탐지하는 데 매우 효과적임을 시사함

5.4 Case Studies

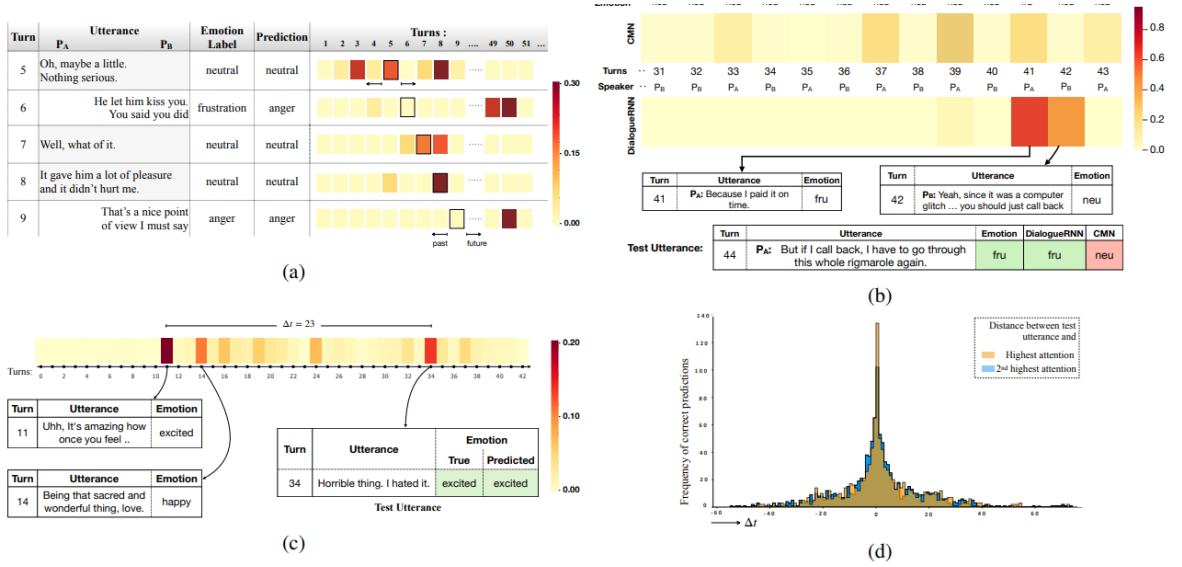

Figure 3: (a) Illustration of the β attention over emotion representations e_t; (b) Comparison of attention scores over utterance history of CMN and DialogueRNN (α attention). (c) An example of long-term dependency among utterances. (d) Histogram of Δt = distance between the target utterance and its context utterance based on β attention scores.

- 이전 발화에 대한 의존성(DialogueRNN)
  - 주의 집중 메커니즘
    - DialogueRNN은 글로벌 GRU(GRUG)의 출력에 대한 주의 모듈을 포함함
    - 이 모델의 주의는 CMN 모델에 비해 더 집중적으로 작용함
    - CMN은 희석된 주의 점수를 제공하여 오분류를 초래하는 반면, DialogueRNN은 더 집중된 주의를 통해 정확도를 높이고자 함
  - 감정 변화 탐지
    - 예를 들어 테스트 발화에서 PA의 감정이 중립에서 좌절로 변하는 경우, DialogueRNN은 PA와 PB가 각각 말한 41번과 42번 턴에 주의를 기울여 이러한 감정 변화를 정확히 예측함
    - CMN은 이러한 의존성을 포착하지 못해 잘못된 감정을 예측할 수 있음
- 미래 발화에 대한 의존성(BiDialogueRNN+Att)
  - 감정 상태 간의 상호 의존성
    - 여성(PA)이 처음에는 중립 상태인 반면, 남성(PB)은 화가 난 상태로 대화를 이어나가는 상황
    - 여성의 감정 주의는 그녀의 중립 상태 동안에 국한되며, 특정 턴들은 과거와 미래의 발화에 주의를 기울임
    - 이는 미래 발화를 통한 감정 상태 간의 상호 의존성을 나타냄
- 먼 맥락에 대한 의존성
  - 장기 감정 의존성
    - IEMOCAP 테스트 세트에서 정확한 예측을 한 경우, 테스트 발화와 가장 관련이 높은 발화 사이의 상대적 거리 분포를 요약함
    - 이는 대화에서 특정 감정 톤을 유지하고 자주 감정 변화가 발생하지 않는 경우에 주로 나타남

5.5 Error Analysis

- 관련 감정 간의 교차 예측
  - 유사 감정 오류
    - '행복' 감정을 '흥분' 클래스로 잘못 분류하는 경향이 있음
    - '화남'과 '좌절' 감정도 서로 잘못 분류되는 경우가 많음

        - 해당 감정 쌍 사이의 미묘한 차이로 인해 구별하기 어렵기 때문이라고 추정됨
- 높은 오류율을 보이는 클래스
    - 중립 클래스 오류
        - 중립 클래스는 높은 비율의 거짓 긍정(false-positives)을 보임
        - 고려된 감정들 중에서 중립 클래스가 분포에서 차지하는 비율이 많기 때문이라고 추정됨
- 대화 수준에서의 오류
    - 감정 변화에서의 오류
        - 같은 당사자의 이전 턴에서 감정이 변할 때 오류가 발생하는 경우가 많음
        - 테스트 세트에서 이러한 감정 변화를 모델이 정확히 예측하는 비율은 47.5%에 불과함
        - 정 변화가 없는 경우의 성공률 69.2%에 비해 낮은 것으로 나타남
- 개선의 여지
    - 감정 변화의 복잡성
        - 대화에서 감정 변화는 숨겨진 역학에 의해 지배되는 복잡한 현상임
        - 이러한 경우 추가 개선은 여전히 연구에서 과제로 남아있음

5.6 Ablation Study

Party State	Emotion GRU	F1
-	+	55.56
+	-	57.38
+	+	59.89

Table 4: Ablated DialogueRNN for IEMOCAP dataset.

- 핵심 구성 요소의 영향
    - 파티 상태의 중요성
        - 파티 상태 없이 모델을 평가했을 때 성능이 4.33% 감소함

        - 파티 상태가 대화 참여자들의 감정과 관련된 유용한 문맥 정보를 추출하는 데 도움이 된다는 것을 시사함
    - 감정 GRU의 영향
        - 감정 GRU 없이 모델을 평가했을 때 성능이 2.51% 감소함
        - 파티 상태만큼은 아니지만 중요한 요소임을 나타냄
        - 감정 GRU의 부재로 인해, 다른 참여자들의 상태에서 오는 문맥 흐름이 선행 발화의 감정 표현을 통해 전달되지 않는 것으로 보임

6. Conclusion

- **RNN 기반 신경망 구조**: 대화 중 감정 탐지를 위한 새로운 접근 방식을 제시
- **발화자의 특성 고려**: 각 발화를 처리할 때 발화자의 특성을 고려하여 더 정교한 문맥 정보를 제공함
- **우수한 성능**: 두 개의 다른 데이터셋에서 텍스트와 멀티모달 설정 모두에서 현재 최신 기술을 능가하는 성능을 보여줌
- **다자간 설정 확장 가능**: 두 명 이상의 발화자가 있는 다자간 설정으로의 확장 가능성을 가지고 있으며, 이는 향후 연구의 주제가 될 것으로 보임

논문에 대한 의견 및 의문점(꼭지)

➡️ 위 논문에서는 기본적으로 대화의 참여자를 2명으로 고정해놓고 실험을 진행했는데, 만약 대화 참여자가 3명 이상으로 늘어날 경우 모델의 성능이 어떻게 달라질지에 대해서 알아보고 싶음. 또한 감정 라벨의 종류를 더 다양하게 하는 방향으로 모델을 개선하면 더욱 효과적일 것으로 생각함.