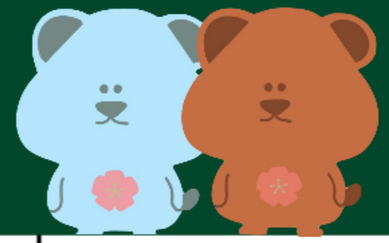
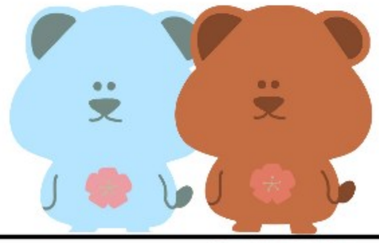


Dialogue RNN: An Attentive RNN for Emotion Detection in Conversations

고급심화 이서연

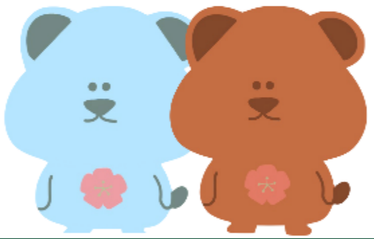
EWHA EURON

목차



EWHA EURON

0. Abstract & Background



EWHA EURON

#0. Abstract

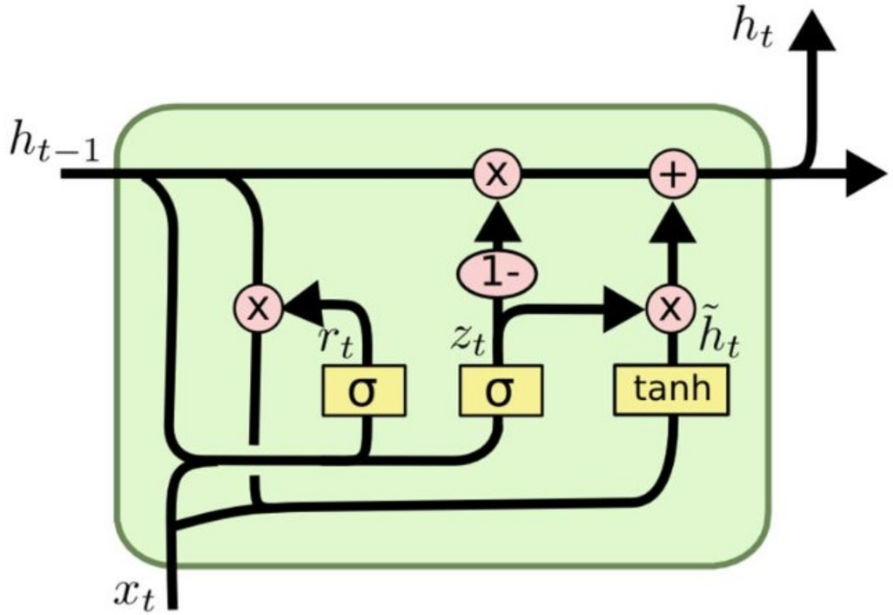
- **Emotion Detection in Conversations**

  - 대화 내에서 감정 감지하는 것은 여러 어플리케이션에서 필수적임
  - 현재 시스템은 대화에서 각 발화자를 개별적으로 구분하지 않음.

    ⇒ 이 논문에서는 대화 전반에서 RNN을 통해 개별 발화자를 추적하고 이를 감정 분류에 사용함

#0. Background

GRU (Gated Recurrent Unit)



- LSTM의 구조는 단순화시키면서 성능은 유지시킴
- 순차 데이터를 처리하는 데 사용됨
- **Reset gate**와 **Update gate**로 구성됨
  - **Reset gate:** 현재 상태에서 얼마나 이전 상태의 정보를 유지할지 결정
    (0~1: 1에 가까울 수록 이전 상태 유지)
  - **Update gate:** 이전 상태의 정보와 새로운 정보 사이의 균형을 결정
    (0~1: 1에 가까울 수록 이전 상태를 우선시)

출처: https://itrepo.tistory.com/40#Work-Flow-of-GRU

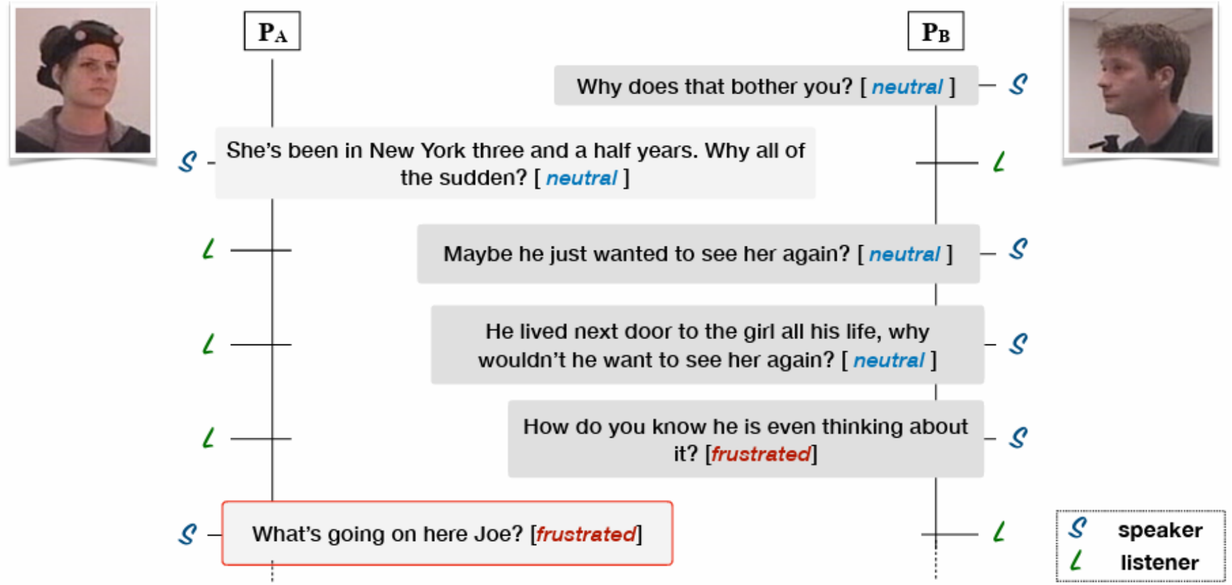
1. Introduction

EWHA EURON

#1. Introduction

- **대화에서 감정과 관련된 주요 3가지 측면**
  - 화자
  - 이전 발화의 문맥
  - 감정

⇒ 개별적으로 모델링하여 성능을 높임



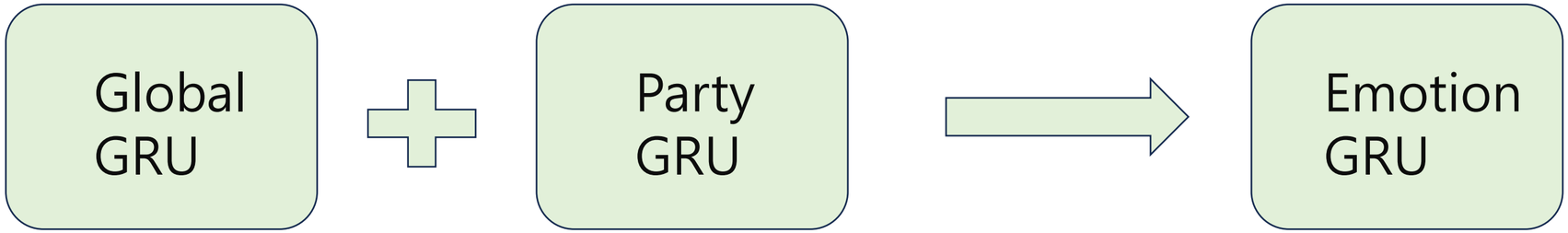
P_A의 감정이 P_B의 이전 발화에 영향을 받음

#1. Introduction

- **주요 3가지 측면 모델링**

  - **Global GRU:** 대화의 전반적인 문맥 파악하고 업데이트
  - **Party GRU:** 대화의 개별 화자 상태 모델링하고 업데이트
  - **Emotion GRU:** 발화의 감정 표현 디코딩 후 감정 분류에 사용

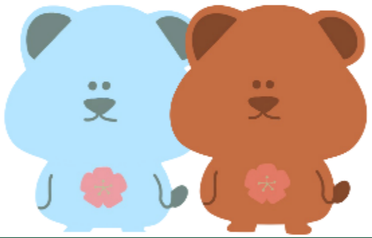
  ⇒Dialogue RNN에서는 3가지 GRU가 재귀적으로 연결됨



대화의 전반적인 문맥과
개별 party state 업데이트

해당 발화의
감정 표현 해독

2. Related Work



EWHA EURON

#2. Related Work

- 감정 인식은 다양한 분야에서 주목을 받고 있음.
  - 감정과 얼굴 표정간의 상관 관계
  - 음성 정보와 시각적 신호의 통합
  - 텍스트 기반의 감정 인식
  - Multi-modal 설정에서의 맥락 정보 활용

- 감정이 시간에 따라 변화하는 동적인 특성은 대화 참여자들 간의 상호 작용에 의해 영향을 받음.

  ⇒ 시간의 흐름에 따른 대화를 재현하기 위해 RNN 도입

  ⇒ 2개의 memory network를 통해 화자 간 상호작용 가능케 함

3. Methodology

#3.1 Problem Definition

○ 감정 Label을 예측해야 함 (**Happy, Sad, Neutral, Angry, Excited, Frustrated**)

○ P1, P2,…Pm: 대화 속 화자

○ S 함수: 발화 Ut와 발화 속 당사자를 매핑

⇒ S(Ut): 발화 Ut에 속하는 당사자를 나타냄

○ 발화 표현은 D차원 벡터로 표현되며, **Feature Extractor** 에 의해 얻을 수 있음

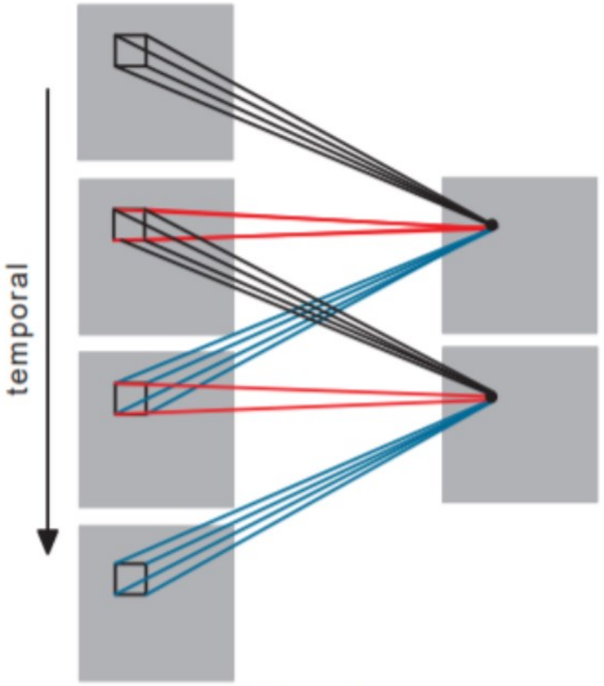
#3.2 Unimodal Feature Extraction

- **Conversational Memory Networks (CMN)** 에서 사용된 특징 추출 과정과 동일한 방법을 사용해서 특징을 추출
  - **텍스트 특징 추출 (Textual Feature Extraction) :** CNN을 사용하여 각 발화에서 텍스트 특징을 추출하는 과정
    - 각 발화를 n-gram 특징으로 변환하여 텍스트 표현을 얻음
  - **청각적, 시각적 특징 추출 (Audio and Visual Feature Extraction):**
    - **시각적 특징 추출:** 3D-CNN 사용
    - **청각적 특징 추출:** openSMILE 사용

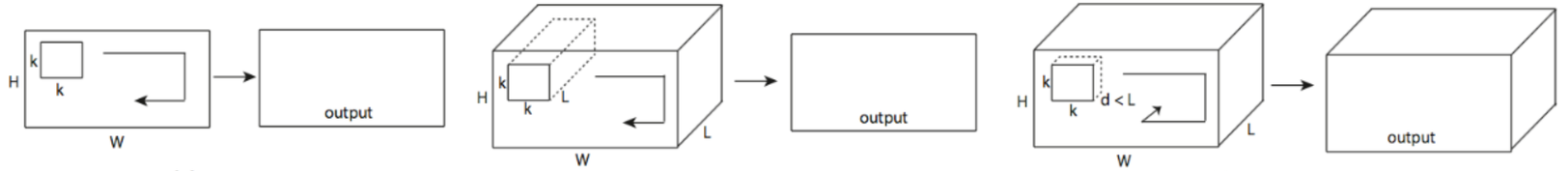
#3.2 Unimodal Feature Extraction

3D-CNN

(a) 2D convolution



(b) 3D convolution



(a) 2D convolution (b) 2D convolution on multiple frames (c) 3D convolution

◯ 2D Conv Net과 비교하여, **3D Conv Net**은 3D convolution과 3D pooling operations으로 인해 **temporal information**을 모델링 가능

◯ Convolution 필터가 모두 **3차원**

◯ 필터에 의해 생성된 feature map도 **3차원**

⇒ 연속된 프레임의 **temporal한 학습**이 가능해짐

출처: https://m.blog.naver.com/khm159/222027509486

출처: https://jay.tech.blog/2017/02/02/3d-convolutional-networks/

#3.2 Unimodal Feature Extraction

openSMILE

○ **오디오 신호의 feature**를 추출하고 음성 및 음악 신호를 분류하는 데 사용됨

○ 음성 콘텐츠를 추출하는 자동 음성 인식과 달리 주어진 음성
또는 음악 세그먼트의 특성을 인식

⇒ 화자의 감정 ,나이, 성별 및 성격, 우울증과 같은 특성

#3.3 Our Model

대화에서 감정 감지의 주요 3요소

○ **화자 :** 각 party는 말할 때마다 변하는 party state를 사용해서 모델링됨

⇒발화 내에서 감정의 변화 추적

○ **이전 발화의 문맥:** global state를 사용해서 모델링됨

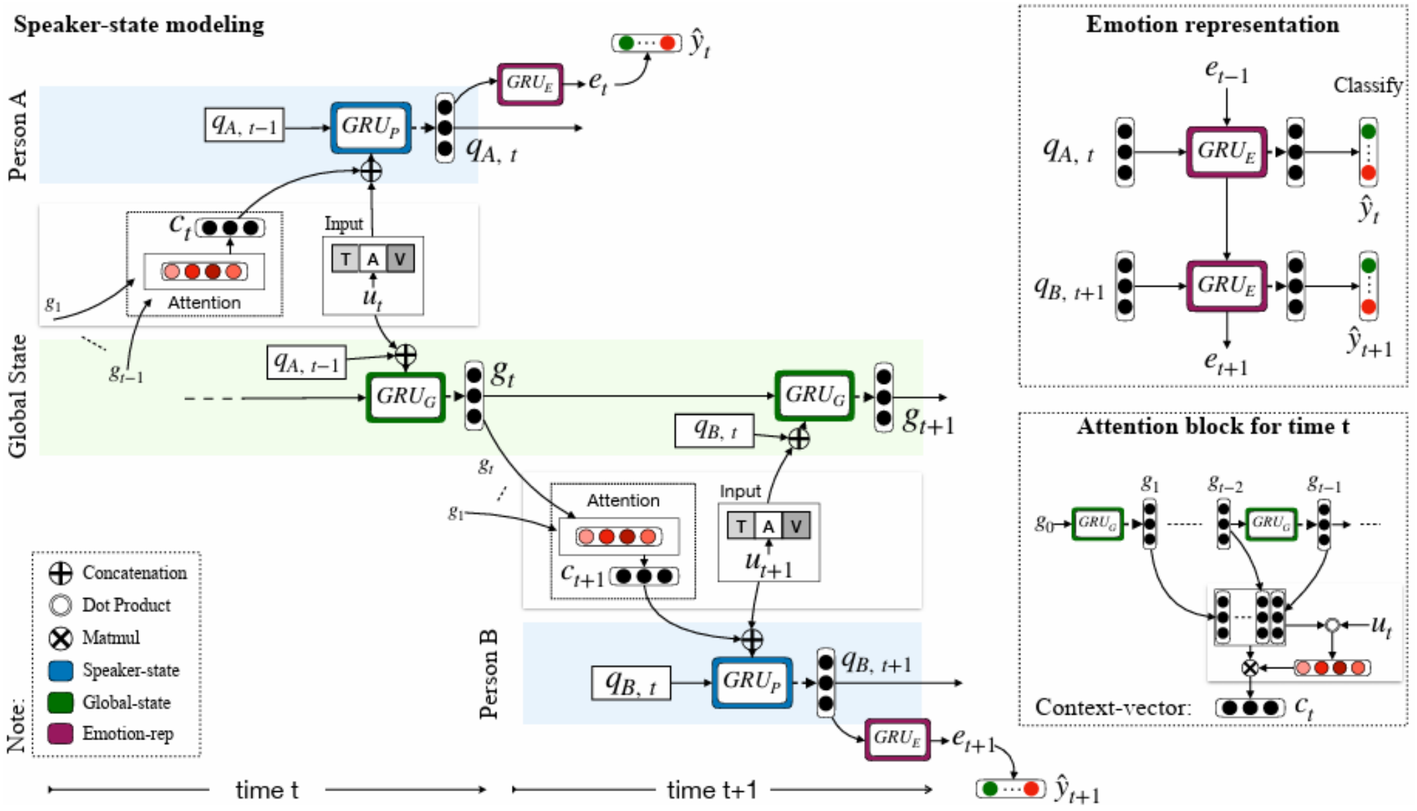
⇒ global state: 이전 발화와 party state가 함께 인코딩 되어 맥락 표현에 사용됨

○ **이전 발화의 숨겨진 감정:** 현재 party state와 이전 화자의 state를 문맥으로 사용하여 감정 표현을 추론

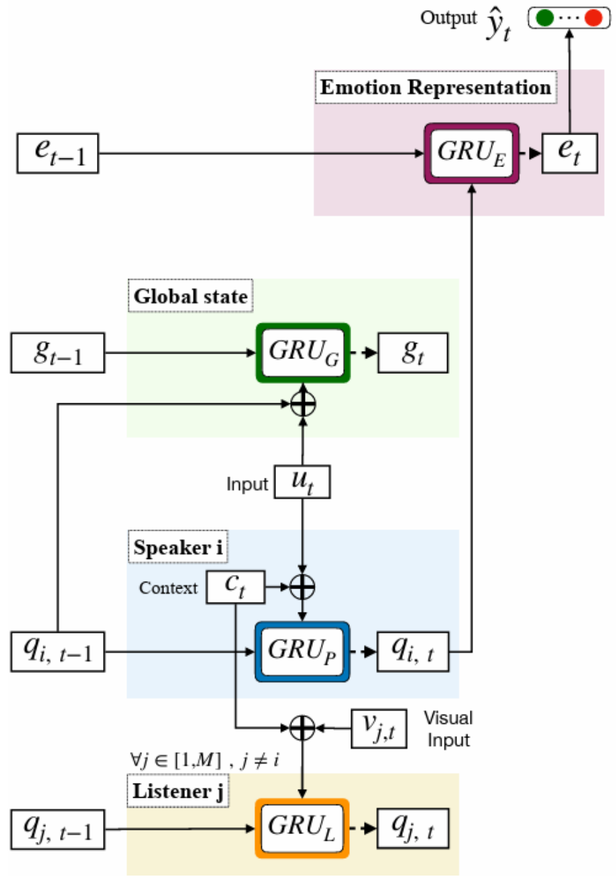
⇒ 이 감정 표현은 최종 감정 분류에 사용됨

#3.3 Our Model

전체적인 Dialogue RNN 구조



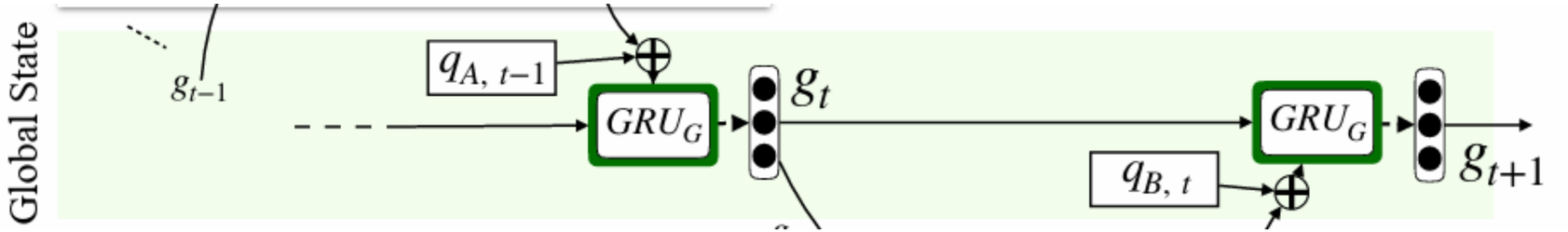
업데이트 체계도



GRU 셀을 통해 state와 representation 업데이트 진행

#3.3 Our Model

Global State (Global GRU)



○ 발화와 화자의 상태를 공동으로 인코딩하여 발화의 문맥 파악

⇒ 화자나 발화 간의 종속성이 생겨 향상된 문맥 표현을 얻을 수 있음

○ 현재 발화 Ut는 화자의 상태를 qs(ut) t−1 → qs(ut) t로 변화시킴

⇒ 이런 변화를 GRU 셀을 통해 포착함

#3.3 Our Model

Party State (Party GRU)

○ 고정된 사이즈의 벡터를 사용해서 개별 화자의 상태를 추적함

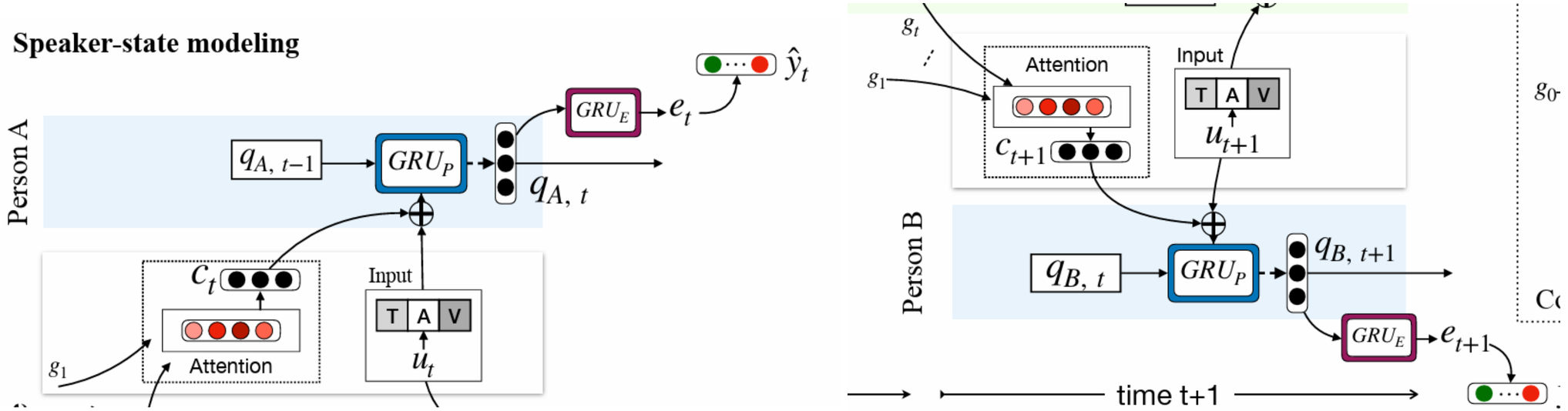
⇒ 주요 목적은 각 발화의 화자를 인식 후 처리하는 것

○ 현재 들어오는 발화 Ut에서 화자인지 청취자인지 참가자의 역할을 업데이트함

⇒ 이는 감정 분류에 영향을 미침

#3.3 Our Model

Speaker Update (Speaker GRU)



○ 화자는 이전 대화의 문맥을 바탕으로 응답함

⇒ 다음 순서와 같이 문맥 파악함

#3.3 Our Model

Speaker Update (Speaker GRU)

$$\alpha = \text{softmax}(u_t^T W_\alpha [g_1, g_2, \ldots, g_{t-1}])$$

- g: 이전 발화를 대표하는 global state
- α: attention score로, Ut와 감정적으로 관련성 높을수록 값이 커짐

$$c_t = \alpha [g_1, g_2, \ldots, g_{t-1}]^T$$

- 문맥 벡터 Ct는 이전 global state를 α 로 pooling함으로써 계산됨

$$q_{s(u_t),t} = GRU_{\mathcal{P}}(q_{s(u_t),t-1}, (u_t \oplus c_t))$$

- Ut와 GRU 셀을 사용해서 현재 화자 상태를 새로운 화자상태로 업데이트

  qs (U_t-1) ⇒ qs (U_t)

#3.3 Our Model

Listener Update

○ 화자의 발화로 인한 청취자의 상태 변화를 모델링

1. 청취자의 상태가 **변하지 않는** 메커니즘

$$\forall i \neq s(u_t), q_{i,t} = q_{i,t-1}.$$

2. 청취자의 시각적 단서를 통해 청취자 상태 업데이트하는 메커니즘

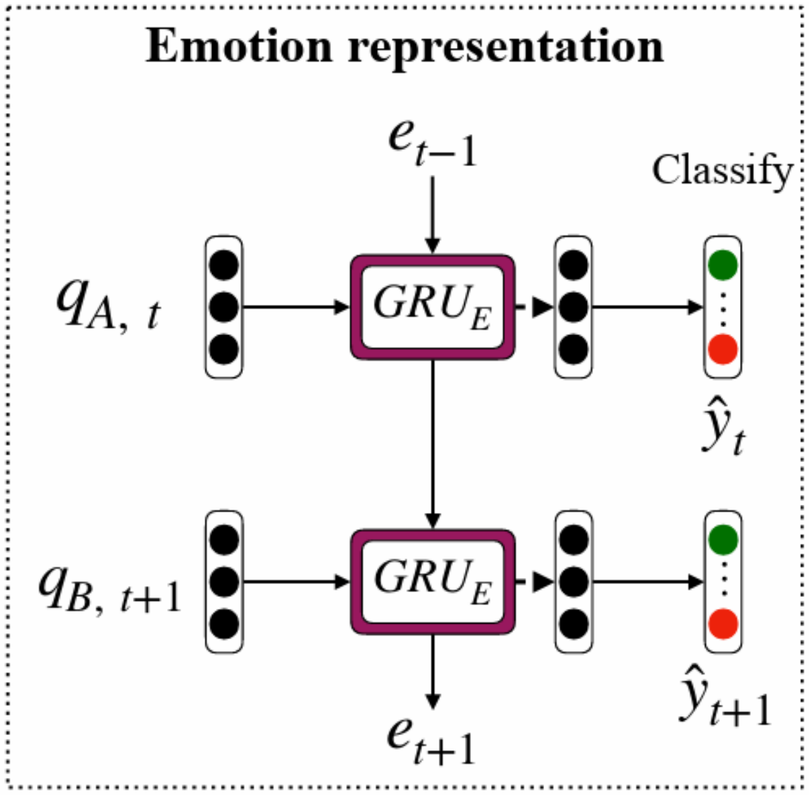
$$\forall i \neq s(u_t), q_{i,t} = GRU_{\mathcal{L}}(q_{i,t-1}, (v_{i,t} \oplus c_t))$$

⇒ 발화하면 상태 qi를 모든 이전 발화와 관련된 정보를 담고 있는 Ct를 통해 업데이트

#3.3 Our Model

Emotion Representation (Emotion GRU)

$$e_t = GRU_{\mathcal{E}}(e_{t-1}, q_{s(u_t),t}),$$



○ 화자 상태 qs(ut),t와 이전 발화의 e(t-1)로부터 감정 표현 et를 추론

○ e(t-1)은 상대방의 파인 튜닝된 감정 문맥 정보를 et에 보낸다

→ 화자 상태와 상대방 상태의 연결을 설정

⇒ party GRU와 global GRU는 인코더와 비슷한 역할

⇒ emotion GRU는 디코더와 비슷한 역할

#3.3 Our Model

Emotion Classification

$$
\begin{aligned}
l_t &= \mathrm{ReLU}(W_l e_t + b_l), \\
\mathcal{P}_t &= \mathrm{softmax}(W_{smax} l_t + b_{smax}), \\
\hat{y}_t &= \underset{i}{\mathrm{argmax}}(\mathcal{P}_t[i]),
\end{aligned}
$$

○ 2 계층의 퍼셉트론으로 구성됨

○ 마지막 Softmax layer는 6개의 감정 클래스의 확률을 계산

#3.3 Our Model

Training

$$L = -\frac{1}{\sum_{s=1}^{N} c(s)} \sum_{i=1}^{N} \sum_{j=1}^{c(i)} \log \mathcal{P}_{i,j}[y_{i,j}] + \lambda \|\theta\|_2 ,$$

○ loss를 측정하기 위해 L2 정규화를 **categorical cross-entropy**로 진행함

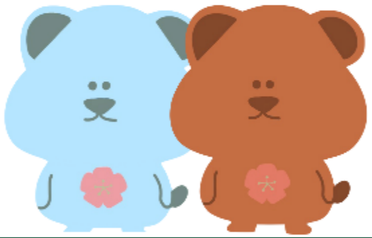
⇒ multi-class 에서 하나의 클래스로 분류

N: 대화의 수
c(i): 발화의 수
Pi,j: 대화 i의 발화 j의 감정 레이블 확률
yi,j: 대화 i의 발화 j의 예상 클래스 레이블
stochastic gradient descent 기반 Adam optimizer 사용

#3.4 Dialogue RNN Variants (변형)

- **Dialogue RNN + Listener State Update:** 발화자의 상태 qs(ut) 를 기반으로 청취자의 상태 업데이트
- **Bidirectional Dialogue RNN:** 두 RNN이 입력 시퀀스에서의 forward와 backward pass를 통해 대화의 과거, 미래 정보가 포함되어 감정 분류를 위해 더 좋은 문맥 제공함
- **Dialogue RNN + attention:** 각 et에 대해 attention이 적용되어 이는 과거와 미래의 발화에 관련성 제공
- **Bidirectional Dialogue RNN + Emotional attention:** 대화의 모든 감정 표현을 통해 대화의 다른 발화에서 문맥 파악

4. Experiments



EWHA EURON

#4.1 Datasets

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

8:2 비율로 train set과 test set을 나눔

◯ **IEMOCAP:** 10명의 발화자와 쌍방향 대화 진행 & 8명만 훈련 세트에 속함.

"happy", "sad", "neutral", "angry", "excited", "frustrated" 중 하나 감정 레이블 가짐

◯ **AVEC:** 사람과 인공 지능 에이전트 간의 대화. 각 대화는 4가지 감정 속성 중 하나 가짐

"valence" (감정의 긍정성/부정성), "arousal" (감정의 활동성/정적성), "expectancy" (감정의 예측 가능성), 그리고 "power" (감정의 강도).

#4.2 Comparison with CMN

○ **CMN:** 대화 기록에서 발화 문맥을 모델링하기 위해 두 개의 서로 다른 GRU를 사용.
현재 발화를 두 개의 서로 다른 메모리 네트워크에 각각 쿼리로 공급하여 발화 표현을 얻음.

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$_l$	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**

○ **IEMOCAP:** Dialogue RNN이 CMN보다 2.77%의 정확도와 3.76%의 F1점수로 평균적으로 성능이 좋다.

○ **AVEC:** 모든 네 가지 속성에 대해 Dialogue RNN이 CMN보다 MAE가 크게 낮고 Pearson 상관 계수(r)가 더 높아 성능이 좋다.

#4.2 Comparison with CMN (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**

Dialogue RNN이 CMN보다 multimodal features를 활용하여 감정 분류 작업에서 우수한 성능을 보임.

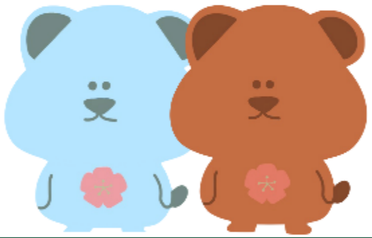
#4.3 Dialogue RNN vs. Dialogue RNN Variants

◯ **Dialogue RNNt**: 청취자 상태 업데이트는 일반적으로 Dialogue RNN보다 성능이 좋지 않지만 happy 레이블에서는 우수한 성능을 보임

◯ **Bi-Dialogue RNN:** 미래 발화에서 문맥을 파악하기 때문에 Dialogue RNN보다 성능이 향상됨. 평균적으로 두 데이터 셋에서 모두 Dialogue RNN보다 성능 좋음.

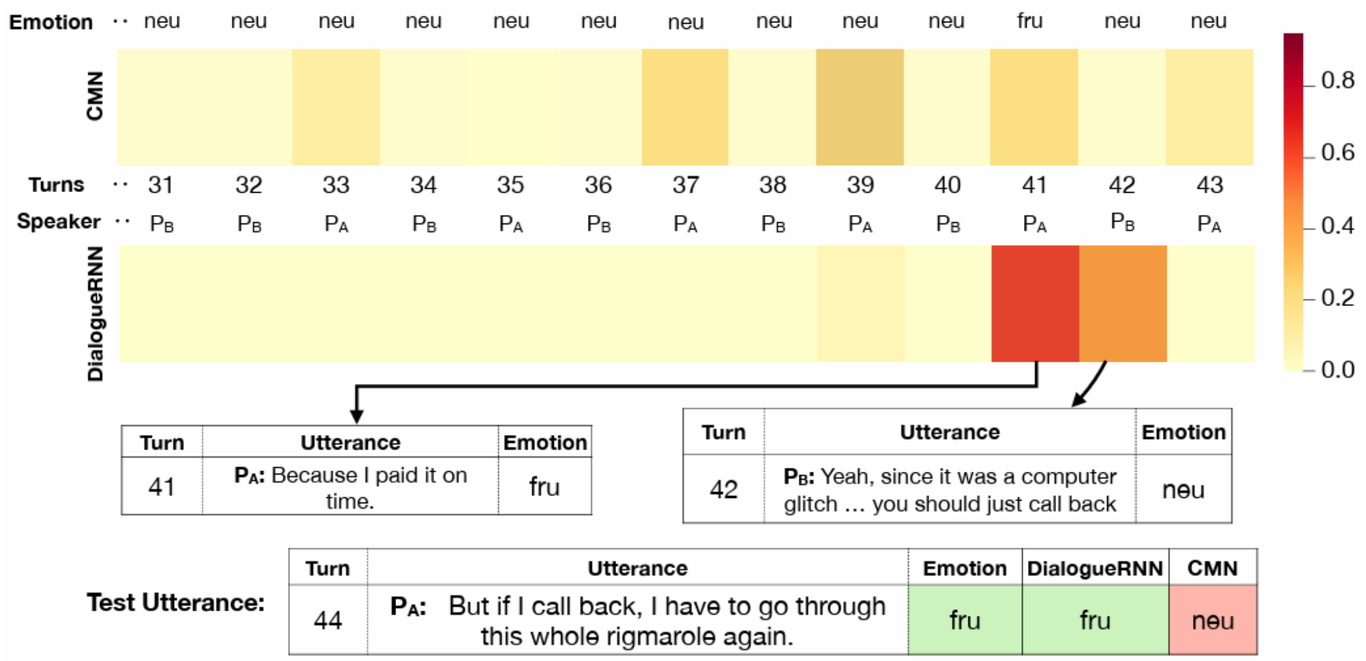
◯ **Bi-Dialogue RNN+ Att n**: Bi-Dialogue-RNN에서 감정 표현을 생성한 후, 최종 감정 표현을 생성하는 방식으로, 이는 다른 모든 방법보다 우수한 성능을 보임.

5. Discussions



EWHA EURON

#5.1 Dependency on preceding utterances (Dialogue RNN)



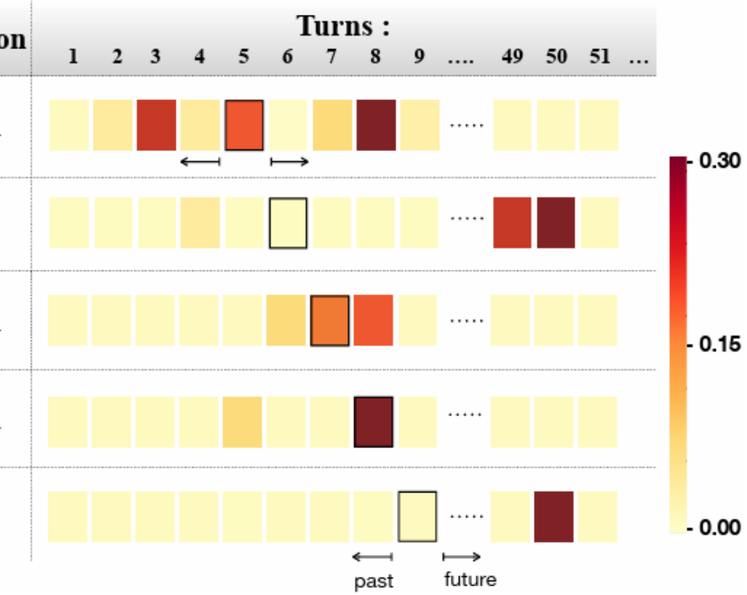
◯ Dialogue RNN의 attention이 CMN보다 더 집중되어 있음

◯ **Dialogue RNN**: PA와 PB의 41번과 42번의 발화에 집중하여 감정 변화를 정확하게 예측 (neutral → frustrated)

◯ **CMN:** 종속성을 포착하지 못하고 neutral로 잘못 예측

#5.2 Dependency on future utterances (BiDialogue RNN + Att)

Turn	Utterance P_A / P_B	Emotion Label	Prediction
5	Oh, maybe a little. Nothing serious.	neutral	neutral
6	He let him kiss you. You said you did	frustration	anger
7	Well, what of it.	neutral	neutral
8	It gave him a lot of pleasure and it didn't hurt me.	neutral	neutral
9	That's a nice point of view I must say	anger	anger

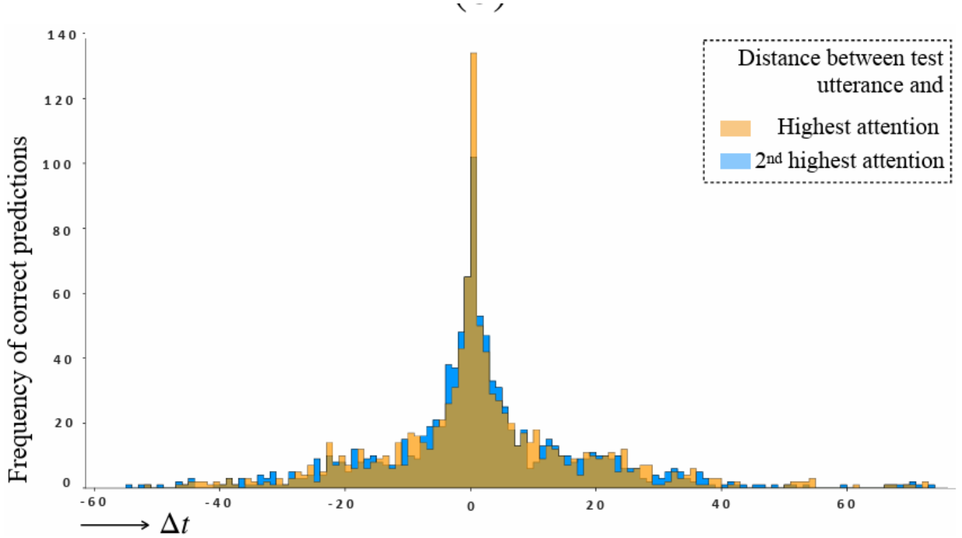


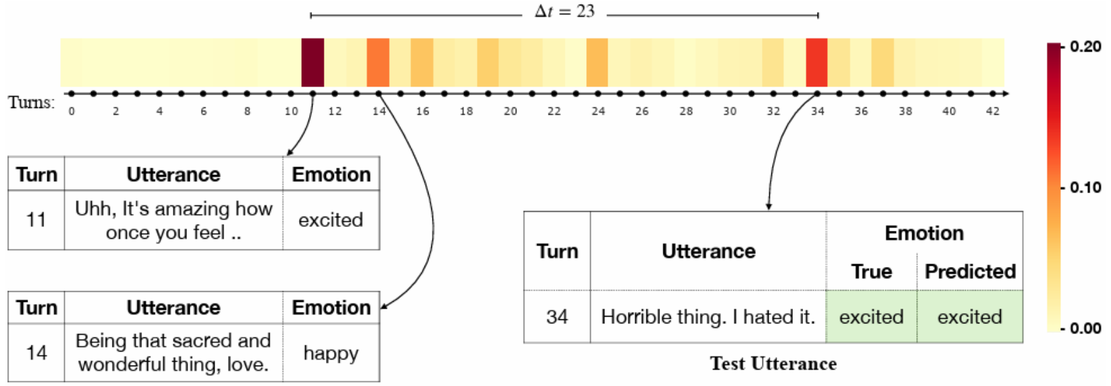
○ 대화의 세그먼트에서 감정 표현에 대한 attention을 시각화

○ 미래 발화에 대한 고려도 포함됨

⇒ 미래 발화와 과거 발화 간의 감정 상태 간의 상호 의존성을 보여줌

#5.3 Dependency on distant context





Turn	Utterance	Emotion
11	Uhh, It's amazing how once you feel ..	excited

Turn	Utterance	Emotion
14	Being that sacred and wonderful thing, love.	happy

Turn	Utterance	Emotion	
		True	Predicted
34	Horrible thing. I hated it.	excited	excited

○ local 컨텍스트 내에서 가장 높은 의존성이 있는 경향

○ 발화의 상당 부분(약 18%)은 자신과 20~40턴 떨어진 발화에 집중

⇒ **장기간의 감정적 종속성 강조**

○ **장기간의 문맥 종속성 사례**

○ 전반적으로 happy한 가운데 "Horrible thing. I hated it" 발화 등장

⇒ 명확히 하기 위해 과거 발화 (11,14)에 집중하여 excited로 알맞게 예측

#5.4 Error Analysis

◯ **관련된 감정들** 사이에서 잘못된 예측이 많음

◯ Happy – Excited

◯ Angry - Frustrated

⇒ 감정들 간의 미묘한 차이로 인해 명확한 구별이 어려움

◯ **Neutral** 클래스에 대한 많은 **False-Positive**

⇒ 다른 감정에 비해 자주 등장

⇒ 특히 이전 발화에서 감정 변화가 없는 경우에 더 많은 오류가 발생

#5.5 Ablation Study

party state와 Emotion GRU의 도입

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

◯ party state 없이 4.33% 성능 감소

⇒ party state: party의 감정과 관련된 문맥 추출에 도움을 줌

◯ Emotion GRU 없이 2.51% 성능 감소

⇒ 이전 발화의 감정 표현만으로는 상대방의 상태에 관한 문맥이 전달되지 않음

6. Conclusion

EWHA EURON

#6 Conclusion

대화에서 감정 탐지를 위한 RNN 기반의 neural architecture

CMN과 달리 발화자의 특성을 고려한 세밀한 문맥

◯ **Party state**: speaker 인식

◯ **Global state**: speaker-specific utterance representation 역할

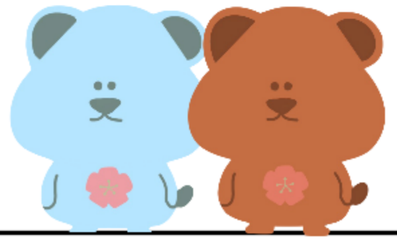
◯ **Emotion Representation**: Party state와 Global state를 통해 추론

⇒ 이를 통해 최종 감정 분류

◯ **Party GRU**를 통해 **Global GRU**의 컨텍스트와 함께 정보 인코딩

◯ **Emotion GRU**가 받은 정보를 바탕으로 감정 분류 수행

THANK YOU



EWHA EURON