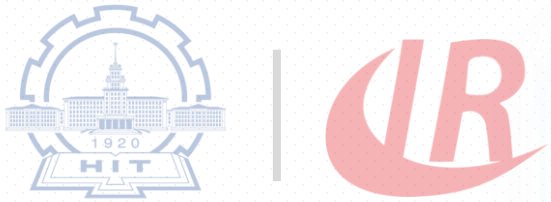


基于迭代情绪交互网络的对话情绪识别

汇报人：陆鑫

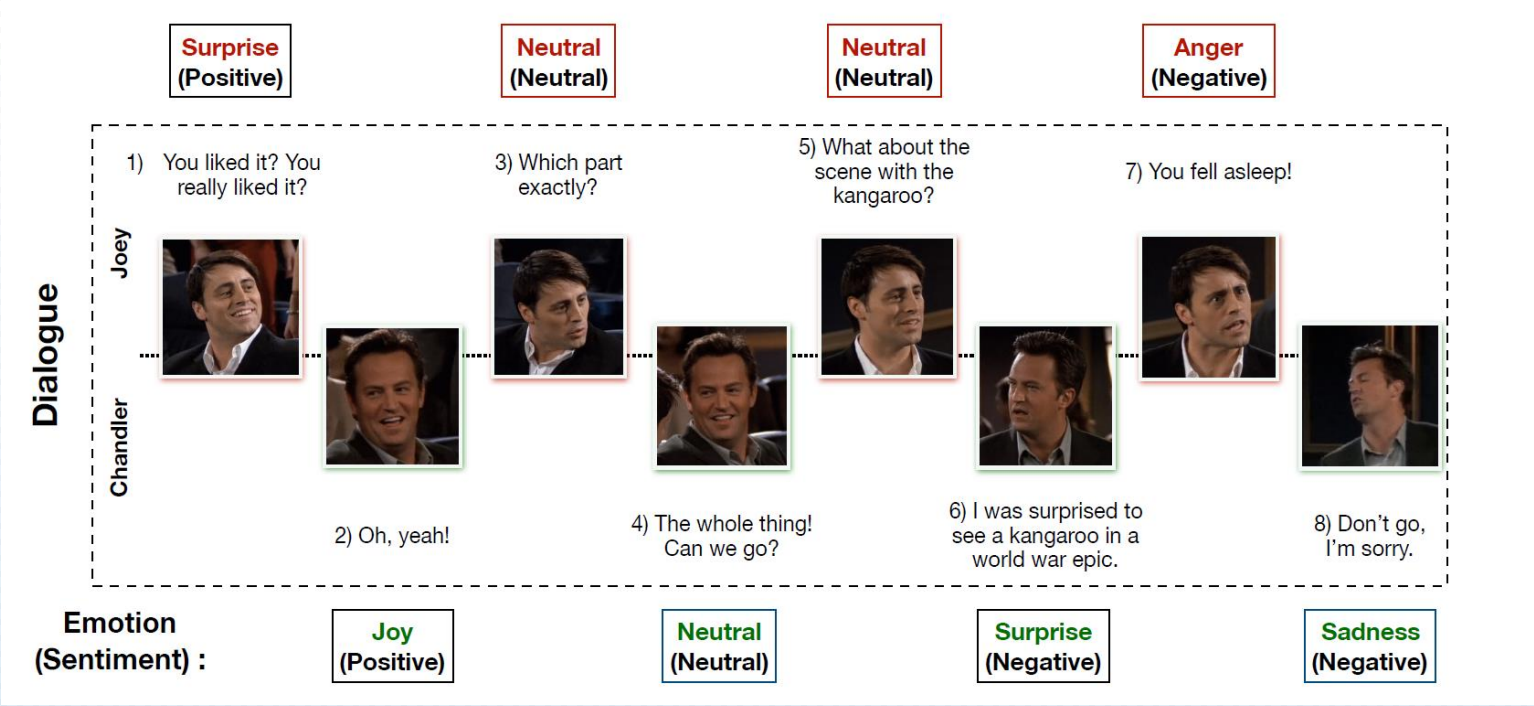
2020年04月09日

对话情绪识别

给定一段对话的上下文，识别对话中所有话语的情绪

对话情绪识别

给定一段对话的上下文，识别对话中所有话语的情绪



MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. Soujanya Poria et al. ACL 2019.

应用场景

客户服务质检

检测用户负面情绪和客服服务态度

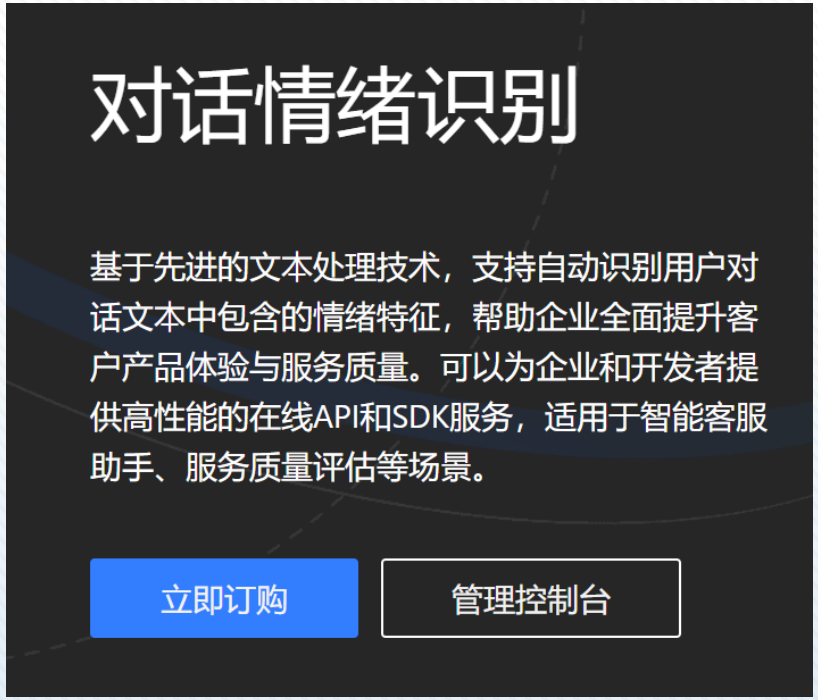
情感聊天机器人

具有情感抚慰能力的情感聊天机器人

相关AI云服务

百度AI开放平台

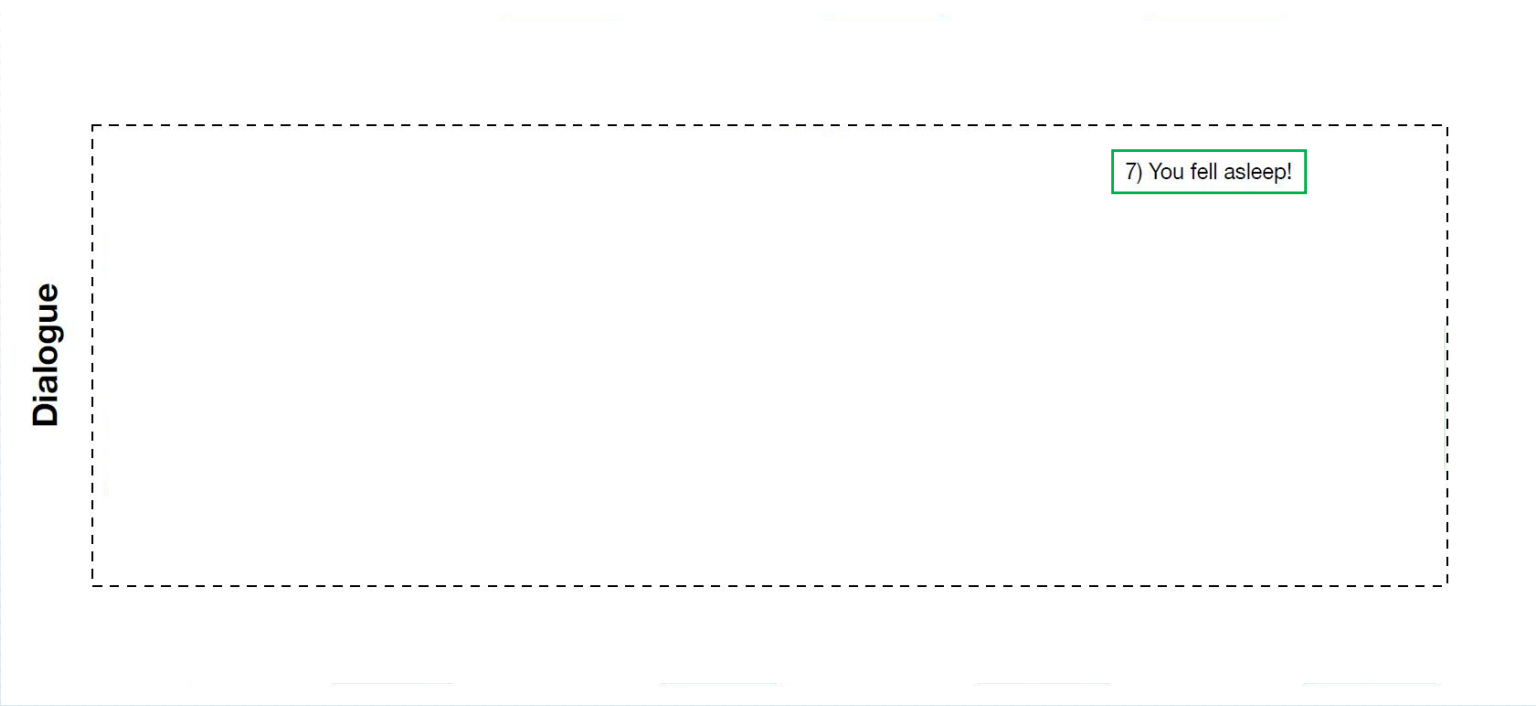
移动云服务



关键问题

- 对话情绪识别 ≠ 句子级情绪识别

  对话中存在很多单句没有的要素和现象，相关工作通常围绕这些展开

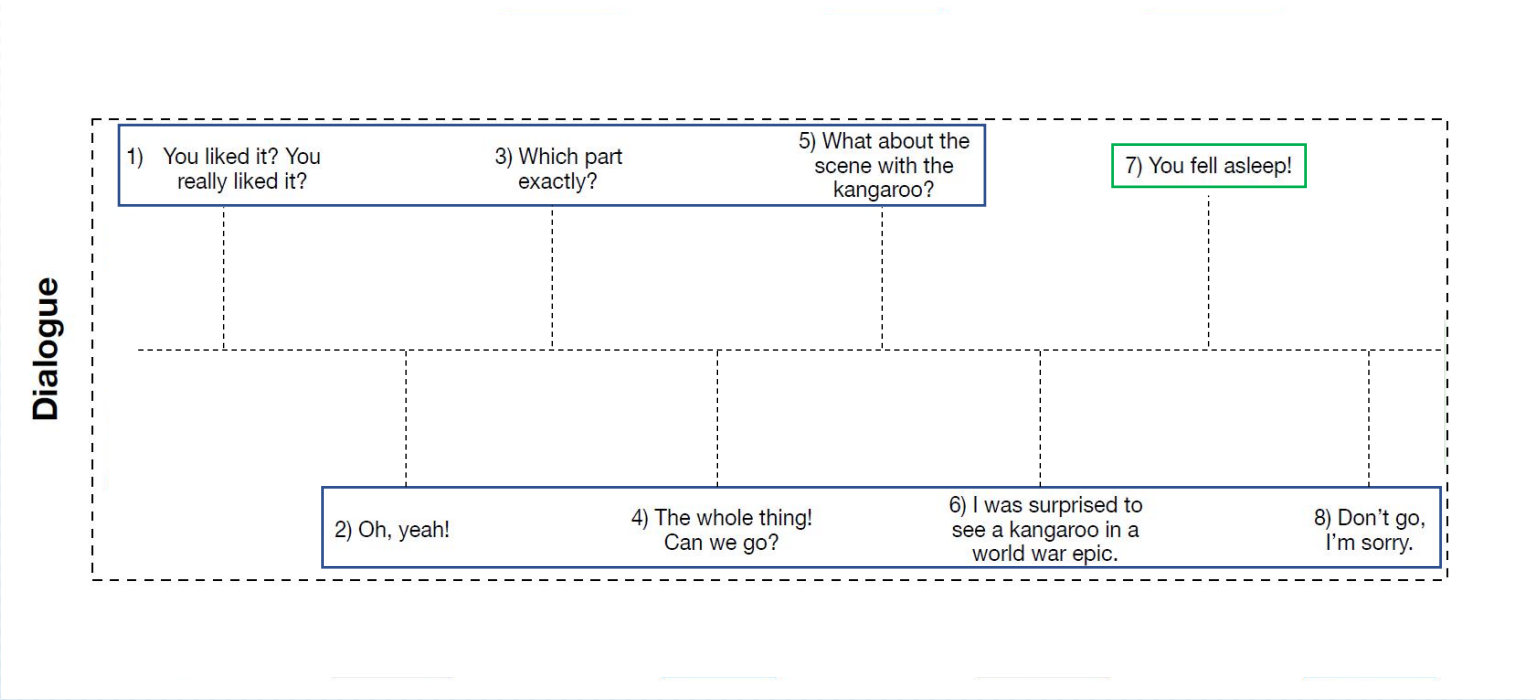


MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. Soujanya Poria et al. ACL 2019.

关键问题

- 对话情绪识别 ≠ 句子级情绪识别

  对话中存在很多单句没有的要素和现象，相关工作通常围绕这些展开

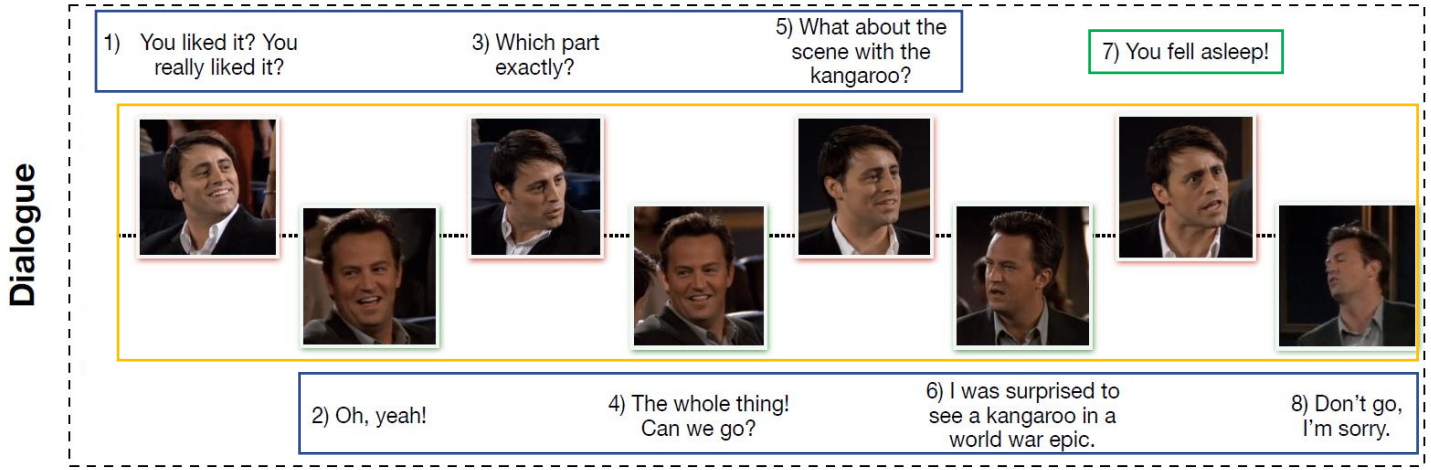


MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. Soujanya Poria et al. ACL 2019.

关键问题

- 对话情绪识别 ≠ 句子级情绪识别

  对话中存在很多单句没有的要素和现象，相关工作通常围绕这些展开

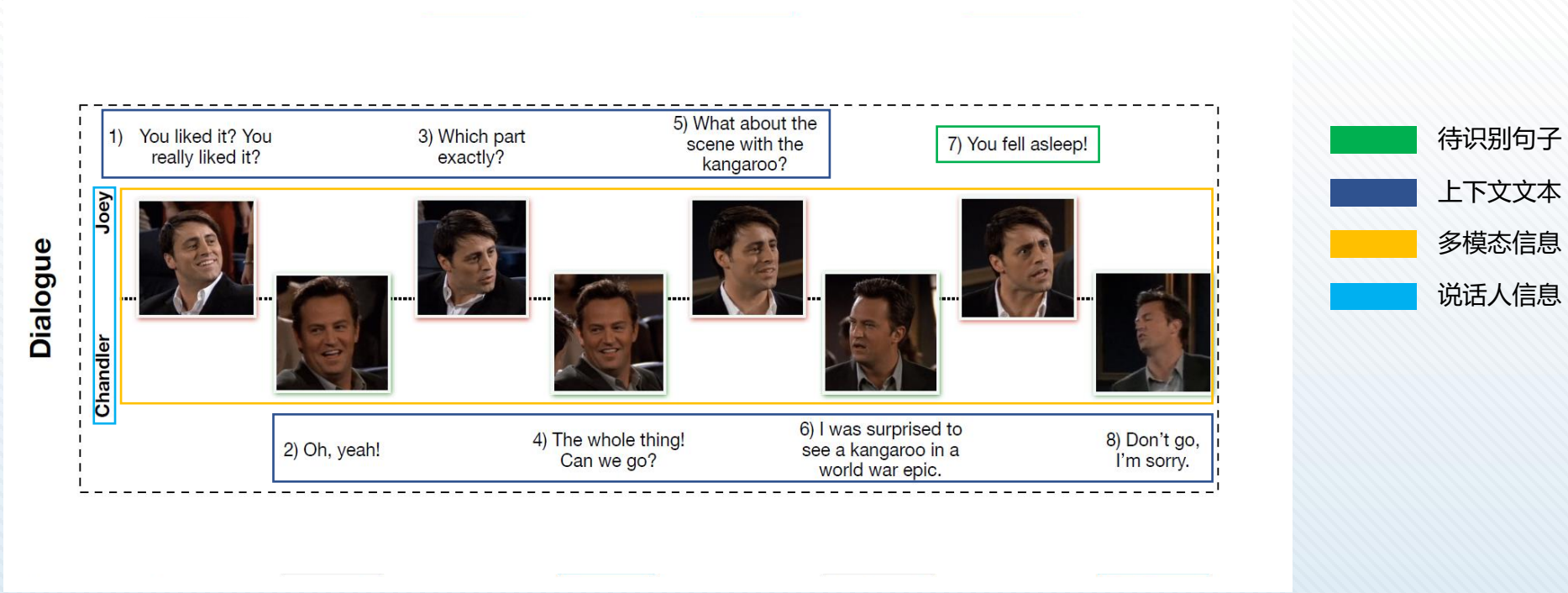


MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. Soujanya Poria et al. ACL 2019.

关键问题

- 对话情绪识别 ≠ 句子级情绪识别

  对话中存在很多单句没有的要素和现象，相关工作通常围绕这些展开

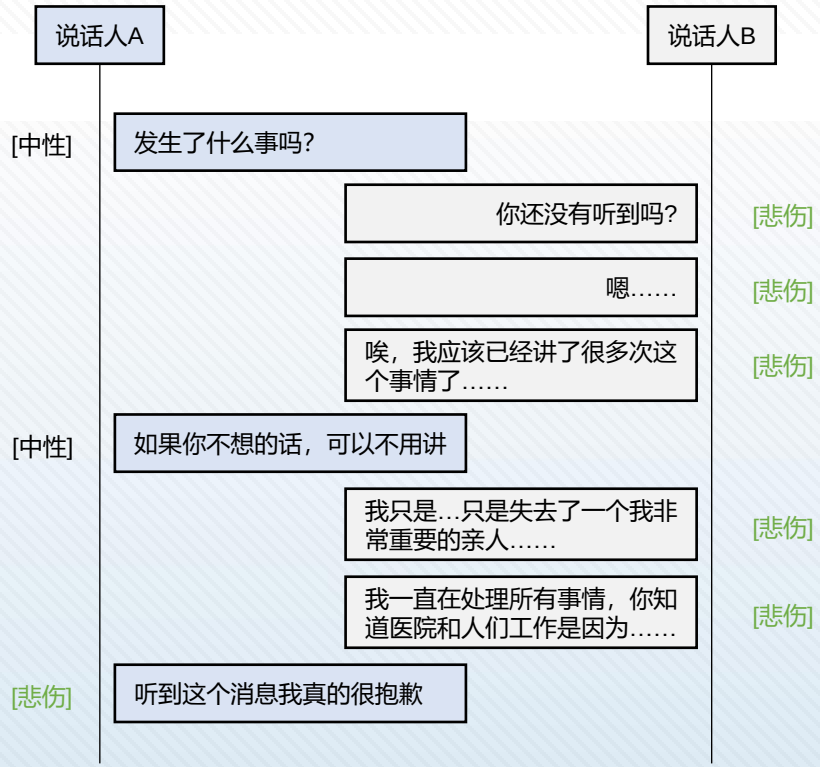


MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. Soujanya Poria et al. ACL 2019.

工作动机

对话中存在情绪交互

对话中话语的情绪之间是存在相互影响的，建模情绪交互对情绪识别有益

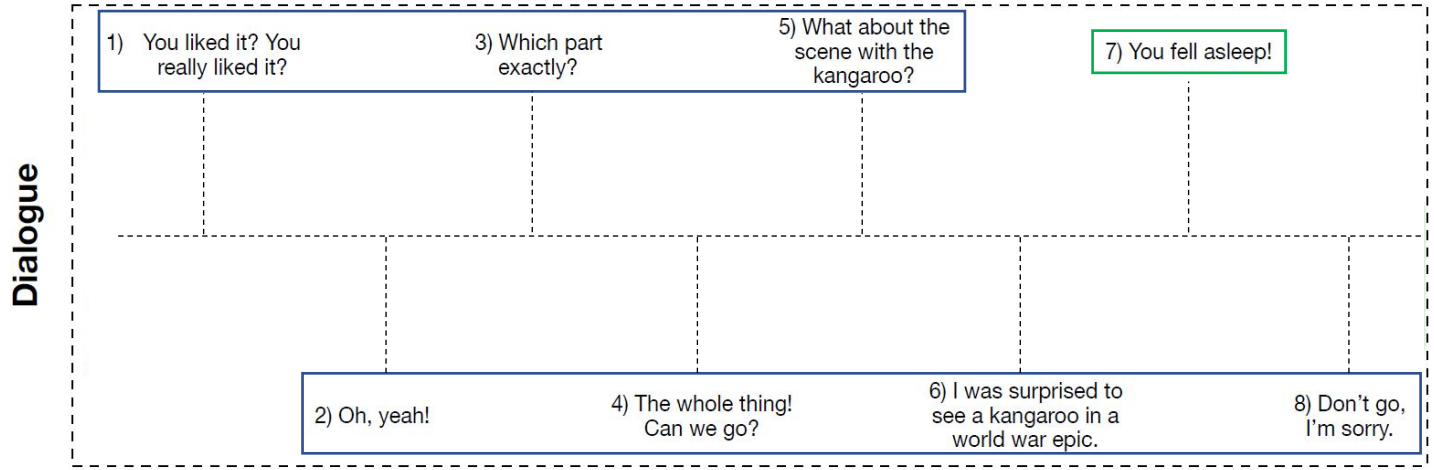


IEMOCAP: interactive emotional dyadic motion capture database. Carlos Busso et al. Language Resources and Evaluation.

工作动机

已有工作做法

已有工作通常仅建模对话上下文，对情绪交互进行隐式建模

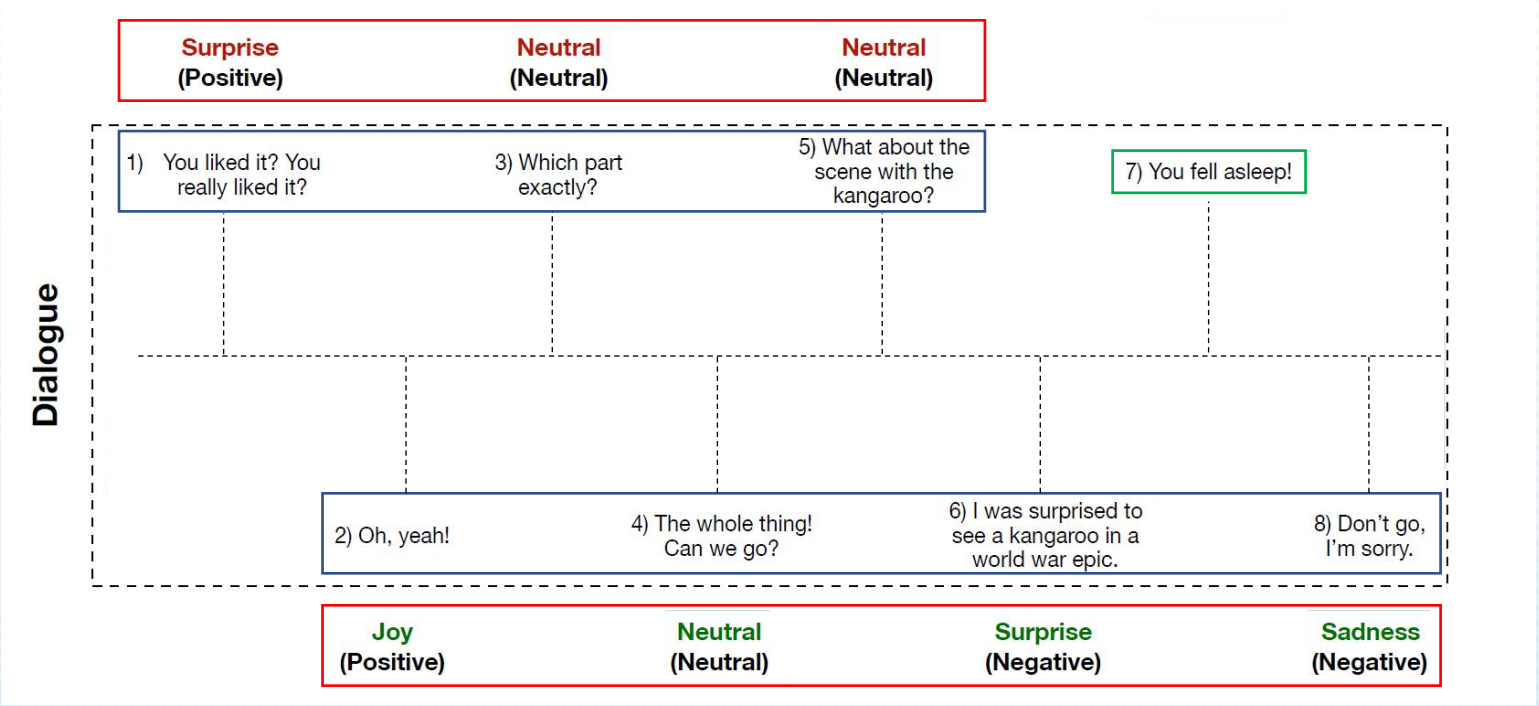


MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. Soujanya Poria et al. ACL 2019.

工作动机

我们工作做法

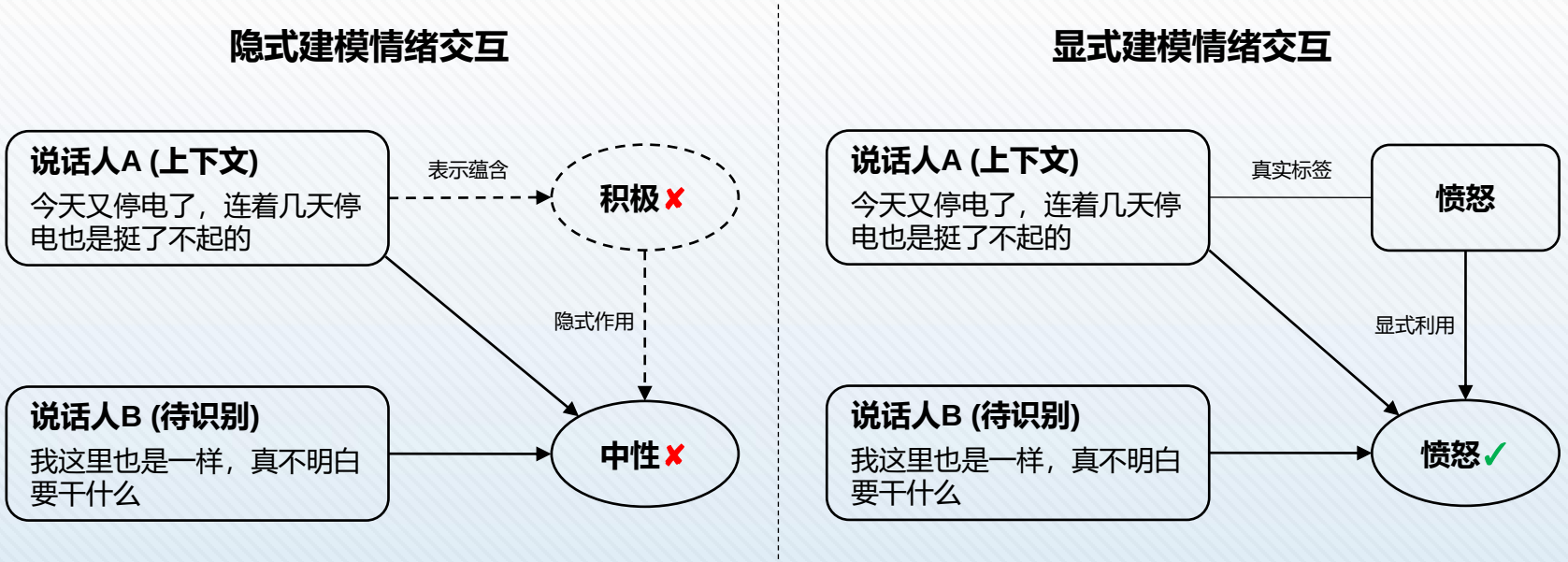
我们考虑引入话语的情绪标签，可以提供显式和更准确的情绪交互



MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. Soujanya Poria et al. ACL 2019.

隐式建模情绪交互 vs 显式建模情绪交互

- 隐式建模常被语言中的复杂表达所干扰，导致情绪交互变得不可靠
- 显式建模则利用了上下文精确的情绪信息，使得话语的情绪判断不易受到干扰



工作动机

隐式建模情绪交互 vs 显式建模情绪交互

- 隐式建模常被语言中的复杂表达所干扰，导致情绪交互变得不可靠
- 显式建模则利用了上下文精确的情绪信息，使得话语的情绪判断不易受到干扰
- 显式建模存在的困难：
  - 情绪标签仅能在训练阶段获得，在测试阶段是不可能事先得到并作为输入的

工作动机

隐式建模情绪交互 vs 显式建模情绪交互

- 隐式建模常被语言中的复杂表达所干扰，导致情绪交互变得不可靠
- 显式建模则利用了上下文精确的情绪信息，使得话语的情绪判断不易受到干扰
- 显式建模存在的困难：
  - 情绪标签仅能在训练阶段获得，在测试阶段是不可能事先得到并作为输入的
- 显式建模的可行方案：
  - 我们放宽了对情绪标签完全准确的要求，假设存在部分噪声的情绪标签也可以使情绪识别受益
  - 我们认为情绪标签精度的不断提升也可以使情绪识别的性能不断增强，并为此设计了迭代提升的模型结构
  - 我们在后面的分析实验中也证实了这个假设的合理性

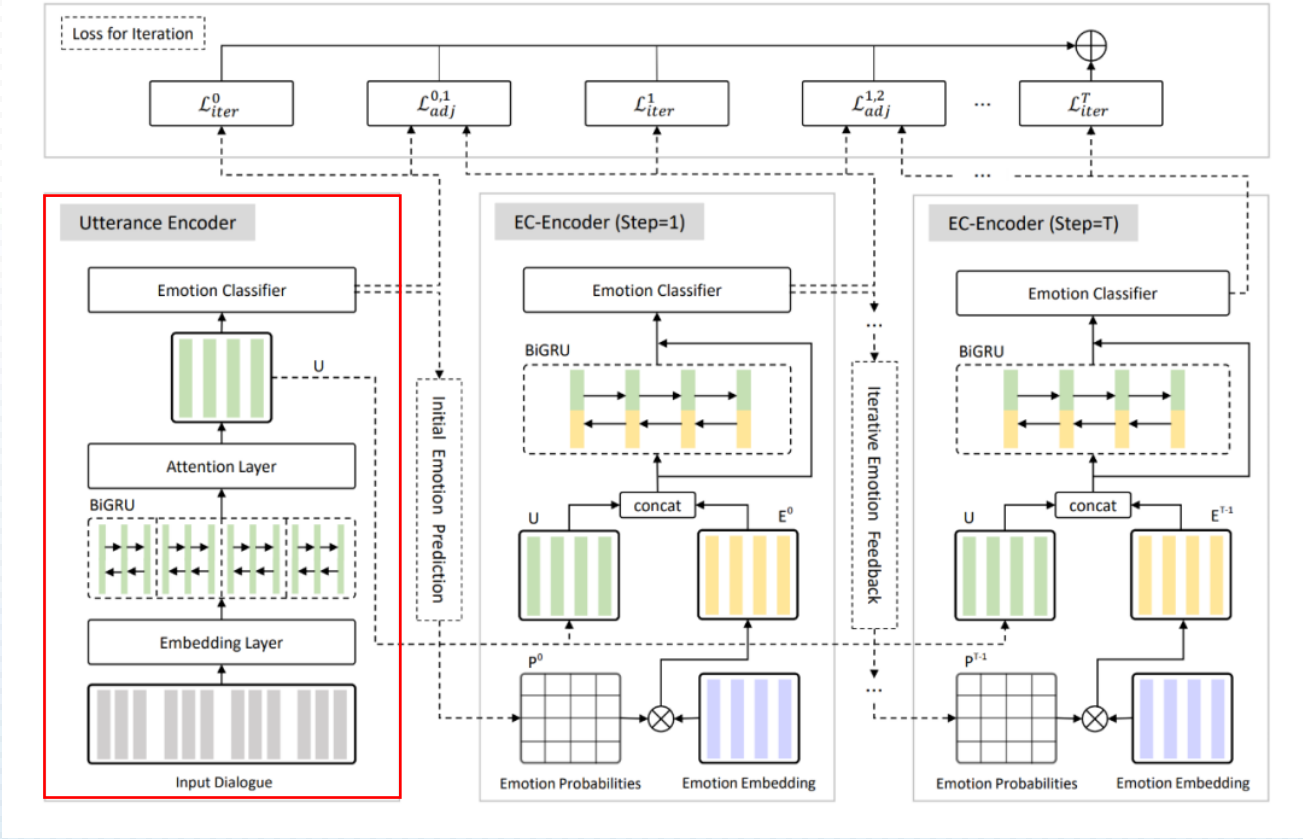
迭代情绪交互网络

话语级别编码器

$$\mathbf{h}_i = [\overrightarrow{\mathbf{h}_i}; \overleftarrow{\mathbf{h}_i}]$$

$$\alpha_i = \frac{\exp(\mathbf{h}_i^\top \mathbf{W}_u)}{\sum_j \exp(\mathbf{h}_j^\top \mathbf{W}_u)}$$

$$\mathbf{u} = \sum_{i=1}^{M} \alpha_i \mathbf{h}_i$$



迭代情绪交互网络

- 话语级别编码器
- 情绪交互上下文编码器

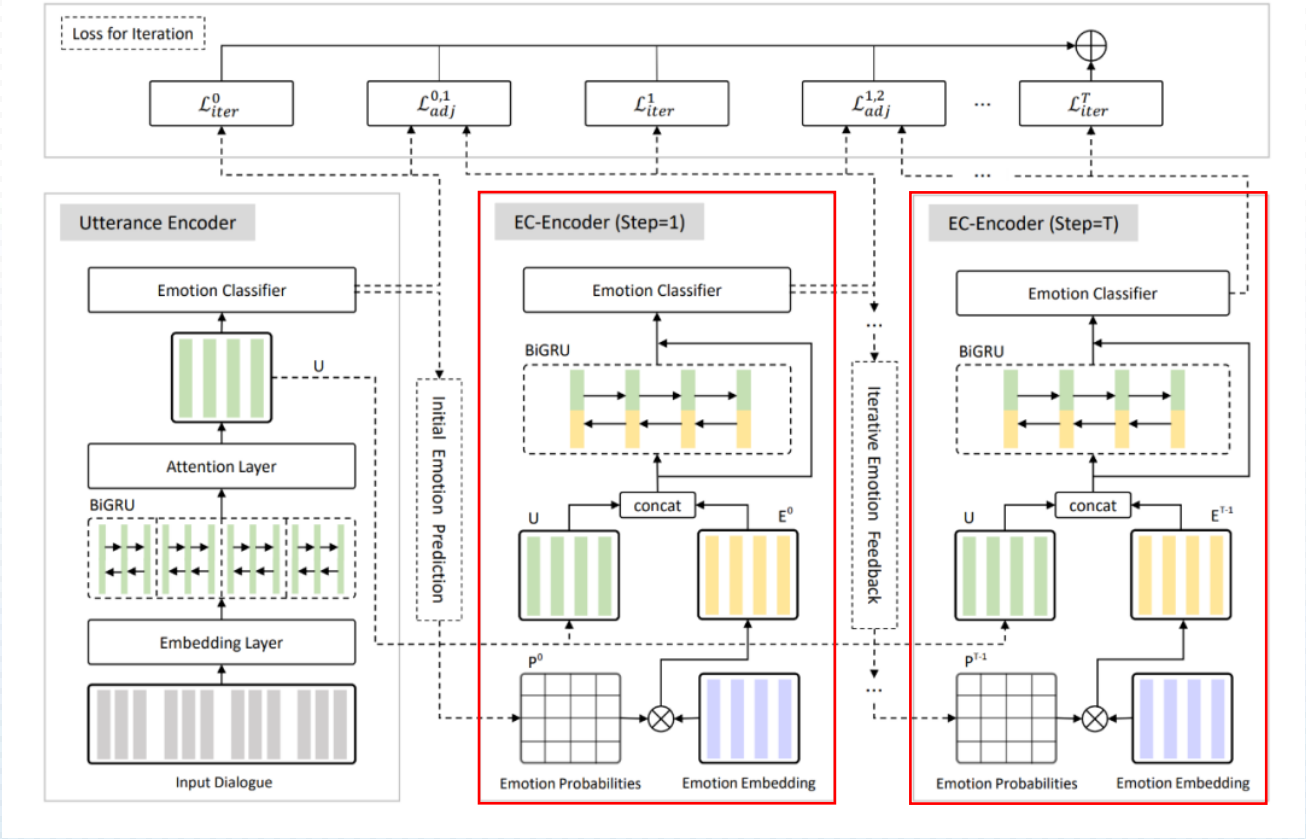
$$\mathbf{e}_i = \sum_{j=1}^{|L|} p_i^j \mathbf{x}_j$$

$$\overrightarrow{\mathbf{h}_i} = \overrightarrow{\mathbf{GRU}}([\mathbf{u}_i; \mathbf{e}_i], \overrightarrow{\mathbf{h}_{i-1}})$$

$$\overleftarrow{\mathbf{h}_i} = \overleftarrow{\mathbf{GRU}}([\mathbf{u}_i; \mathbf{e}_i], \overleftarrow{\mathbf{h}_{i+1}})$$

$$\mathbf{h}_i = \overrightarrow{\mathbf{h}_i} + \overleftarrow{\mathbf{h}_i} + [\mathbf{u}_i; \mathbf{e}_i]$$

$$\mathbf{p}_i' = \mathrm{softmax}(\mathbf{W}_e \mathbf{h}_i + \mathbf{b}_e)$$



迭代情绪交互网络

- 话语级别编码器
- 情绪交互上下文编码器
- 迭代提升机制

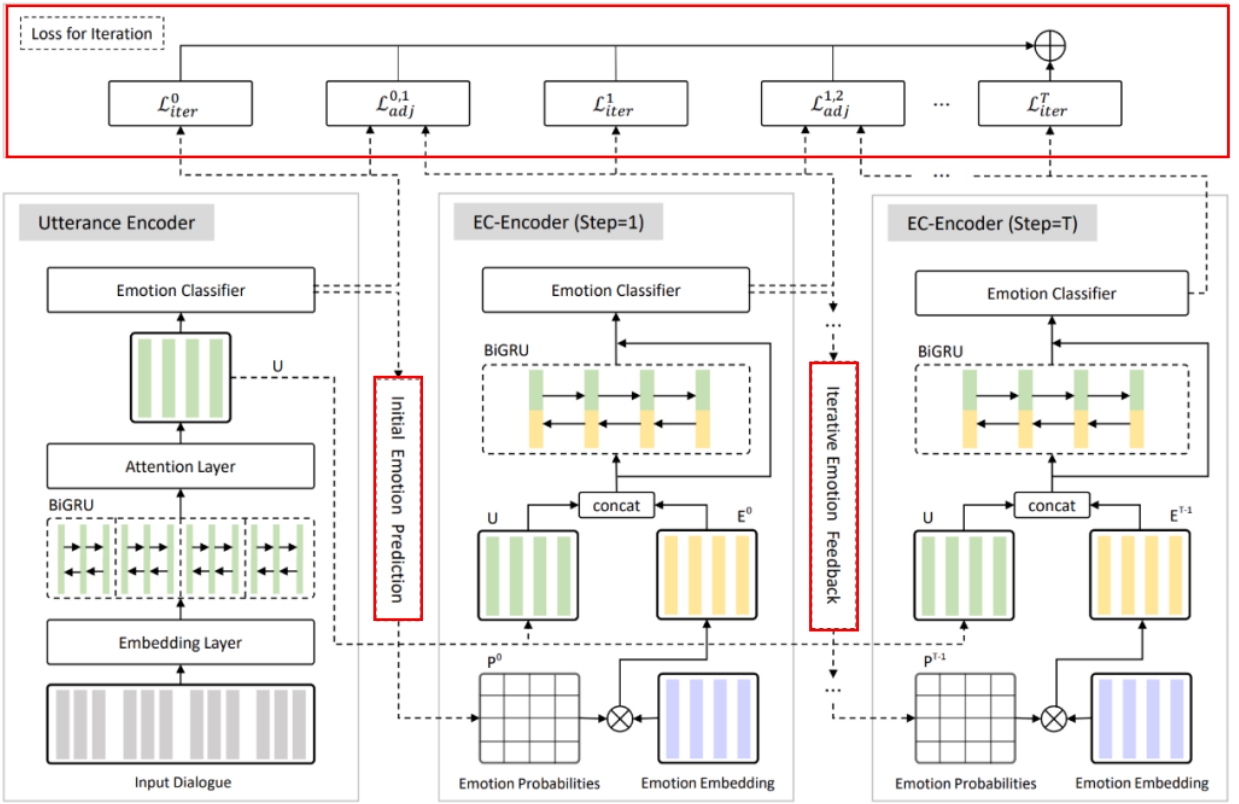
$$\mathbf{p}_i^0 = \text{softmax}(\mathbf{W}_p\mathbf{u}_i + \mathbf{b}_p)$$

$$\mathbf{P}^i = \text{EC-Encoder}(\mathbf{P}^{i-1}, \mathbf{U})$$

$$\mathcal{L}_{iter}^i = -\frac{1}{N_a}\sum_{j=1}^{N_a}\sum_{k=1}^{|L|} y_{j,k}\log(p_{j,k}^i)$$

$$\mathcal{L}_{adj}^{i,\,i+1} = \frac{1}{N_a}\sum_{j=1}^{N_a}\sum_{k=1}^{|L|} y_{j,k}\max(0, p_{j,k}^i - p_{j,k}^{i+1})$$

$$\mathcal{L} = \frac{1}{T+1}\sum_{i=0}^{T}\mathcal{L}_{iter}^i + \lambda * \frac{1}{T}\sum_{i=0}^{T-1}\mathcal{L}_{adj}^{i,\,i+1}$$



实验设置

实验数据集

- IEMOCAP：包含 152段对话，7,433个话语，标注了6个情绪类别
- MELD：包含 1,433段对话，13,708个话语，标注了7个情绪类别

对比模型

- CNN：句子级别TextCNN模型
- C-LSTM：Soujanya Poria等人发表在ACL 2017上的工作
- DialogueRNN：Navonil Majumder等人发表在AAAI 2019上的工作
- DialogueGCN：Deepanway Ghosal等人发表在EMNLP 2019上的工作

评价指标

- Weight-F1：加权平均F1-Score

IEMOCAP数据集结果

- 我们方法取得最好的结果，并在多数小类上结果最优

Model	Joy	Sadness	Neutral	Anger	Excited	Frustrate	w-Avg.
CNN	32.91	50.41	52.33	55.24	46.84	54.51	50.15
C-LSTM	30.66	69.86	55.15	58.52	55.93	60.74	57.01
C-LSTM + CRF	35.71	69.59	56.43	62.44	50.34	60.23	56.98
DialogueRNN	38.74	76.08	58.26	**63.10**	68.75	60.37	62.15
DialogueGCN	51.87	76.76	56.76	62.26	**72.71**	58.04	63.16
Our Approach	**53.17**	**77.19**	**61.31**	61.45	69.23	**60.92**	**64.37**

MELD数据集结果

- 我们方法取得最好的结果，并在多数小类上结果最优

Model	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	w-Avg.
CNN	77.24	50.54	0.32	22.28	54.19	2.86	42.88	58.48
C-LSTM	76.47	50.17	0.92	**26.51**	55.62	9.65	46.77	59.33
C-LSTM + CRF	76.42	50.22	1.48	26.29	55.58	8.51	46.96	59.29
DialogueRNN	76.23	49.59	0.00	26.33	54.55	0.81	46.76	58.73
DialogueGCN	76.02	46.37	0.98	24.32	53.62	1.22	43.03	57.52
Our Approach	**77.52**	**53.65**	**3.31**	23.62	**56.63**	**19.38**	**48.88**	**60.72**

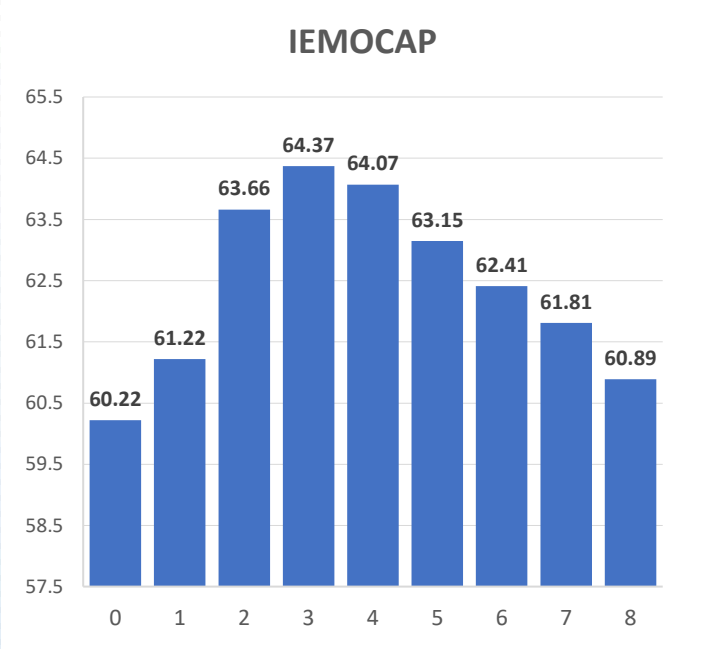
情绪交互有效性分析

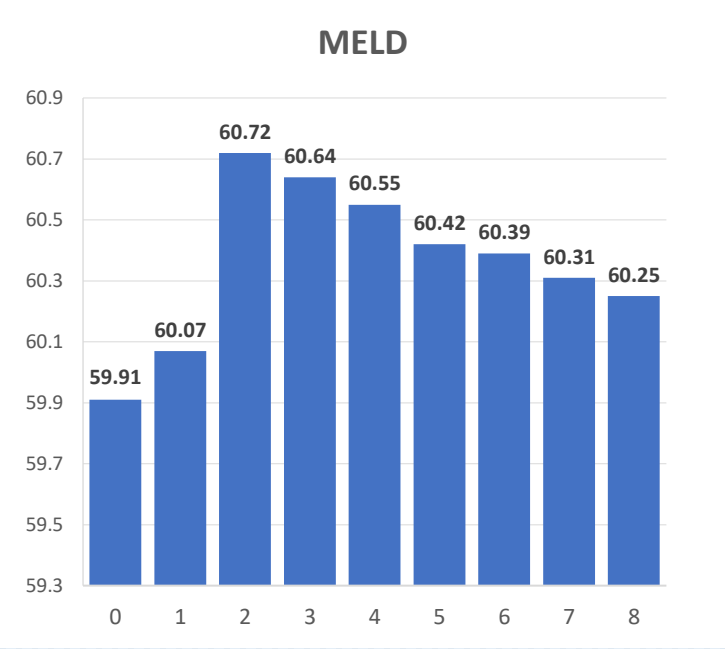
- No Label代表仅建模上下文的隐式建模方法，这种情况性能最差
- Gold Label代表输入真实标签的显式建模方法，这种情况性能最好
- 我们方法的性能介于两者之间，说明有效保留了显式建模方法的性能优势

Model	IEMOCAP	MELD
No Label	60.22	59.91
Our Approach, iter=1	61.22	60.07
Our Approach, iter=2	63.66	**60.72**
Our Approach, iter=3	**64.37**	60.64
Gold Label	66.75	62.28

最大迭代轮数影响分析

- 随着最大迭代轮数的增加，性能有一个先上升后下降的趋势





迭代修正行为分析

- 迭代过程内部的指标越来越好，说明迭代轮次之间的修正现象存在
- 由错改对的修改占多数，说明确实是在进行有意义的标签修正

Model	Step	wF1	Step -> Step	R -> W	W -> R	W -> W
IEMOCAP (iter=3)	Step 1	61.97	Step 1 -> Step 2	27.84%	**46.25%**	25.91%
	Step 2	63.71				
	Step 3	**64.37**	Step 2 -> Step 3	27.81%	**47.78%**	24.41%
MELD (iter=2)	Step 1	60.45	Step 1 -> Step 2	32.53%	**39.86%**	27.61%
	Step 2	**60.72**				

案例分析

- 第一轮预测结果存在较多错误，在第二轮得到修复
- 两个错误预测的纠正，与上下文内容和上下文情绪密切相关

No.	Speaker	Utterance	Step=1	Step=2	Gold
1	Chandler	What are you doing tonight?	neutral	neutral	neutral
2	Joey	Huh? Uh.	neutral	neutral	neutral
3	Chandler	Dude. Dude.	neutral	neutral	surprise
4	Joey	Oh, Sorry. Uh, I've got those plans with Phoebe, why?	neutral	neutral	neutral
5	Chandler	Oh really? Uh, Monica said she had a date at 9:00.	surprise	surprise	surprise
6	Joey	What? Tonight?	surprise	surprise	surprise
7	Chandler	That's what Monica said.	neutral	neutral	neutral
8	Joey	After she gave me that big speech?	neutral	surprise	surprise
9	Joey	She goes and makes a date on the same night she has plans with me?	neutral	anger	anger
10	Joey	I think she's trying to pull a fast one on Big Daddy.	anger	anger	anger

总结与未来计划

工作总结

- 我们的工作显式建模上下文情绪交互，相比隐式建模情绪交互可靠性更好
- 我们提出了迭代情绪交互模型，实现了情绪预测性能的逐步提升
- 我们的方法在两个数据集上实现了最好的结果，并对有效性进行了实验分析

未来计划

- 融合上下文的多模态信息
- 对话情感生成中引入对话情绪识别

感 谢 聆 听


Co-GAT: A Co-Interactive Graph Attention Network for Joint Dialog Act Recognition and Sentiment Classification

Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, Ting Liu

Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, China

1 Task Definition

Task Definition

- Problem Formulation
  - Dialog Act Recognition:
    - $X = \{u_1, u_2, \ldots, u_T\}$ (a dialog consists of a sequence of T utterances)
    - $Y = \{y_1^d, y_2^d, \ldots, y_T^d\}$
  - Sentiment Classification:
    - $X = \{u_1, u_2, \ldots, u_T\}$ (a dialog consists of a sequence of T utterances)
    - $Y = \{y_1^s, y_2^s, \ldots, y_T^s\}$
  - Example:

Speaker	Utterance	DA Label	Sentiment Label
User A	they are as tired of social media as I am .	Statement	Negative
User B	yes ! i don't get it . everyone i talk to about facebook - - everyone - - hates it , but none of them will take action .	Agreement	Negative

2 Motivation

Motivation

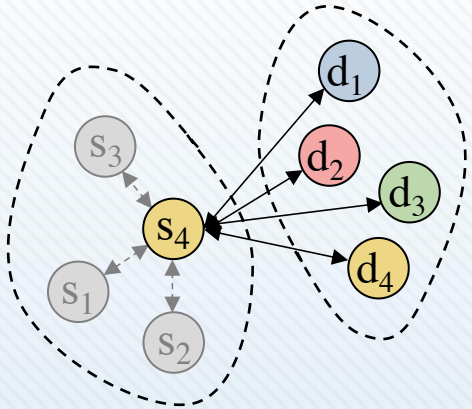
- The two factors that contribute to the dialog act recognition and sentiment prediction.
  - mutual interaction information
  - contextual information

- Example

Speaker	Utterance	DA Label	Sentiment Label
User A	they are as tired of social media as I am .	Statement	Negative
User B	yes ! i don't get it . everyone i talk to about facebook - - everyone - - hates it , but none of them will take action .	Agreement	Negative

Previous joint model work

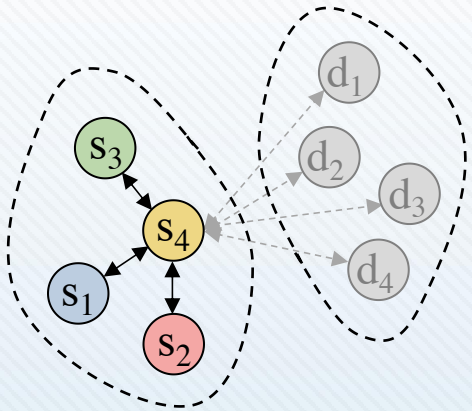
- Focus on mutual interaction information
  - Cerisara et al. (2018) proposed a multi-task framework to jointly model the two tasks
  - implicitly extract the shared mutual interaction information
  - fail to effectively capture the contextual information.



Cerisara, C.; Jafaritazehjani, S.; Oluokun, A.; and Le, H. T. Multi-task dialog act and sentiment recognition on mastodon. In Proc. of COLING 2018.

Previous joint model work

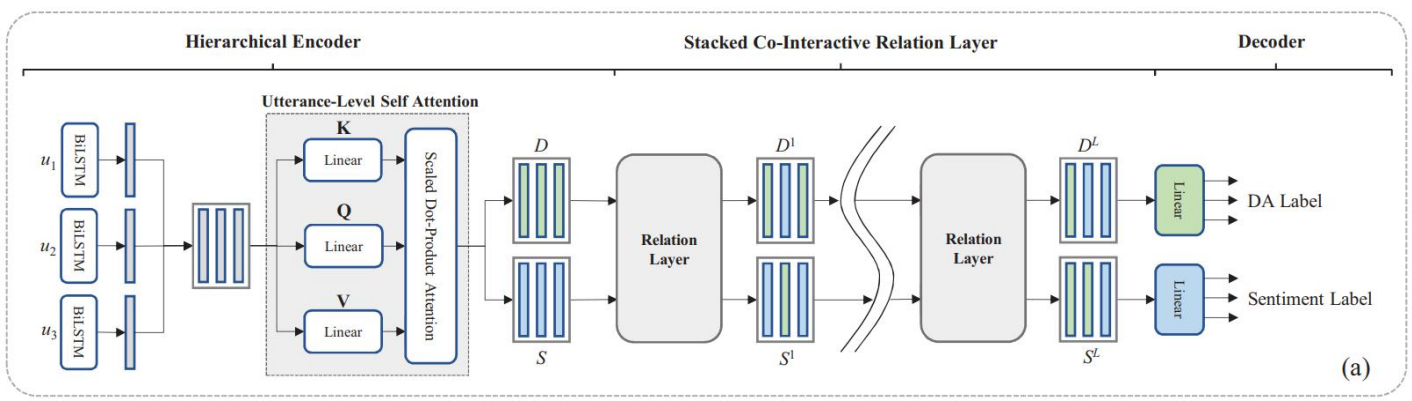
- Focus on contextual information
  - Kim and Kim (2018) explicitly leveraged the previous act information to guide the current DA prediction, which captures the contextual information.
  - ignored the mutual interaction information



Kim, M.; and Kim, H. Integrated neural network model for identifying speech acts, predicators, and sentiments of dialogue utterances. Pattern Recognition Letters 2018.

Previous joint model work – Pipeline Method

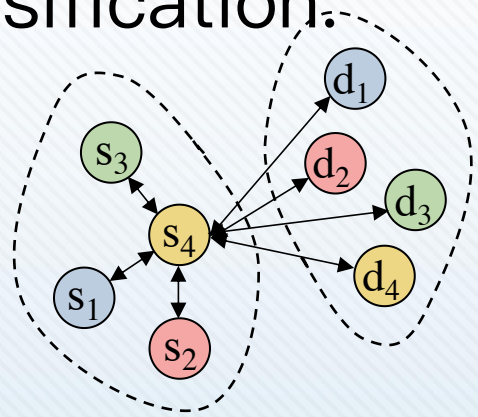
- Qin et al. (2020a) propose a pipeline method (DCR–Net) to incorporate the two types of information.
  - a hierarchical encoder is proposed to capture the contextual information
  - a relation layer to consider the mutual interaction information
  - two information are modeled separately



Libo Qin, Wanxiang Che, Yangming Li, Mingheng Ni, Ting Liu. DCR–Net: A Deep Co–Interactive Relation Network for Joint Dialog Act Recognition and Sentiment Classification. AAAI 2020.

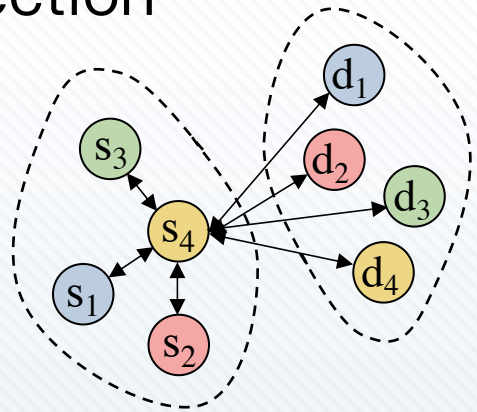
Motivation – Problem

- Problem: Can we simultaneously model the mutual interaction and contextual information in a unified framework to fully incorporate them?
- We propose a Co-Interactive Graph Attention Network (Co-GAT) for joint dialog act recognition and sentiment classification.



Motivation

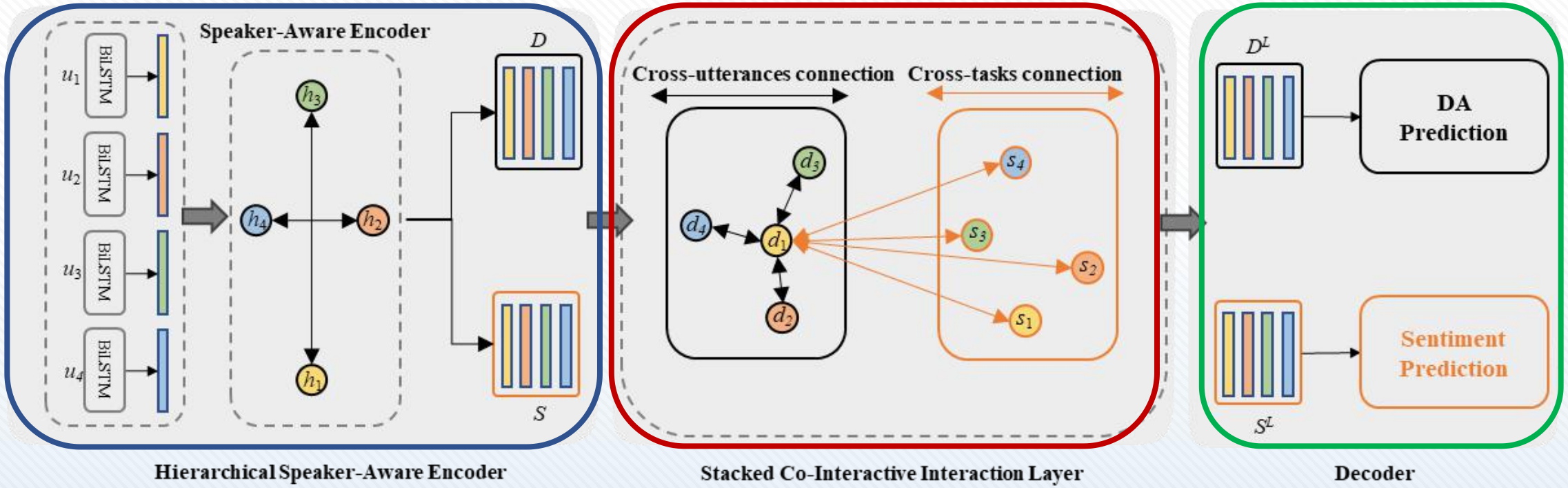
- In Co-Interactive graph, we perform a dual-connection
  - cross-utterances connection
  - cross-tasks connection



- Answer: each utterance node can be updated simultaneously with the contextual information and mutual interaction information.

3 Framework

Framework


Speaker-Aware Encoder
BiLSTM
BiLSTM
BiLSTM
BiLSTM
u_1
u_2
u_3
u_4
h_3
h_4
h_2
h_1
D
S
Cross-utterances connection
Cross-tasks connection
d_3
d_4
d_1
d_2
s_4
s_3
s_2
s_1
D^L
DA Prediction
S^L
Sentiment Prediction
Hierarchical Speaker-Aware Encoder
Stacked Co-Interactive Interaction Layer
Decoder

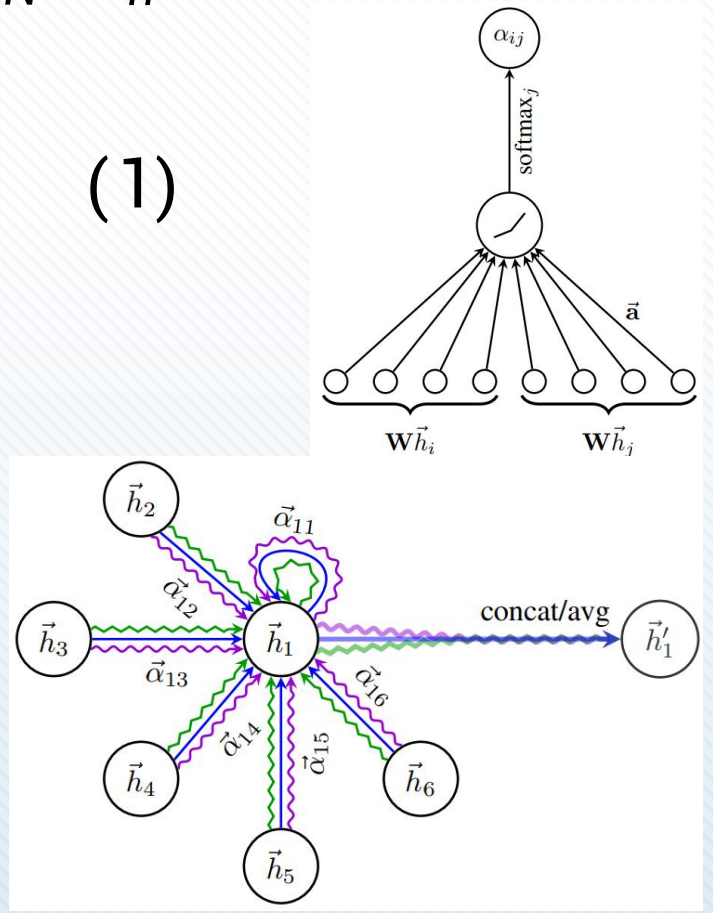
Vanilla Graph Attention Network

$$\vec{H} = \{\vec{h}_1, \ldots, \vec{h}_N\}, \vec{h}_n \in \mathbb{R}^F \xrightarrow{GAT} \vec{H}' = \{\vec{h}'_1, \ldots, \vec{h}'_N\}, \vec{h}'_n \in \mathbb{R}^{F'}$$

$$\mathcal{F}(\vec{h}_i, \vec{h}_j) = \text{LeakyReLU}(\mathbf{a}^\top [\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j]) \quad (1)$$

$$\alpha_{ij} = \frac{\exp(\mathcal{F}(\vec{h}_i, \vec{h}_j))}{\sum_{j' \in \mathcal{N}_i} \exp(\mathcal{F}(\vec{h}_i, \vec{h}_{j'}))} \quad (2)$$

$$\vec{h}'_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}\vec{h}_j\right) \quad (3)$$



Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. ICLR 2018.

Hierarchical Speaker–Aware Encoder

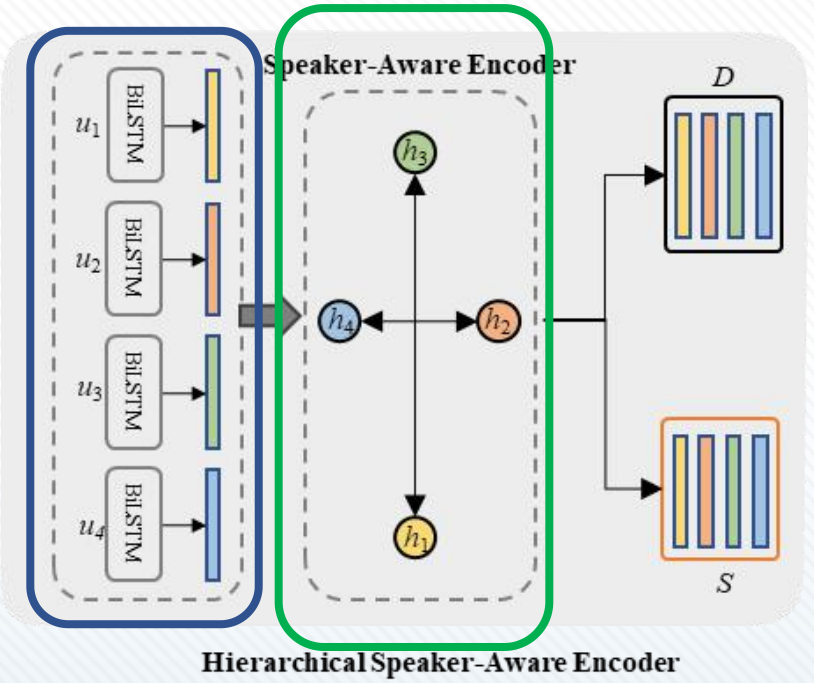
Utterance Encoder with Bi–LSTM

$$C = (u_1, \ldots, u_N) \qquad H = \{\boldsymbol{h}_1^t, \boldsymbol{h}_2^t, \ldots, \boldsymbol{h}_n^t\}$$

$$\boldsymbol{h}_i^t = \mathrm{LSTM}(\phi^{emb}(w_i^t), \boldsymbol{h}_{i-1}^t), t \in [1, n],$$

$$\boldsymbol{h}_i^t = \mathrm{LSTM}(\phi^{emb}(w_i^t), \boldsymbol{h}_{i-1}^t), t \in [n, 1],$$

$$\boldsymbol{h}_i^t = [\boldsymbol{h}_i^t, \boldsymbol{h}_i^t]$$

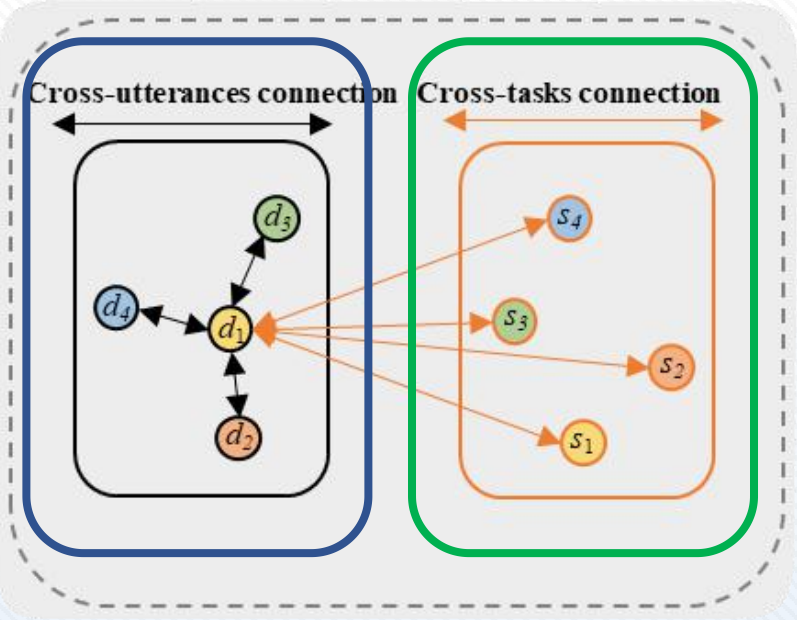


Speaker–Lever Encoder

- Vertices: Each utterance in the conversation is represented as a vertex.
- Edges: vertex i and vertex j should be connected if they belong to the same speaker.

Stacked Co-Interactive Graph Layer

- Vertices: 2N nodes in the graph
  - N nodes for sentiment classification task
  - N nodes for dialog act recognition task
- Edges
  - Cross-utterances connection
    - node i connects to its context utterance node to take the contextual information into account
  - Cross-tasks connection
    - node i connects to all another task node to explicitly leverage the mutual interaction information



Stacked Co-Interactive Interaction Layer

Decoder and Training

Decoder

$$y_t^d = \mathrm{softmax}\left(W^d d_t^{L'} + b_d\right) \quad (4)$$

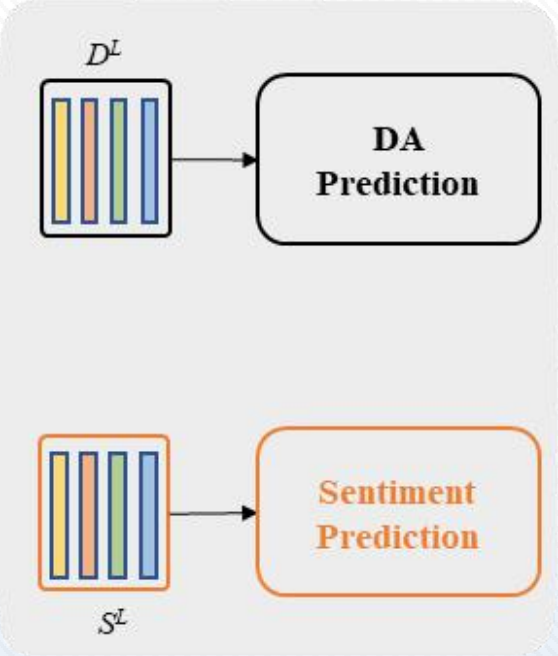
$$y_t^s = \mathrm{softmax}\left(W^s s_t^{L'} + b_s\right) \quad (5)$$

Joint Training

$$\mathcal{L}_1 \triangleq -\sum_{i=1}^{N}\sum_{j=1}^{N_S} \hat{y}_i^{(j,s)} \log\left(y_i^{(j,s)}\right) \quad (6)$$

$$\mathcal{L}_2 \triangleq -\sum_{i=1}^{N}\sum_{j=1}^{N_D} \hat{y}_i^{(j,d)} \log\left(y_i^{(j,d)}\right) \quad (7)$$

$$\mathcal{L}_\theta = \mathcal{L}_1 + \mathcal{L}_2 \quad (8)$$



Decoder

4 Experiments

Dataset

- **Mastodon** (Cerisara et al. 2018)
  - The Mastodon dataset consists of 269 dialogues for a total of 1,075 utterances in training dataset and the test dataset is a corpus of 266 dialogues for a total of 1,142 utterances.

- **Dailydialog** (Li et al. 2017)
  - For Dailydialog dataset, we adopt the standard split from the original dataset employing 11,118 dialogues for training, 1,000 for validating, and 1,000 for testing.

Cerisara, C.; Jafaritazehjani, S.; Oluokun, A.; and Le, H. T. Multi-task dialog act and sentiment recognition on mastodon. In Proc. of COLING 2018.
Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Proc. of IJCNLP 2017.

Results

Model	Mastodon						Dailydialog					
	SC			DAR			SC			DAR		
	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)
HEC (Kumar et al. 2018)	-	-	-	56.1	55.7	56.5	-	-	-	77.8	76.5	77.8
CRF-ASN (Chen et al. 2018)	-	-	-	55.1	53.9	56.5	-	-	-	76.0	75.6	78.2
CASA (Raheja and Tetreault 2019)	-	-	-	56.4	57.1	55.7	-	-	-	78.0	76.5	77.9
DialogueRNN (Majumder et al. 2019)	41.5	42.8	40.5	-	-	-	40.3	37.7	44.5	-	-	-
DialogueGCN (Ghosal et al. 2019)	42.4	43.4	41.4	-	-	-	43.1	44.5	41.8	-	-	-
JointDAS (Cerisara et al. 2018)	37.6	41.6	36.1	53.2	51.9	55.6	31.2	28.8	35.4	75.1	74.5	76.2
IIIM (Kim and Kim 2018)	39.4	40.1	38.7	54.3	52.2	56.3	33.0	28.5	38.9	75.7	74.9	76.5
DCR-Net + Co-Attention (Qin et al. 2020a)	45.1	47.3	43.2	58.6	56.9	60.3	45.4	40.1	56.0	79.1	79.0	79.1
Our model	**48.1***	**53.2***	**44.0***	**60.5***	**60.6***	**60.4**	**51.0***	**45.3***	**65.9***	**79.4**	78.1	**81.0***

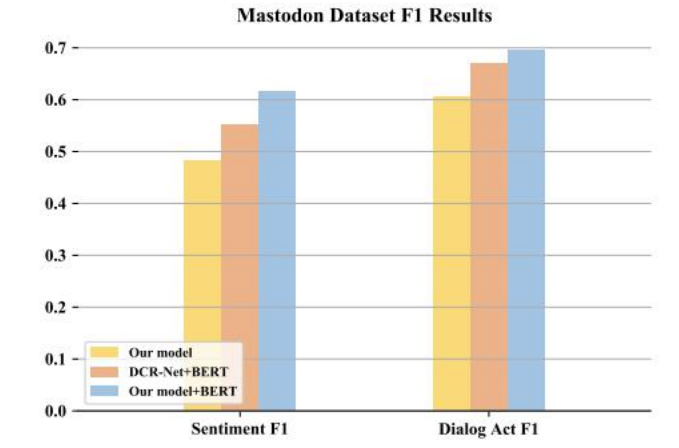
- Our framework outperforms the state-of-the-art dialog act recognition and sentiment classification models which trained in separate task in all metrics on two datasets.
- We obtain large improvements compared with the state-of-the-art joint models.

Ablation

Model	Mastodon		Dailydialog	
	SC (F1)	DAR (F1)	SC (F1)	DAR (F1)
without cross-tasks connection	46.1	58.1	49.7	78.2
without cross-utterances connection	44.9	58.7	48.1	78.2
separate modeling	46.7	58.4	45.6	78.3
co-attention mechanism	46.5	59.4	46.2	79.1
without speaker information	46.4	59.0	47.6	79.2
Our model	**48.1**	**60.5**	**51.0**	**79.4**

- Analysis
  - Effectiveness of the Mutual Interaction Information
  - Effectiveness of the Contextual Information
  - Simultaneous Modeling vs. Separate Modeling
  - Effectiveness of Speaker Information

Results on pretrained models



Model	Mastodon					
	SC			DAR		
	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)
RoBERTa+Linear	55.7	54.4	59.7	61.6	61.8	61.4
Co-GAT+RoBERTa	61.3	58.8	64.3	66.1	64.8	67.5
XLNet+Linear	58.7	60.9	56.6	62.6	61.8	63.4
Co-GAT+XLNet	65.9	65.8	66.1	67.5	66.0	69.2

- Analysis
  - The BERT–based model performs remarkably well and outperforms the baseline (DCR–Net + BERT)
  - Our contribution from Co–GAT does not fully overlap with contextualized word representations (Roberta, XLNet)

5 Conclusion

Conclusion

- We make the first attempt to simultaneously incorporate contextual information and mutual interaction information for DAR and SC.
- We propose a co-interactive graph attention network, achieving to model simultaneously incorporate contextual information and mutual interaction information.
- Our model achieves the SOTA performance and is also beneficial when combined with pre-trained models (BERT, Roberta, XLNet)

Paper&Code

Paper

Code

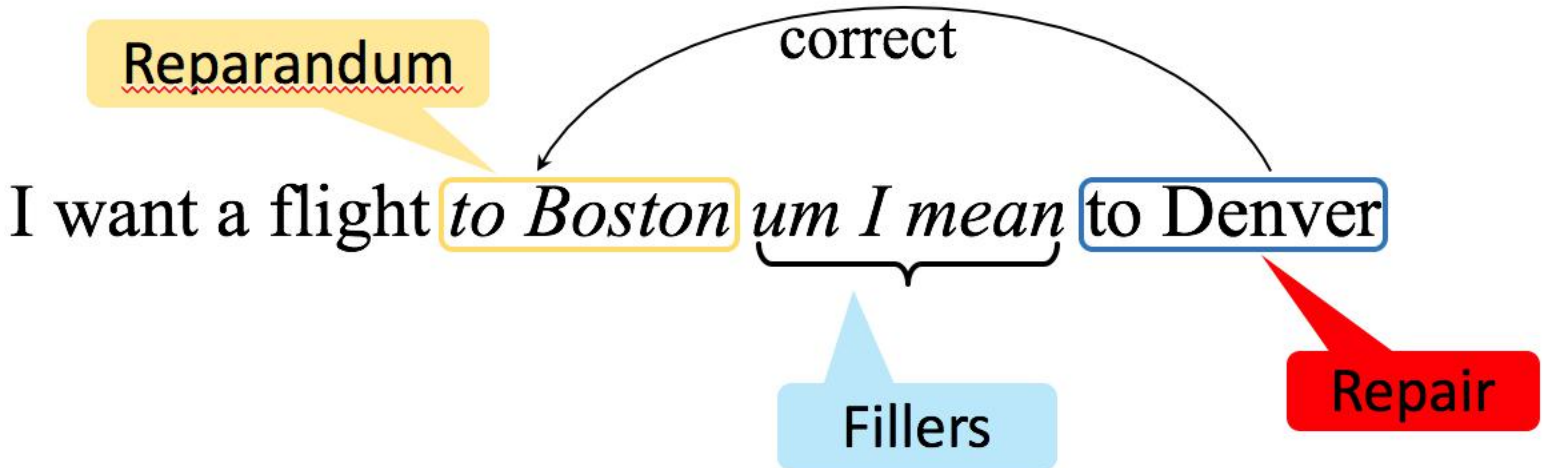
Thank you!

Combining Self-Training and Self-Supervised Learning for Unsupervised Disfluency Detection

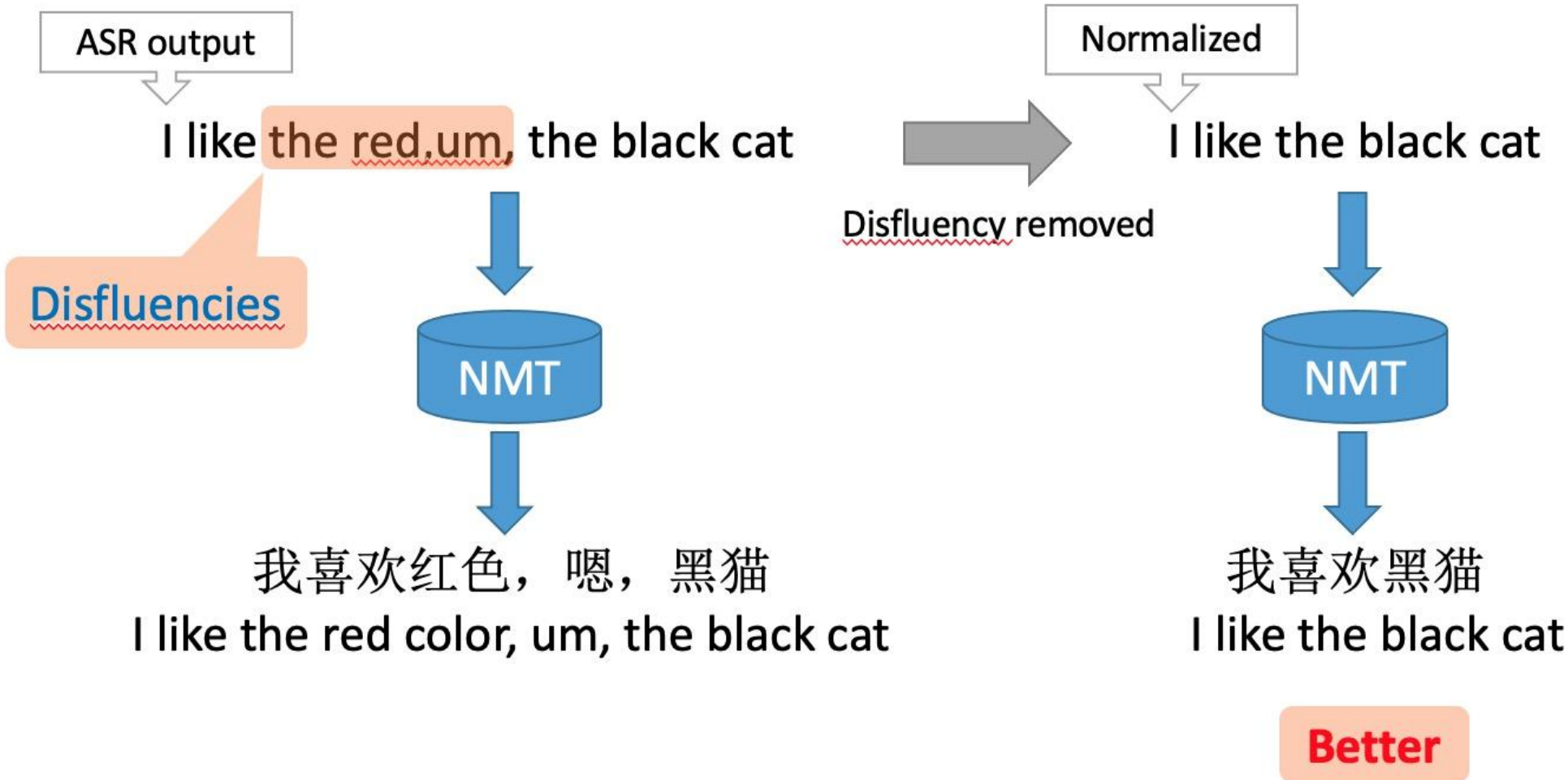
Shaolei Wang, Zhongyuan Wang, Wanxiang Che, Ting Liu

School of Computer Science and Technology
Harbin Institute of Technology, Harbin, China

Disfluency Detection



Disfluency Effect on Machine Translation



Our Motivations

- Unsupervised Disfluency Detection
  - Utilize Self-Training method as the main framework

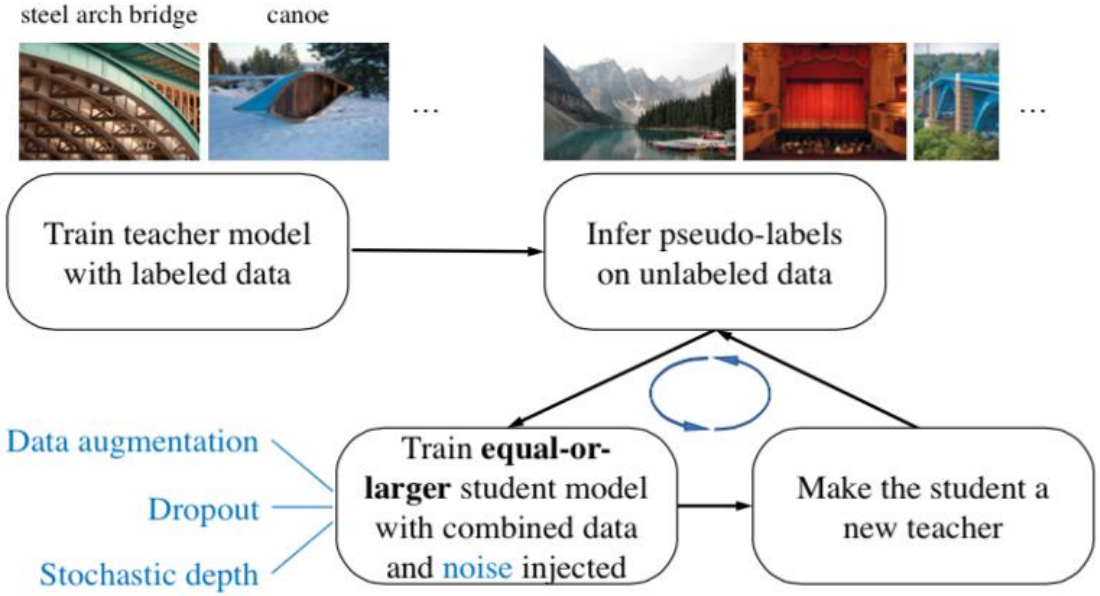


Figure 1: Illustration of the Noisy Student Training. (All shown images are from ImageNet.)

Our Motivations

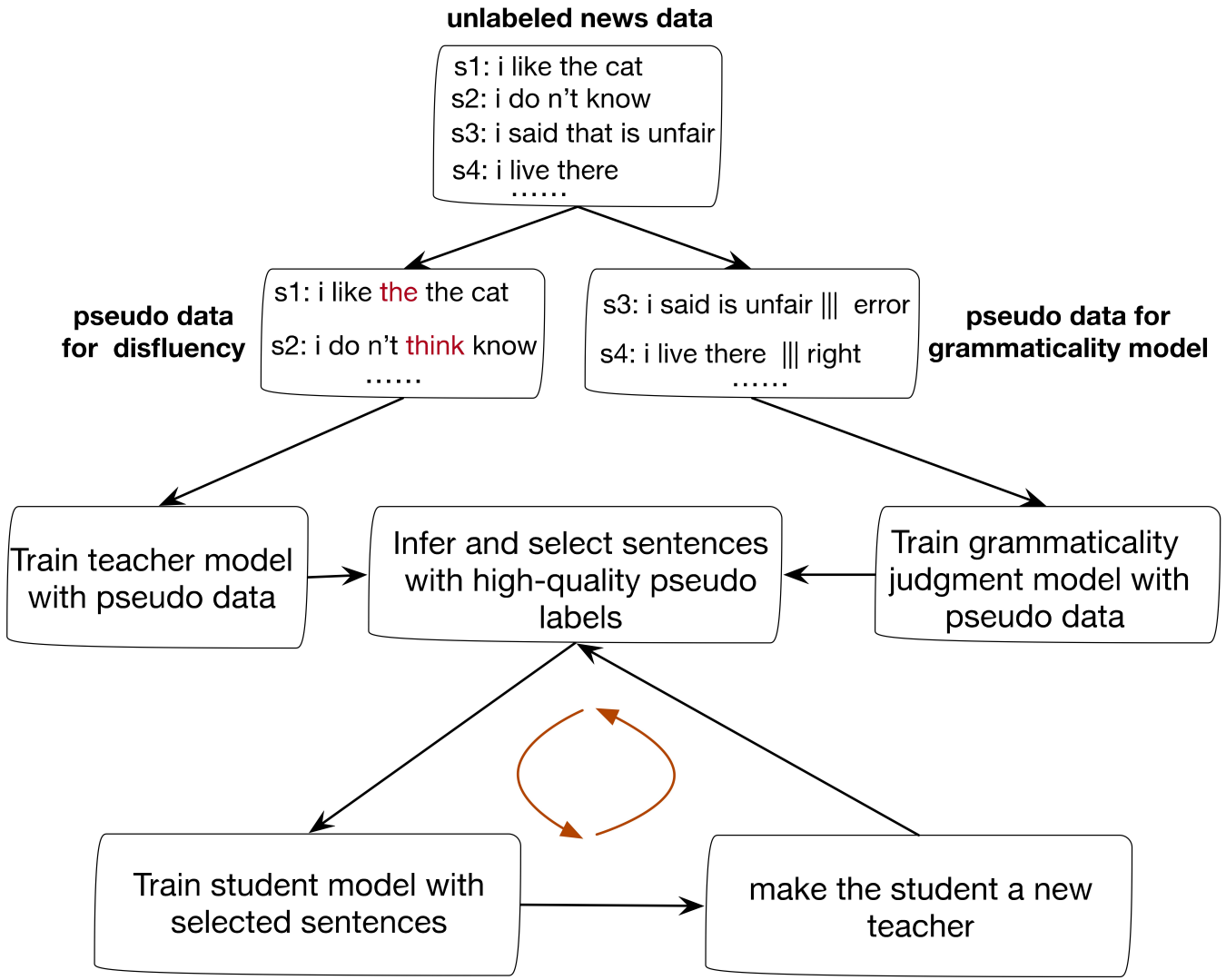
- Unsupervised Disfluency Detection
  - Utilize Self-Training method as the main framework
  - Utilize self-supervised learning method to train weak model as the teacher model

Our Motivations

Unsupervised Disfluency Detection

- Utilize Self-Training method as the main framework
- Utilize self-supervised learning method to train weak model as the teacher model
- Utilize self-supervised learning method to train a sentence grammaticality judgment model to help select sentences with high-quality pseudo labels

Our Model


unlabeled news data
s1: i like the cat
s2: i do n't know
s3: i said that is unfair
s4: i live there
.....
pseudo data for disfluency
s1: i like the the cat
s2: i do n't think know
.....
s3: i said is unfair ||| error
s4: i live there ||| right
.....
pseudo data for grammaticality model
Train teacher model with pseudo data
Infer and select sentences with high-quality pseudo labels
Train grammaticality judgment model with pseudo data
Train student model with selected sentences
make the student a new teacher

Experimental Setting

- English ELECTRA-Base model as initialization
- Dataset
  - Valid data: English Switchboard corpus
  - Test data:
    - English Switchboard corpus
    - CallHome
    - SCOTUS
    - FCIC
  - Pre-training dataset: News Discussions from WMT2017
  - Dataset for self-training: Fisher Speech Transcripts

Experiment results

- Experiment results on on the Switchboard dev set

Method	P	R	F1
Transition-based	91.9	85.1	88.4
BERT-Base fine-tuning	92.2	89.8	90.9
ELECTRA-Small fine-tuning	91.6	89.5	90.5
ELECTRA-Base fine-tuning	92.9	91.2	92.0
Teacher fine-tuning	92.5	92.1	**92.3**
Unsupervised teacher	86.8	62.0	72.3
Our unsupervised	90.2	89.1	**89.6**

Experiment results

- Comparison with the previous state-of-the-art methods

Method	P	R	F1
UBT (Wu et al., 2015)	90.3	80.5	85.1
Bi-LSTM (Zayats et al., 2016)	91.8	80.6	85.9
NCM (Lou and Johnson, 2017)	-	-	86.8
Transition-based (Wang et al., 2017)	91.1	84.1	87.5
Self-supervised(Wang et al., 2019)	93.4	87.3	90.2
Self-training(Lou and Johnson, 2020)	87.5	93.8	90.6
EGBC(Bach and Huang, 2019)	95.7	88.3	**91.8**
Our Method	88.2	87.8	**88.0**

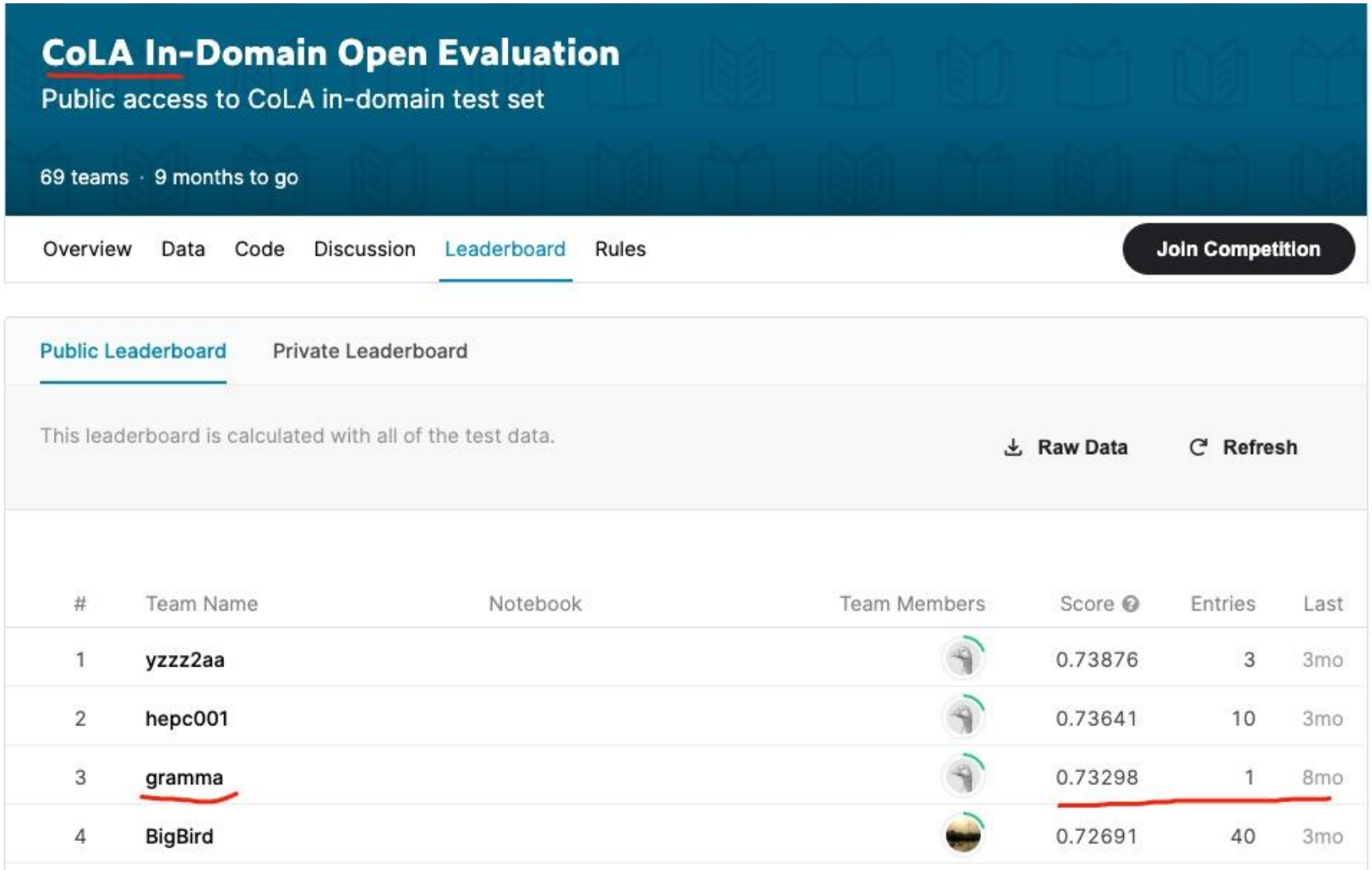
Performance on Cross-domain Data

Method	CallHome	SCOTUS	FCIC
Unsupervised teacher	45.7	63.9	43.2
ELECTRA-Base	60.9	79.4	62.8
Teacher fine-tuning	63.7	81.9	64.3
Pattern-match	65.2	79.9	66.1
Our unsupervised	60.2	80.3	63.3

Ablation Test

Method	SWBD	CallHome	SCOTUS	FCIC
Teacher	72.3	45.7	63.9	43.2
No-select	83.4	55.9	70.4	56.1
Select	89.6	60.2	80.3	63.3

Power of grammaticality judgment model



CoLA In-Domain Open Evaluation

Public access to CoLA in-domain test set

69 teams · 9 months to go

Overview Data Code Discussion Leaderboard Rules **Join Competition**

Public Leaderboard Private Leaderboard

This leaderboard is calculated with all of the test data. **Raw Data** **Refresh**

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	yzzz2aa			0.73876	3	3mo
2	hepc001			0.73641	10	3mo
3	gramma			0.73298	1	8mo
4	BigBird			0.72691	40	3mo

Conclusion

- We explore unsupervised disfluency detection by combining self-training and self-supervised learning

- We showed that it is possible to completely remove the need of human-annotated data and train a high-performance disfluency detection system in a completely unsupervised manner.


SCIR
Research Center for Social Computing and Information Retrieval
理解语言 认知社会

Thank you!