Speaker-Guided Encoder-Decoder Framework for Emotion Recognition in Conversation

Yinan Bao[1,2], Qianwen Ma[1,2], Lingwei Wei[1,2], Wei Zhou[1,*] and Songlin Hu[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
{baoyinan, maqianwen, weilingwei, zhouwei, husonglin}@iie.ac.cn

2022.6.29 • ChongQing

—— IJCAI 2022

Reported by Sijin Liu

1.Introduction

2.Method

3.Experiments

Introduction

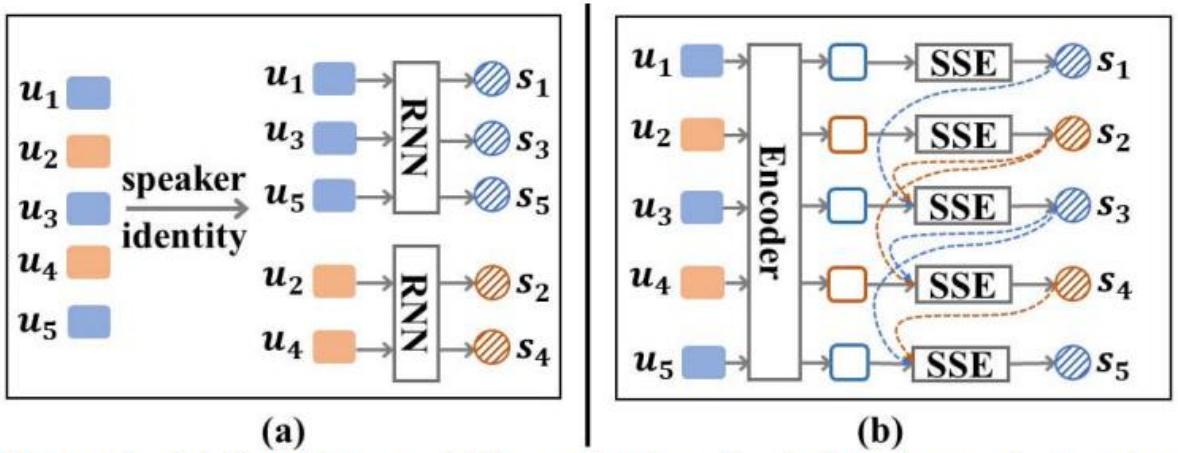


Figure 1: (a) Speaker modeling scheme of existing dynamic speaker-specific modeling methods, which is based on speaker identity to explore intra-speaker dependency. (b) Our novel speaker modeling scheme, which uses the speaker state encoder (SSE) to explore intra- and inter-speaker dependencies dynamically. u_i means the i_{th} utterance in a dialogue and s_i means the corresponding speaker state vector of u_i.

Sequences composed of separate utterances attaching to the same speaker are disjunctive and incomplete, which hinder the comprehension of the context.

Overreliance on the speaker identity hinders the modeling of dynamic inter-speaker dependency. In this way, the model can't explore dynamic interactions between different speakers.

Method

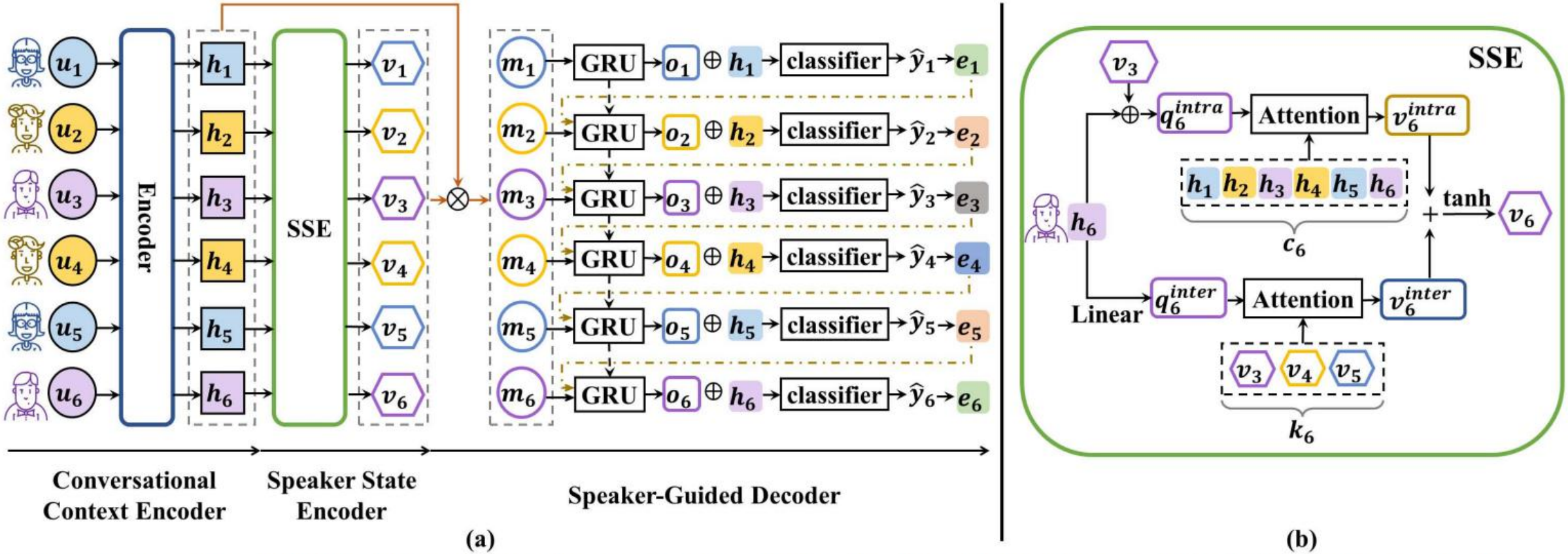
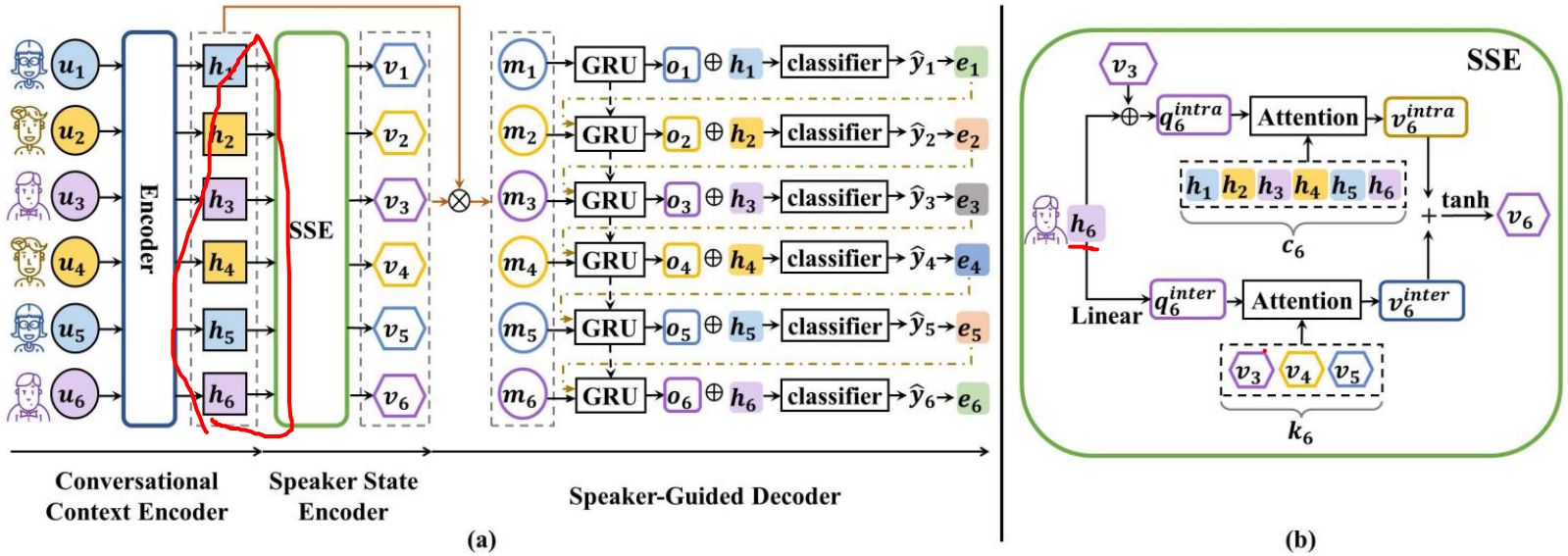


Figure 2: An overview of SGED. SSE represents the speaker state encoder. As shown in the right part of the figure, we take u_6 as an example to show how to obtain the speaker state vector of it.

Method



Conversational Context Encoder

$$C = [u_1, ..., u_N]$$
$$\mathbf{H} = Encoder(C), \quad (1)$$

Speaker State Encoder

Intra-Speaker Dependency Modeling

$$\mathbf{q}_i^{intra} = \mathbf{W}_q^{intra}[\mathbf{v}_{\psi(u_i)}||\mathbf{h}_i] + \mathbf{b}_q^{intra}, \quad (2)$$

where $\mathbf{v}_{\psi(u_i)} \in \mathbb{R}^{h\times 1}$ means the last previous speaker state vector of $s_{\phi(u_i)}$; $||$ means the concatenation operation; $\mathbf{W}_q^{intra} \in \mathbb{R}^{h\times 2h}$ and $\mathbf{b}_q^{intra} \in \mathbb{R}^1$ are model parameters.

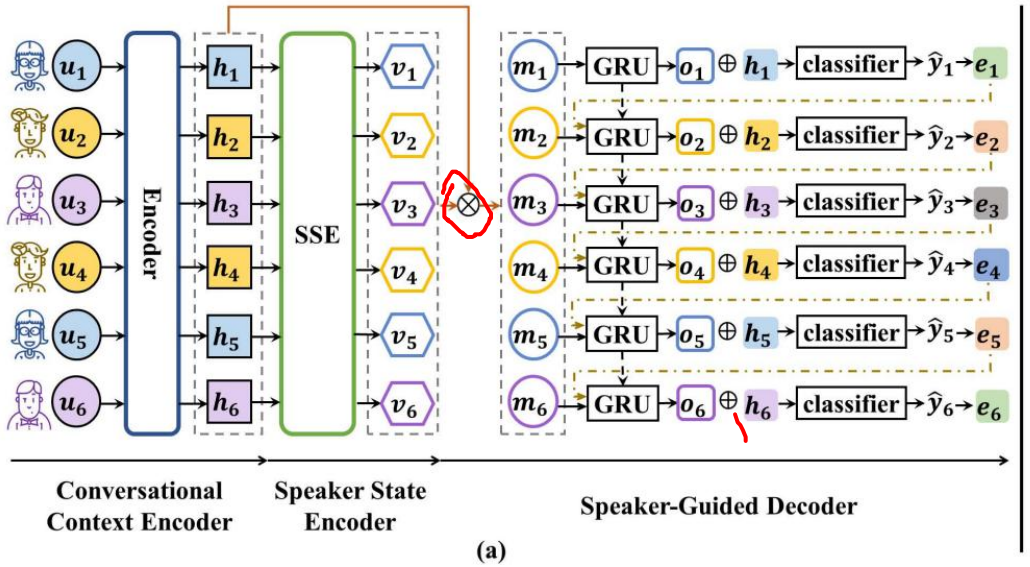
$$\alpha_i^{intra} = \mathbf{softmax}(\mathbf{W}_1(\mathbf{q}_i^{intra} \odot \mathbf{c}_i) + \mathbf{b}_1), \quad (3)$$
$$\mathbf{v}_i^{intra} = \alpha_i^{intra}\mathbf{c}_i^T,$$

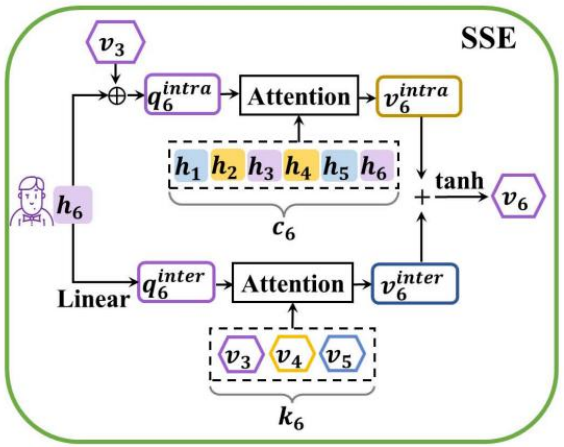
Inter-Speaker Dependency Modeling

$$\mathbf{q}_i^{inter} = \mathbf{W}_q^{inter}\mathbf{h}_i + \mathbf{b}_q^{inter},$$
$$\alpha_i^{inter} = \mathbf{softmax}(\mathbf{W}_2(\mathbf{q}_i^{inter} \odot \mathbf{k}_i) + \mathbf{b}_2), \quad (4)$$
$$\mathbf{v}_i^{inter} = \alpha_i^{inter}\mathbf{k}_i^T,$$

$$\mathbf{v}_i = \mathbf{tanh}(\mathbf{v}_i^{intra} + \mathbf{v}_i^{inter}), \quad (5)$$

Method





Speaker-Guided Decoder

$$\mathbf{m}_i = \mathbf{ReLU}(\mathbf{v}_i \odot (\mathbf{W}_m \mathbf{h}_i + \mathbf{b}_m)^T), \tag{6}$$

$$\mathbf{o}_i = \overrightarrow{\mathbf{GRU}}([\mathbf{m}_i || \mathbf{e}_{i-1}], \mathbf{o}_{i-1}). \tag{7}$$

$$\begin{aligned} \mathbf{z}_i &= \mathbf{ReLU}(\mathbf{W}_o[\mathbf{h}_i || \mathbf{o}_i^T] + \mathbf{b}_o), \\ \mathcal{P}_i &= \mathbf{softmax}(\mathbf{W}_z \mathbf{z}_i + \mathbf{b}_z), \\ \hat{y}_i &= \underset{g}{\operatorname{argmax}}\, \mathcal{P}_i[g], \\ \mathbf{e}_i &= \mathbf{E}[\hat{y}_i], \end{aligned} \tag{8}$$

$$\mathcal{L}(\theta) = -\sum_{j=1}^{M} \sum_{t=1}^{N_j} log \mathcal{P}_{j,t}[y_{j,t}], \tag{9}$$

Experiments

Dataset	Conversations			Utterances		
	Train	Val	Test	Train	Val	Test
IEMOCAP	120		31	5810		1623
MELD	1038	114	280	9989	1109	2610
EmoryNLP	713	99	85	9934	1344	1328

Table 1: The statistics of three datasets.

Model	MELD	EmoryNLP	IEMOCAP
KET	58.18	33.95	59.56
COSMIC	65.21	38.11	65.28
DialogueRNN	57.03	-	62.75
+ RoBERTa	63.61	37.44	64.76
DialogueCRN	58.39	-	66.20
+ RoBERTa	63.42	38.91	66.46
RoBERTa	62.88	37.78	63.38
SGED + RoBERTa	**63.34**	**38.47**	**64.11**
bc-LSTM+att	-	-	-
+ RoBERTa	62.95	38.28	64.51
SGED + bc-LSTM+att	**63.37**	**38.89**	**65.03**
DialogueGCN	58.10	-	64.18
+ RoBERTa	63.02	38.10	64.91
SGED + DialogueGCN	**64.55**	**39.73**	**65.90**
DAG-ERC	63.65	39.02	68.03
DAG-ERC*	63.39	38.84	67.45
SGED + DAG-ERC*	**65.46**	**40.24**	**68.53**

Table 2: Overall performance on the three datasets. We choose weighted-average F1 to evaluate each method. DAG-ERC* means that we use the one-layer DAG-ERC.

Experiments

Method	EmoryNLP	IEMOCAP
SGED + DAG-ERC*	**40.24**	**68.53**
w/o **SSE** - intra-speaker	39.87 (↓0.37)	68.05 (↓0.48)
w/o **SSE** - inter-speaker	39.86 (↓0.38)	67.72 (↓0.81)
w/o **SSE**	39.17 (↓1.07)	67.67 (↓0.86)
w/o **SGD**	38.96 (↓1.28)	67.30 (↓1.23)
w/o **SSE + SGD**	38.84 (↓1.40)	67.45 (↓1.08)

Table 3: Results of ablation study on EmoryNLP (multi-party) and IEMOCAP (two-party) dataset. We use the one-layer DAG-ERC as the conversational context encoder of our framework.

Experiments

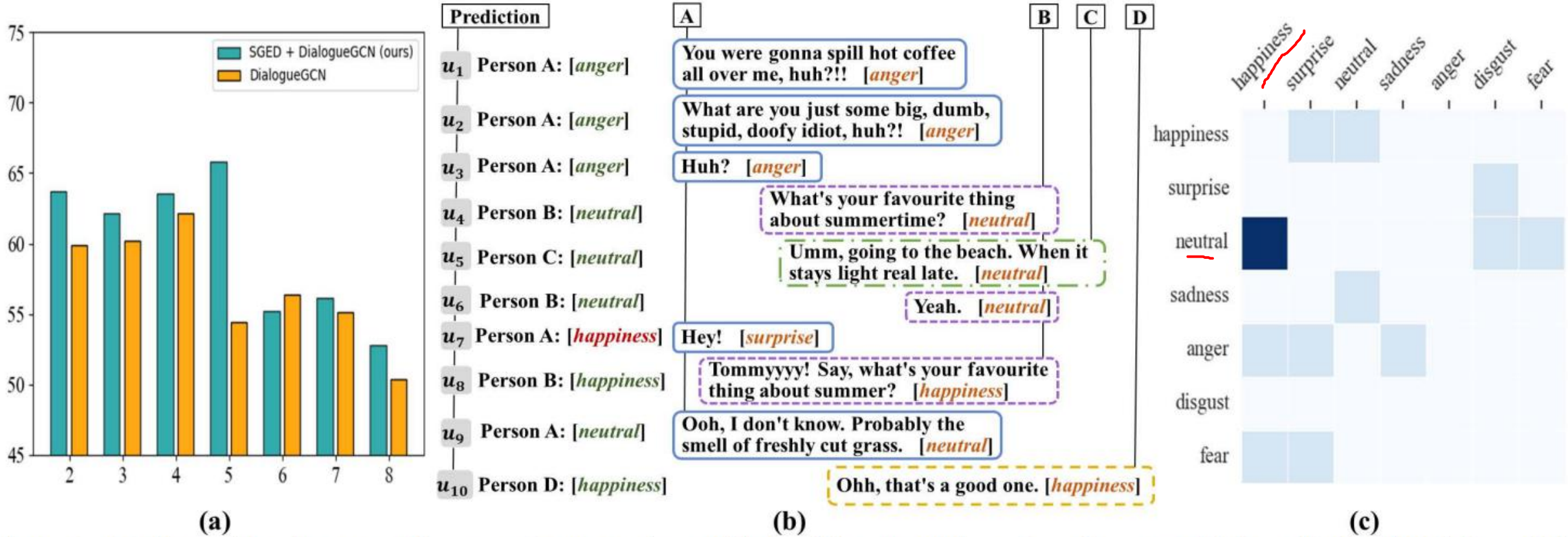


Figure 3: (a) The weighted-average F1 score of conversations with the different number of speakers on MELD, achieving by DialogueGCN and our SGED framework which takes DialogueGCN as the conversational context encoder. (b) An abridged dialogue from MELD in which our SGED framework recognizes most of the emotions correctly. (c) The confusion matrix of conversations with 6 speakers excluding the diagonal on MELD, is achieved by our SGED framework which takes DialogueGCN as the conversational context encoder.