Upgrade Open in app

Published in DAIR.AI · Follow

elvis · Follow
Mar 25, 2019 · 4 min read · Listen

DialogueRNN: Emotion Classification in Conversation

Perform emotion classification in the context of a conversation.

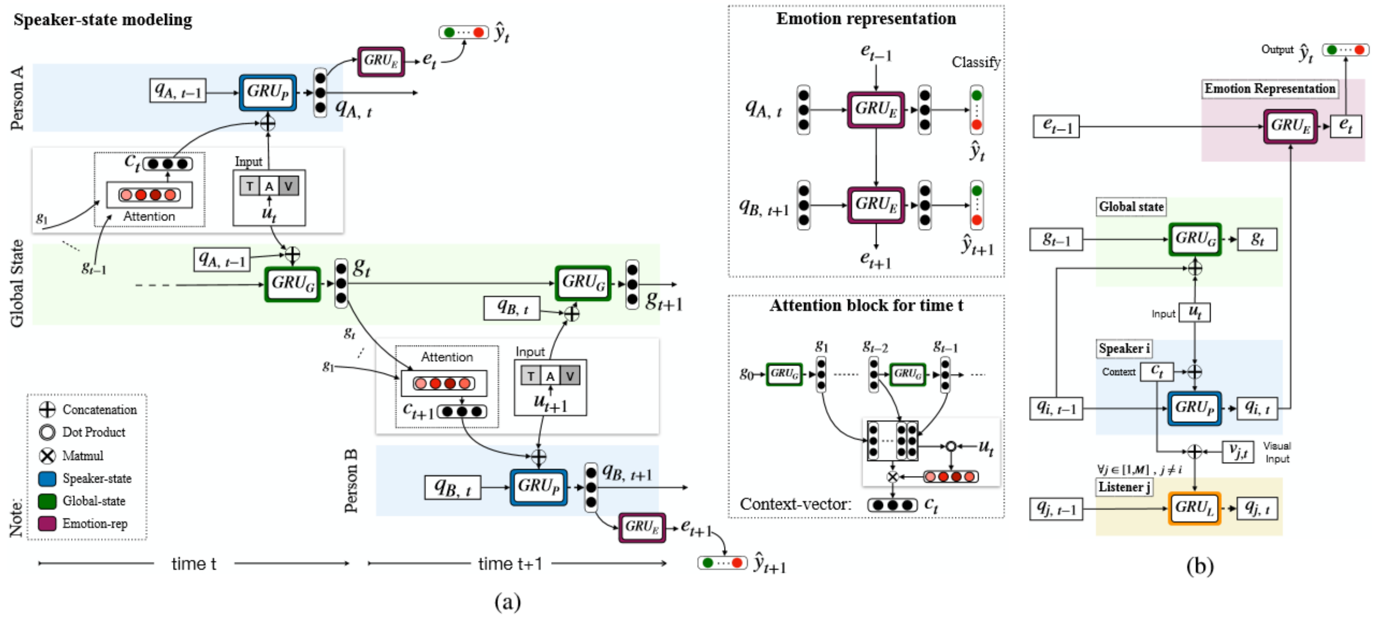
Welcome to another paper review!

DialogueRNN is a method that aims to perform emotion classification of utterances in the context of a conversation. Many applications can benefit from such type of analysis such as understanding the emotional context and interchange in *debates* and *social media threads*.

Previous methods do not pay attention to individuals' emotional states. The proposed emotion detection model considers individual speakers by focusing on three different aspects: *the speaker*, the *context of preceding utterances*, and the *emotion from preceding utterances*. The idea is that these three aspects are important to accurately predict the emotion of the utterance.

Model

Utterances for a party (i.e., individual) are represented through textual features obtained from a convolutional neural network. As utterances come in a *multimodal setting*, audio and visual features are also extracted using 3D-CNN and openSMILE, respectively. The network is trained at the utterance level with the target emotion labels.



The proposed model (called DialogueRNN), illustrated in the figure above, determines the final emotion of the utterance through the following *factors*:

- **Party state** — models the parties' emotion dynamics through the conversations. The basic idea behind the party state is to ensure that the model is *aware of the speaker of each utterance* in the conversation.
- **Global state** — models the context of an utterance in the dialogue, given by jointly encoding preceding utterances and the party state. Note that attention mechanism is applied to the global state to provide improved context representation. This state basically serves as the *speaker-specific utterance representation*.
- **Emotion representation** — inferred through party state and preceding speaker's states as context (global state). This representation is used to perform the final emotion classification via a softmax layer.

applied an attention mechanism. The role of the attention mechanism is that it assigns higher attention scores to the utterances that are emotionally relevant to the current utterance.

Overall, the speaker update encodes — via the Party GRU (shown in blue) — the information on the current utterance along with its context from the Global GRU (shown in green). All this information is important for performing the final emotion classification, which is performed by the emotion GRU (shown in maroon). Note that the current emotion classification also relies on the previous emotion-relevant information as well.

Variants

Several variants of the DialogueRNN model are proposed and compared in this study:

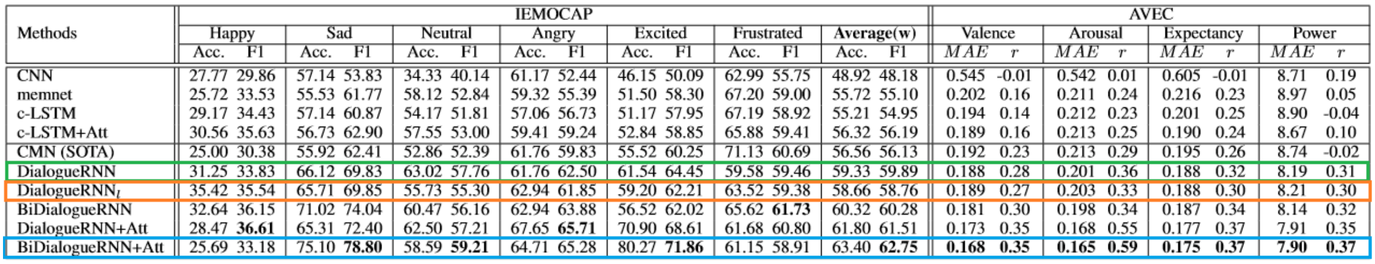
- **DialogueRNN_l** — considers an extra listener state (defined at the end of this post) while a speaker utters.
- **BiDialogueRNN** — a bidirectional RNN architecture is used instead
- **DialogueRNN+Att** — attention is applied over all surrounding emotion representations
- **BiDialogueRNN+Att** — similar to the previous model but considers a bidirectional RNN instead

Other baselines are also proposed which you can refer to in the paper.

Results

Two datasets are used for all experiments: IEMOCAP and AVEC. Both datasets contain interactions between multiple parties.

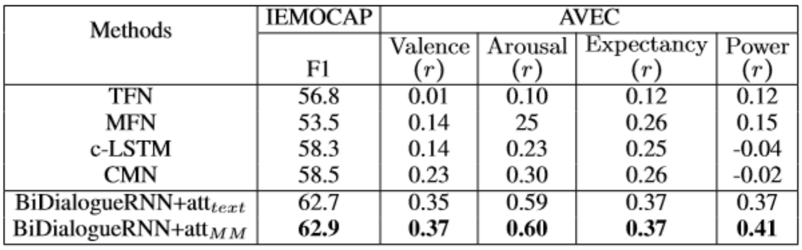
From the table below, we can observe that DialogueRNN (highlighted in green) outperforms all baselines and the state-of-the-art model (CMN) on both datasets. Note that these results are only using the *text modality*.



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**

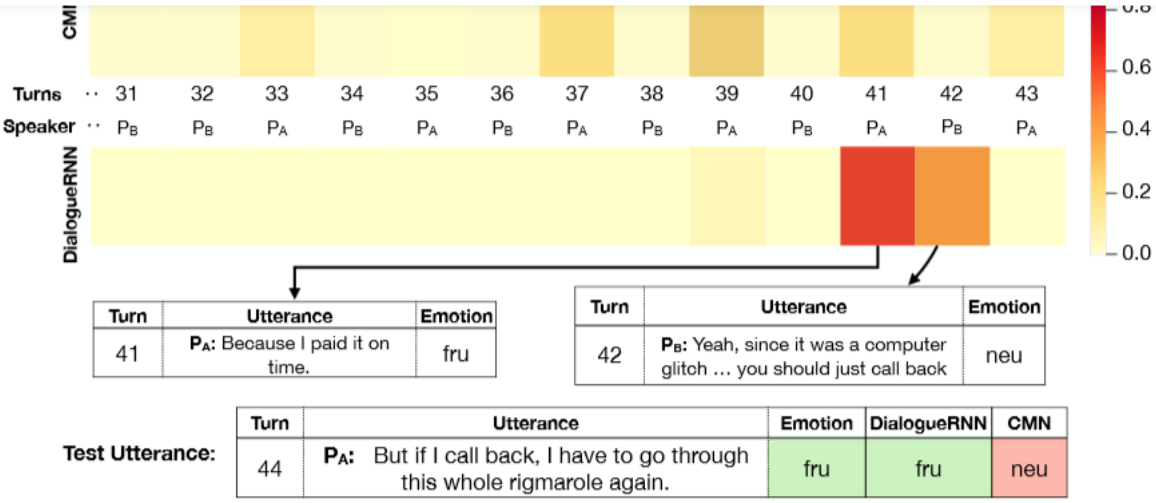
We can also observe in the table above that the *listener component* (model highlighted in orange) doesn't improve the model's performance. In general, the other variants were found to perform well, especially the BiDialogueRNN+Att, which in general produced the better results.

As shown in the table below, the proposed model, DialogueRNN, also significantly outperforms other models in the multimodal setting (using a fusion of modalities).



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**

As a case study, we can observe from the attention figure below that DialogueRNN correctly anticipates the emotion of *frustration* (labeled Turn 44) using the preceding context (41 and 42). For the CMN model, this was found not to be the case.



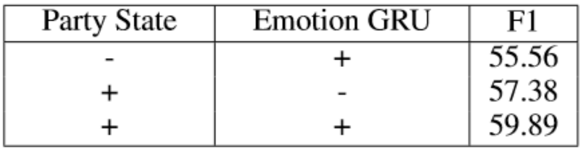
Turn	Utterance	Emotion
41	PA: Because I paid it on time.	fru

Turn	Utterance	Emotion
42	PB: Yeah, since it was a computer glitch … you should just call back	neu

Test Utterance:

Turn	Utterance	Emotion	DialogueRNN	CMN
44	PA: But if I call back, I have to go through this whole rigmarole again.	fru	fru	neu

An important *ablation study* was conducted to observe the importance of Emotion GRU and Party State components. We can see from the table below that the absence of part state decreases performance. In fact, it can be observed that the party state seems to be more important than Emotion GRU.



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

***Listener update** changes the state of the listener based on the current speaker utterance. Visual cues are used to represent this information. However, authors found via the experiments that this update has no effect in a conversation as a silent party has no influence in a conversation.*

Reference

DialogueRNN: An Attentive RNN for Emotion Detection in Conversations — [Paper] | [Code]