2022-03-18 Paper Introduction

Journal of Information Fusion 65(2021), IF=12.975, Cited=29

Conversational transfer learning for emotion recognition

Davamanyu Hazarika, Soujanya Poria, Roger Zimmerman, Rada Mihalcea

Takato Hayashi

Abstract

- Recognizing emotions in conversations is a challenging task due to the presence of contextual dependencies governed by self- and inter-personal influences.
- However, purely supervised strategies demand large amounts of annotated data, which is lacking in most of available corpora in this task.
- This paper proposed an approach, TL-ERC, where we pre-train a hierarchical dialogue model on multi-turn conversations (source) and then transfer its parameters to a conversational emotion classifier (target).
- TL-ERC improves in performance and robustness against limited training data. This model also achieves better validation performances in significantly fewer epochs.

Introduction

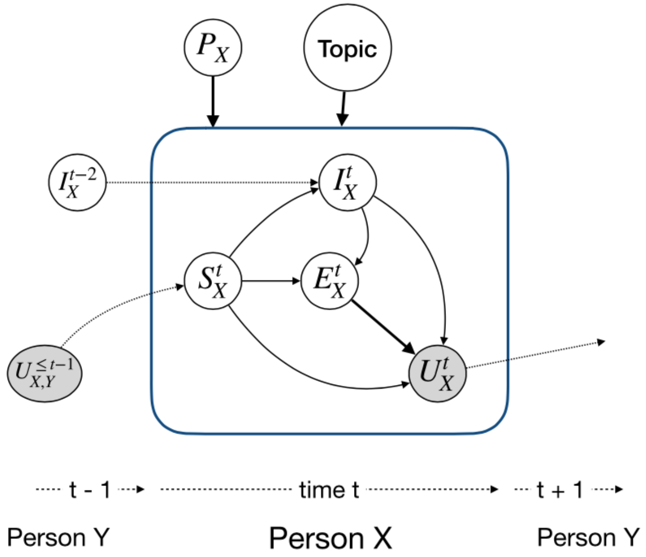
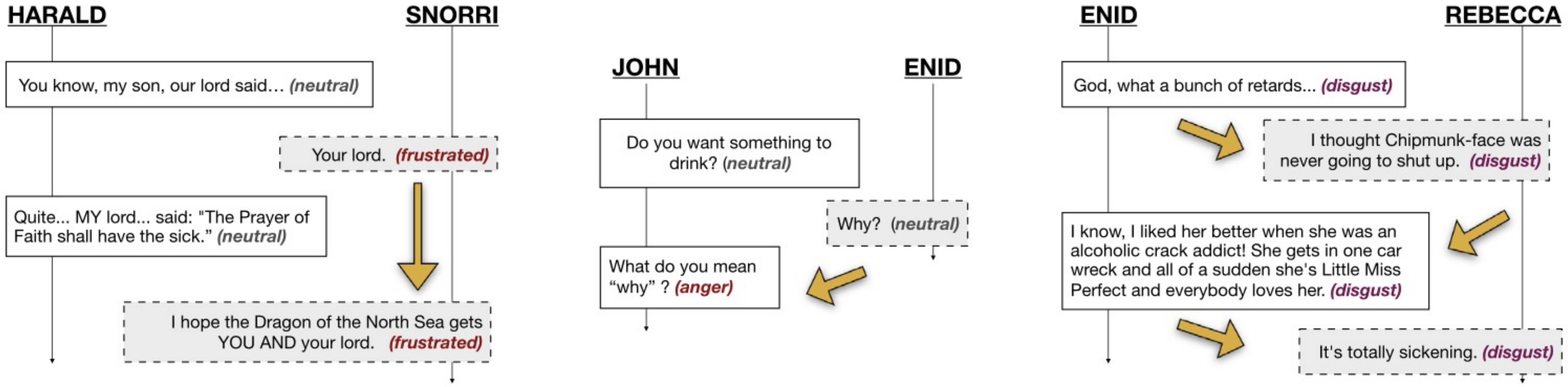


Fig. 1. Dyadic conversation—between person X and Y—are governed by interactions between several latent factors. Emotions are a crucial component in this generative process. In the illustration, P represents the personality of the speaker; S represents speaker-state; I denotes the intent of the speaker; E refers to the speaker's emotional state, and U refers to the observed utterance. Speaker personality and the topic always condition upon the variables. At turn t, the speaker conceives several pragmatic concepts such as argumentation logic, viewpoint, and inter-personal relationship - which we collectively represent using the speaker-state S [6]. Next, the intent I of the speaker gets formulated based on the current speaker-state and previous intent of the same speaker (at $t-2$). These two factors influence the emotional feeling of the speaker, which finally manifests as the spoken utterance [1].

- Several works in the literature have indicated that emotional goals and influences act as latent controllers in dialogues [1, 2]

- Poria et al [3] demonstrated the interplay of several factors, such as the topic of the conversation, speakers' personality, argumentation-logic, viewpoint, and intent, which modulate the emotional state of the speaker and finally lead to an utterance.

Introduction

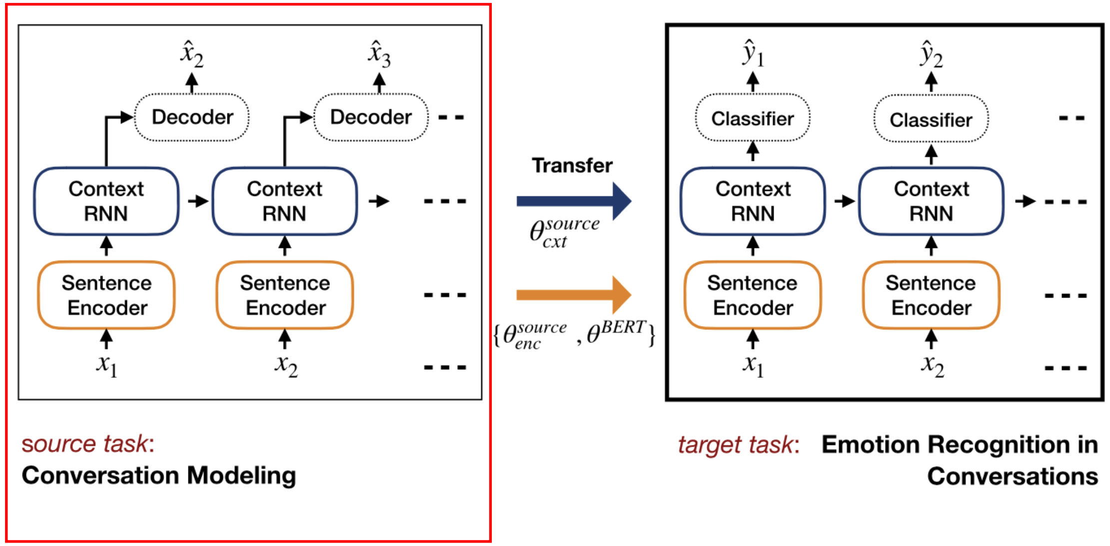


(a) (b) (c)

- (a) illustrates the presence of *emotional inertia* which occur thought self-influences in emotional states. The character *Snorri* maintains a frustrated emotional state by not being affected/influenced by the other speakers.
- conversation (b) and (c) demonstrate the role if inter-speaker influences in emotional transition across turns.
- In (b), the character *Josh* is triggered for an *emotional shift* due to influenced based on his counterpart responses.
- (c) demonstrates the effect of *mirroring* which often arises due to topical agreement between speakers.

Methodology

➢ Source : generative conversation modeling



- To perform the generative task of conversation modeling, we use the Hierarchical Recurrent Encoder-Decoder (HRED) architecture. HRED is a classic framework for seq2seq conversational response generation that models conversations in a hierarchical fashion.

- For a given conversation context with sentences $x_1, \cdots, x_t$, HRED generates the response x_{t+1} as follow:

1. Sentence encoder : It encodes each sentence in the context using an encoder RNN, such that,

$$h_t^{enc} = f_\theta^{enc}(x_t, h_{t-1}^{enc})$$

Methodology

2. Context encoder : The sentence representations are then fed into a context RNN that models the conversational context until time step t as

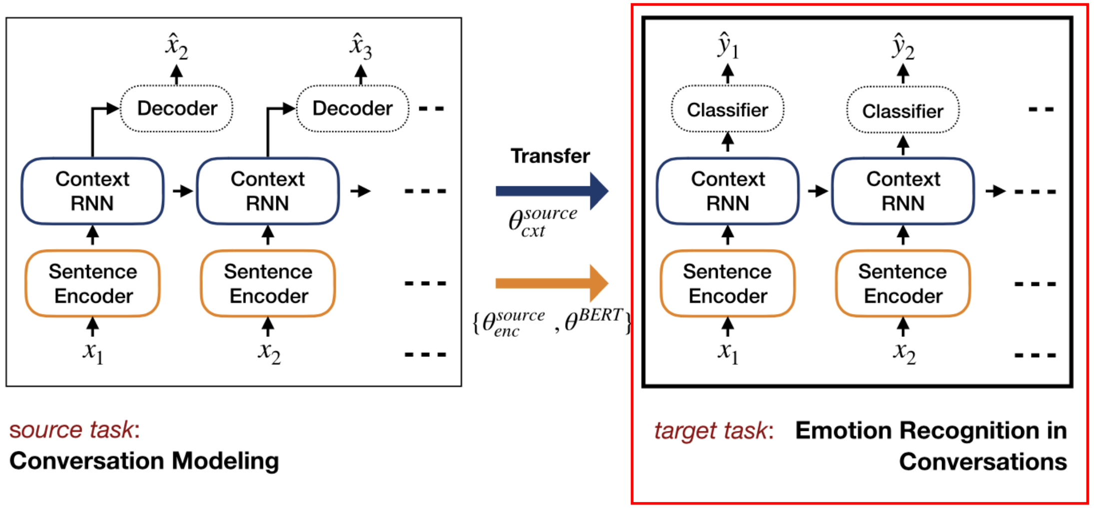
$$h_t^{ctx} = f_\theta^{ctx}(h_t^{enc}, h_{t-1}^{cxt})$$

3. Sentence decoder : Finally, an auto-regressive decoder RNN generates sentence x_{t+1}conditioned on h_t^{ctx}, i.e.,

$$\begin{aligned} p_\theta\left(x_{t+1} \middle| x_{\le t}\right) &= f_\theta^{dec}\left(x \middle|\ h_t^{cxt}\ \right) \\ &= \prod_i f_\theta^{dec}\left(x_{t+1,i} \middle|\ h_t^{cxt}, x_{t+1,<i}\right) \end{aligned}$$

- With the ith conversation being a sequence of utterances $C_i = \left[x_{i,1}, \cdots, x_{i,n_i}\right]$, HRED trains all the conversations in the dataset together by using the maximum likelihood estimation objective $argmax_\theta = \sum_i log\, p_\theta(C_i)$
- We call the parameters associated with Sentence encoder as θ_{enc}^{source}, the parameters associated with Context encoder as θ_{ctx}^{source}, the parameters associated with Sentence decoder as θ_{dec}^{source}.

➢ Target : Emotion Recognition in conversation



- The input for this task is also a conversation C with constituent utterances $[x_{i,1}, \cdots, x_{i,n_i}]$. Each x_i is associated with an emotion label $y_i \in \mathbb{Y}$.

1. Sentence encoding

    - To encode each utterance in the conversation, this paper use BERT, with its parameters represented as θ^{BERT}. BERT is chosen over the HRED sentence encoder (θ^{source}_{enc}) as its provides better performance. Hidden vector of the first token [CLS] across the considered transformer layers and mean-pool them is used as final sentence representation.

2. Context encoding

- A similar context encoder RNN is used as the source HRED model with the option to transfer the learned parameter θ_{ctx}^{source}. The context RNN transforms it as follows :

$$\begin{aligned}
\mathbf{z}_t &= \sigma\left(V^z \mathbf{h}_t^{enc} + W^z \mathbf{h}_{t-1}^{cxt} + \mathbf{b}^z\right) \\
\mathbf{r}_t &= \sigma\left(V^r \mathbf{h}_t^{enc} + W^r \mathbf{h}_{t-1}^{cxt} + \mathbf{b}^r\right) \\
\mathbf{v}_t &= \tanh\left(V^h \mathbf{h}_t^{enc} + W^h \left(\mathbf{h}_{t-1}^{cxt} \otimes \mathbf{r}_t\right) + \mathbf{b}^h\right) \\
\mathbf{h}_t^{cxt} &= (1 - \mathbf{z}_t) \otimes \mathbf{v}_t + \mathbf{z}_t \otimes \mathbf{h}_{t-1}^{cxt} \\
\mathbf{h}_t^{cxt} &= \tanh\left(W^p \mathbf{h}_t^{cxt} + \mathbf{b}^p\right)
\end{aligned}$$

- Here, $\{V^{Z,r,h}, W^{Z,r,h}, \boldsymbol{b}^{Z,r,h}\}$ are parameters for the RNN function and $\{W^p, \boldsymbol{b}^p\}$ are additional parameters of a dense layers. For our setup, adhering to size considerations, we consider our transfer parameters to be $\theta_{ctx}^{source} = \{W^{Z,r,h,p}, \boldsymbol{b}^{Z,r,h,p}\}$.

3. Classification

- For each turn in the conversation, the output from the context RNN is projected to the label-space, which provides the predicted emotion for the associated utterance. Similar to HRED, we train for all the utterances in the conversation together using the standard Cross Entropy loss. For regression targets, we utilize the Mean Square Error (MSE) loss, instead.

Datasets

Dataset			Dataset splits		
			Train	Validation	Test
Source	Cornell	#D	66,477	8310	8310
		#U	244,030	30,436	30,247
	Ubuntu	#D	898,142	18,920	19,560
		#U	6,893,060	135,747	139,775
Target	IEMOCAP	#D	120		31
		#U	5810		1623
Target	SEMAINE	#D	58		22
		#U	4386		1430
	Dailydialog	#D	11,118	1000	1000
		#U	87,170	7740	8069

	Iemocap		Dailydialog		
	Train/val	Test	Train	val	Test
hap	504	144	11,182	684	1019
sad	839	245	969	79	102
neu	1324	384	72,143	7108	6321
ang	933	170	827	77	118
exc	742	299	–	–	–
frus	1468	381	–	–	–
surp	–	–	1600	107	116
fear	–	–	146	11	17
disg	–	–	303	3	47

➢ Source task

- Cornell movie dialog corpus is a popular collection of fictional conversations extracted from movie scripts. In this dataset, conversations are sampled from a diverse set of 617 movies leading to over 83k dialogues.

- Ubuntu dialog corpus is a larger corpus with around 1 million dialogues, which, like the Cornell corpus, comprises of unstructured multi-turn dialogues based on Ubuntu chat logs (Internet Relay Chat).

Datasets

- Source task

- Primarily, this research consider the textual modality of a small-sized multimodal dataset IEMOCAP. Each conversational video is segmented into utterances and annotated with the following emotion labels: *anger, happiness, sadness, neutral, excitement, and frustration.*

- This research also analyze results on a moderately-sized emotional dialogue dataset DailyDialog with labeled emotions: *anger, happiness, sadness, surprise, fear disgust and no_emotion*. Unlike spoken utterances in IEMOCAP, the conversations are chat-based based on daily life topics.

- Finally, this research choose a regression-based dataset SEMAINE with labeled *valence, arousal, power, and expectancy*, which is a video-based corpus of human-agent emotional interactions.

- Metrics

- For ERC, this research use weighted-F-score metric for the classification tasks on IEMOCAP and DailyDialog. For DailyDialog, this research remove *no_emotion* class from the F-score calculations due to its high majority. For the regression task on SEMAINE, we take the Pearson correlation coefficient (r) as its metric. This research also provide the average best epoch (BE) on which the least validation losses are observed. A lower BE represents the model's ability to reach optimum performance in lesser training epochs.

Model variants and baselines

Variant	Initial weight		Model description
	$sent_{enc}$	cxt_{enc}	
(1)	–	–	Sentence encoders – *randomly* initialized. Context encoders – *randomly* initialized.
(2)	θ^{BERT}	–	Sentence encoders – BERT parameters. Context encoders – *randomly* initialized.
(3)	θ^{BERT}	$\theta_{cxt}^{ubuntu/cornell}$	**TL-ERC** Sentence encoders – BERT parameters. Context encoders – initialized from generative models pre-trained on Ubuntu/Cornell corpus.

- This research experiment on different variants of TL-ERC based on the parameter initialization procedure.
- Next, to compare TL-ERC with the existing literature, this research select some prior state-of-the-art models evaluated on the target datasets:

CNN, Memmet, C-LSTM, C-LSTM+Att, CMN, DialogueRNN

Result and Analysis

Variant	Initial weights		Dataset: IEMOCAP							
			10%		25%		50%		100%	
	sent_{enc}	cxt_{enc}	*F*-score	BE	*F*-score	BE	*F*-score	BE	*F*-score	BE
(1)	–	–	23.2 $_{\pm 0.4}$	48.4	41.6 $_{\pm 0.8}$	72.5	48.4 $_{\pm 0.3}$	75.1	53.8 $_{\pm 0.3}$	13.8
(2)	θ^{BERT}	–	32.4 $_{\pm 1.1}$	**11.0**	41.9 $_{\pm 0.5}$	**8.0**	49.2 $_{\pm 1.0}$	**6.3**	55.1 $_{\pm 0.6}$	**5.0**
(3)	θ^{BERT}	θ^{ubuntu}_{cxt}	35.7 $_{\pm 1.1}$	14.2	45.9 $_{\pm 2.0}$	11.2	**53.1** $_{\pm 0.7}{}^{\dagger}$	7.8	**58.8** $_{\pm 0.5}{}^{\dagger}$	5.4
		$\theta^{cornell}_{cxt}$	**36.3** $_{\pm 1.1}{}^{\dagger}$	17.0	**46.0** $_{\pm 0.5}{}^{\dagger}$	11.2	50.9 $_{\pm 1.5}$	8.2	58.5 $_{\pm 0.8}$	**5.0**

Variant	Initial weights		Dataset: DailyDialog			
			10%		100%	
	sent_{enc}	cxt_{enc}	*F*-score	BE	*F*-score	BE
(1)	–	–	33.5 $_{\pm 2.2}$	12.3	45.3 $_{\pm 1.9}$	7.9
(2)	θ^{BERT}	–	37.5 $_{\pm 1.8}$	**2.6**	47.4 $_{\pm 1.2}$	**2.4**
(3)	θ^{BERT}	θ^{ubuntu}_{cxt}	37.7 $_{\pm 3.1}$	3.1	47.1 $_{\pm .76}$	**2.4**
		$\theta^{cornell}_{cxt}$	**38.5** $_{\pm 1.5}{}^{\dagger}$	3.2	**48.0** $_{\pm 1.8}{}^{\dagger}$	**2.4**

Variant	Initial weights		Dataset: SEMAINE							
			DV		DA		DP		DE	
	sent_{enc}	cxt_{enc}	*r*	BE	*r*	BE	*r*	BE	*r*	BE
(1)	–	–	0.14	**4**	0.27	6.2	0.18	12.8	−0.03	287.4
(2)	θ^{BERT}	–	0.64	13.8	0.36	7.8	0.33	4.8	−0.03	23
(3)	θ^{BERT}	θ^{ubuntu}_{cxt}	**0.66**	10.2	0.41	**6**	0.34	3.8	−0.03	23
		$\theta^{cornell}_{cxt}$	0.65	10.2	**0.42**	8.8	**0.35**	**3.4**	−0.029	**22.7**

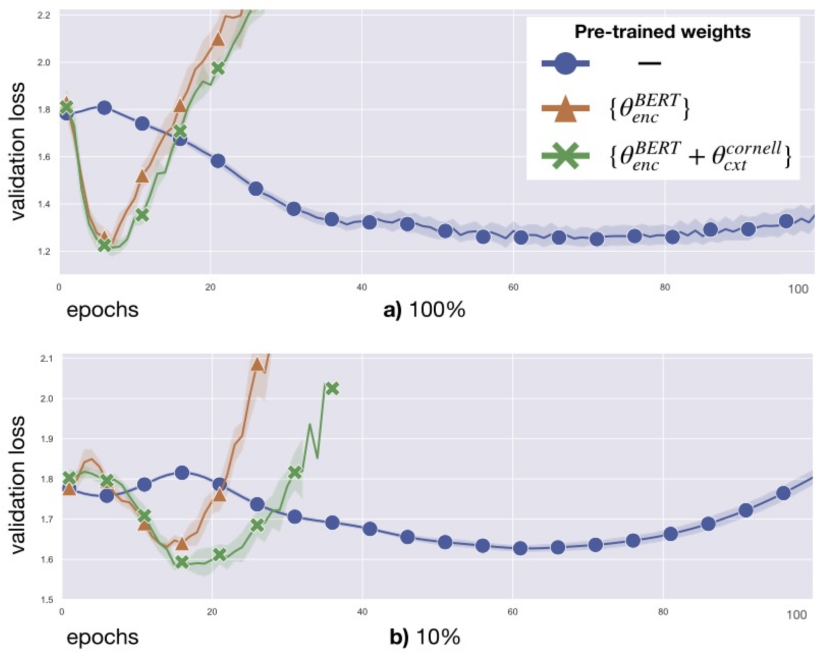
- In both datasets of IEMOCAP and DailyDialog, results indicate clear and statistically significant improvements of the models that use pre-trained weights over the randomly initialized variant.

- Similar trends are observed in the regression task based on the SE- MAINE corpus. For *valence, arousal*, and *power* dimensions, the improvement is significant. For *expectation*, the performance is marginally better but at a much lesser BE, indicating faster generalization.

- Result also indicate that the pre-trained models are significantly more robust against limited training resources compared to models trained from scratch.

Result and Analysis

Variant	Initial weight		Dataset: **IEMOCAP**							
			10%				50%			
	sent_{enc}	cxt_{enc}	split_1^*	split_2	split_3	split_4	split_1^*	split_2	split_3	split_4
(1)	–	–	23.2 ±0.4	31.5 ±0.6	25.0 ±1.7	8.8 ±1.1	48.4 ±0.3	48.5 ±1.3	49.1 ±0.9	51.3 ±0.5
(2)	θ^{BERT}	–	32.4 ±1.1	31.6 ±1.2	30.5 ±0.8	23.65 ±1.3	49.2 ±1.0	49.0 ±0.7	48.8 ±0.9	51.4 ±0.6
(3)	θ^{BERT}	θ_{cxt}^{ubuntu}	35.7 ±1.1	32.0 ±1.1	**39.0** ±0.2	**24.90** ±3.0	**53.1** ±0.7	53.2 ±1.3	52.9 ±1.9	54.2 ±0.8
		$\theta_{cxt}^{cornell}$	**36.3** ±1.1	**34.2** ±0.8	35.7 ±0.5	24.70 ±1.2	50.9 ±1.5	**54.3** ±0.8	**53.5** ±0.6	**55.4** ±1.0



- Effect of bias in random splits is investigated. the relative performance within each split follows similar trends of improvement for TL-based models.
- The trace of the validation loss indicates that the presence of weight initialization leads to faster convergence in terms of the best validation loss.

Result and Analysis

Initial weight		Dataset: IEMOCAP	
		10%	100%
sent_{enc}	cxt_{enc}	*F*-score	*F*-score
–	–	23.2 ± 0.4	53.8 ± 0.3
$\theta_{enc}^{cornell}$	–	26.3 ± 0.9	54.9 ± 0.3
	$\theta_{cxt}^{cornell}$	27.5 ± 1.3	55.1 ± 0.9
θ_{enc}^{ubuntu}	–	24.6 ± 0.9	53.2 ± 0.5
	θ_{cxt}^{ubuntu}	23.3 ± 0.8	53.7 ± 0.9
θ^{BERT}	–	32.4 ± 1.1	55.1 ± 0.6
	θ_{cxt}^{ubuntu}	35.7 ± 1.1	**58.8** $\pm$ **0.5**
	$\theta_{cxt}^{cornell}$	**36.3** $\pm$ **1.1**	58.5 ± 0.8

	Iemocap	SEMAINE			
		DV	DA	DP	DE
Models	*F*-score	*r*	*r*	*r*	*r*
CNN	48.1	−0.01	0.01	−0.01	0.19
Memnet	55.1	0.16	0.24	0.23	0.05
c-LSTM	54.9	0.14	0.23	0.25	−0.04
c-LSTM + Att	56.1	0.16	0.25	0.24	0.10
CMN	56.1	0.23	0.29	0.26	−0.02
DialogueRNN	**59.8**	0.28	0.36	0.32	**0.31**
TL-ERC	58.8	**0.66**	**0.42**	**0.35**	−0.02

- It is conducted a comparative study between the performance of models initialized with HRED-based sentence encoders (θ_{enc}^{source}) versus the BERT encoders (θ^{BERT}). Results demonstrate that BERT provides better representations, which leads to better performance.

- It is provided the results for various baselines. As seen, our proposed TL-ERC comfortably outperforms both non-contextual and contextual baselines.