DialogueRNN: An Attentive RNN for Emotion Detection in Conversations

Navonil Majumder[†≡], Soujanya Poria[‡≡], Devamanyu Hazarika[Φ],
Rada Mihalcea[▽], Alexander Gelbukh[†], Erik Cambria[‡]
[†]Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico
[‡]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[Φ]School of Computing, National University of Singapore, Singapore
[▽]Computer Science & Engineering, University of Michigan, Ann Arbor, USA
[≡] means authors contributed equally

Abstract

Emotion detection in conversations is a necessary step for a number of applications, including opinion mining over chat history, social media threads, debates, argumentation mining, understanding consumer feedback in live conversations, etc. Currently systems do not treat the parties in the conversation individually by adapting to the speaker of each utterance. In this paper, we describe a new method based on recurrent neural networks that keeps track of the individual party states throughout the conversation and uses this information for emotion classification. Our model outperforms the state of the art by a significant margin on two different datasets.

1 Introduction

Emotion detection in conversation attracts increasing attention of the community due to its applications in many important tasks such as opinion mining over chat history and social media threads in YouTube, Facebook, Twitter, etc. In this paper, we present a method based on recurrent neural network (RNN) that can cater to these needs by processing the huge amount of available conversational data.

Current systems, including the state of the art (Hazarika et al. 2018), do not distinguish different parties in a conversation in a meaningful way. They are not aware of the speaker of a given utterance. In contrast, we model individual party with party states, as the conversation flows, basing on the utterance, the context, and current party state. Our model is based on the assumption that there are three major aspects relevant to the emotion in a conversation: the speaker, the context from the preceding utterances, and the emotion of the preceding utterances. These three aspects are not necessarily independent, but their separate modeling significantly outperforms the state of the art (Table 2). In dyadic conversations, the parties have distinct roles. Hence, to extract the context, it is crucial to consider the preceding turns of both speaker and listener at a given moment (Fig. 1).

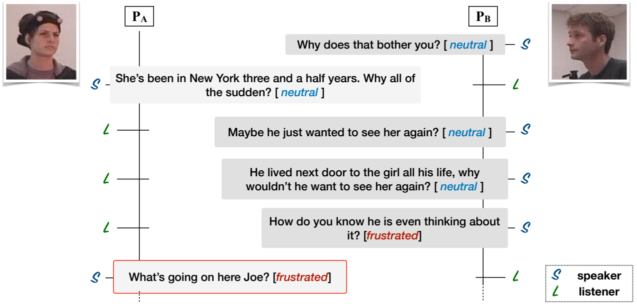
Our DialogueRNN employs three gated recurrent units (GRU) (Chung et al. 2014) to model these aspects. The incoming utterance is fed into two GRUs called global GRU and party GRU to update the context and party state, respectively. The global GRU encodes corresponding party information while encoding an utterance. Attending over this GRU gives contextual representation that has information of all preceding utterances by different parties in the conversation. The speaker state depends on this context through attention and the speaker's previous state. This ensures that at time t, the speaker state directly gets information from the speaker's previous state and global GRU which has information on the preceding parties. Finally, the updated speaker state is fed into the emotion GRU to decode the emotion representation of the given utterance, which is used for emotion classification. At time t, emotion GRU cell gets the emotion representation of $t-1$ and speaker state of t.

The emotion GRU, along with the global GRU, plays a pivotal role in inter-party relation modeling. On the other hand, party GRU models relation between two sequential states of the same party. In DialogueRNN, all these three different types of GRUs are connected in a recurrent manner. We believe that DialogueRNN outperforms state-of-the-art contextual emotion classifiers such as (Hazarika et al. 2018; Poria et al. 2017) because of better context representation.

The rest of the paper is organized as follows: Section 2 discusses related work; Section 3 provides detailed description of our model; Sections 4 and 5 present the experimental results; finally, Section 6 concludes the paper.

2 Related Work

Emotion recognition has attracted attention in various fields such as natural language processing, psychology, cognitive science, and so on (Picard 2010). Ekman (1993) found correlation between emotion and facial cues. Datcu and



Why does that bother you? [*neutral*]

She's been in New York three and a half years. Why all of the sudden? [*neutral*]

Maybe he just wanted to see her again? [*neutral*]

He lived next door to the girl all his life, why wouldn't he want to see her again? [*neutral*]

How do you know he is even thinking about it? [*frustrated*]

What's going on here Joe? [*frustrated*]

Figure 1: In this dialogue, P_A's emotion changes are influenced by the behavior of P_B.

Rothkrantz (2008) fused acoustic information with visual cues for emotion recognition. Alm, Roth, and Sproat (2005) introduced text-based emotion recognition, developed in the work of Strapparava and Mihalcea (2010). Wöllmer et al. (2010) used contextual information for emotion recognition in multimodal setting. Recently, Poria et al. (2017) successfully used RNN-based deep networks for multimodal emotion recognition, which was followed by other works (Chen et al. 2017; Zadeh et al. 2018a; 2018b).

Reproducing human interaction requires deep understanding of conversation. Ruusuvuori (2013) states that emotion plays a pivotal role in conversations. It has been argued that emotional dynamics in a conversation is an interpersonal phenomenon (Richards, Butler, and Gross 2003). Hence, our model incorporates inter-personal interactions in an effective way. Further, since conversations have a natural temporal nature, we adopt the temporal nature through recurrent network (Poria et al. 2017).

Memory networks (Sukhbaatar et al. 2015) has been successful in several NLP areas, including question answering (Sukhbaatar et al. 2015; Kumar et al. 2016), machine translation (Bahdanau, Cho, and Bengio 2014), speech recognition (Graves, Wayne, and Danihelka 2014), and so on. Thus, Hazarika et al. (2018) used memory networks for emotion recognition in dyadic conversations, where two distinct memory networks enabled inter-speaker interaction, yielding state-of-the-art performance.

3 Methodology

3.1 Problem Definition

Let there be M parties/participants $p_1, p_2, \dots, p_M$ ($M = 2$ for the datasets we used) in a conversation. The task is to predict the emotion labels (*happy*, *sad*, *neutral*, *angry*, *excited*, and *frustrated*) of the constituent utterances $u_1, u_2, \dots, u_N$, where utterance u_t is uttered by party $p_{s(u_t)}$, while s being the mapping between utterance and index of its corresponding party. Also, $u_t \in \mathbb{R}^{D_m}$ is the utterance representation, obtained using feature extractors described below.

3.2 Unimodal Feature Extraction

For a fair comparison with the state-of-the-art method, conversational memory networks (CMN) (Hazarika et al. 2018), we follow identical feature extraction procedures.

Textual Feature Extraction We employ convolutional neural networks (CNN) for textual feature extraction. Following Kim (2014), we obtain n-gram features from each utterance using three distinct convolution filters of sizes 3, 4, and 5 respectively, each having 50 feature-maps. Outputs are then subjected to max-pooling followed by rectified linear unit (ReLU) activation. These activations are concatenated and fed to a 100 dimensional dense layer, which is regarded as the textual utterance representation. This network is trained at utterance level with the emotion labels.

Audio and Visual Feature Extraction Identical to Hazarika et al. (2018), we use 3D-CNN and openSMILE (Eyben, Wöllmer, and Schuller 2010) for visual and acoustic feature extraction, respectively.

3.3 Our Model

We assume that the emotion of an utterance in a conversation depends on three major factors:

1. the speaker.
2. the context given by the preceding utterances.
3. the emotion behind the preceding utterances.

Our model DialogueRNN,[1] shown in Fig. 2a, models these three factors as follows: each party is modeled using a *party state* which changes as and when that party utters an utterance. This enables the model to track the parties' emotion dynamics through the conversations, which is related to the emotion behind the utterances. Furthermore, the context of an utterance is modeled using a *global state* (called global, because of being shared among the parties), where the preceding utterances and the party states are jointly encoded for context representation, necessary for accurate party state representation. Finally, the model infers emotion representation from the party state of the speaker along with the preceding speakers' states as context. This emotion representation is used for the final emotion classification.

We use GRU cells (Chung et al. 2014) to update the states and representations. Each GRU cell computes a hidden state defined as $h_t = GRU_*(h_{t-1}, x_t)$, where x_t is the current input and h_{t-1} is the previous GRU state. h_t also serves as the current GRU output. We provide the GRU computation details in the supplementary. GRUs are efficient networks with trainable parameters: $W_{*,\{h,x\}}^{\{r,z,c\}}$ and $b_*^{\{r,z,c\}}$.

We model the emotion representation of the current utterance as a function of the emotion representation of the previous utterance and the state of the current speaker. Finally, this emotion representation is sent to a softmax layer for emotion classification.

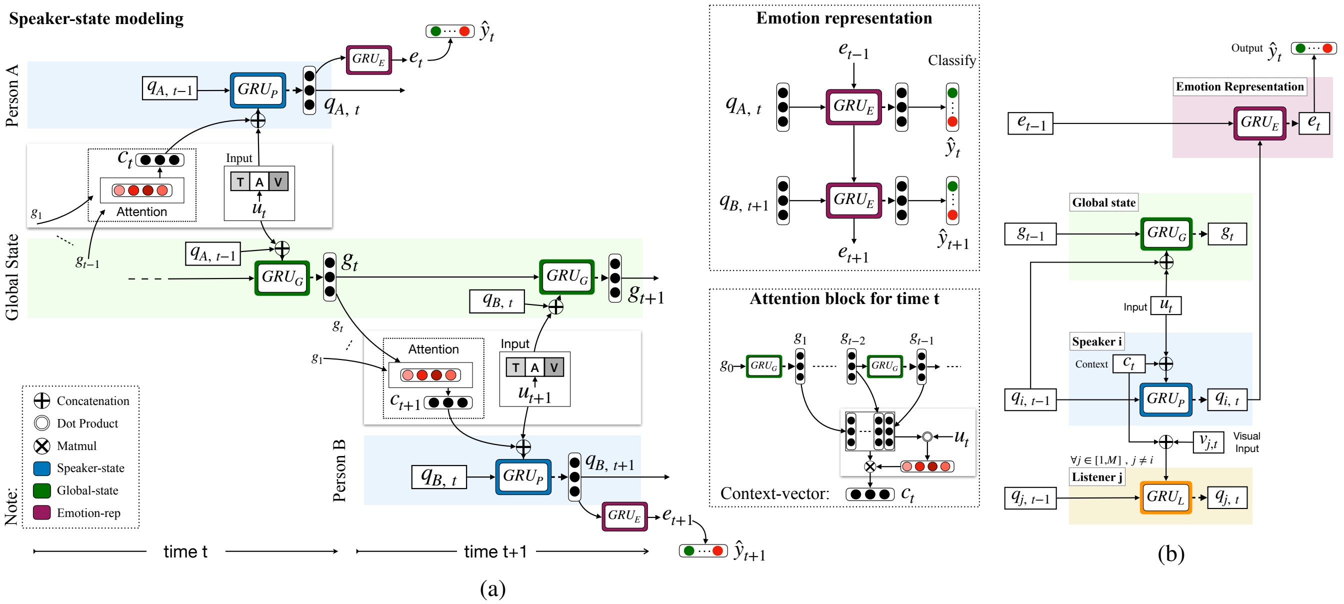
Global State (Global GRU) Global state aims to capture the context of a given utterance by jointly encoding utterance and speaker state. Each state also serves as speaker-specific utterance representation. Attending on these states facilitates the inter-speaker and inter-utterance dependencies to produce improved context representation. The current utterance u_t changes the speaker's state from $q_{s(u_t),t-1}$ to $q_{s(u_t),t}$. We capture this change with GRU cell $GRU_{\mathcal{G}}$ with output size $D_{\mathcal{G}}$, using u_t and $q_{s(u_t),t-1}$:

$$g_t = GRU_{\mathcal{G}}(g_{t-1}, (u_t \oplus q_{s(u_t),t-1})), \tag{1}$$

where $D_{\mathcal{G}}$ is the size of global state vector, $D_{\mathcal{P}}$ is the size of party state vector, $W_{\mathcal{G},h}^{\{r,z,c\}} \in \mathbb{R}^{D_{\mathcal{G}} \times D_{\mathcal{G}}}$, $W_{\mathcal{G},x}^{\{r,z,c\}} \in \mathbb{R}^{D_{\mathcal{G}} \times (D_m + D_{\mathcal{P}})}$, $b_{\mathcal{G}}^{\{r,z,c\}} \in \mathbb{R}^{D_{\mathcal{G}}}$, $q_{s(u_t),t-1} \in \mathbb{R}^{D_{\mathcal{P}}}$, $g_t, g_{t-1} \in \mathbb{R}^{D_{\mathcal{G}}}$, $D_{\mathcal{P}}$ is party state size, and $\oplus$ represents concatenation.

Party State (Party GRU) DialogueRNN keeps track of the state of individual speakers using fixed size vectors $q_1, q_2, \dots, q_M$ through out the conversation. These states are representative of the speakers' state in the conversation, relevant to emotion classification. We update these states based

[1] Implementation available at https://github.com/senticnet/conv-emotion



(a)

(b)

Figure 2: (a) DialogueRNN architecture. (b) Update schemes for global, speaker, listener, and emotion states for t^{th} utterance in a dialogue. Here, Person i is the speaker and Persons $j \in [1, M]$ and $j \neq i$ are the listeners.

on the current (at time t) role of a participant in the conversation, which is either speaker or listener, and the incoming utterance u_t. These state vectors are initialized with null vectors for all the participants. The main purpose of this module is to ensure that the model is aware of the speaker of each utterance and handle it accordingly.

Speaker Update (Speaker GRU): Speaker usually frames the response based on the context, which is the preceding utterances in the conversation. Hence, we capture context c_t relevant to the utterance u_t as follows:

$$\alpha = \text{softmax}(u_t^T W_\alpha [g_1, g_2, \ldots, g_{t-1}]), \tag{2}$$

$$\text{softmax}(x) = [e^{x_1}/\Sigma_i e^{x_i}, e^{x_2}/\Sigma_i e^{x_i}, \ldots], \tag{3}$$

$$c_t = \alpha [g_1, g_2, \ldots, g_{t-1}]^T, \tag{4}$$

where $g_1, g_2, \ldots, g_{t-1}$ are preceding $t-1$ global states ($g_i \in \mathbb{R}^{D_\mathcal{G}}$), $W_\alpha \in \mathbb{R}^{D_m \times D_\mathcal{G}}$, $\alpha^T \in \mathbb{R}^{(t-1)}$, and $c_t \in \mathbb{R}^{D_\mathcal{G}}$. In Eq. (2), we calculate attention scores α over the previous global states representative of the previous utterances. This assigns higher attention scores to the utterances emotionally relevant to u_t. Finally, in Eq. (4) the context vector c_t is calculated by pooling the previous global states with α.

Now, we employ a GRU cell $GRU_\mathcal{P}$ to update the current speaker state $q_{s(u_t),t-1}$ to the new state $q_{s(u_t),t}$ based on incoming utterance u_t and context c_t using GRU cell $GRU_\mathcal{P}$ of output size $D_\mathcal{P}$

$$q_{s(u_t),t} = GRU_\mathcal{P}(q_{s(u_t),t-1}, (u_t \oplus c_t)), \tag{5}$$

where $W_{\mathcal{P},h}^{\{r,z,c\}} \in \mathbb{R}^{D_\mathcal{P} \times D_\mathcal{P}}$, $W_{\mathcal{P},x}^{\{r,z,c\}} \in \mathbb{R}^{D_\mathcal{P} \times (D_m + D_\mathcal{G})}$, $b_\mathcal{P}^{\{r,z,c\}} \in \mathbb{R}^{D_\mathcal{P}}$, and $q_{s(u_t),t}, q_{s(u_t),t-1} \in \mathbb{R}^{D_\mathcal{P}}$. This encodes the information on the current utterance along with its context from the global GRU into the speaker's state $q_{s(u_t)}$, which helps in emotion classification down the line.

Listener Update: Listener state models the listeners' change of state due to the speaker's utterance. We tried two listener state update mechanisms:

- Simply keep the state of the listener unchanged, that is

$$\forall i \neq s(u_t), q_{i,t} = q_{i,t-1}. \tag{6}$$

- Employ another GRU cell $GRU_\mathcal{L}$ to update the listener state based on listener visual cues (facial expression) $v_{i,t}$ and its context c_t, as

$$\forall i \neq s(u_t), q_{i,t} = GRU_\mathcal{L}(q_{i,t-1}, (v_{i,t} \oplus c_t)), \tag{7}$$

  where $v_{i,t} \in \mathbb{R}^{D_\mathcal{V}}$, $W_{\mathcal{L},h}^{\{r,z,c\}} \in \mathbb{R}^{D_\mathcal{P} \times D_\mathcal{P}}$, $W_{\mathcal{L},x}^{\{r,z,c\}} \in \mathbb{R}^{D_\mathcal{P} \times (D_\mathcal{V} + D_\mathcal{G})}$, and $b_\mathcal{L}^{\{r,z,c\}} \in \mathbb{R}^{D_\mathcal{P}}$. Listener visual features of party i at time t $v_{i,t}$ are extracted using the model introduced by Arriaga, Valdenegro-Toro, and Plöger (2017), pretrained on FER2013 dataset, where feature size $D_\mathcal{V} = 7$.

The simpler first approach turns out to be sufficient, since the second approach yields very similar result while increasing number of parameters. This is due to the fact that a listener becomes relevant to the conversation only when he/she speaks. In other words, a silent party has no influence in a conversation. Now, when a party speaks, we update his/her state q_i with context c_t which contains relevant information on all the preceding utterances, rendering explicit listener state update unnecessary. This is shown in Table 2.

Emotion Representation (Emotion GRU) We infer the emotionally relevant representation e_t of utterance u_t from the speaker's state $q_{s(u_t),t}$ and the emotion representation of the previous utterance e_{t-1}. Since context is important to the emotion of the incoming utterance u_t, e_{t-1} feeds fine-tuned emotionally relevant contextual information from other the party states $q_{s(u_{<t}),<t}$ into the emotion representation e_t. This establishes a connection between the speaker state and the other party states. Hence, we model e_t with a GRU cell ($GRU_{\mathcal{E}}$) with output size $D_{\mathcal{E}}$ as

$$e_t = GRU_{\mathcal{E}}(e_{t-1}, q_{s(u_t),t}), \tag{8}$$

where $D_{\mathcal{E}}$ is the size of emotion representation vector, $e_{\{t,t-1\}} \in \mathbb{R}^{D_{\mathcal{E}}}$, $W_{\mathcal{E},h}^{\{r,z,c\}} \in \mathbb{R}^{D_{\mathcal{E}} \times D_{\mathcal{E}}}$, $W_{\mathcal{E},x}^{\{r,z,c\}} \in \mathbb{R}^{D_{\mathcal{E}} \times D_{\mathcal{P}}}$, and $b_{\mathcal{E}}^{\{r,z,c\}} \in \mathbb{R}^{D_{\mathcal{E}}}$.

Since speaker state gets information from global states, which serve as speaker-specific utterance representation, one may claim that this way the model already has access to the information on other parties. However, as shown in the ablation study (Section 5.6) emotion GRU helps to improve the performance by directly linking states of preceding parties. Further, we believe that speaker and global GRUs ($GRU_{\mathcal{P}}$, $GRU_{\mathcal{G}}$) jointly act similar to an encoder, whereas emotion GRU serves as a decoder.

Emotion Classification We use a two-layer perceptron with a final softmax layer to calculate $c = 6$ emotion-class probabilities from emotion representation e_t of utterance u_t and then we pick the most likely emotion class:

$$l_t = \text{ReLU}(W_l e_t + b_l), \tag{9}$$

$$\mathcal{P}_t = \text{softmax}(W_{smax} l_t + b_{smax}), \tag{10}$$

$$\hat{y_t} = \underset{i}{\text{argmax}}(\mathcal{P}_t[i]), \tag{11}$$

where $W_l \in \mathbb{R}^{D_l \times D_{\mathcal{E}}}$, $b_l \in \mathbb{R}^{D_l}$, $W_{smax} \in \mathbb{R}^{c \times D_l}$, $b_{smax} \in \mathbb{R}^c$, $\mathcal{P}_t \in \mathbb{R}^c$, and $\hat{y_t}$ is the predicted label for utterance u_t.

Training We use categorical cross-entropy along with L2-regularization as the measure of loss (L) during training:

$$L = -\frac{1}{\sum_{s=1}^{N} c(s)} \sum_{i=1}^{N} \sum_{j=1}^{c(i)} \log \mathcal{P}_{i,j}[y_{i,j}] + \lambda \left\|\theta\right\|_2, \tag{12}$$

where N is the number of samples/dialogues, $c(i)$ is the number of utterances in sample i, $\mathcal{P}_{i,j}$ is the probability distribution of emotion labels for utterance j of dialogue i, $y_{i,j}$ is the expected class label of utterance j of dialogue i, λ is the L2-regularizer weight, and θ is the set of trainable parameters where

$$\theta = \{W_\alpha, W_{\mathcal{P},\{h,x\}}^{\{r,z,c\}}, b_{\mathcal{P}}^{\{r,z,c\}}, W_{\mathcal{G},\{h,x\}}^{\{r,z,c\}}, b_{\mathcal{G}}^{\{r,z,c\}}, W_{\mathcal{E},\{h,x\}}^{\{r,z,c\}},$$
$$b_{\mathcal{E}}^{\{r,z,c\}}, W_l, b_l, W_{smax}, b_{smax}\}.$$

We used stochastic gradient descent based Adam (Kingma and Ba 2014) optimizer to train our network. Hyperparameters are optimized using grid search (values are added to the supplementary material).

3.4 DialogueRNN Variants

We use DialogueRNN (Section 3.3) as the basis for the following models:

DialogueRNN + Listener State Update (DialogueRNN$_l$): This variant updates the listener state based on the the resulting speaker state $q_{s(u_t),t}$, as described in Eq. (7).

Bidirectional DialogueRNN (BiDialogueRNN): Bidirectional DialogueRNN is analogous to bidirectional RNNs, where two different RNNs are used for forward and backward passes of the input sequence. Outputs from the RNNs are concatenated in sequence level. Similarly, in BiDialogueRNN, the final emotion representation contains information from both past and future utterances in the dialogue through forward and backward DialogueRNNs respectively, which provides better context for emotion classification.

DialogueRNN + attention (DialogueRNN+Att): For each emotion representation e_t, attention is applied over all surrounding emotion representations in the dialogue by matching them with e_t (Eqs. (13) and (14)). This provides context from the relevant (based on attention score) future and preceding utterances.

Bidirectional DialogueRNN + Emotional attention (BiDialogueRNN+Att): For each emotion representation e_t of BiDialogueRNN, attention is applied over all the emotion representations in the dialogue to capture context from the other utterances in dialogue:

$$\beta_t = \text{softmax}(e_t^T W_\beta [e_1, e_2, \ldots, e_N]), \tag{13}$$

$$\tilde{e}_t = \beta_t [e_1, e_2, \ldots, e_N]^T, \tag{14}$$

where $e_t \in \mathbb{R}^{2D_{\mathcal{E}}}$, $W_\beta \in \mathbb{R}^{2D_{\mathcal{E}} \times 2D_{\mathcal{E}}}$, $\tilde{e}_t \in \mathbb{R}^{2D_{\mathcal{E}}}$, and $\beta_t^T \in \mathbb{R}^N$. Further, $\tilde{e}_t$ are fed to a two-layer perceptron for emotion classification, as in Eqs. (9) to (11).

4 Experimental Setting

4.1 Datasets Used

We use two emotion detection datasets IEMOCAP (Busso et al. 2008) and AVEC (Schuller et al. 2012) to evaluate DialogueRNN. We partition both datasets into train and test sets with roughly 80/20 ratio such that the partitions do not share any speaker. Table 1 shows the distribution of train and test samples for both dataset.

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

Table 1: Dataset split ((train + val) / test ≈ 80%/20%).

IEMOCAP: IEMOCAP (Busso et al. 2008) dataset contains videos of two-way conversations of ten unique speakers, where only the first eight speakers from session one to four belong to the train-set. Each video contains a single dyadic dialogue, segmented into utterances. The utterances are annotated with one of six emotion labels, which are happy, sad, neutral, angry, excited, and frustrated.

AVEC: AVEC (Schuller et al. 2012) dataset is a modification of SEMAINE database (McKeown et al. 2012) containing interactions between humans and artificially intelligent agents. Each utterance of a dialogue is annotated with four real valued affective attributes: valence ($[-1, 1]$), arousal ($[-1, 1]$), expectancy ($[-1, 1]$), and power ($[0, \infty)$). The annotations are available every 0.2 seconds in the original database. However, in order to adapt the annotations to our need of utterance-level annotation, we averaged the attributes over the span of an utterance.

4.2 Baselines and State of the Art

For a comprehensive evaluation of DialogueRNN, we compare our model with the following baseline methods:

c-LSTM (Poria et al. 2017): Biredectional LSTM (Hochreiter and Schmidhuber 1997) is used to capture the context from the surrounding utterances to generate context-aware utterance representation. However, this model does not differentiate among the speakers.

c-LSTM+Att (Poria et al. 2017): In this variant attention is applied applied to the c-LSTM output at each timestamp by following Eqs. (13) and (14). This provides better context to the final utterance representation.

TFN (Zadeh et al. 2017): This is specific to multimodal scenario. Tensor outer product is used to capture inter-modality and intra-modality interactions. This model does not capture context from surrounding utterances.

MFN (Zadeh et al. 2018a): Specific to multimodal scenario, this model utilizes multi-view learning by modeling view-specific and cross-view interactions. Similar to TFN, this model does not use contextual information.

CNN (Kim 2014): This is identical to our textual feature extractor network (Section 3.2) and it does not use contextual information from the surrounding utterances.

Memnet (Sukhbaatar et al. 2015): As described in Hazarika et al. (2018), the current utterance is fed to a memory network, where the memories correspond to preceding utterances. The output from the memory network is used as the final utterance representation for emotion classification.

CMN (Hazarika et al. 2018): This state-of-the-art method models utterance context from dialogue history using two distinct GRUs for two speakers. Finally, utterance representation is obtained by feeding the current utterance as query to two distinct memory networks for both speakers.

4.3 Modalities

We evaluated our model primarily on textual modality. However, to substantiate efficacy of our model in multimodal scenario, we also experimented with multimodal features.

5 Results and Discussion

We compare DialogueRNN and its variants with the baselines for textual data in Table 2. As expected, on average DialogueRNN outperforms all the baseline methods, including the state-of-the-art CMN, on both of the datasets.

5.1 Comparison with the State of the Art

We compare the performance of DialogueRNN against the performance of the state-of-the-art CMN on IEMOCAP and AVEC datasets for textual modality.

IEMOCAP As evidenced by Table 2, for IEMOCAP dataset, our model surpasses the state-of-the-art method CMN by 2.77% accuracy and 3.76% f1-score on average. We think that this enhancement is caused by the fundamental differences between CMN and DialogueRNN, which are

1. party state modeling with $GRU_{\mathcal{P}}$ in Eq. (5),
2. speaker specific utterance treatment in Eqs. (1) and (5),
3. and global state capturing with $GRU_{\mathcal{G}}$ in Eq. (1).

Since we deal with six unbalanced emotion labels, we also explored the model performance for individual labels. DialogueRNN outperforms the state-of-the-art method CMN in five out of six emotion classes by significant margin. For *frustrated* class, DialogueRNN lags behind CMN by 1.23% f1-score. We think that DialogueRNN may surpass CMN using a standalone classifier for *frustrated* class. However, it can be observed in Table 2 that some of the other variants of DialogueRNN, like BiDialogueRNN has already outperformed CMN for *frustrated* class.

AVEC DialogueRNN outperforms CMN for valence, arousal, expectancy, and power attributes; see Table 2. It yields significantly lower mean absolute error (MAE) and higher Pearson correlation coefficient (r) for all four attributes. We believe this to be due to the incorporation of party state and emotion GRU, which are missing from CMN.

5.2 DialogueRNN vs. DialogueRNN Variants

We discuss the performance of different DialogueRNN variants on IEMOCAP and AVEC datasets for textual modality.

DialogueRNN$_l$: Following Table 2, using explicit listener state update yields slightly worse performance than regular DialogueRNN. This is true for both IEMOCAP and AVEC datasets in general. However, the only exception to this trend is for *happy* emotion label for IEMOCAP, where DialogueRNN$_l$ outperforms DialogueRNN by 1.71% f1-score. We surmise that, this is due to the fact that a listener becomes relevant to the conversation only when he/she speaks. Now, in DialogueRNN, when a party speaks, we update his/her state q_i with context c_t which contains relevant information on all the preceding utterances, rendering explicit listener state update of DialogueRNN$_l$ unnecessary.

BiDialogueRNN: Since BiDialogueRNN captures context from the future utterances, we expect improved performance from it over DialogueRNN. This is confirmed in Table 2, where BiDialogueRNN outperforms DialogueRNN on average on both datasets.

DialogueRNN+Attn: DialogueRNN+Attn also uses information from the future utterances. However, here we take information from both past and future utterances by matching them with the current utterance and calculating attention score over them. This provides relevance to emotionally important context utterances, yielding better performance than

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**

Table 2: Comparison with the baseline methods for textual modality; Acc. = Accuracy, MAE = Mean Absolute Error, r = Pearson correlation coefficient; bold font denotes the best performances. Average(w) = Weighted average.

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**

Table 3: Comparison with the baselines for trimodal (T+V+A) scenario. BiDialogueRNN+att$_{MM}$ = BiDialogueRNN+att in multimodal setting.

BiDialogueRNN. The improvement over BiDialogueRNN is 1.23% f1-score for IEMOCAP and consistently lower MAE and higher r in AVEC.

BiDialogueRNN+Attn: Since this setting generates the final emotion representation by attending over the emotion representation from BiDialogueRNN, we expect better performance than both BiDialogueRNN and DialogueRNN+Attn. This is confirmed in Table 2, where this setting performs the best in general than any other methods discussed, on both datasets. This setting yields 6.62% higher f1-score on average than the state-of-the-art CMN and 2.86% higher f1-score than vanilla DialogueRNN for IEMOCAP dataset. For AVEC dataset also, this setting gives the best performance across all the four attributes.

5.3 Multimodal Setting

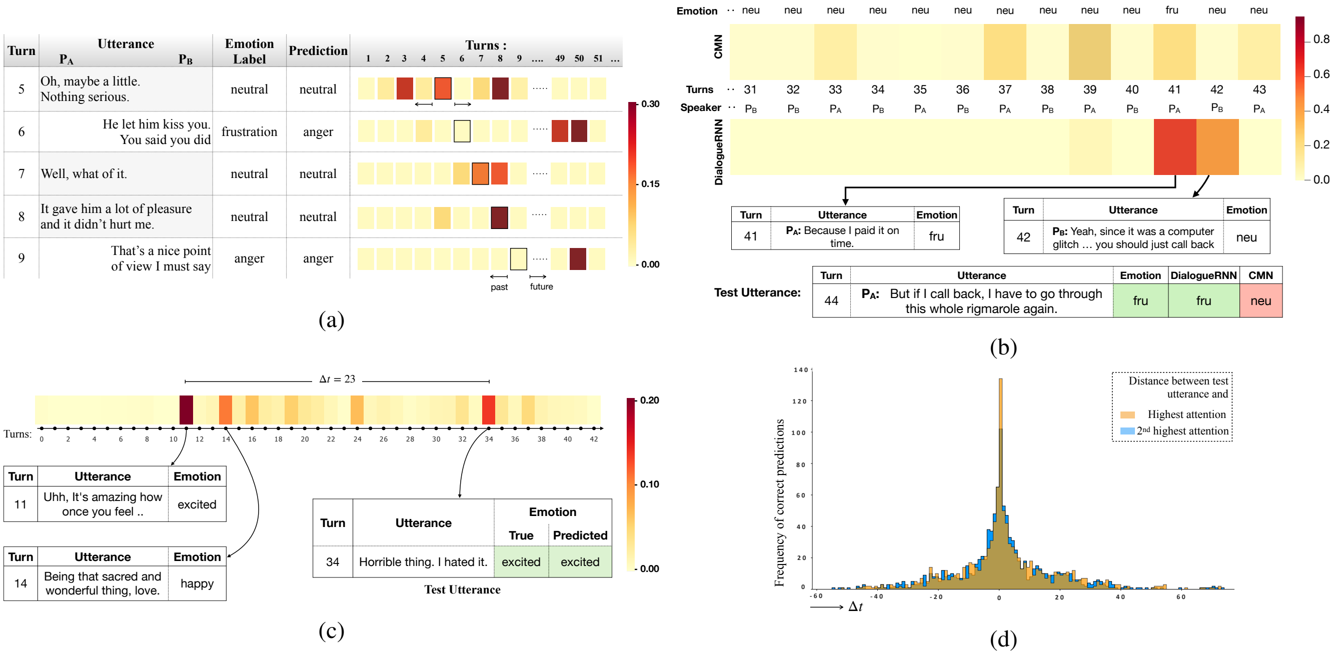
As both IEMOCAP and AVEC dataset contain multimodal information, we have evaluated DialogueRNN on multimodal features as used and provided by Hazarika et al. (2018). We use concatenation of the unimodal features as a fusion method by following Hazarika et al. (2018), since fusion mechanism is not a focus of this paper. Now, as we can see in Table 3, DialogueRNN significantly outperforms the strong baselines and state-of-the-art method CMN.

5.4 Case Studies

Dependency on preceding utterances (DialogueRNN) One of the crucial components of DialogueRNN is its attention module over the outputs of global GRU ($GRU_\mathcal{G}$). Figure 3b shows the α attention vector (Eq. (2)) over the history of a given test utterance compared with the attention vector from the CMN model. The attention of our model is more focused compared with CMN: the latter gives diluted attention scores leading to misclassifications. We observe this trend of focused-attention across cases and posit that it can be interpreted as a confidence indicator. Further in this example, the test utterance by P_A (turn 44) comprises of a change in emotion from *neutral* to *frustrated*. DialogueRNN anticipates this correctly by attending to turn 41 and 42 that are spoken by P_A and P_B, respectively. These two utterances provide self and inter-party influences that trigger the emotional shift. CMN, however, fails to capture such dependencies and wrongly predicts *neutral* emotion.

Dependency on future utterances (BiDialogueRNN+Att) Fig. 3a visualizes the β (Eq. (13)) attention over the emotion representations e_t for a segment of a conversation between a couple. In the discussion, the woman (P_A) is initially at a neutral state, whereas the man (P_B) is angry throughout. The figure reveals that the emotional attention of the woman is localized to the duration of her neutral state (turns 1-16 approximately). For example, in the dialogue, turns $5, 7$, and 8 strongly attend to turn 8. Interestingly, turn 5 attends to both past (turn 3) and future (turn 8) utterances. Similar trend across other utterances establish inter-dependence between emotional states of future and past utterances. The beneficial consideration of future utterances through $GRU_\mathcal{E}$ is also apparent through turns $6, 9$. These utterances focus on the distant future (turn $49, 50$) where the man is at an enraged state, thus capturing emotional correlations across time. Although, turn 6 is misclassified by our model, it still manages to infer a related emotional state (*anger*) against the correct state (*frustrated*). We analyze more of this trend in section 5.5.

Dependency on distant context For all correct predictions in the IEMOCAP test set in Fig. 3d we summarize the distribution over the relative distance between test utterance and (2^{nd}) highest attended utterance – either in the history or future – in the conversation. This reveals a decreasing trend with the highest dependence being within the local context. However, a significant portion of the test utterances (~ 18%), attend to utterances that are 20 to 40 turns away from themselves, which highlights the important role of long-term emotional dependencies. Such cases primarily occur in conversations that maintain a specific affective tone and do not incur frequent emotional shifts. Fig. 3c demonstrates a case of long-term context dependency. The presented conversation maintains a *happy* mood throughout the dialogue. Although the 34^{th} turn comprising the sentence *Horrible thing. I hated it.* seems to be a negative expres-



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

(a)

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

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

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

(b)

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

(c)

(d)

Figure 3: (a) Illustration of the β attention over emotion representations e_t; (b) Comparison of attention scores over utterance history of CMN and DialogueRNN (α attention). (c) An example of long-term dependency among utterances. (d) Histogram of Δt = distance between the target utterance and its context utterance based on β attention scores.

sion, when seen with the global context, it reveals the *excitement* present in the speaker. To disambiguate such cases, our model attends to distant utterances in the past (turn 11, 14) which serve as prototypes of the emotional tonality of the overall conversation.

5.5 Error Analysis

A noticeable trend in the predictions is the high level of cross-predictions amongst related emotions. Most of the misclassifications by the model for *happy* emotion are for *excited* class. Also, *anger* and *frustrated* share misclassifications amongst each other. We suspect this is due to subtle difference between those emotion pairs, resulting in harder disambiguation. Another class with high rate of false-positives is the *neutral* class. Primary reason for this could be its majority in the class distribution over the considered emotions.

At the dialogue level, we observe that a significant amount of errors occur at turns having a change of emotion from the previous turn of the same party. Across all the occurrences of these *emotional-shifts* in the testing set, our model correctly predicts 47.5% instances. This stands less as compared to the 69.2% success that it achieves at regions of no *emotional-shift*. Changes in emotions in a dialogue is a complex phenomenon governed by latent dynamics. Further improvement of these cases remain as an open area of research.

5.6 Ablation Study

The main novelty of our method is the introduction of party state and emotion GRU ($GRU_{\mathcal{E}}$). To comprehensively study the impact of these two components, we remove them one at a time and evaluate their impact on IEMOCAP.

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

Table 4: Ablated DialogueRNN for IEMOCAP dataset.

As expected, following Table 4, party state stands very important, as without its presence the performance falls by 4.33%. We suspect that party state helps in extracting useful contextual information relevant to parties' emotion.

Emotion GRU is also impactful, but less than party state, as its absence causes performance to fall by only 2.51%. We believe the reason to be the lack of context flow from the other parties' states through the emotion representation of the preceding utterances.

6 Conclusion

We have presented an RNN-based neural architecture for emotion detection in a conversation. In contrast to the state-of-the-art method, CMN, our method treats each incoming utterance taking into account characteristics of the speaker, which gives finer context to the utterance. Our model outperforms the current state of the art on two distinct datasets in both textual and multimodal setting. Our method is designed to be scalable to multi-party setting with more than two speakers, though we could not test it due to unavailability of a multi-party conversation dataset with emotion labels. This is left to our future work.

References

Alm, C. O.; Roth, D.; and Sproat, R. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 579–586. Association for Computational Linguistics.

Arriaga, O.; Valdenegro-Toro, M.; and Plöger, P. 2017. Real-time convolutional neural networks for emotion and gender classification. *CoRR* abs/1710.07557.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42(4):335–359.

Chen, M.; Wang, S.; Liang, P. P.; Baltrušaitis, T.; Zadeh, A.; and Morency, L.-P. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 163–171. ACM.

Chung, J.; Gülçehre, Ç.; Cho, K.; and Bengio, Y. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *CoRR* abs/1412.3555.

Datcu, D., and Rothkrantz, L. 2008. Semantic audio-visual data fusion for automatic emotion recognition. *Euromedia'2008*.

Ekman, P. 1993. Facial expression and emotion. *American psychologist* 48(4):384.

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, 1459–1462. ACM.

Graves, A.; Wayne, G.; and Danihelka, I. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.-P.; and Zimmermann, R. 2018. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2122–2132. New Orleans, Louisiana: Association for Computational Linguistics.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Kim, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kingma, D. P., and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980.

Kumar, A.; Irsoy, O.; Ondruska, P.; Iyyer, M.; Bradbury, J.; Gulrajani, I.; Zhong, V.; Paulus, R.; and Socher, R. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning*, 1378–1387.

McKeown, G.; Valstar, M.; Cowie, R.; Pantic, M.; and Schroder, M. 2012. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing* 3(1):5–17.

Picard, R. W. 2010. Affective computing: From laughter to ieee. *IEEE Transactions on Affective Computing* 1(1):11–17.

Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; and Morency, L.-P. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 873–883. Vancouver, Canada: Association for Computational Linguistics.

Richards, J. M.; Butler, E. A.; and Gross, J. J. 2003. Emotion regulation in romantic relationships: The cognitive consequences of concealing feelings. *Journal of Social and Personal Relationships* 20(5):599–620.

Ruusuvuori, J. 2013. Emotion, affect and conversation. *The handbook of conversation analysis* 330–349.

Schuller, B.; Valster, M.; Eyben, F.; Cowie, R.; and Pantic, M. 2012. AVEC 2012: The Continuous Audio/Visual Emotion Challenge. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, 449–456. New York, NY, USA: ACM.

Strapparava, C., and Mihalcea, R. 2010. Annotating and identifying emotions in text. In *Intelligent Information Access*. Springer. 21–38.

Sukhbaatar, S.; Szlam, A.; Weston, J.; and Fergus, R. 2015. End-to-end Memory Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, 2440–2448. Cambridge, MA, USA: MIT Press.

Wöllmer, M.; Metallinou, A.; Eyben, F.; Schuller, B.; and Narayanan, S. S. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *INTERSPEECH 2010*.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114. Copenhagen, Denmark: Association for Computational Linguistics.

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018a. Memory Fusion Network for Multi-view Sequential Learning. In *AAAI Conference on Artificial Intelligence*, 5634–5641.

Zadeh, A.; Liang, P. P.; Poria, S.; Vij, P.; Cambria, E.; and Morency, L.-P. 2018b. Multi-attention recurrent network for human communication comprehension. In *AAAI Conference on Artificial Intelligence*, 5642–5649.

Hyperparameter	DialogueRNN	BiDialogueRNN	DialogueRNN+Att	BiDialogueRNN+Att
$D_{\mathcal{G}}$	300	150	150	150
$D_{\mathcal{P}}$	400	150	150	150
$D_{\mathcal{E}}$	400	100	100	100
D_l	200	100	100	100
lr	0.0001	0.0001	0.0001	0.0001
λ	0.00001	0.00001	0.00001	0.00001

Table 5: Hyperparameter values for DialogueRNN variants; lr = learning rate.

7 Supplementary Material

7.1 GRU Details

We use GRU cells are defined as $h_t = GRU_*(h_{t-1}, x_t)$, where:

$$r_t = \sigma(W^r_{*,h} h_{t-1} + W^r_{*,x} x_t + b^r_*), \tag{15}$$

$$z_t = \sigma(W^z_{*,h} h_{t-1} + W^z_{*,x} x_t + b^z_*), \tag{16}$$

$$c_t = \tanh(W^c_{*,h}(r_t \circ h_{t-1}) + W^c_{*,x} x_t + b^c_*), \tag{17}$$

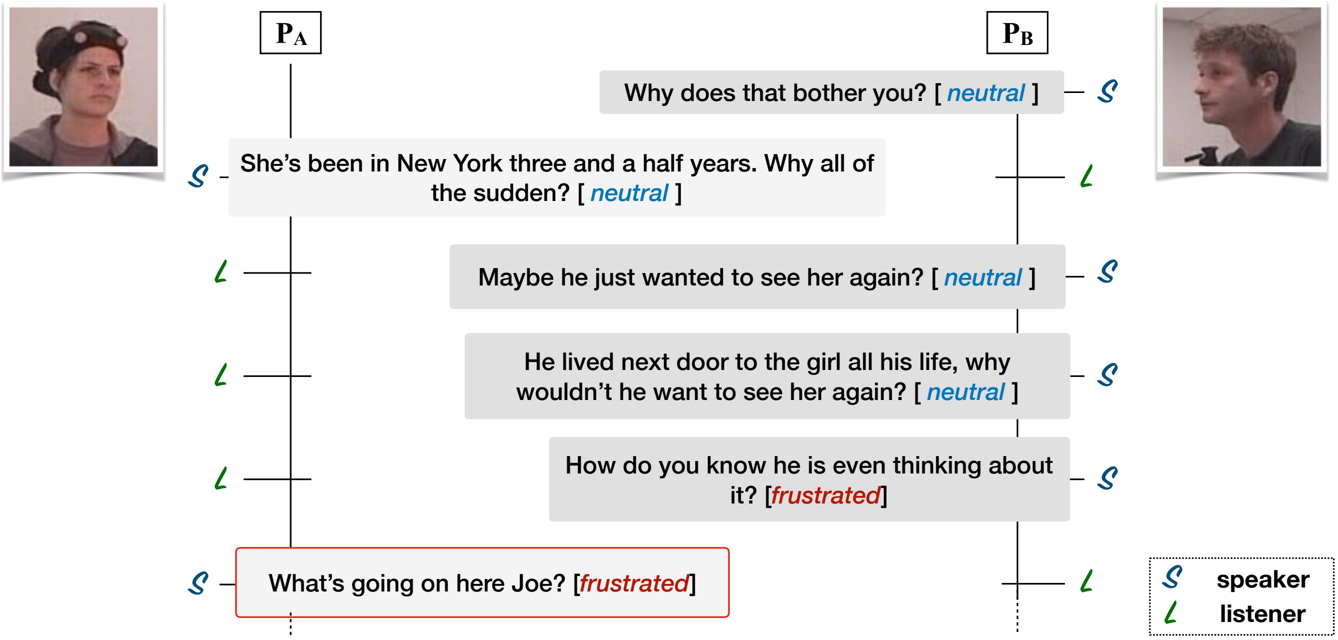
$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ c_t. \tag{18}$$

x_t is the current input, h_{t-1} is the previous GRU output, and h_t is the current GRU output. $W^{\{r,z,c\}}_{*,\{h,x\}}$ and $b^{\{r,z,c\}}_*$ are GRU parameters. r_t, z_t are refresh gate and update gate respectively. c_t is the candidate output. $\circ$ stands for hadamard product.

Algorithm 1 DialogueRNN algorithm

```
procedure DialogueRNN($U$, $S$)    ▷ $U$=utterances in the conversation, $S$=speakers
    Initialize the participant states with null vector:
    for i:[1,$M$] do
        $q_{0,i} \leftarrow 0$
    Set the initial global and emotional state as null vector:
    $g_0 \leftarrow 0$
    $e_0 \leftarrow 0$
    Pass the dialogue through RNN:
    for t:[1,$N$] do
        $e_t, g, q_t \leftarrow$ DialogueCell($e_{t-1}, g, q_{t-1}, U_t, S_t$)
    return $e$

procedure DialogueCell($\_e$, $g$, $\_q$, $u$, $s$)
    Update global state:
    $g_t \leftarrow GRU_{\mathcal{G}}(g_{t-1}, u \oplus \_q_s)$
    Get context from preceding global states:
    $c \leftarrow Attention([g_1, g_2, \ldots, g_{t-1}], u)$
    Update participant states:
    for i:[1,$M$] do
        if $i = s$ then
            Update speaker state:
            $q_s \leftarrow GRU_{\mathcal{P}}(\_q_s, u \oplus c)$
        else
            Update listener state:
            $q_i \leftarrow \_q_i$
    Update emotion representation:
    $e \leftarrow GRU_{\mathcal{E}}(\_e, q_s)$
    return $e, g, q$
```



Why does that bother you? [*neutral*]

She's been in New York three and a half years. Why all of the sudden? [*neutral*]

Maybe he just wanted to see her again? [*neutral*]

He lived next door to the girl all his life, why wouldn't he want to see her again? [*neutral*]

How do you know he is even thinking about it? [*frustrated*]

What's going on here Joe? [*frustrated*]

Figure 4: In this dialogue, P_A changes emotion influenced by the behavior of P_B.

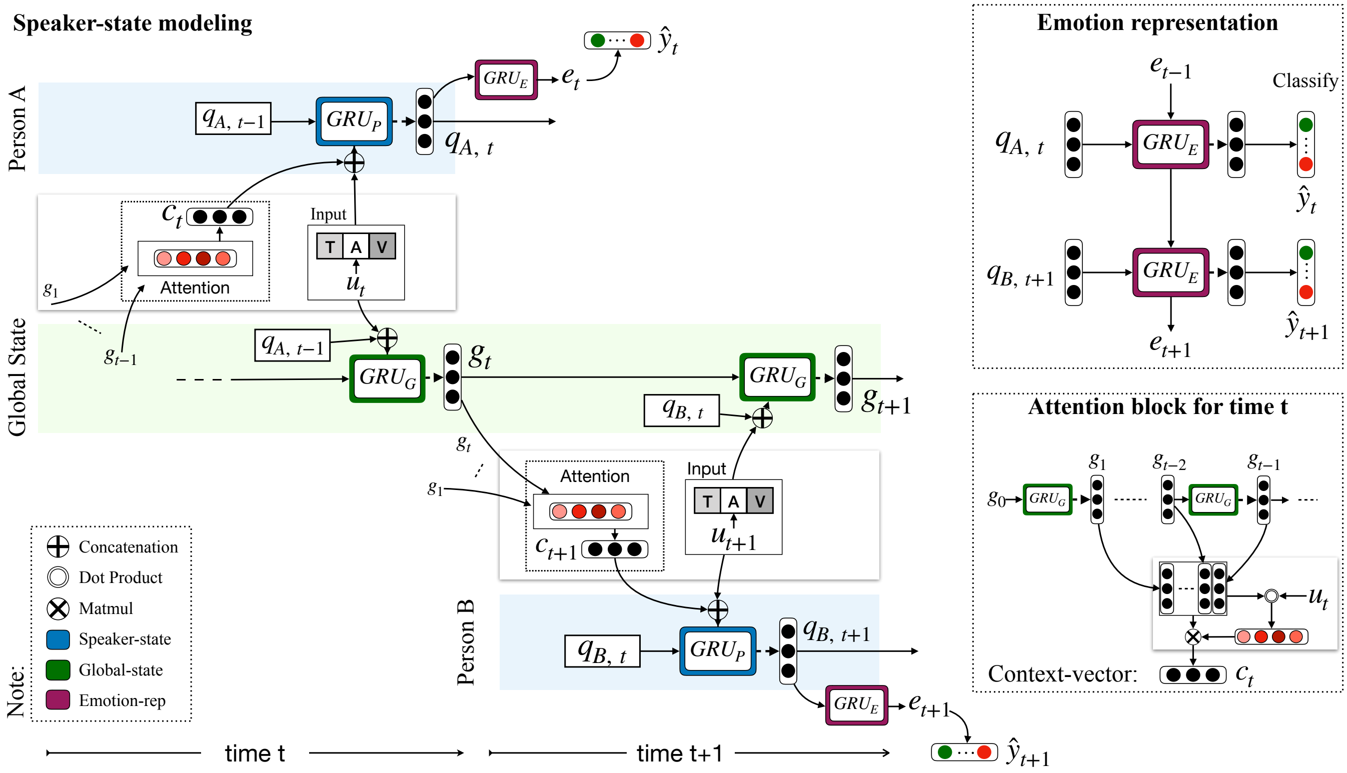


Figure 5: DialogueRNN architecture.

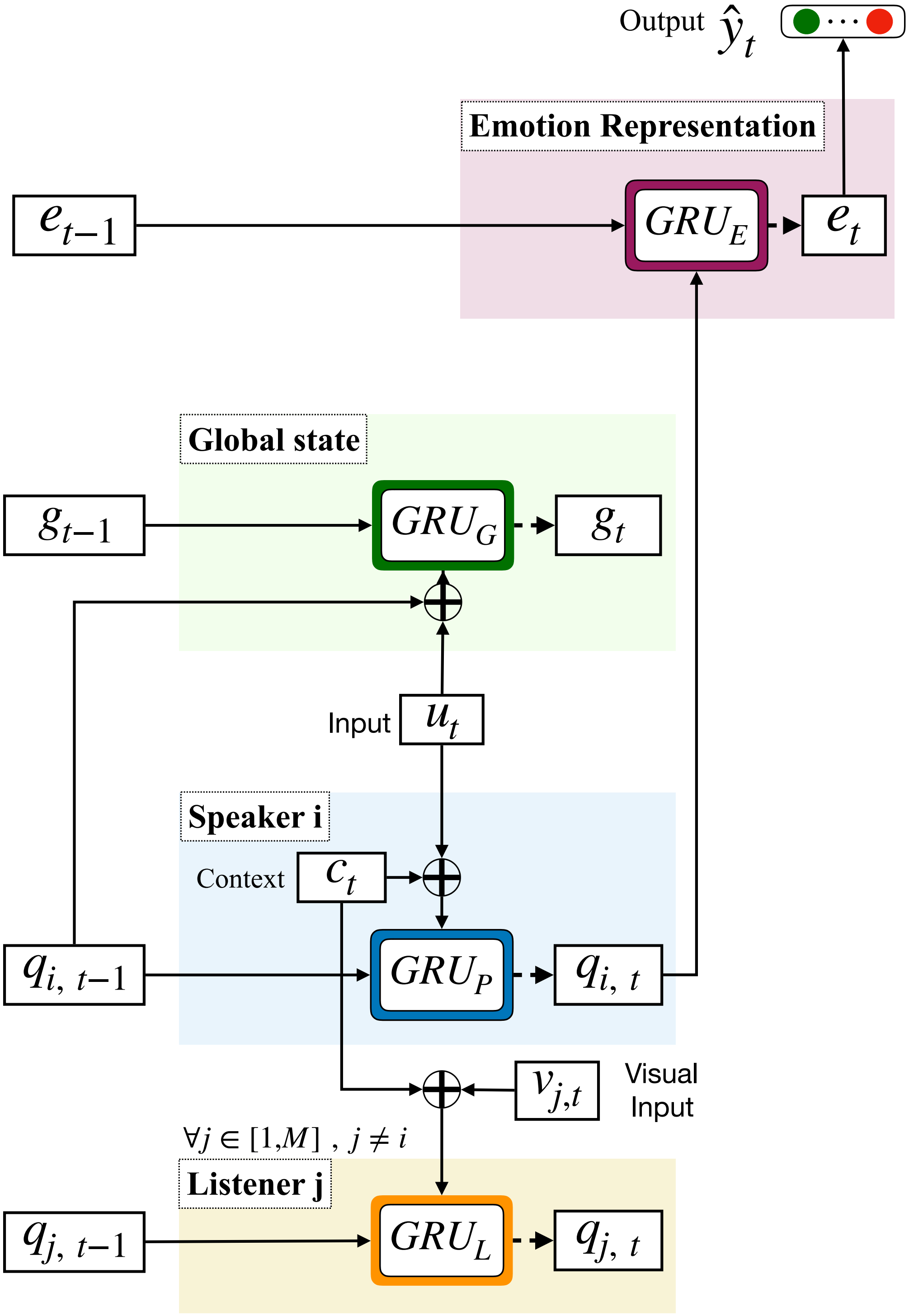
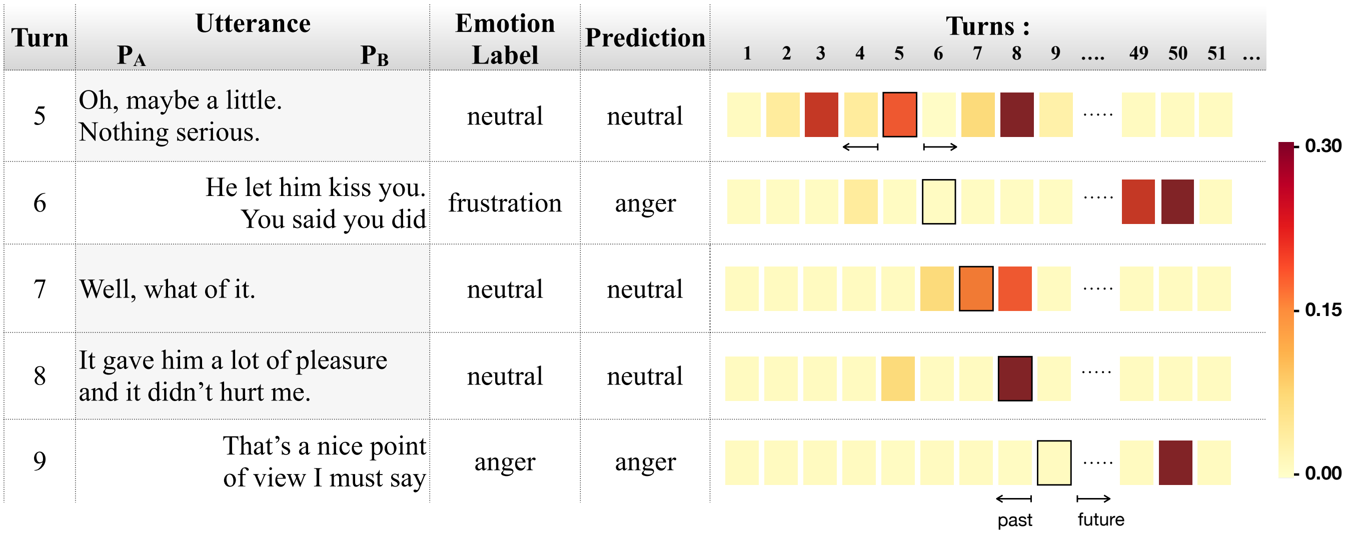
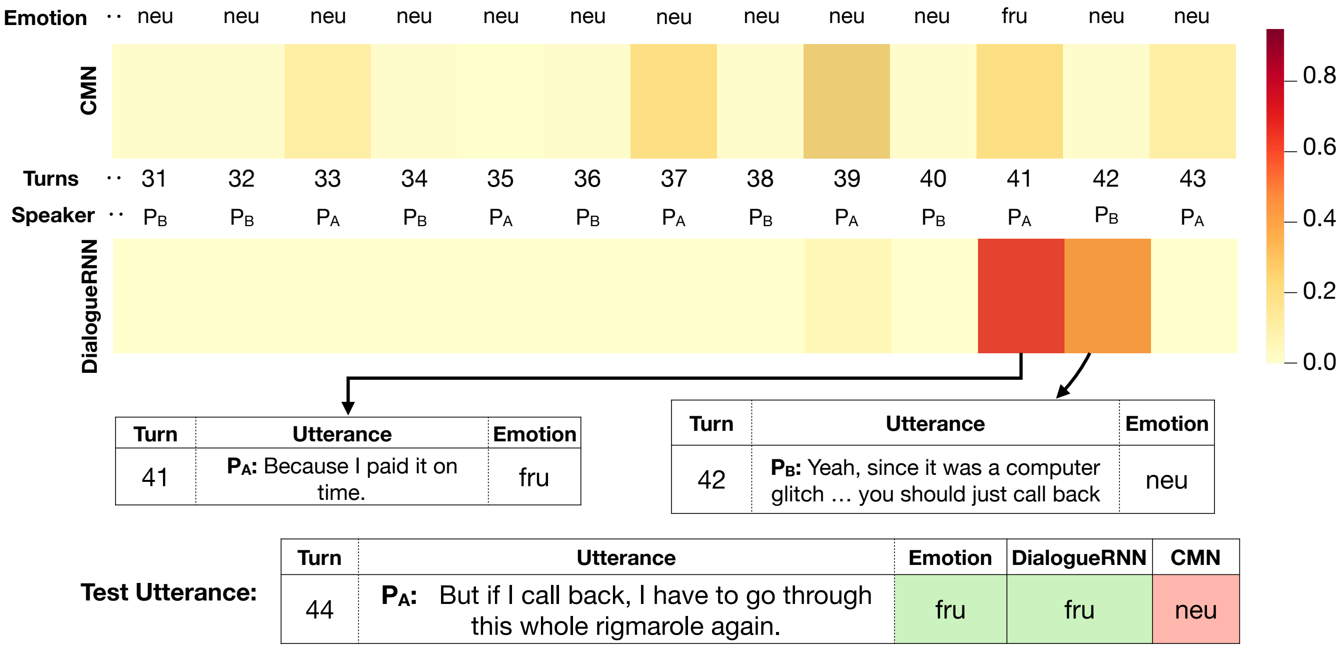


Figure 6: Update schemes for global, speaker, listener and emotion states for t^{th} utterance in a dialogue. Here, Person i is the speaker and Persons $j \in [1, M]$ and $j \neq i$ are the listeners.



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

Figure 7: Illustration of the β attention over emotion representations e_t.



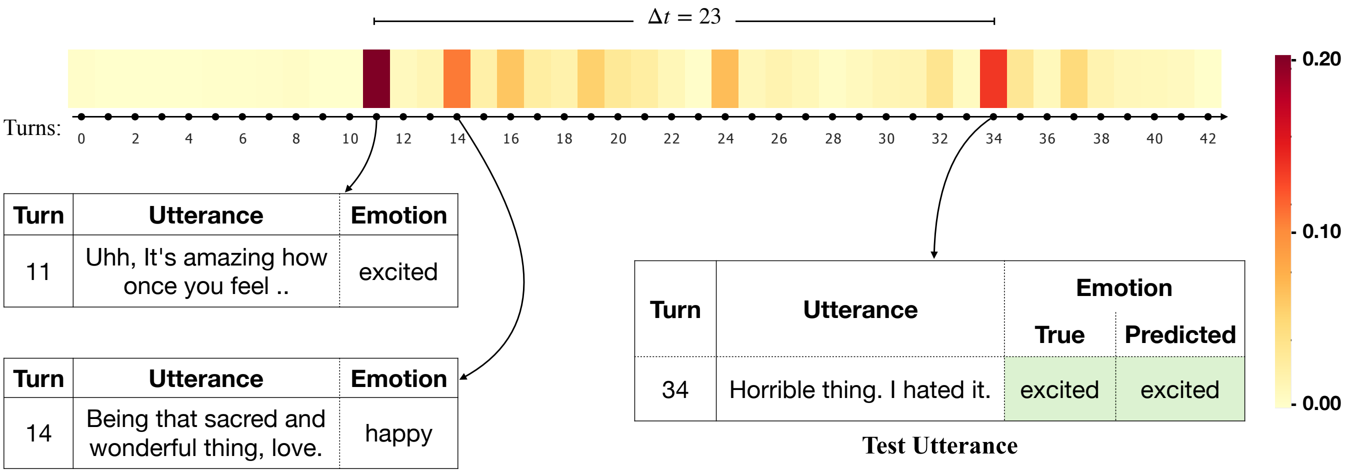
Turn	Utterance	Emotion
41	**PA:** Because I paid it on time.	fru

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

Figure 8: Comparison of attention scores over utterance history of CMN and DialogueRNN (α attention).



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

Figure 9: An example of long-term dependency among utterances.

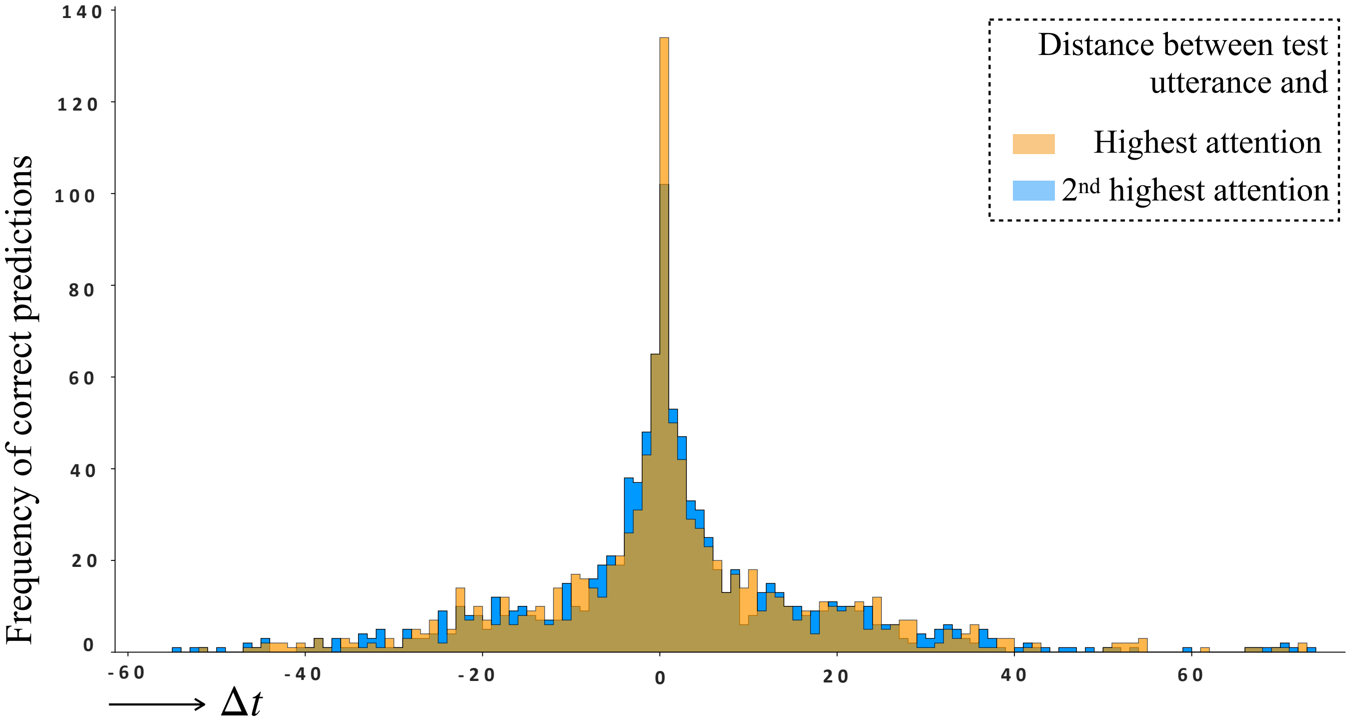


Figure 10: Histogram of Δt = distance between the target utterance and its context utterance based on β attention scores.