DARER: Dual-task Temporal Relational Recurrent Reasoning Network for Joint Dialog Sentiment Classification and Act Recognition

Bowen Xing[1] **and Ivor W. Tsang**[2,1]
[1]Australian Artificial Intelligence Institute, University of Technology Sydney, Australia
[2]Centre for Frontier Artificial Intelligence Research, A*STAR, Singapore
bwxing714@gmail.com, ivor_tsang@ihpc.a-star.edu.sg

Abstract

The task of joint dialog sentiment classification (DSC) and act recognition (DAR) aims to simultaneously predict the sentiment label and act label for each utterance in a dialog. In this paper, we put forward a new framework which models the explicit dependencies via integrating *prediction-level interactions* other than semantics-level interactions, more consistent with human intuition. Besides, we propose a speaker-aware temporal graph (SATG) and a dual-task relational temporal graph (DRTG) to introduce *temporal relations* into dialog understanding and dual-task reasoning. To implement our framework, we propose a novel model dubbed DARER, which first generates the context-, speaker- and temporal-sensitive utterance representations via modeling SATG, then conducts recurrent dual-task relational reasoning on DRTG, in which process the estimated label distributions act as key clues in prediction-level interactions. Experiment results show that DARER outperforms existing models by large margins while requiring much less computation resource and costing less training time. Remarkably, on DSC task in Mastodon, DARER gains a relative improvement of about 25% over previous best model in terms of F1, with less than 50% parameters and about only 60% required GPU memory.

1 Introduction

Dialog sentiment classification (DSC) and dialog act recognition (DAR) are two challenging tasks in dialog systems (Ghosal et al., 2021). DSC aims to predict the sentiment label of each utterance in a dialog, while DAR aims to predict the act label. Recently, researchers have discovered that these two tasks are correlative and they can assist each other (Cerisara et al., 2018; Kim and Kim, 2018).

Utterances	Act	Sentiment
u_a: I highly recommend it. Really awesome progression and added difficulty	Statement	Positive
u_b: I never have.	Disagreement	Negative

Table 1: A dialog snippet from the Mastodon dataset.

An example is shown in Table 1. To predict the sentiment of u_b, besides its *semantics*, its Disagreement act *label* and the Positive sentiment *label* of its *previous* utterance (u_a) can provide useful references, which contribute a lot when humans do this task. This is because the Disagreement act label of u_b denotes it has the opposite opinion with u_a, and thus u_b tends to have a Negative sentiment label, the opposite one with u_a (Positive). Similarly, the opposite sentiment labels of u_b and u_a are helpful to infer the Disagreement act label of u_b. In this paper, we term this process as dual-task reasoning, where there are three key factors: 1) the semantics of u_a and u_b; 2) the temporal relation between u_a and u_b; 3) u_a's and u_b's labels for another task.

In previous works, different models are proposed to model the correlations between the two tasks. (Cerisara et al., 2018) propose a multi-task model in which the two tasks share a single encoder. (Kim and Kim, 2018; Qin et al., 2020; Li et al., 2020; Qin et al., 2021) try to model the semantics-level interactions of the two tasks. The framework of previous models is shown in Fig. 1 (a). For dialog understanding, Co-GAT (Qin et al., 2021) applies graph attention network (GAT) (Velickovic et al., 2018) over an undirected disconnected graph which consists of isolated speaker-specific full-connected subgraphs. Therefore, it suffers from the issue that the inter-speaker interactions cannot be modeled, and the temporal relations between utterances are omitted. For dual-task reasoning, on the one hand, previous works only consider the parameter sharing and semantics-level interactions, while the label information is not integrated into the dual-task interactions. Consequently, the explicit dependencies between the two tasks cannot be captured and previous dual-task reasoning processes are incon-

Findings of the Association for Computational Linguistics: ACL 2022, pages 3611 - 3621
May 22-27, 2022 ©2022 Association for Computational Linguistics

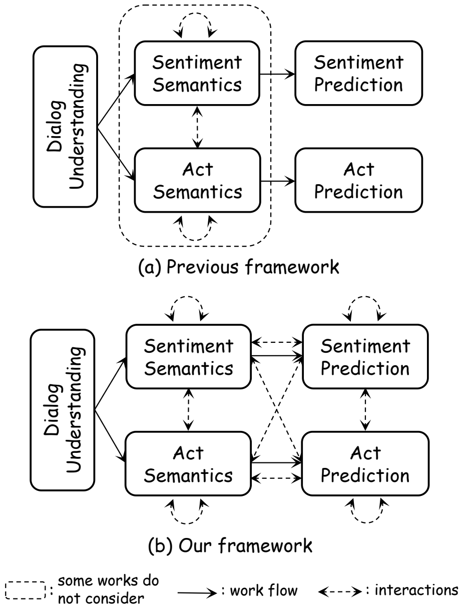

Figure 1: Illustration of previous framework and ours.

sistent with human intuition, which leverages the label information as crucial clues. On the other hand, previous works do not consider the temporal relations between utterances in dual-task reasoning, while in which they play a key role.

In this paper, we try to address the above issues by introducing temporal relations and leveraging label information. To introduce temporal relations, we design a **s**peaker-**a**ware **t**emporal **g**raph (SATG) for dialog understanding, and a **d**ual-task **r**easoning **t**emporal **g**raph (DRTG) for dual-task relational reasoning. Intuitively, different speakers' semantic states will change as the dialog goes, and these semantic state transitions trigger different sentiments and acts. SATG is designed to model the speaker-aware semantic states transitions, which provide essential indicative semantics for both tasks. Since the temporal relation is a key factor in dual-task reasoning, DRTG is designed to integrate inner- and inter-task temporal relations, making the dual-task reasoning process more rational and effective.

To leverage label information, we propose a new framework, as shown in Fig. 1 (b). Except for semantics-level interactions, it integrates several kinds of prediction-level interactions. First, self-interactions of sentiment predictions and act predictions. In both tasks, there are prediction-level correlations among the utterances in a dialog. In the DSC task, the sentiment state of each speaker tends to be stable until the utterances from others trigger the changes (Ghosal et al., 2019; Wang et al., 2020). In the DAR task, there are different patterns (e.g., Questions-Inform and Directives-Commissives) reflecting the interactions between act labels (Li et al., 2017). Second, interactions between the predictions and semantics. Intuitively, the predictions can offer feedback to semantics, which can rethink then reversely help revise the predictions. Third, prediction-prediction interactions between DSC and DAR, which model the explicit dependencies. However, since our objective is to predict the labels of both tasks, there is no ground-truth label available for prediction-level interactions. To this end, we design a recurrent dual-task reasoning mechanism that leverages the label distributions estimated in the previous step as prediction clues of the current step for producing new predictions. In this way, the label distributions of both tasks are gradually improved along the step.

To implement our framework, we propose a novel **D**ual-t**A**sk temporal **R**elational r**E**current **R**easoning Network (DARER), which includes three modules. The Dialog Understanding module conducts relation-specific graph transformations (RSGT) over SATG to generate context-, speaker- and temporal-sensitive utterance representations. The Initial Estimation module outputs the initial label information fed to the Recurrent Dual-task Reasoning module, in which RSGT operates on DRTG to conduct dual-task relational reasoning. Moreover, we design logic-heuristic training objectives to force DSC and DAR to prompt each other in the recurrent dual-task reasoning process gradually. Experiments on public datasets show that DARER significantly outperforms existing models. And further improvements can be obtained by utilizing pre-trained language models as the utterance encoder. Besides, compared with the previous best model, DARER reduces the number of parameters, required GPU memory, and training time.

The source code of DARER is publicly available at https://github.com/XingBowen714/DARER.

2 Methodology

Given a dialog consisting of N utterances: $\mathcal{D} = (u_1, u_2, ..., u_N)$, our objective is to predict both the dialog sentiment labels $Y^S = y_1^s, ..., y_N^s$ and the dialog act labels $Y^A = y_1^a, ..., y_N^a$ in a single run. Before delving into the details of DARER, we start with our designed SATG and DRTG.

r_{ij}	1	2	3	4	5	6	7	8
$I_s(i)$	1	1	1	1	2	2	2	2
$I_s(j)$	1	1	2	2	1	1	2	2
$pos(i,j)$	$>$	$\leq$	$>$	$\leq$	$>$	$\leq$	$>$	$\leq$

Table 2: All relation types in SATG (assume there are two speakers). $I_s(i)$ indicates the speaker node i is from. $pos(i, j)$ indicates the relative position of node i and j.

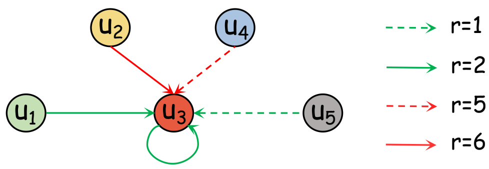

Figure 2: An example of SATG. u_1, u_3 and u_5 are from speaker 1 while u_2 and u_4 are from speaker 2. w.l.o.g, only the edges directed into u_3 node are illustrated.

r'_{ij}	1	2	3	4	5	6	7	8	9	10	11	12
$I_t(i)$	S	S	S	S	S	S	A	A	A	A	A	A
$I_t(j)$	S	S	S	A	A	A	S	S	S	A	A	A
$pos(i,j)$	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$

Table 3: All relation types in DRTG. $I_t(i)$ indicates that node i is a sentiment (S) node or act (A) node.

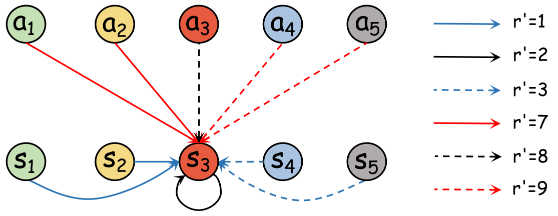

Figure 3: An example of DRTG. s_i and a_i respectively denote the node of DAC task and DAR task. w.l.o.g, only the edges directed into s_3 are illustrated.

2.1 Speaker-aware Temporal Graph

We design a SATG to model the information aggregation between utterances in a dialog. Formally, SATG is a complete directed graph denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$. In this paper, the nodes in $\mathcal{G}$ are the utterances in the dialog, i.e., $|\mathcal{V}| = N, \mathcal{V} = (u_1, ..., u_N)$, and the edge $(i, j, r_{ij}) \in \mathcal{E}$ denotes the information aggregation from u_i to u_j under the relation $r_{ij} \in \mathcal{R}$. Table 2 lists the definitions of all relation types in $\mathcal{R}$. In particular, there are three kinds of information conveyed by r_{ij}: the speaker of u_i, the speaker of u_j, and the relative position of u_i and u_j. Naturally, the utterances in a dialog are chronologically ordered, so the relative position of two utterances denotes their temporal relation. An example of SATG is shown in Fig. 2. Compared with previous dialog graph structure (Qin et al., 2020, 2021), our SATG has two main advancements. First, as a complete directed graph, SATG can model both the intra- and inter-speaker semantic interactions. Second, incorporating temporal information, SATG can model the transitions of speaker-aware semantic states as the dialog goes on, which benefit both tasks.

2.2 Dual-task Reasoning Temporal Graph

We design a DRTG to provide an advanced platform for dual-task relational reasoning. It is also a complete directed graph that consists of $2N$ dual nodes: N sentiment nodes and N act nodes. The definitions of all relation types in $\mathcal{R}'$ are listed in Table 3. Intuitively, when predicting the label of a node, the information of its dual node plays a key role, so we emphasize the temporal relation of '=' rather than merge it with '<' like SATG. Specifically, the relation r'_{ij} conveys three kinds of information: the task of n_i, the task of n_j and the temporal relation between n_i and n_j. An example of DRTG is shown in Fig. 3. Compared with previous dual-task graph structure (Qin et al., 2020, 2021), our DRTG has two major advancements. First, the temporal relations in DRTG can make the DTR-RSGT capture the the temporal information, which are essential for dual-task reasoning, while this cannot be achieved by the co-attention (Qin et al., 2020) or graph attention network (Qin et al., 2021) operating on their non-temporal graphs. Second, in DRTG , the information aggregated into a node is decomposed by different relations that correspond to individual contributions, rather than only depending on the semantic similarity measured by the attention mechanisms.

2.3 DARER

The network architecture of our proposed DARER is shown in Fig. 4. It consists of three modules, and we introduce their details next.

2.3.1 Dialog Understanding

Utterance Encoding In previous works, BiLSTM (Hochreiter and Schmidhuber, 1997) is widely adopted as the utterance encoder to generate the initial utterance representation: $H = (h_0, ..., h_N)$. In this paper, besides BiLSTM, we also study the effect of different pre-trained language model (PTLM) encoders in Sec. 3.6.

BiLSTM: We apply the BiLSTM over the word

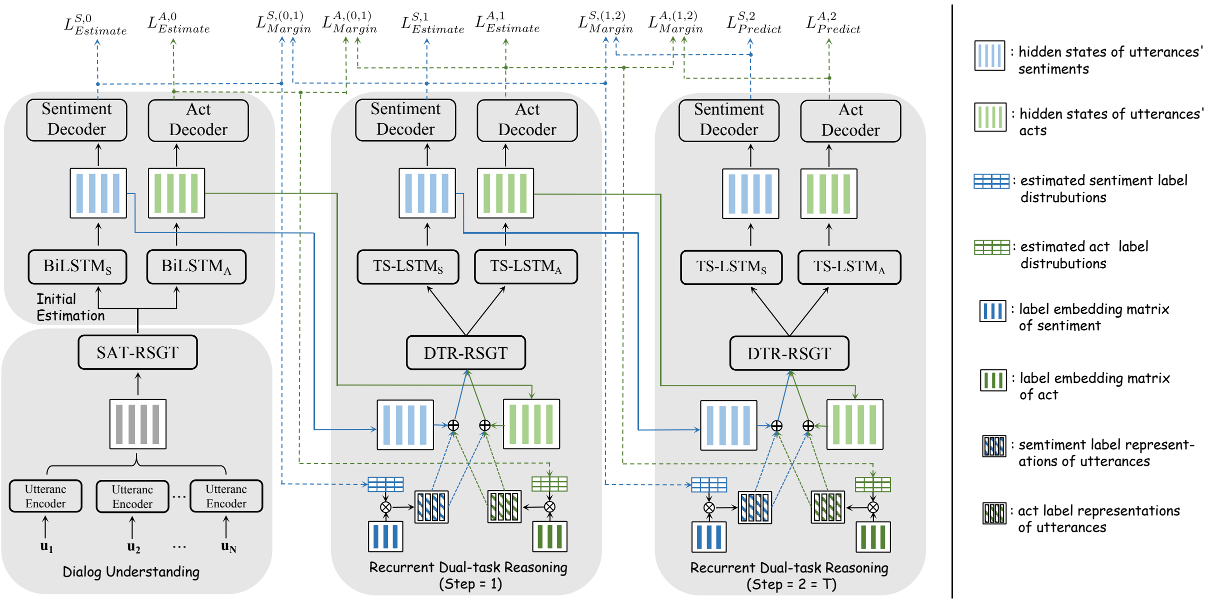
$L^{S,0}_{Estimate}$
$L^{A,0}_{Estimate}$
$L^{S,(0,1)}_{Margin}$
$L^{A,(0,1)}_{Margin}$
$L^{S,1}_{Estimate}$
$L^{A,1}_{Estimate}$
$L^{S,(1,2)}_{Margin}$
$L^{A,(1,2)}_{Margin}$
$L^{S,2}_{Predict}$
$L^{A,2}_{Predict}$
Sentiment Decoder
Act Decoder
BiLSTM_S
BiLSTM_A
Initial Estimation
SAT-RSGT
Utteranc Encoder
Utteranc Encoder
Utteranc Encoder
u_1
u_2
u_N
Dialog Understanding
TS-LSTM_S
TS-LSTM_A
DTR-RSGT
Recurrent Dual-task Reasoning (Step = 1)
Recurrent Dual-task Reasoning (Step = 2 = T)
: hidden states of utterances' sentiments
: hidden states of utterances' acts
: estimated sentiment label distrubutions
: estimated act label distrubutions
: label embedding matrix of sentiment
: label embedding matrix of act
: sentiment label representations of utterances
: act label representations of utterances

Figure 4: The network architecture of our proposed DARER model. Without loss of generality, the step number T in this illustration is set 2.

embeddings of u_t to capture the inner-sentence dependencies and temporal relationships among the words, producing a series of hidden states $H_{u,i} = (h^0_{u,i}, ..., h^{l_i}_{u,i})$, where l_i is the length of u_i. Then we feed $H_{u,i}$ into a max-pooling layer to get the representation for each u_i.
PTLM: We separately feed each utterance into the PTLM encoder and take the output hidden state of the `[CLS]` token as the utterance representation.

Speaker-aware Temporal RSGT To capture the inter- and intra-speaker semantic interactions and the speaker-aware temporal dependencies between utterances, we conduct Speaker-aware Temporal relation-specific graph transformations (SAT-RSGT) inspired from (Schlichtkrull et al., 2018) over SATG. The information aggregation of SAT-RSGT can be formulated as:

$$\hat{h}_i = W_1 h^0_i + \sum_{r\in\mathcal{R}} \sum_{j\in\mathcal{N}^r_i} \frac{1}{|N^r_i|} W^r_1 h^0_j \tag{1}$$

where W_1 is self-transformation matrix and W^r_1 is relation-specific matrix. Now we obtain the context-, speaker- and temporal-sensitive utterance representations: $\hat{H} = (\hat{h_0}, ..., \hat{h_N})$.

2.3.2 Initial Estimation

To obtain task-specific utterances representations, we separately apply two BiLSTMs over $\hat{H}$ to obtain the utterance hidden states for sentiments and acts respectively: $H^0_s = \text{BiLSTM}_\text{S}(\hat{H})$, $H^0_a = \text{BiLSTM}_\text{A}(\hat{H})$, where $H^0_s = \{h^0_{s,i}\}^N_{i=1}$ and $H^0_a = \{h^0_{a,i}\}^N_{i=1}$. Then H^0_s and H^0_a are separately fed into Sentiment Decoder and Act Decoder to produce the initial estimated label distributions:

$$\begin{aligned} P^0_S &= \{P^0_{S,i}\}^N_{i=1},\ P^0_A = \{P^0_{A,i}\}^N_{i=1} \\ P^0_{S,i} &= softmax(W^s_d h^0_{a,i} + b^s_d) \\ &= [p^0_{s,i}[0], ..., p^0_{s,i}[k], ..., p^0_{s,i}(|\mathcal{C}_s|-1)] \\ P^0_{A,i} &= softmax(W^a_d h^0_{s,i} + b^a_d) \\ &= [p^0_{a,i}[0], ..., p^0_{a,i}[k], ..., p^0_{a,i}(|\mathcal{C}_a|-1)] \end{aligned} \tag{2}$$

where W^*_d and b^*_d are weight matrices and biases, $\mathcal{C}_s$ and $\mathcal{C}_a$ are sentiment class set and act class set.

2.3.3 Recurrent Dual-task Reasoning

At step t, the recurrent dual-task reasoning module takes two streams of inputs: 1) hidden states $H^{t-1}_s \in \mathbb{R}^{N\times d}$ and $H^{t-1}_a \in \mathbb{R}^{N\times d}$; 2) label distributions $P^{t-1}_S \in \mathbb{R}^{N\times|\mathcal{C}_s|}$ and $P^{t-1}_A \in \mathbb{R}^{N\times|\mathcal{C}_a|}$.

Projection of Label Distribution To achieve the prediction-level interactions, we should represent the label information in vector form to let it participate in calculations. We use P^{t-1}_S and P^{t-1}_A to respectively multiply the sentiment label embedding matrix $M^e_s \in \mathbb{R}^{|\mathcal{C}_s|\times d}$ and the act label embedding matrix $M^e_a \in \mathbb{R}^{|\mathcal{C}_a|\times d}$, obtaining the sentiment label representations $E^t_S = \{e^t_{s,i}\}^N_{i=1}$ and act label representations $E^t_A = \{e^t_{a,i}\}^N_{i=1}$. In particular, for each utterance, its sentiment label representation and act label representation are computed as:

$$e^t_{s,i} = \sum_{k=0}^{|\mathcal{C}_s|-1} p^{t-1}_{s,i}[k] \cdot v^k_s$$
$$e^t_{a,i} = \sum_{k'=0}^{|\mathcal{C}_a|-1} p^{t-1}_{a,i}[k'] \cdot v^{k'}_a \qquad (3)$$

where v^k_s and $v^{k'}_a$ are the label embeddings of sentiment class k and act class k', respectively.

Dual-task Reasoning RSGT To achieve the self- and mutual-interactions between the semantics and predictions, for each node in DRTG, we superimpose its corresponding utterance's label embeddings of both tasks on its hidden state:

$$\hat{h}^t_{s,i} = h^{t-1}_{s,i} + e^t_{s,i} + e^t_{a,i}$$
$$\hat{h}^t_{a,i} = h^{t-1}_{a,i} + e^t_{s,i} + e^t_{a,i} \qquad (4)$$

Thus the representation of each node contains the task-specific semantic features and both tasks' label information, which are then incorporated into the relational reasoning process to achieve semantics-level and prediction-level interactions.

The obtained $\hat{\mathbf{H}}^{\mathbf{t}}_{\mathbf{s}}$ and $\hat{\mathbf{H}}^{\mathbf{t}}_{\mathbf{a}}$ both have N vectors, respectively corresponding to the N sentiment nodes and N act nodes on DRTG. Then we feed them into the Dual-task Reasoning relation-specific graph transformations (DTR-RSGT) conducting on DRTG. Specifically, the node updating process of DTR-RSGT can be formulated as:

$$\overline{h}^t_i = W_2 \hat{h}^t_i + \sum_{r \in \mathcal{R}'} \sum_{j \in \mathcal{N}^{r'}_i} \frac{1}{|N^{r'}_i|} W^r_2 \hat{h}^t_j \qquad (5)$$

where W_2 is self-transformation matrix and W^r_2 is relation-specific matrix. Now we get $\overline{H}^t_s$ and $\overline{H}^t_a$.

Label Decoding For each task, we use a task-specific BiLSTM (TS-BiLSTM) to generate a new series of hidden states that are more task-specific:

$$H^t_s = \text{TS-BiLSTM}_\text{S}(\overline{H}^t_s)$$
$$H^t_a = \text{TS-BiLSTM}_\text{A}(\overline{H}^t_a) \qquad (6)$$

Besides, as $\overline{H}^t_s$ and $\overline{H}^t_a$ both contains the label information of the two tasks, the two TS-BiLSTMs have another advantage of label-aware sequence reasoning, which has been proven can be achieved by LSTM (Zheng et al., 2017).

Then H^t_S and H^t_A are separately fed to Sentiment Decoder and Act Decoder to produce P^t_S and P^t_A.

2.3.4 Logic-heuristic Training Objective

Intuitively, there are two important logic rules in our DARER. First, the produced label distributions should be good enough to provide useful label information for the next step. Otherwise, noisy label information would be introduced, misleading the dual-task reasoning. Second, both tasks are supposed to learn more and more beneficial knowledge from each other in the recurrent dual-task reasoning process. Scilicet the estimated label distributions should be gradually improved along steps. In order to force DARER to obey these two rules, we propose a constraint loss $L_{Constraint}$ that includes two terms: $L_{Estimate}$ and L_{Margin}, which correspond to the two rules, respectively.

Estimate Loss $L_{Estimate}$ is the cross-entropy loss forcing DARER to provide good enough label distributions for the next step. At step t, for DSC task, $\mathcal{L}^{S,t}_{Estimate}$ is defined as:

$$\mathcal{L}^{S,t}_{Estimate} = \sum_{i=1}^{N} \sum_{k=0}^{|\mathcal{C}_s|-1} y^k_{s,i} \text{log}\left(p^t_{s,i}[k]\right) \qquad (7)$$

Margin Loss L_{Margin} works on the label distributions of two adjacent steps, and it promotes the two tasks gradually learning beneficial knowledge from each other via forcing DARER to produce better predictions at step t than step $t-1$. For DSC task, $\mathcal{L}^{S,(t,t-1)}_{Margin}$ is a margin loss defined as:

$$\mathcal{L}^{S,(t,t-1)}_{Margin} = \sum_{i=1}^{N} \sum_{k=0}^{|\mathcal{C}_s|-1} y^k_{s,i} \max(0, p^{t-1}_{s,i}[k] - p^t_{s,i}[k]) \qquad (8)$$

Constraint loss $L_{Constraint}$ is the weighted sum of $L_{Estimate}$ and L_{Margin}, with a hyper-parameter γ balancing the two kinds of punishments. For DSC task, $\mathcal{L}^S_{Constraint}$ is defined as:

$$\mathcal{L}^S_{Constraint} = \sum_{t=0}^{T-1} \mathcal{L}^{S,t}_{Estimate} + \gamma * \sum_{t=1}^{T} \mathcal{L}^{S,(t,t-1)}_{margin} \qquad (9)$$

Final Training Objective The total loss for DSC task ($\mathcal{L}^S$) is the sum of $\mathcal{L}^S_{Constraint}$ and $\mathcal{L}^S_{Prediction}$:

$$\mathcal{L}^S = \mathcal{L}^S_{Prediction} + \mathcal{L}^S_{Constraint} \qquad (10)$$

where $\mathcal{L}^S_{Prediction}$ is the cross-entropy loss of the produced label distributions at the final step T:

$$\mathcal{L}^S_{Prediction} = \sum_{i=1}^{N} \sum_{k=0}^{|\mathcal{C}_s|-1} y_{s,i} \log\left(p^T_{s,i}[k]\right) \qquad (11)$$

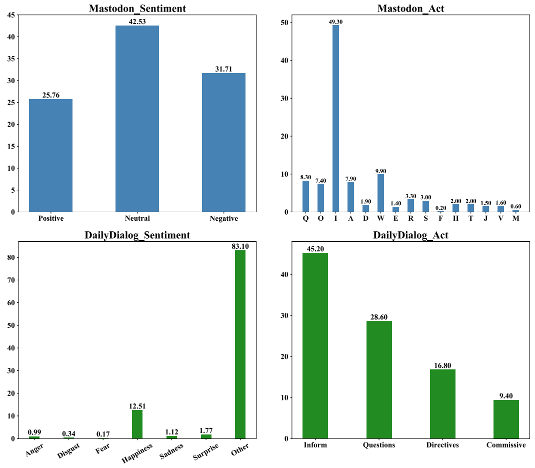

Figure 5: Illustration of class distributions.

The total loss of DAR task ($\mathcal{L}^A$) can be derivated similarly like eqs. (7) to (11).

The final training objective of DARER is the sum of the total losses of the two tasks:

$$\mathcal{L} = \mathcal{L}^S + \mathcal{L}^A \tag{12}$$

3 Experiments

3.1 Datasets and Metrics

Dataset. We conduct experiments on two publicly available dialogue datasets: Mastodon[1] (Cerisara et al., 2018) and Dailydialog[2] (Li et al., 2017). The Mastodon dataset includes 269 dialogues for training and 266 dialogues for testing. And there are 3 sentiment classes and 15 act classes. Since there is no official validation set, we follow the same partition as Qin et al. (2021). Finally, there are 243 dialogues for training, 26 dialogues for validating, and 266 dialogues for testing. As for Dailydialog dataset, we adopt the official train/valid/test/ split from the original dataset (Li et al., 2017): 11,118 dialogues for training, 1,000 for validating, and 1,000 for testing. And there are 7 sentiment classes and 4 act classes. The class distributions of the two tasks on the two datasets are illustrated in Fig. 5.

Evaluation Metrics. Following previous works (Cerisara et al., 2018; Qin et al., 2020, 2021), on Dailydialog dataset, we adopt macro-average Precision (P), Recall (R), and F1 for the two tasks, while on Mastodon dataset, we ignore the neutral sentiment label in DSC task and for DAR task we adopt the average of the F1 scores weighted by the prevalence of each dialogue act.

3.2 Implement Details and Baselines

DARER is trained with Adam optimizer with the learning rate of $1e^{-3}$ and the batch size is 16. We exploit 300-dimensional Glove vectors for the word embeddings, and the dimension of hidden states (label embeddings) is 128 for Mastodon and 256 for DailyDialog. The step number T for recurrent dual-task reasoning is set to 3 for Mastodon and 1 for DailyDialog. The coefficient γ is set to 3 for Mastodon and $1e^{-4}$ for DailyDialog. To alleviate overfitting, we adopt dropout, and the ratio is 0.2 for Mastodon and 0.3 for DailyDialog. For all experiments, we pick the model performing best on validation set then report the average results on test set based on three runs with different random seeds. The epoch number is 100 for Mastodon and 50 for DailyDialog. All computations are conducted on an NVIDIA RTX 6000 GPU.

We compare our model with: JointDAS (Cerisara et al., 2018), IIIM (Kim and Kim, 2018), DCR-Net (Co-Attention) (Qin et al., 2020), BCDCN (Li et al., 2020) and Co-GAT (Qin et al., 2021).

3.3 Main Results

Table 4 lists the experiment results on the test sets of the two datasets. We can observe that:

1. Our DARER significantly outperforms all baselines, achieving new state-of-the-art (SOTA). In particular, over Co-GAT, the existing SOTA, DARER achieves an absolute improvement of 13.1% in F1 score on DSC task in Mastodon, a relative improvement of over 1/4. The satisfying results of DARER come from (1) our framework integrates not only semantics-level interactions but also prediction-level interactions, thus captures explicit dependencies other than implicit dependencies; (2) our SATG represents the speaker-aware semantic states transitions, capturing the important basic semantics benefiting both tasks; (3) our DRTG provides a rational platform on which more effective dual-task relational reasoning is conducted. (4) the advanced architecture of DARER allows DSC and DAR to improve each other in the recurrent dual-task reasoning process gradually.

2. DARER shows more prominent superiority on DSC task than DAR task. We surmise the probable reason is that generally, act label is more complicated to deduce than sentiment label in dual-task reasoning. For instance, it is easy to infer u_i's Negative label on DSC given u_i's Agreement label on DAR and u_{i-1}'s Negative label on DSC.

[1] https://github.com/cerisara/DialogSentimentMastodon

[2] http://yanran.li/dailydialog

Models	Mastodon						DailyDialog					
	DSC			DAR			DSC			DAR		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
JointDAS	36.1	41.6	37.6	55.6	51.9	53.2	35.4	28.8	31.2	76.2	74.5	75.1
IIIM	38.7	40.1	39.4	56.3	52.2	54.3	38.9	28.5	33.0	76.5	74.9	75.7
DCR-Net	43.2	47.3	45.1	60.3	56.9	58.6	56.0	40.1	45.4	79.1	79.0	79.1
BCDCN	38.2	62.0	45.9	57.3	61.7	59.4	55.2	45.7	48.6	80.0	80.6	80.3
Co-GAT	44.0	53.2	48.1	60.4	60.6	60.5	65.9	45.3	51.0	81.0	78.1	79.4
Co-GAT*	45.40 ±2.31	48.11 ±2.91	46.47 ±0.37	62.55 ±0.46	58.66 ±1.71	60.54 ±1.10	58.04 ±0.84	44.65 ±0.36	48.82 ±0.22	79.14 ±0.40	79.71 ±0.16	79.39 ±0.14
DARER	**56.04**† ±0.85	**63.33**† ±0.30	**59.59**† ±0.70	**65.08**‡ ±1.25	**61.88**† ±0.37	**63.43**† ±0.85	**59.96**‡ ±1.25	**49.51**† ±1.33	**53.42**† ±0.18	**81.39**† ±0.55	**80.80**‡ ±0.43	**81.06**† ±0.04

Table 4: Experiment results. * denotes we reproduce the results using official code. ± denotes standard deviation. † denotes that our DARER significantly outperforms Co-GAT with $p < 0.01$ under t-test and ‡ denotes $p < 0.05$.

Variants	Mastodon		DailyDialog	
	DSC	DAR	DSC	DAR
DARER	**59.59**	**63.43**	**53.42**	**81.39**
w/o Label Embeddings	56.76	62.15	50.64	79.87
w/o Harness Loss	56.22	61.99	49.94	79.76
w/o SAT-RSGT	57.37	62.96	50.25	80.52
w/o DTR-RSGT	56.69	61.69	50.11	79.76
w/o TS-LSTMs	56.30	61.49	51.61	80.33
w/o Tpl Rels in SATG	58.23	62.21	50.99	80.70
w/o Tpl Rels in DRTG	57.22	62.15	50.52	80.28

Table 5: Results of ablation experiments on F1 score.

Reversely, given the label information that u_i and u_{i-1} are both negative on DSC, it is hard to infer the act label of u_i because there are several act labels possibly satisfying this case, e.g., Disagreement, Agreement, Statement.

3. DARER's improvements on DailyDialog are smaller than those on Mastodon. We speculate this is caused by the extremely unbalanced sentiment class distribution on DailyDialog. From Fig. 5 we can find that over 83% utterances do not express sentiment, while the act labels are rich and varied. This hinders DARER from learning valuable correlations between the two tasks.

3.4 Ablation Study

We conduct ablation experiments to study each component of DARER. Table 5 lists the results.

(1) Removing **label embeddings** causes prediction-level interactions cannot be achieved. The sharp drops of results prove that our method of leveraging label information to achieve prediction-level interactions effectively improves dual-task reasoning via capturing explicit dependencies. (2) Without **harness loss**, the two logic rules can hardly be met, so there is no constrain forcing DSC and DAR to gradually prompt each other, resulting in the dramatic decline of performances. (3) As the core of Dialog Understanding, **SAT-RSGT** captures speaker-aware semantic states transitions, which provides essential basic task-free knowledge for both tasks. Without it, some essential indicative semantics would be lost, then the results decrease. (4) The worst results of 'w/o DTR-RSGT' prove that **DTR-RSGT** is the core of DARER, and it plays the vital role of conducting dual-task relational reasoning over the semantics and label information. (5) The significant results decrease of 'w/o TS-LSTMs' prove that **TS-LSTMs** also plays an important role in DARER by generating task-specific hidden states for both tasks and have some capability of sequence label-aware reasoning. (6) Removing of the **temporal relations** (Tpl Rels) in SATG or DRTG causes distinct results decline. This can prove the necessity and effectiveness of introducing temporal relations into dialog understanding and dual-task reasoning.

3.5 Impact of Step Number T

The performances of DARER over different T are plotted in Fig. 6. $T = 0$ denotes the output of Initial Estimation module is regarded as final predictions. We can find that appropriately increasing T brings results improvements. Particularly, with T increasing from 0 to 1, the results increase sharply. This verifies that the Initial Estimation module can provide useful label information for dual-task reasoning. Furthermore, DARER can learn beneficial mutual knowledge from recurrent dual-task reason-

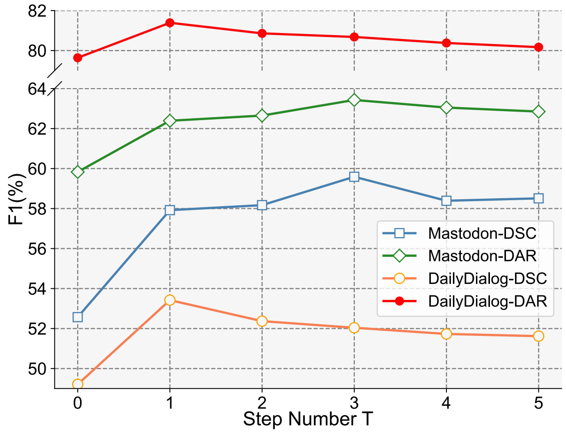

Figure 6: Performances of DARER over different T.

ing in which DSC and DAR prompt each other. Generally, when T surpasses a certain point, the performances declines slightly. The possible reason is that after the peak, more dual-task interactions cause too much deep information fusion of the two tasks, leading to the loss of some important task-specific features and overfitting.

3.6 Effect of Pre-trained Language Model

	Models	Mastodon					
		DSC			DAR		
		P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
BERT	+ Linear	61.79	61.09	60.60	70.20	67.49	68.82
	+ Co-GAT	66.03	58.13	61.56	70.66	67.62	69.08
	+ DARER	65.98	67.39	**66.42**	73.82	71.67	**72.73**
RoBERTa	+ Linear	57.83	60.54	57.83	62.49	61.93	62.20
	+ Co-GAT	61.28	57.25	58.26	66.46	64.01	65.21
	+ DARER	61.36	67.27	**63.66**	70.87	68.68	**69.75**
XLNet	+ Linear	61.42	67.80	63.35	67.31	63.04	65.09
	+ Co-GAT	64.01	65.30	63.71	67.19	64.09	65.60
	+ DARER	68.05	69.47	**68.66**	72.04	69.63	**70.81**

Table 6: Results based on different PTLM encoders.

In this section, we study the effects of three PTLM encoders: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019), which replace the BiLSTM utterance encoder in DARER. We adopt the base versions of the PTLMs implemented in PyTorch by Wolf et al. (2020). In our experiments, the whole models are trained by AdamW optimizer with the learning rate of $1e^{-5}$ and the batch size is 16. And the PTLMs are fine-tuned in the training process. Results are listed in Table 6. We can find that since single PTLM encoders are powerful in language understanding, they obtain promising results even without any interactions between utterances or the two tasks. Nevertheless, stacking DARER on PTLM encoders further obtains around 5% absolute improvements on F1. This is because our DARER achieves prediction-level interactions and integrates temporal relations, which complement the high-quality semantics grasped by PTLM encoders. In contrast, Co-GAT only models the semantics-level interactions, whose advantages are diluted by PTLM. Consequently, based on PTLM encoders, Co-GAT brings much less improvement than our DARER.

3.7 Computation Efficiency

Models	Number of Parameters	Training Time per Epoch	GPU Memory	Avg. F1
Co-GAT	6.93M	2.35s	2007MB	53.66%
DARER	2.50M	2.20s	1167MB	61.51%
Improve	**-63.92%**	**-6.38%**	**-41.85%**	**14.63%**

Table 7: Comparison with SOTA on different aspects.

In practical application, in addition to the performance, the number of parameters, the time cost, and GPU memory required are important factors. Taking Mastodon as the testbed, we compare our DARER with the up-to-date SOTA (Co-GAT) on these factors, and results are shown in Table 7. Avg. F1 denotes the average of the F1 scores on the two tasks. Remarkably, although our DARER surpasses SOTA on Avg. F1 by 14.6%, it cut the number of parameters and required GPU memory by about 1/2. This is due to the parameter sharing mechanism in DARER. Moreover, our DARER costs less time for training. Therefore, it is proven that our DARER is more efficient in practical application.

4 Related Works

Dialog Sentiment Classification (Hazarika et al., 2018; Ghosal et al., 2019; Zhong et al., 2019; Jiao et al., 2020; Zhu et al., 2021; Shen et al., 2021) and Dialog Act Recognition (Inui et al., 2001; Raheja and Tetreault, 2019; Shang et al., 2020; Saha et al., 2020) are both utterance-level classification tasks. Recently, it has been found that these two tasks are correlative, and they can work together to indicate the speaker's more comprehensive intentions (Kim and Kim, 2018). With the development of well-annotated corpora, (Li et al., 2017; Cerisara et al., 2018), in which both the act label and sentiment label of each utterance are provided, several models have been proposed to tackle the joint dialog sentiment classification and act recognition task.

Cerisara et al. (2018) propose a multi-task framework based on a shared encoder that implicitly models the dual-task correlations. Kim and Kim (2018) integrate the identifications of dialog acts, predictors and sentiments into a unified model. To explicitly model the mutual interactions between the two tasks, Qin et al. (2020) propose a stacked co-interactive relation layer and Li et al. (2020) propose a context-aware dynamic convolution network to capture the crucial local context. More recently, Qin et al. (2021) propose Co-GAT, which applies graph attentions on a fully-connected undirected graph consisting of two groups of nodes corresponding to the two tasks, respectively.

This work is different from previous works on three aspects. First, we model the inner- and inter-speaker temporal dependencies for dialog understanding. Second, we model the cross- and self-task temporal dependencies for dual-task reasoning; Third, we achieve prediction-level interactions in which the estimated label distributions act as important and explicit clues other than semantics.

5 Conclusion and Future Work

In this paper, we present a new framework that integrates prediction-level interactions to leverage estimated label distribution as explicit and important clues other than implicit semantics. Besides, we design the SATG and DRTG to introduce temporal relations into dialog understanding and dual-task reasoning. Moreover, we propose a novel model named DARER to allow temporal information, label information, and semantics to work together to let DSC and DAR gradually promote each other, which is further forced by the proposed logic-heuristic training objective. Experimental results demonstrate the superiority of our method, which not only surpasses previous models on performances by a large margin but also significantly economizes computation resources.

Our work brings two insights for dialog understanding and multi-task reasoning in dialog systems: (1) exploiting the temporal relations between utterances for reasoning; (2) leveraging estimated label distributions to capture explicit correlations;. In the future, we will apply our method to other multi-task learning scenarios in dialog systems.

Acknowledgements

This work was supported by Australian Research Council Grant (DP180100106 and DP200101328). Ivor W. Tsang was also supported by A*STAR Centre for Frontier AI Research (CFAR).

References

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. 2018. Multi-task dialog act and sentiment recognition on mastodon. In *COLING*, pages 745–754. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1435–1449, Online. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP/IJCNLP (1)*, pages 154–164. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 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)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

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

N. Inui, T. Ebe, B. Indurkhya, and Y. Kotani. 2001. A case-based natural language dialogue system using dialogue act. In *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, volume 1, pages 193–198 vol.1.

Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *AAAI*, pages 8002–8009. AAAI Press.

Minkyoung Kim and Harksoo Kim. 2018. Integrated neural network model for identifying speech acts, predicators, and sentiments of dialogue utterances. *Pattern Recognition Letters*, 101:1–5.

Jingye Li, Hao Fei, and Donghong Ji. 2020. Modeling local contexts for joint dialogue act recognition and sentiment classification with bi-channel dynamic convolutions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 616–626, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Libo Qin, Wanxiang Che, Yangming Li, Minheng Ni, and Ting Liu. 2020. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8665–8672. AAAI Press.

Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2021. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13709–13717.

Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3727–3733, Minneapolis, Minnesota. Association for Computational Linguistics.

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multimodal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372, Online. Association for Computational Linguistics.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *ESWC*, pages 593–607.

Guokan Shang, Antoine Tixier, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2020. Speaker-change aware CRF for dialogue act classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 450–464, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560, Online. Association for Computational Linguistics.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195, 1st virtual meeting. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1227–1236, Vancouver, Canada. Association for Computational Linguistics.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in*

Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 165–176, Hong Kong, China. Association for Computational Linguistics.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582, Online. Association for Computational Linguistics.