Chongqing University of Technology

ATAI
Advanced Technique of Artificial Intelligence

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

Bowen Xing1 and Ivor W. Tsang2,1
1Australian Artificial Intelligence Institute, University of Technology Sydney, Australia
2Centre for Frontier Artificial Intelligence Research, A*STAR, Singapore
bwxing714@gmail.com, ivor_tsang@ihpc.a-star.edu.sg

2022. 3. 17 • ChongQing

ACL2022

Leibniz-Institut für Sozialwissenschaften

Reported by Sijin Liu

Artificial

1.Introduction

2.Method

3.Experiments

Introduction

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.

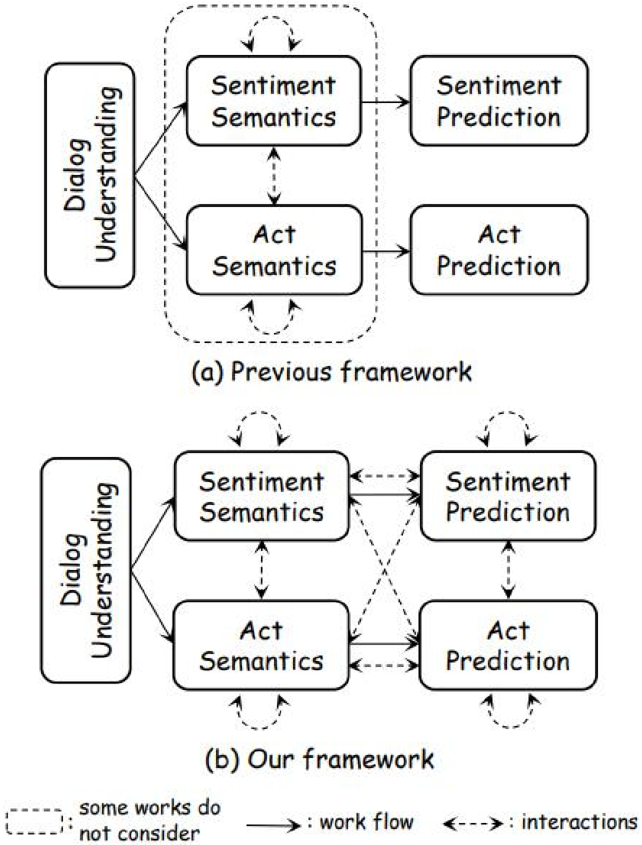


Figure 1: Illustration of previous framework and ours.

- Previous works only consider the parameter sharing and semantics-level interactions, while the label information is not integrated into the dual-task interactions.
- 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.

Method

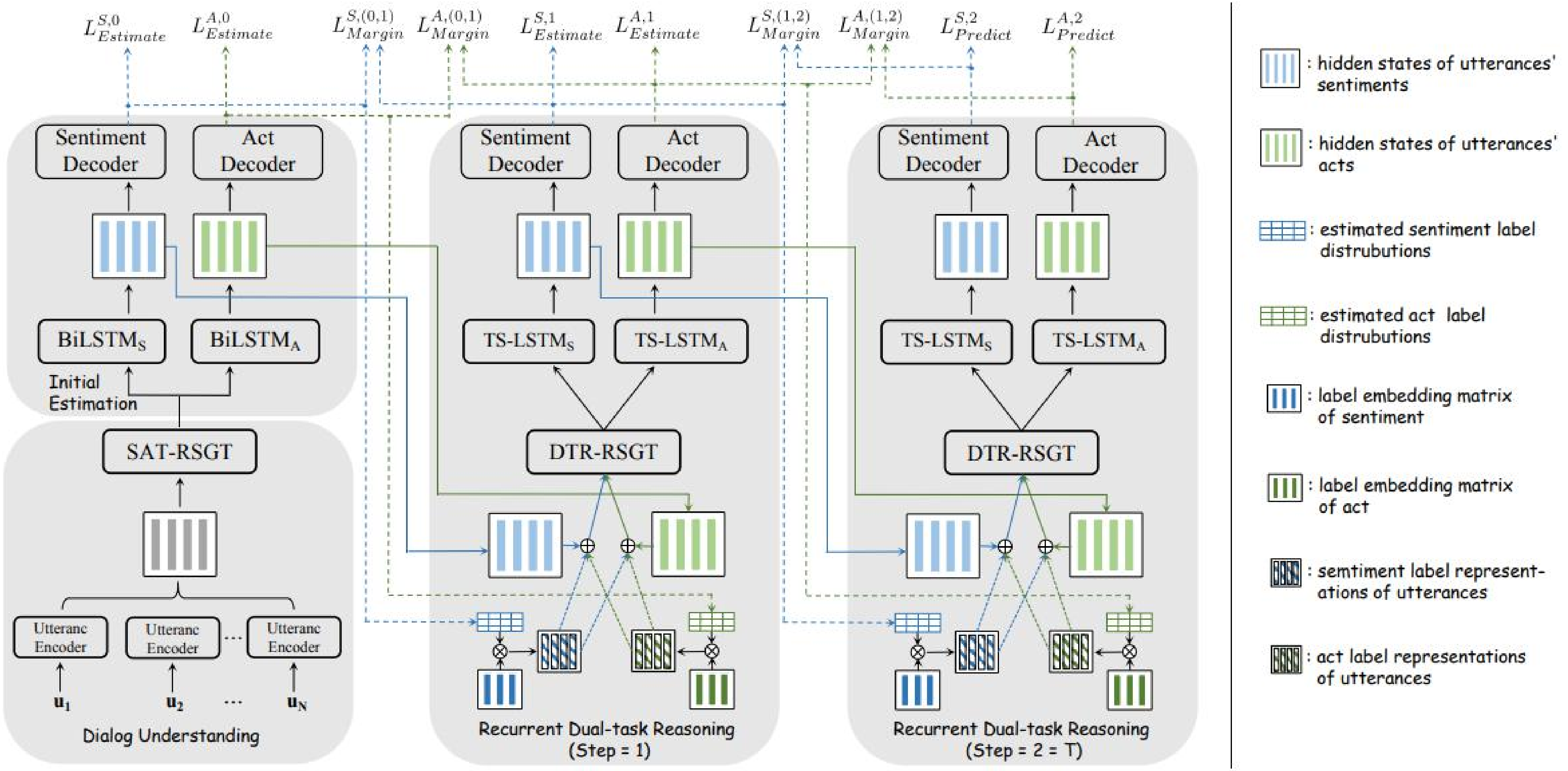
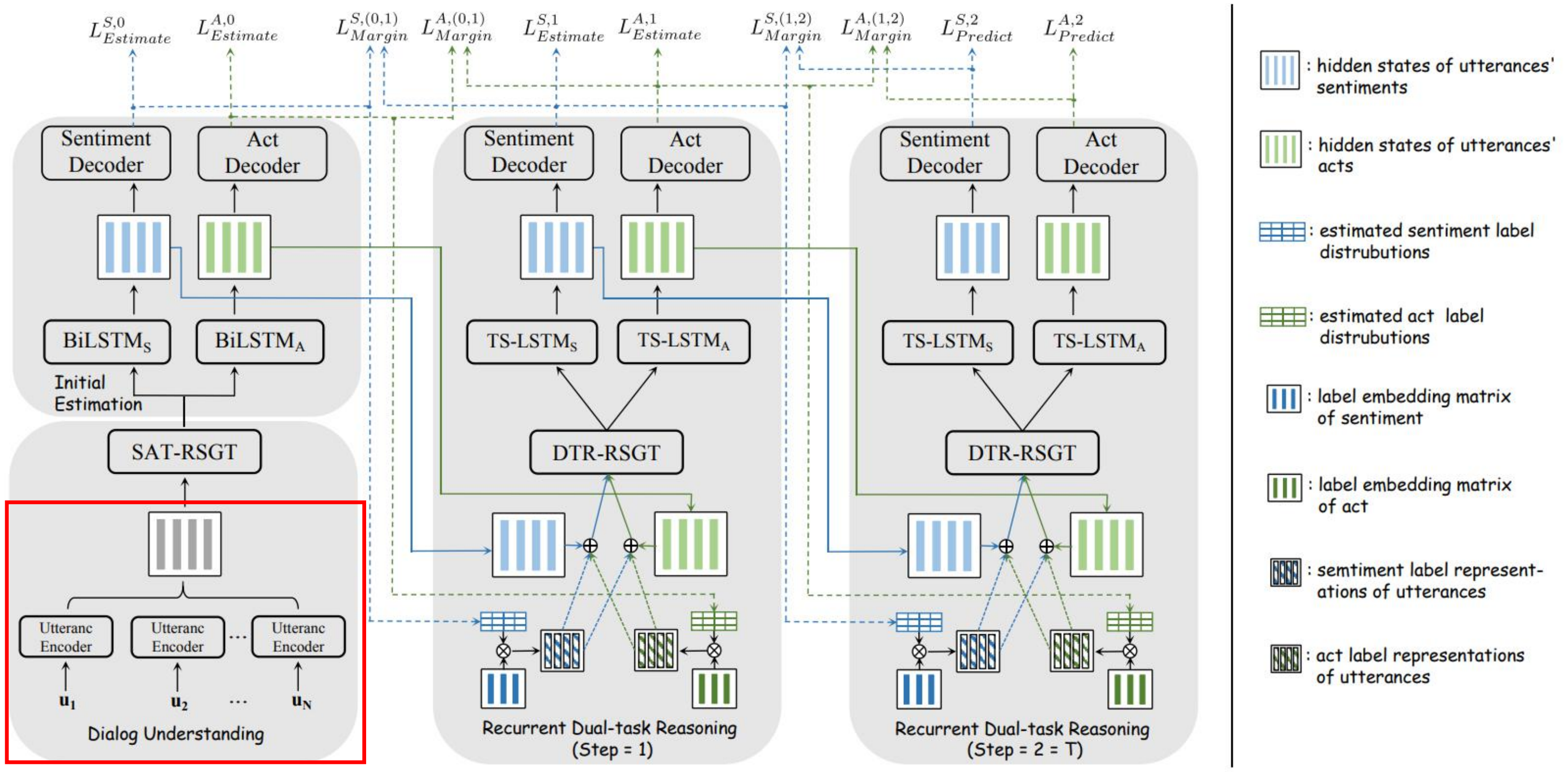


Figure 4: The network architecture of our proposed DARER model. Without loss of generality, the step number T in this illustration is set 2.

Method



Problem Definition

$$\mathcal{D} = \{u_1, u_2, ..., u_N\}$$

$$\text{dialog sentiment labels } Y^S = y_1^s, ..., y_N^s$$

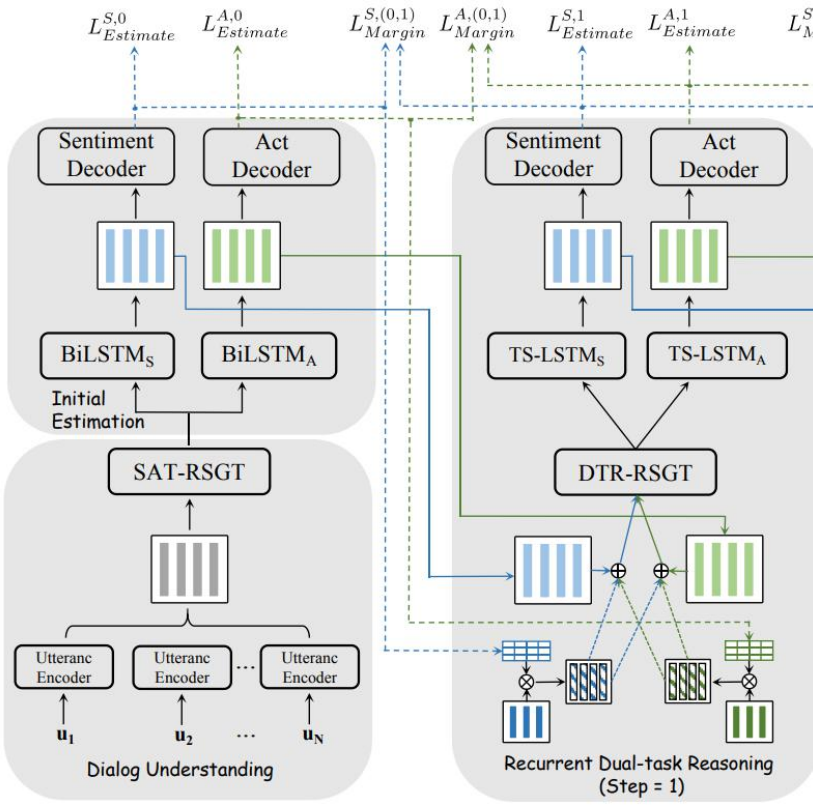
$$\text{dialog act labels } Y^A = y_1^a, ..., y_N^a$$

Utterance Encoder

$$H = (h_0, ..., h_N)$$

$$H_{u,i} = (h_{u,i}^0, ..., h_{u,i}^{l_i})$$

Method



$$M_i^l = \sum_{j \in \mathcal{N}_i} \alpha_{ij} W_{r_{ij}}^l H_j^l, \tag{5}$$

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.

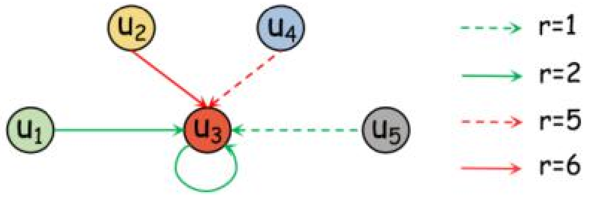


Figure 2: An example of SATG. u_1, u_3 and u_5 are from speaker 1 while u_2 and u_4 are from speaker 2. w.l.o.g, only the edges directed into u_3 node are illustrated.

Speaker-aware Temporal relation-specific graph transformations

Speaker-aware Temporal RSGT

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$$

$$\hat{h}_i = W_1 h_i^0 + \sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} \frac{1}{|\mathcal{N}_i^r|} W_1^r h_j^0 \tag{1}$$

$$\hat{H} = (\hat{h}_0, \ldots, \hat{h}_N)$$

Initial Estimation

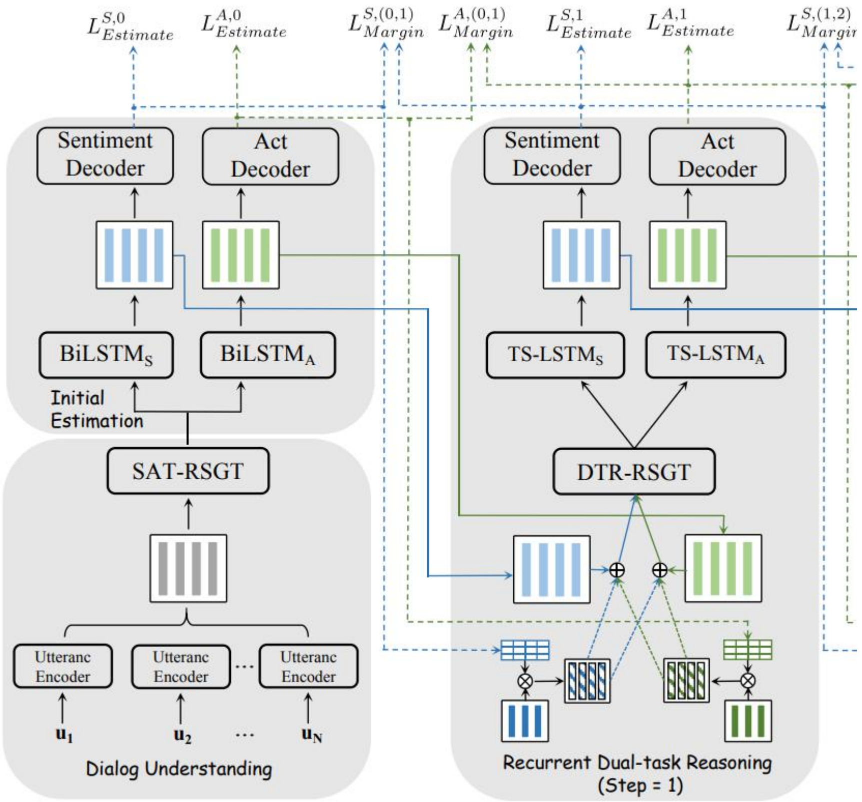
$$H_s^0 = \mathrm{BiLSTM_S}(\hat{H})$$
$$H_a^0 = \mathrm{BiLSTM_A}(\hat{H})$$

where $H_s^0 = \{h_{s,i}^0\}_{i=1}^N$ and $H_a^0 = \{h_{a,i}^0\}_{i=1}^N$

$$P_S^0 = \{P_{S,i}^0\}_{i=1}^N,\ P_A^0 = \{P_{A,i}^0\}_{i=1}^N$$
$$\begin{aligned} P_{S,i}^0 &= softmax(W_d^s h_{a,i}^0 + b_d^s) \\ &= [p_{s,i}^0[0], \ldots, p_{s,i}^0[k], \ldots, p_{s,i}^0(|\mathcal{C}_s|-1)] \quad (2) \\ P_{A,i}^0 &= softmax(W_d^a h_{s,i}^0 + b_d^a) \\ &= [p_{a,i}^0[0], \ldots, p_{a,i}^0[k], \ldots, p_{a,i}^0(|\mathcal{C}_a|-1)] \end{aligned}$$

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.

Method



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.

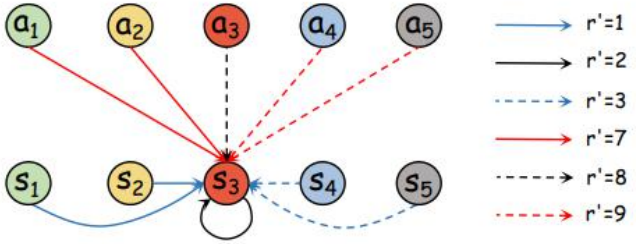


Figure 3: An example of DRTG. s_i and a_i respectively denote the node of DAC task and DAR task. w.l.o.g, only the edges directed into s_3 are illustrated.

Recurrent Dual-task Reasoning

- **Projection of Label Distribution**

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

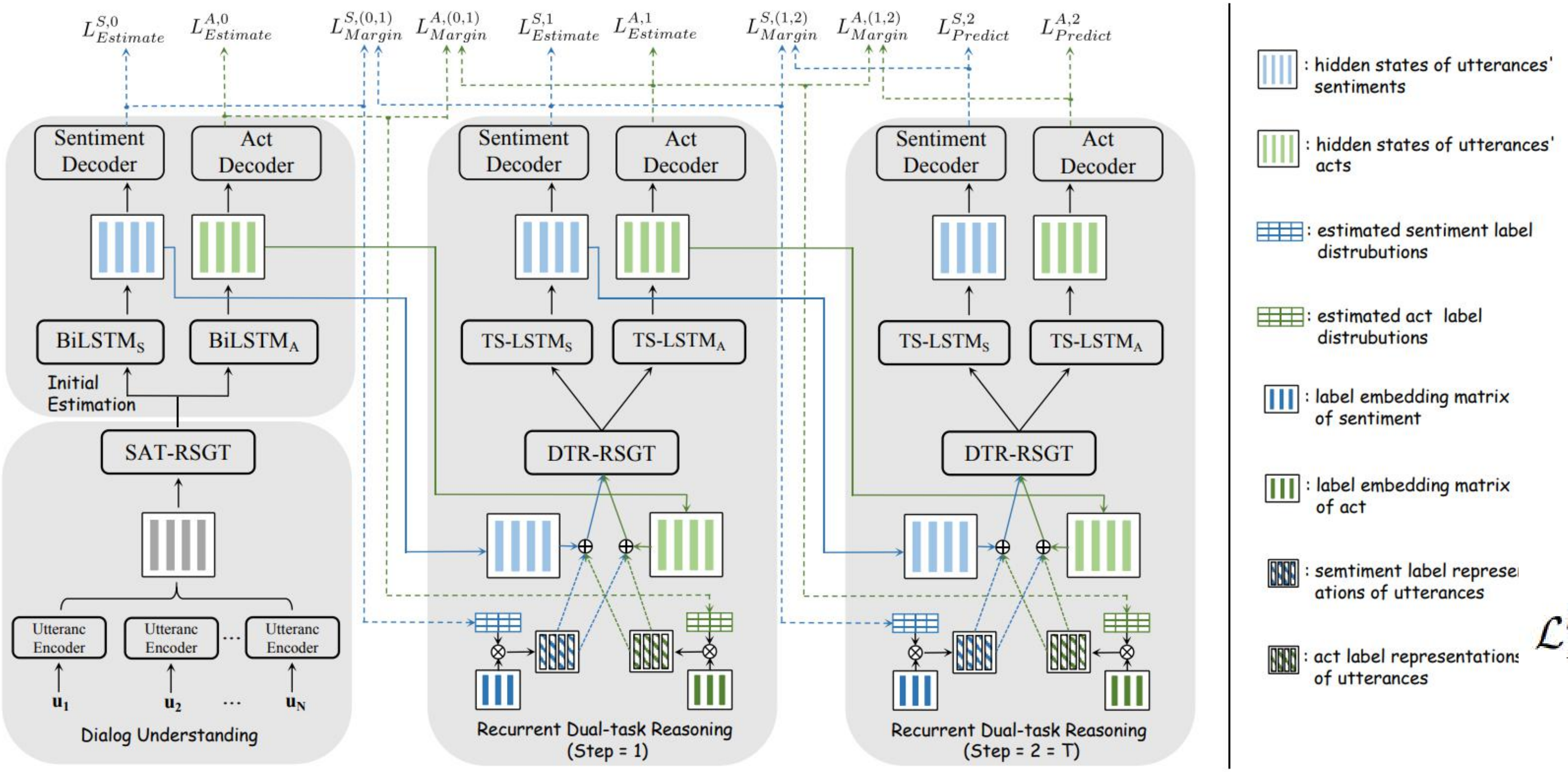
where v^k_s and $v^{k'}_a$ are the label embeddings of sentiment class k and act class k', respectively.

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

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

$$H^t_s = \mathrm{TS\text{-}BiLSTM_S}(\overline{H}^t_s)$$
$$H^t_a = \mathrm{TS\text{-}BiLSTM_A}(\overline{H}^t_a) \quad (6)$$

Method



Estimate Loss

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

Margin Loss

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

Constraint loss

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

Prediction loss

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

Final Training Objective

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

$$\mathcal{L} = \mathcal{L}^{S} + \mathcal{L}^{A} \qquad (12)$$

Experiments

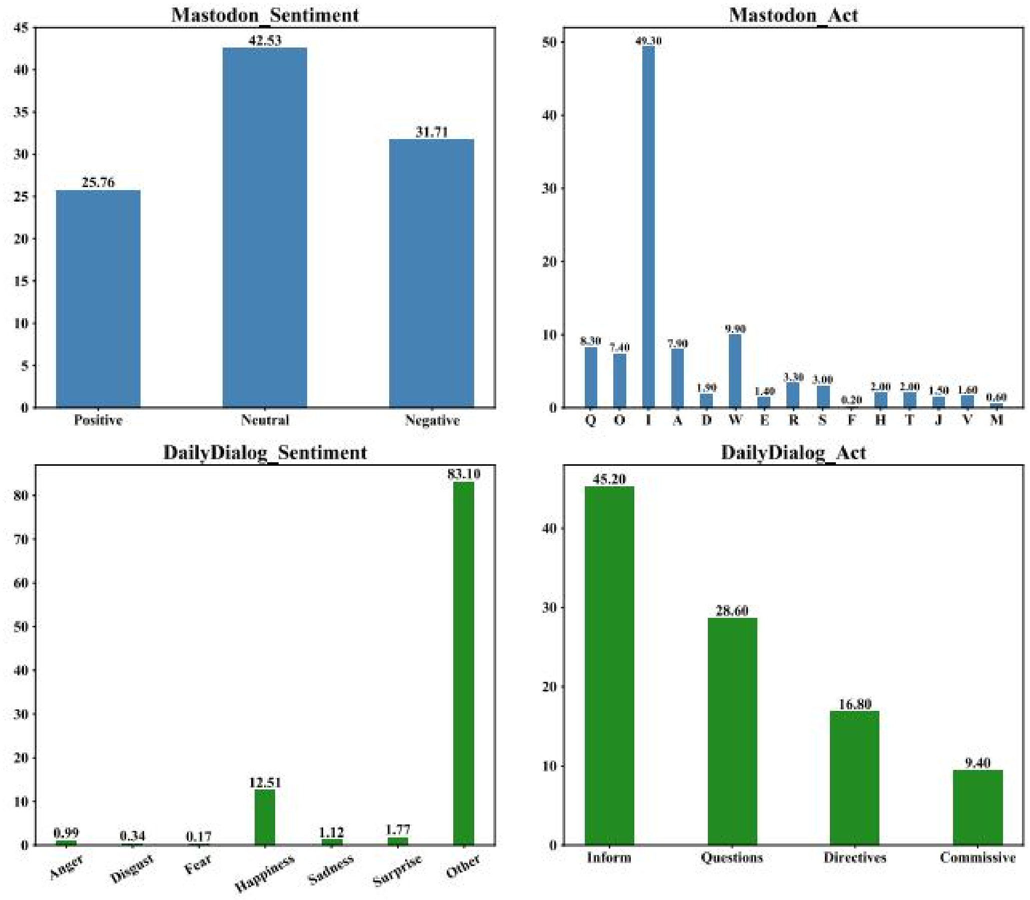


Figure 5: Illustration of class distributions.

Experiments

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 $\pm$2.31	48.11 $\pm$2.91	46.47 $\pm$0.37	62.55 $\pm$0.46	58.66 $\pm$1.71	60.54 $\pm$1.10	58.04 $\pm$0.84	44.65 $\pm$0.36	48.82 $\pm$0.22	79.14 $\pm$0.40	79.71 $\pm$0.16	79.39 $\pm$0.14
DARER	**56.04**† $\pm$0.85	**63.33**† $\pm$0.30	**59.59**† $\pm$0.70	**65.08**‡ $\pm$1.25	**61.88**† $\pm$0.37	**63.43**† $\pm$0.85	**59.96**‡ $\pm$1.25	**49.51**† $\pm$1.33	**53.42**† $\pm$0.18	**81.39**† $\pm$0.55	**80.80**‡ $\pm$0.43	**81.06**† $\pm$0.04

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

Experiments

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.

Experiments

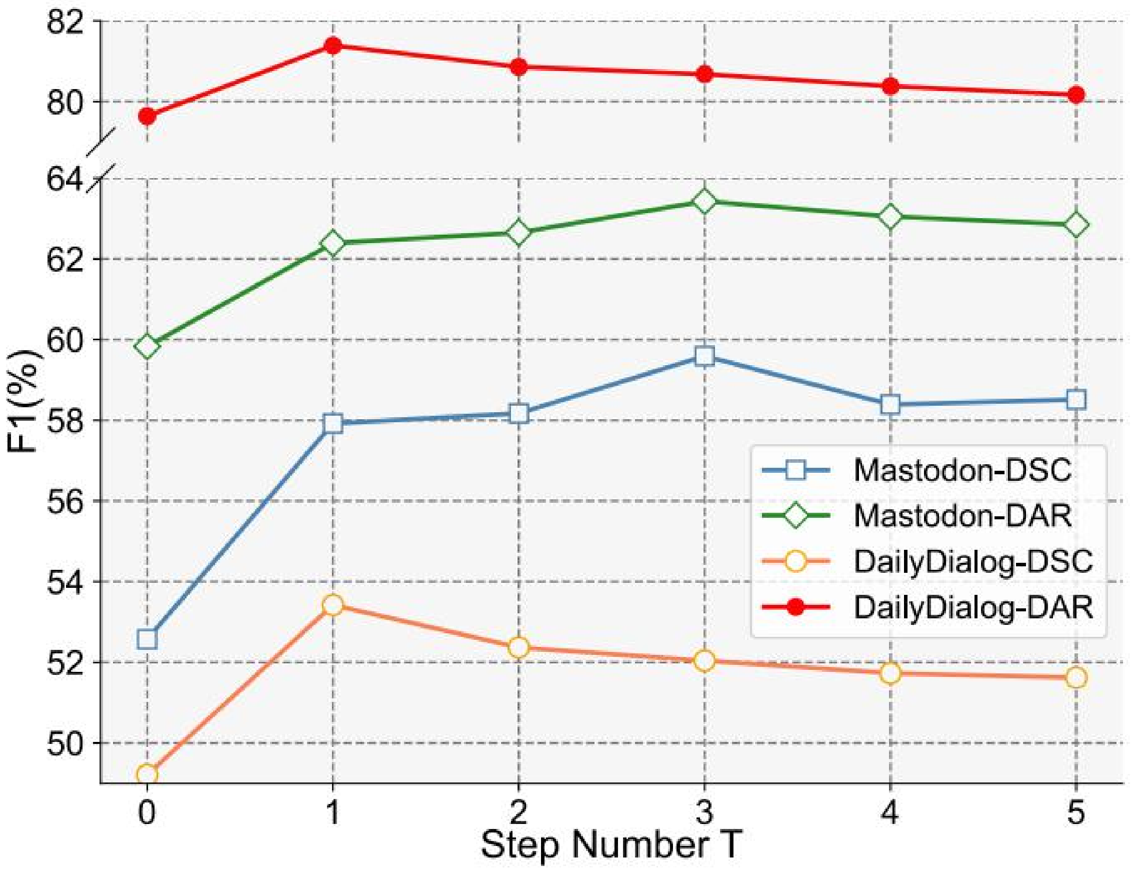


Figure 6: Performances of DARER over different T.

Models		Mastodon DSC P(%)	DSC R(%)	DSC F1(%)	Mastodon DAR P(%)	DAR R(%)	DAR 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.

Experiments

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.