Artificial

Improving Multi-label Malevolence Detection in Dialogues

Yangjun Zhang[1], Pengjie Ren[2*], Wentao Deng[2], Zhumin Chen[2], Maarten de Rijke[1]
[1]University of Amsterdam, [2]Shandong University
[1]{y.zhang6, m.derijke}@uva.nl, [2]{renpengjie, wentao.deng, chenzhumin}@sdu.edu.cn

2022.9.07 • ChongQing

—— ACL 2022

Code: https://github.com/ repozhang/MCRF

Leibniz-Institut für Sozialwissenschaften

Reported by Sijin Liu

1.Introduction

2.Method

3.Experiments

Introduction

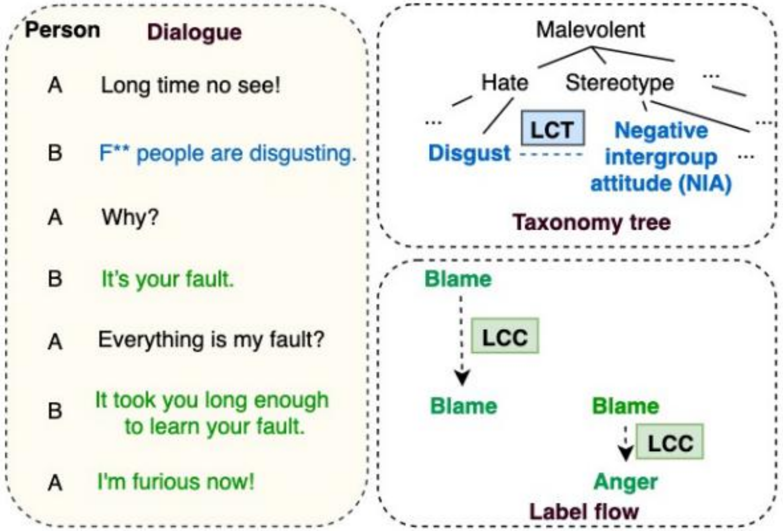


Figure 1: Label correlation in taxonomy (LCT) and label correlation in context (LCC). In terms of LCT, "negative intergroup attitude (NIA)" is correlated with "disgust", which can be reflected by the utterance in blue (LCT). In different turns, "blame" is likely to co-occur with "anger" and "blame", which can be reflected by the utterances in green (LCC).

Conversational Causal Emotion Entailment (C2E2) aims to detect causal utterances for a non-neutral targeted utterance from a conversation.

Causal utterances with different emotions, especially neutral ones (neutral causal utterances occupy 87% of this kind of causes), is still hard to detect even with emotion information. Models are limited in reasoning causal clues and passing them between utterances.

Method

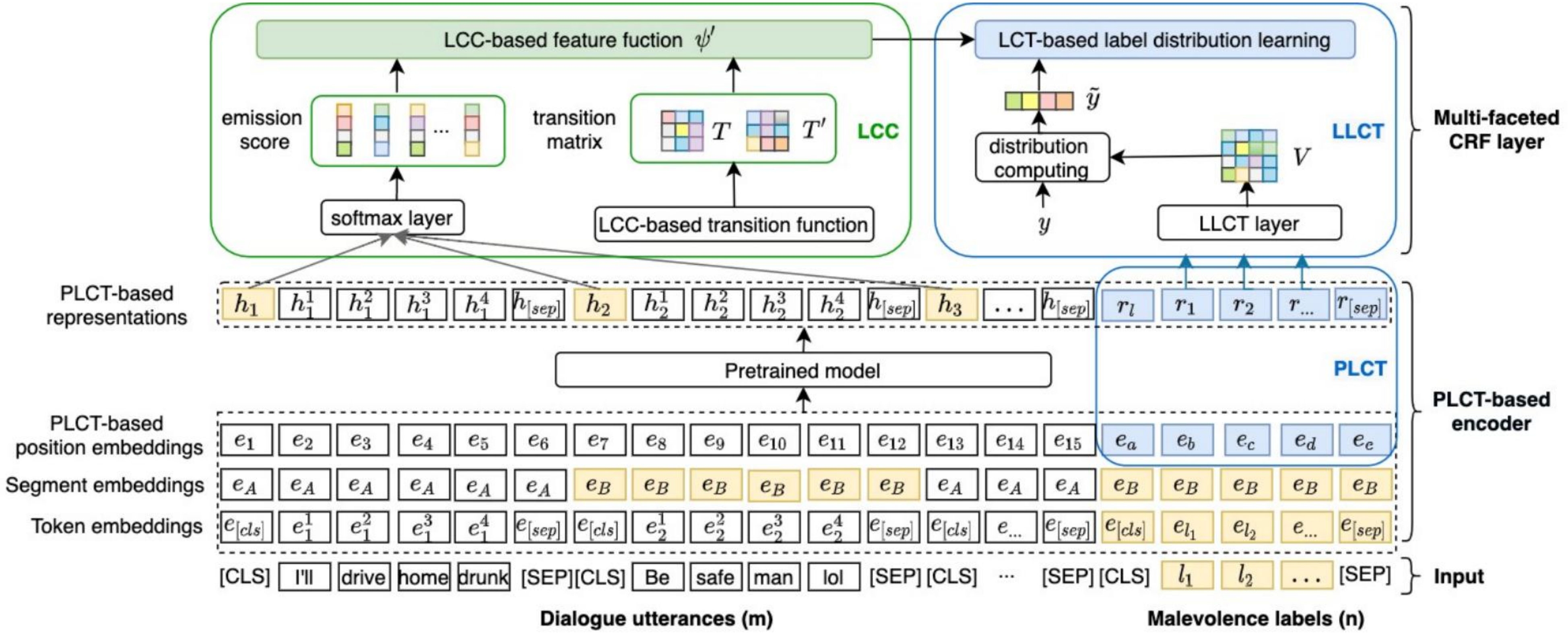


Figure 2: Framework of the proposed *multi-faceted label correlation enhanced CRF* (MCRF) model.

Method

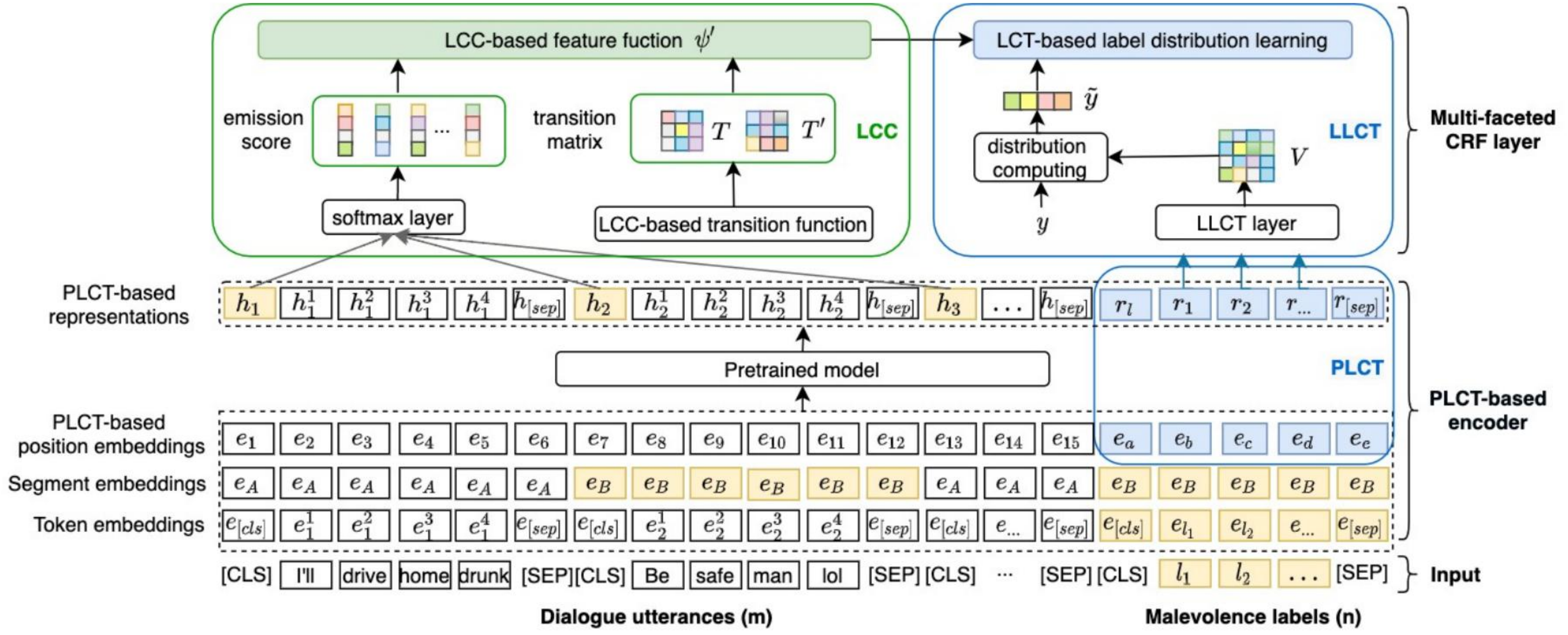


Figure 2: Framework of the proposed *multi-faceted label correlation enhanced CRF* (MCRF) model.

Task Definition

Given a dialogue that contains m utterances, $x = [x_1, x_2, \ldots, x_i, \ldots, x_m]$ and x_i is the i-th utterance in the dialogue. $y = [y_1, y_2, \ldots, y_i, \ldots, y_m]$ denotes the label sequence of one dialogue, where $y_i \in \{0, 1\}^n$ is the label for each utterance. $l = \{l_1, l_2, \ldots, l_j, \ldots, l_n\}$ denotes the label set, where l_j is the j-th label,
categories. *Multi-label dialogue malevolence detection* (MDMD) aims to assign the most reliable labels to each x_i. Since there is no large-scale

Method

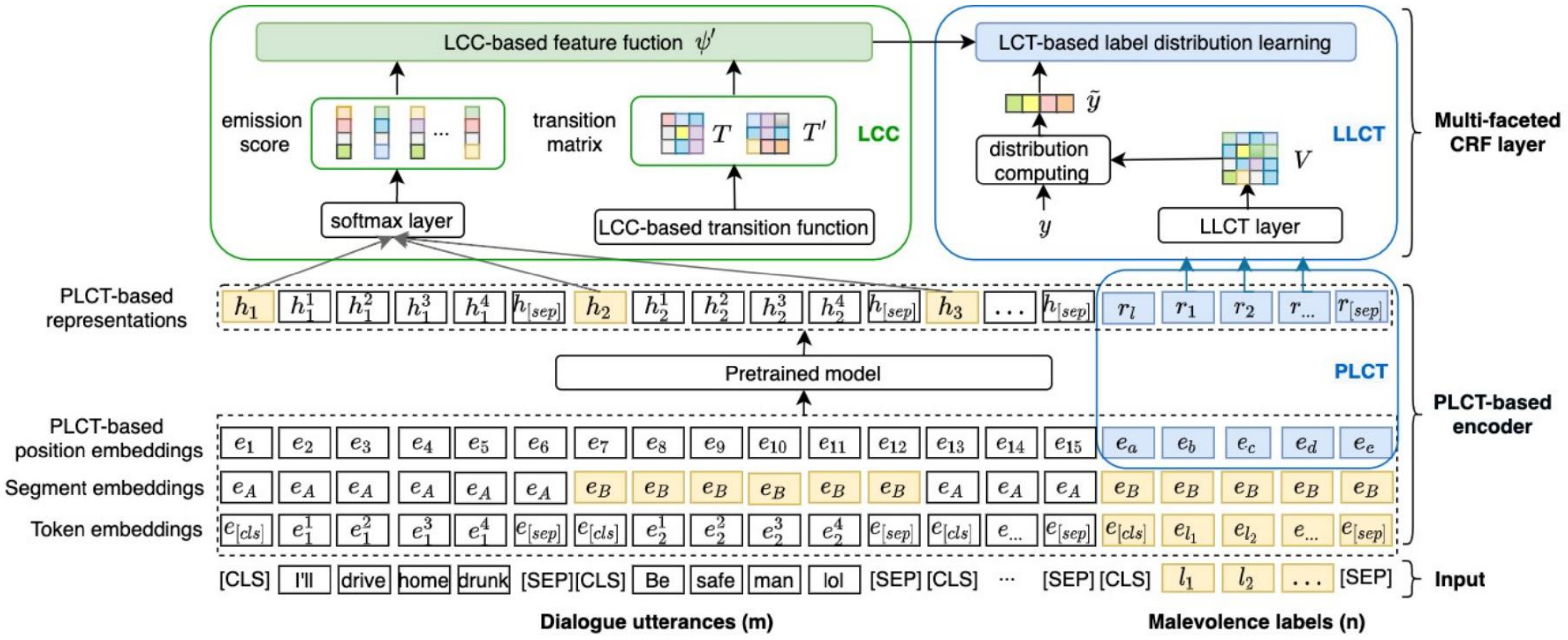


Figure 2: Framework of the proposed *multi-faceted label correlation enhanced CRF* (MCRF) model.

Utterance and label encoder

$$
\begin{aligned}
H, R &= PTM([e(x_i), e(l_j)]), \\
e &= e_{tok} + e_{seg} + e_{pos},
\end{aligned}
\tag{1}
$$

$$H = \{h_1, h_2, \ldots, h_i, \ldots, h_m\}$$

$$R = \{r_1, r_2, \ldots, r_j, \ldots, r_n\}$$

Method

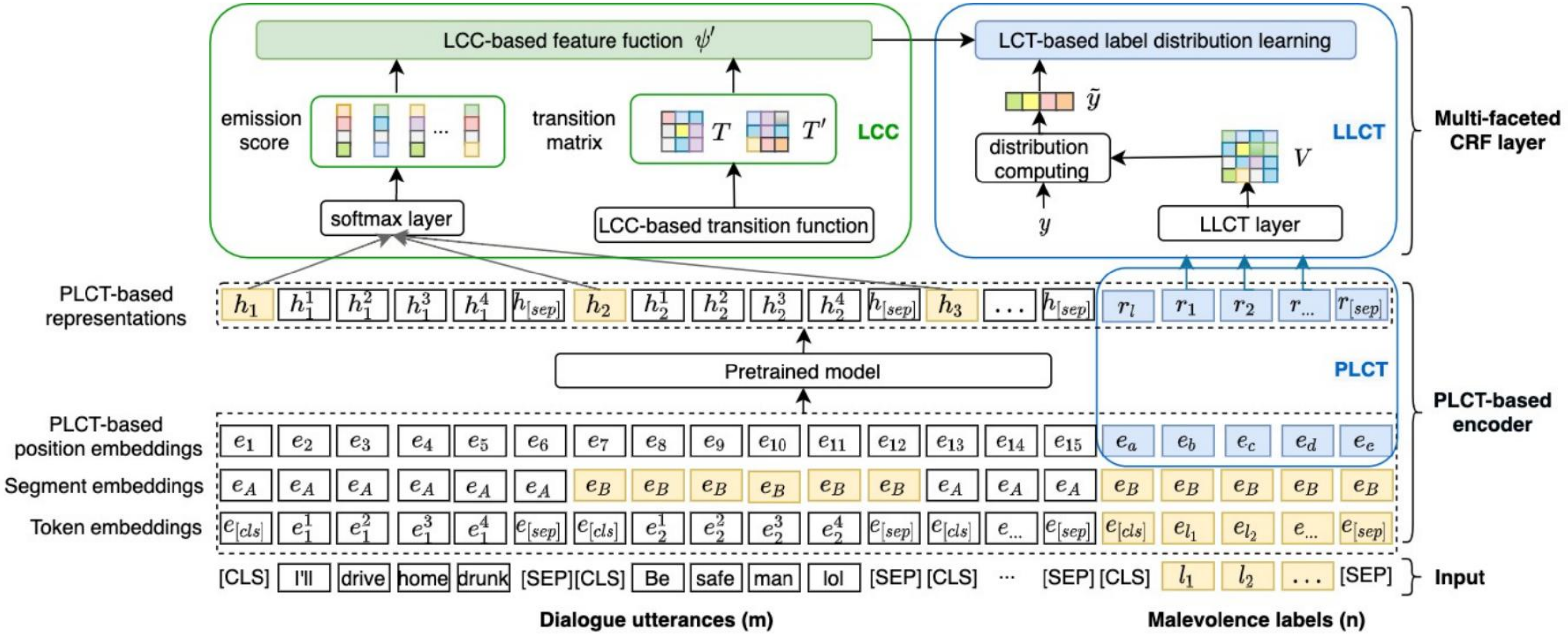


Figure 2: Framework of the proposed *multi-faceted label correlation enhanced CRF* (MCRF) model.

Multi-faceted label correlation

Label correlation in taxonomy

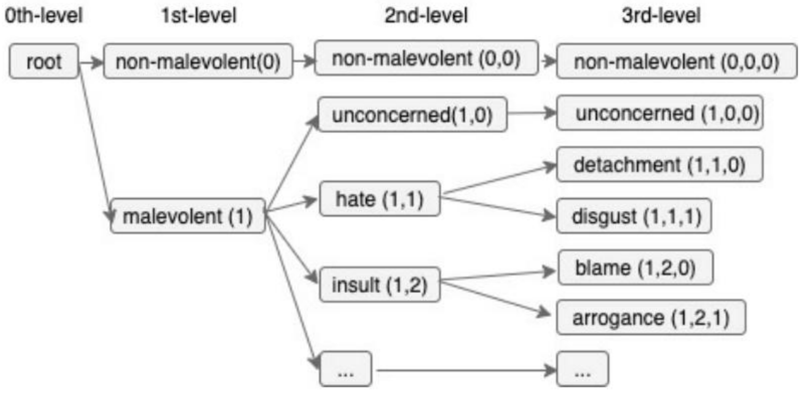


Figure 3: Demonstration of taxonomy tree of labels.

$$V = \frac{1}{2}(\hat{V}_{j,j'} + V'_{j,j'}), \qquad (2)$$

LLCT $\hat{V}_{j,j'} = d(r_j, r_{j'})$ $\quad V'_{j,j'} = d(c_j, c_{j'})$

c_j and $c_{j'}$ are the n-gram bag-of-words vectors of the utterances belong to the j-th and j'-th label, respectively.

Method

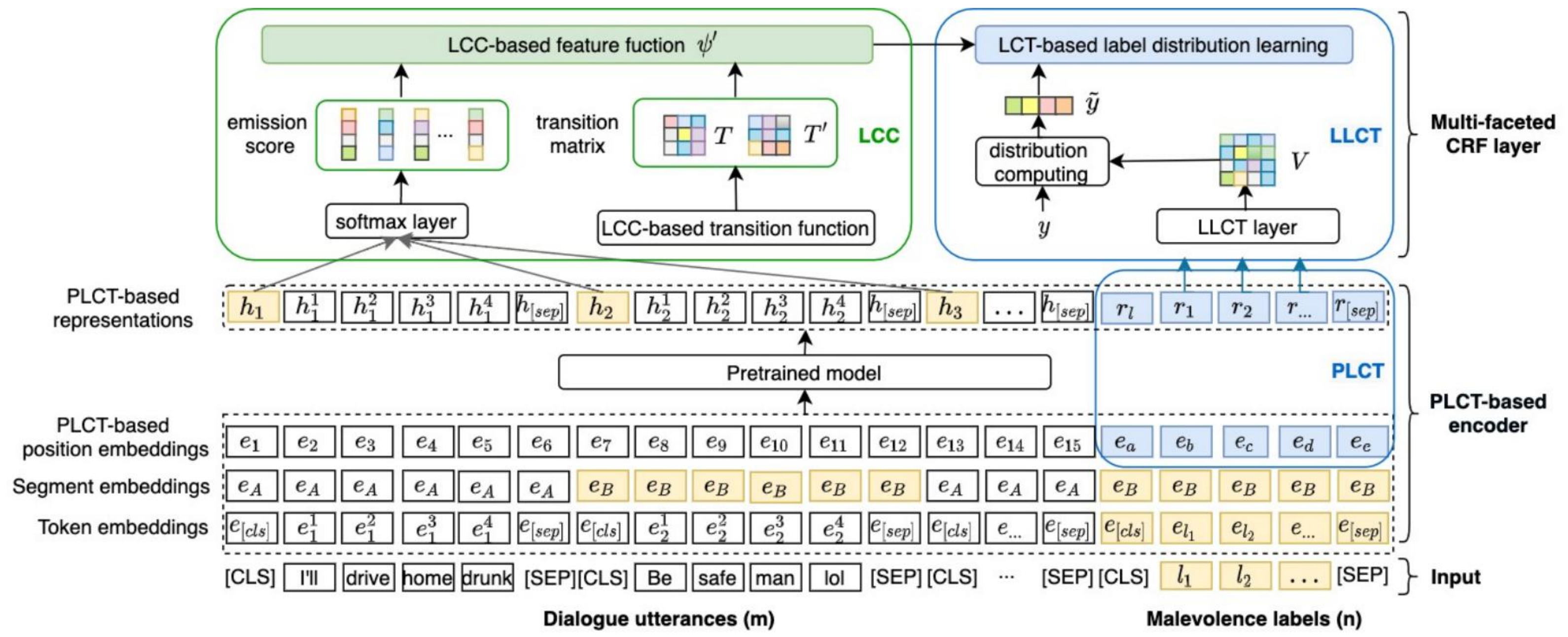


Figure 2: Framework of the proposed *multi-faceted label correlation enhanced CRF* (MCRF) model.

Multi-faceted label correlation

Label correlation in context

$$
\begin{aligned}
t(y_{i-1} = l_j, y_i = l_{j'}) &= T_{(l_j, l_{j'})}, \\
t'(y_{i-2} = l_j, y_i = l_{j'}) &= T'_{(l_j, l_{j'})},
\end{aligned}
\qquad (3)
$$

where l_j and $l_{j'}$ denote the j-th and j'-th labels. T and T' are two $n \times n$ matrices initialized randomly and trained by LCC-based label distribution learning, which is introduced next.

Method

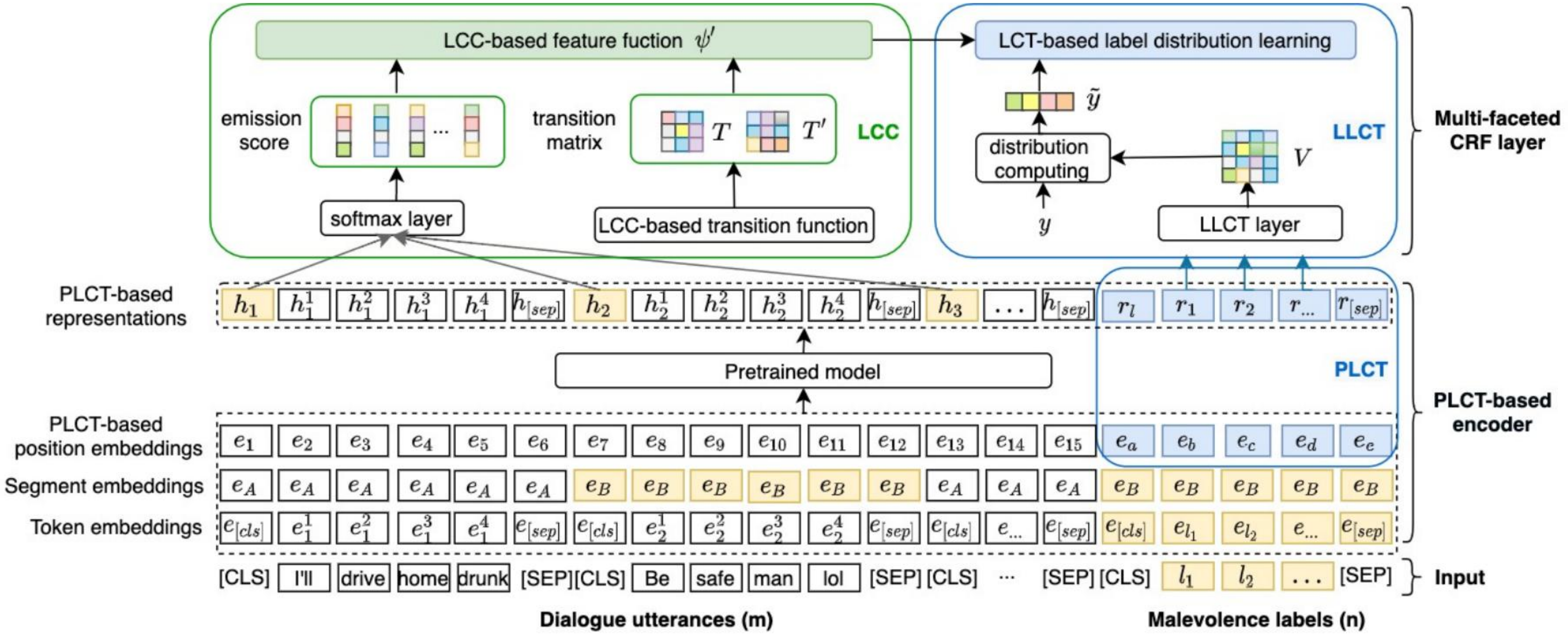


Figure 2: Framework of the proposed *multi-faceted label correlation enhanced CRF* (MCRF) model.

Multi-faceted CRF layer

Given a sequence of utterances, a linear chain CRF can be used to predict the label of an utterance:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_i \psi(x_i, y_i)\right), \quad (4)$$

where Z is a normalization function, and

$$\psi(x, y) = \sum_i s(y_i, x) + \sum_i t(y_{i-1}, y_i), \quad (5)$$

LCC-based feature function

$$s(y_i, x) = \mathrm{softmax}(h_i), \quad (6)$$

$$\psi'(x, y) = \frac{1}{2}\left(\psi(x, y) + \sum_i s(y_i, x) + \sum_i t'(y_{i-2}, y_i)\right), \quad (7)$$

Method

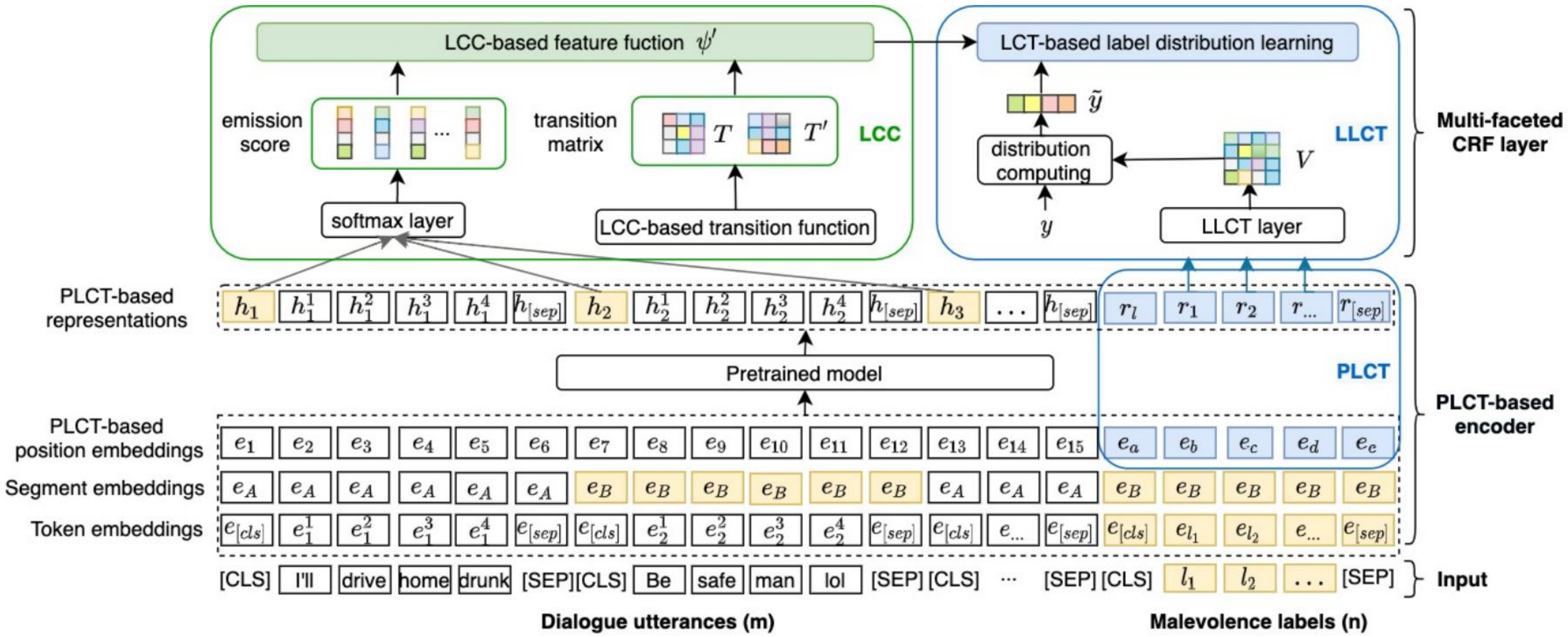


Figure 2: Framework of the proposed *multi-faceted label correlation enhanced CRF* (MCRF) model.

Multi-faceted CRF layer

LCT-based label distribution learning

$$\tilde{y_i} = \lambda V y_i + y_i, \tag{8}$$

where λ denotes how much the original one-hot distribution is redefined and V is the matrix that estimates the LCT in Eq. 2.

Our training objective is the KL-divergence loss except that we replace gold label y with estimated gold label $\tilde{y}$:

$$\mathcal{L} = \sum_{y} q(y|x) \log \frac{q(y|x)}{p(y|x)}, \tag{9}$$

where $q(y|x)$ is the target distribution to learn, we use the probability of $\tilde{y}$ given x for $q(y|x)$; $p(y|x)$ is the predicted distribution.

$$\begin{aligned}\mathcal{L} = &\sum_{i}\sum_{y_i}\{q(y_i|x)\log q(y_i|x)\} \\ &- \sum_{y}\{q(y|x)\psi'(y,x)\} + \log Z(x),\end{aligned} \tag{10}$$

Experiments

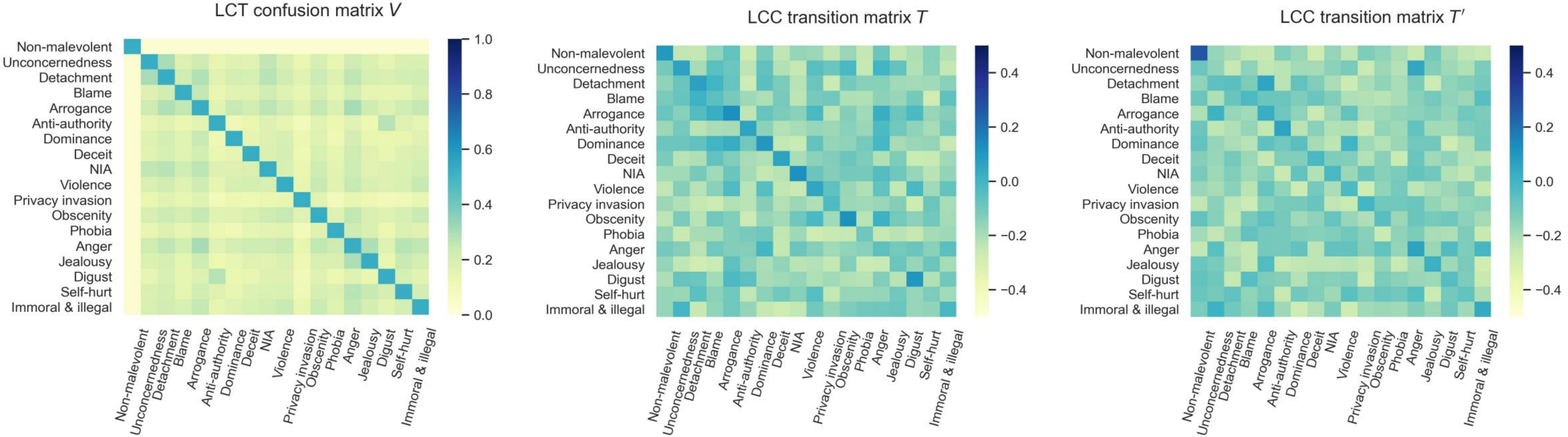
	Malevolent		Non-malevolent		Total
	Valid.	Test	Valid.	Test	
1-label	413	733	2,088	4,276	7,510
2-label	264	574	–	–	838
3-label	22	85	–	–	107
4-label	2	5	–	–	7
Total	701	1,397	2,088	4,276	8,462

Table 1: Statistics of the validation and test sets of MDMD.

Model	Precision	Recall	F1	Jaccard
BERT	67.73	33.59	42.32	37.25
BERT-CRF	69.62	33.57	43.30	40.83
BERT-MCRF	**82.99**	**38.12**	**49.20**	**43.46**

Table 2: Main results of MCRF on the MDMD test set.

Experiments



(a) LCT confusion matrix V.

(b) LCC transition matrix T.

(c) LCC transition matrix T'.

Figure 4: Visualization of LCT and LCC.

Experiments

Model	1-label	2-label	3-label	4-label
BERT	40.16	11.84	11.48	8.00
BERT-CRF	44.02	13.06	11.89	**11.33**
BERT-MCRF	**46.39**	**15.23**	**12.88**	10.00

Table 3: Jaccard scores of different label groups.

Settings	Precision	Recall	F1	Jaccard
LCT ($\lambda = 0$)	83.60	36.78	47.96	42.75
LCT ($\lambda = 1/2$)	84.58	37.04	48.50	42.85
LCT ($\lambda = 1$)	**82.99**	**38.12**	**49.20**	**43.46**
LCT ($\lambda = 2$)	82.28	38.09	49.10	42.98
LCC (T)	84.37	37.08	48.58	43.43
LCC (T')	84.43	35.99	47.10	42.62
LCC (T+T')	**82.99**	**38.19**	**49.20**	**43.46**

Table 4: BERT-MCRF performance w.r.t. different LCT and LCC settings. λ is the hyperparameter in Eq. 8, T and T' are the transition matrices by Eq. 3.

Experiments

Model	Precision	Recall	F1	Jaccard
BERT-MCRF	82.99	38.19	49.20	43.46
−LCC	84.37	37.08	48.58	43.43
−LLCT	83.60	36.78	47.96	42.75
−PLCT	69.34	33.79	43.27	40.86
−LCT	69.87	33.16	42.62	40.83

Table 5: Ablation study results. Note that LCC of different users T is already captured by BERT-CRF, therefore the ablation of LCC keeps T but deletes T'.

Metric	Precision	Recall	F1	Jaccard
BERT	64.14	36.33	43.97	37.02
BERT-CRF	73.05	35.84	44.81	40.89
BERT-MCRF	**80.39**	**39.82**	**50.58**	**42.68**

Table 6: Main results of BERT-MCRF on the multi-label validation set.

User A: S** my head into a brick building while having a steal toe boot r** up my a** would hurt less.
BERT: non-malevolent. **BERT-CRF**: non-malevolent. **BERT-MCRF**: violence, self-hurt. **Gold**: violence, self-hurt.
User A: Mom: I can't believe you haven't seen birdman , Edward Norton is in it ! n Me: I know she gets me. User B: Hope Gasols forgive me when I marry him. User A: Invite me so i can get drunk and be inappropriate.
BERT: non-malevolent; non-malevolent; immoral & illegal. **BERT-CRF**: non-malevolent; non-malevolent; immoral & illegal. **BERT-MCRF**: non-malevolent; non-malevolent; non-malevolent. **Gold**: non-malevolent; non-malevolent; non-malevolent.

Table 7: Case study. Upper: utterances and labels of example 1; bottom: utterances and labels of example 2.