



Co-Evolving Graph Reasoning Network for Emotion-Cause Pair Extraction

Bowen Xing^{1,2,3} and Ivor W. Tsang^{1,2,3}(✉)

¹ AAIL, University of Technology Sydney, Sydney, Australia

`ivor_tsang@cfar.a-star.edu.sg`

² CFAR, Agency for Science, Technology and Research, Singapore, Singapore

³ IHPC, Agency for Science, Technology and Research, Singapore, Singapore

Abstract. Emotion-Cause Pair Extraction (ECPE) aims to extract all emotion clauses and their corresponding cause clauses from a document. Existing approaches tackle this task through multi-task learning (MTL) framework in which the two subtasks provide indicative clues for ECPE. However, the previous MTL framework considers only one round of multi-task reasoning and ignores the reverse feedbacks from ECPE to the subtasks. Besides, its multi-task reasoning only relies on semantics-level interactions, which cannot capture the explicit dependencies, and both the encoder sharing and multi-task hidden states concatenations can hardly capture the causalities. To solve these issues, we first put forward a new MTL framework based on Co-evolving Reasoning. It (1) models the bidirectional feedbacks between ECPE and its subtasks; (2) allows the three tasks to evolve together and prompt each other recurrently; (3) integrates prediction-level interactions to capture explicit dependencies. Then we propose a novel multi-task relational graph (MRG) to sufficiently exploit the causal relations. Finally, we propose a Co-evolving Graph Reasoning Network (CGR-Net) that implements our MTL framework and conducts Co-evolving Reasoning on MRG. Experimental results show that our model achieves new state-of-the-art performance, and further analysis confirms the advantages of our method.

Keywords: Multi-Task Learning · Relational Graph Reasoning · Emotion-Cause Extraction · Natural Language Processing

1 Introduction

Emotion-Cause Pair Extraction (ECPE) is a new while challenging task in the field of natural language processing/artificial intelligence. It aims to automatically extract all emotion clauses and the corresponding cause clauses from a raw document, which is of great value for real-world application [26]. Consider a document “[*In the memory of the students*]₁, [*he often paid the tuition fees for them*]₂, [*which is respectable and touching*]₃.”, the third clause expresses an emotion, which is triggered by the second clause, so these two clauses form an emotion-cause pair. Intuitively, detecting the clauses that express causes and

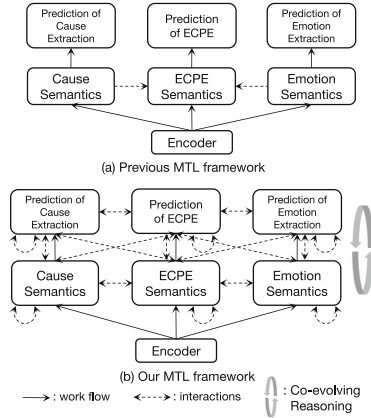


Fig. 1. Comparison of the previous MTL framework and our MTL framework for ECPE.

emotions, namely cause extraction (CE) and emotion extraction (EE), are two subtasks of ECPE. Accordingly, recent models [1, 10, 13, 37] implement the multi-task learning framework shown in Fig. 1(a), introducing CE and EE to provide indicative clues for ECPE.

Although the previous multi-task learning (MTL) framework has made promising progress, based on our observation, it still suffers from several issues which hinder the multi-task reasoning between ECPE and CE/EE. First, previous works only consider one-way messages from CE/EE to ECPE. The predictions of CE/EE may be incorrect due to their unreliable semantics. In this case, the false information from CE/EE may mislead ECPE. Second, the single-round multi-task reasoning process in previous works is not competent, considering that it is hard for machines to understand emotions, causes, and their causalities like humans due to the inherent ambiguity and subtlety of emotions and causes [12]. Third, in previous works, the multi-task reasoning between ECPE and CE/EE is only achieved by implicit semantics-level interactions such as shared encoders. For one thing, this is inconsistent with human intuition and the causal relations between ECPE and CE/EE, both of which are based on predictions or labels; for another, the indicative information conveyed in semantics is implicit and relatively insufficient compared with prediction information.

On account of the above issues, we propose a new MTL framework based on Co-evolving Reasoning as shown in Fig. 1(b). Firstly, in addition to the one-way message from CE/EE to ECPE, our MTL framework also models the reverse feedbacks from ECPE to CE/EE. If ECPE predicts correctly, the indicative information transferred to CE/EE can improve them. And the two improved subtasks can further promote ECPE reversely. If ECPE predicts incorrectly, the defective information transferred to CE/EE can act as their feedback and make them rethink to provide better information for ECPE, which thus can further prompt CE/EE reversely. Secondly, to achieve this virtuous cycle, we design the recurrent multi-task reasoning mechanism. In this way, the knowledge of

the three tasks can gradually evolve together and mutually prompt each other. Thirdly, we propose to exploit the correlations in predictions via introducing two kinds of prediction-level interactions: prediction-prediction and prediction-semantics interactions. In this way, explicit correlative information conveyed by predictions (estimated label distributions) can flow in our MTL framework then facilitate *Co-evolving Reasoning*. And the semantics can get straightforward feedback at each step from the predictions then rethink to improve.

Furthermore, the MTL sequence structures employed in previous works are simply based on shared encoder and multi-task hidden states concatenations, which can hardly capture the causal relations. This motivates us to seek a more effective method to sufficiently exploit the causalities among ECPE and CE/EE. To this end, we design a novel multi-task relational graph (MRG), in which there are three groups of nodes derived from the clauses in the document and corresponding to ECPE, CE, and EE, respectively. Moreover, we design different relation types which correspond to the causal relations between ECPE and CE/EE.

To implement our MTL framework and conduct Co-evolving Reasoning on MRG, we propose a Co-evolving Graph Reasoning Network (CGR-Net), whose core is a multi-task relational graph transformation (MRGT) cell. CGR-Net first generates the task-specific hidden states and produces the initial estimated label distributions. Then the MRGT cell recurrently takes the hidden states and label distributions of the three tasks as input and then updates them in three steps: projection of label distributions, relational local graph transformation, and non-local self-transformation. Finally, the predictions of ECPE at the final step are used to extract the potential emotion-cause pairs. And we design a harness loss based on logical constraints to force the three tasks to gradually promote each other in the virtuous cycle of Co-evolving Reasoning.

In summary, our contributions are three-fold:

- (1) We propose a new MTL framework based on Co-evolving Reasoning and an MRG to exploit the correlations and causal relations sufficiently. To the best of our knowledge, our MTL framework is the first one allowing ECPE and CE/EE to promote each other recurrently, and MRG is the first MTL graph structure for ECPE.
- (2) To implement our MTL framework and conduct Co-evolving Reasoning on MRG, we propose a novel CGR-Net, whose core is a multi-task relational graph transformation cell.
- (3) Experimental results on the benchmark dataset show that our CGR-Net significantly outperforms existing state-of-the-art models. And further analysis proves the effectiveness of different components of CGR-Net and the superiority of MRG.

2 Related Works

Emotion cause extractions (ECE) [4, 9, 14–16, 18, 20, 21, 36] is a long-standing task whose objective is to extract the causes of given emotion expressions in the

document. However, it requires that the emotions must be annotated manually, which constrains the practical application. Therefore, recently [26] propose the ECPE task and a two-step solution while the error propagation may occur from the first step to the second. To this end, recent works propose unified end-to-end models [2, 5, 6, 10, 11, 13, 23, 25] to tackle ECPE in the MTL framework.

[10] propose a model integrating the 2D representation of emotion-cause pair, the interactions, and predictions. [23] tackle ECPE from a ranking perspective and adopt kernel-based relative position embedding for ranking. [38] propose a tagging scheme coding the distance between the emotion clause and cause clause in an emotion-cause pair. Based on this, [13] propose a tag distribution refinement method that adjusts the output label distribution of ECPE using the ones of CE and EE according to a pre-defined rule. However, the refinement method does not participate in model training, only working on the output of evaluation.

More recently, Multi-Granularity Semantic Aware Graph model (MGSAG) [1] incorporates fine-grained and coarse-grained semantic features jointly, aiming to resolve the distance limitation of clause semantics. And the Matrix Capsule-based multi-granularity framework (MaCa) [37] introduces the matrix capsule to obtain more fine-grained features of clause pairs, clustering the relationship of each clause pair.

Different from previous works, we (1) propose Co-evolving Reasoning, which allows the three tasks gradually and sufficiently promote each other; (2) introduce prediction-prediction and prediction-semantics interactions to model the explicit correlations and provide feedback for semantics which can rethink to improve; (3) effectively exploit the casual relations via designing a novel MRG.

3 Methodology

Before delving into MRG and CGR-Net, we first introduce the task formulation in our work.

We cast ECPE as a *tag* classification task and use the cause-centric tagging scheme [38]. Each clause x_i has a two-tuples tag $y_i^t = (y_i^{t,c}, y_i^{t,d}) \in \mathcal{C}_t$, where $y_i^{t,c} \in \{C, O\}$ denotes whether x_i is a cause clause, and $y_i^{t,d} \in \{-\gamma, \dots, -1, 0, 1, \dots, \gamma, \perp\}$ denotes the distance between x_i and its triggered emotion clause, while ‘ \perp ’ always associates with ‘O’, denoting that x_i is a non-cause clause. And γ is a hyperparameter controlling the max span of emotion-cause pairs. Thus ECPE (*tag*) totally has $|\mathcal{C}_t| = 2(\gamma + 1)$ classes.

As for CE (*cause*) and EE (*emotion*), they are both formulated as binary classification tasks: $y_i^c \in \mathcal{C}_c = \{1, 0\}$ and $y_i^e \in \mathcal{C}_e = \{1, 0\}$.

3.1 Constructing a MRG from a Document

In this paper, we design a multi-task relational graph (MRG) $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ to exploit the causalities via modeling the self- and mutual-interactions of the three tasks (*cause*, *tag* and *emotion*). Each clause x_i in document \mathcal{D} derives

Table 1. Relation types in MRG, w.l.o.g. $\gamma = 2$. $I_t(i)$ indicates node i is a cause (c) node, tag (t) node or emotion (e) node. ‘-’ denotes the set of $[-2, -1, 0, 1, 2]$.

r_{ij}	cc	tt	ee	ct	tc	te:-2	te:-1	te:0	te:1	te:2	et:-2	et:-1	et:0	et:1	et:2
$I_t(i)$	c	t	e	c	t	t	t	t	t	t	e	e	e	e	e
$I_t(j)$	c	t	e	t	c	e	e	e	e	e	t	t	t	t	e
$rdis(i, j)$	-	-	-	0	0	-2	-1	0	1	2	-2	-1	0	1	2

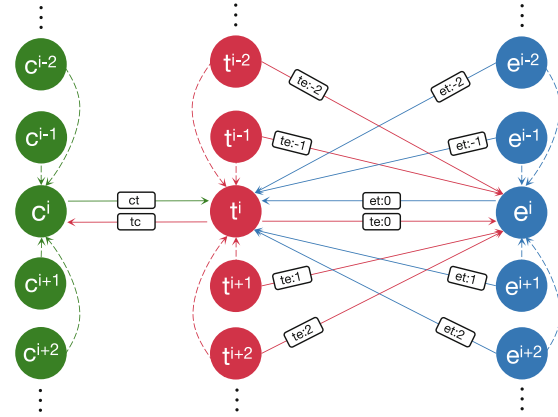


Fig. 2. An example of MRG ($\gamma = 2$). W.L.O.G, only the edges directed into c_i, t_i and e_i are illustrated. And self-loops are not shown for simplification.

three nodes c_i, t_i and e_i , respectively for *cause*, *tag* and *emotion*, thus $|\mathcal{V}| = 3n$. The edge $(i, j, r_{ij}) \in \mathcal{E}$ denotes the information propagation from node i to node j , and $r_{ij} \in \mathcal{R}$ is the relation type of the edge. Note that node i and node j may correspond to different tasks. We define three kinds of rules to determine the connection between two nodes in MRG:

Direction: $(j, i, r_{ji}) \in \mathcal{E}$ if $(i, j, r_{ij}) \in \mathcal{E}$. In MRG, the information propagation between two nodes is bidirectional. This guarantees the bidirectional correlations between *tag* and *cause/emotion*. **Local Connection:** $\forall (i, j, r_{ij}), |rdis(i, j)| \leq \gamma$, where $rdis(i, j)$ denotes the relative distance between the clauses of node i and node j in \mathcal{D} . In general, the probability of two distant clauses having causal relation is relatively small regarding the cohesion and coherence of discourse [7]. Therefore, the edges in MRG are based on local connections, and in this work, we constrain that the relative distance of two connected nodes' clauses in \mathcal{D} ranges from $-\gamma$ to γ , consistent with the span range of the ECPE tag.

Relation Type for Causality: Table 1 lists the relation types in MRG. And an example of MRG is shown in Fig. 2. To capture the self-task local contextual dependencies, we define $r_{ij} = cc$, $r_{ij} = tt$ and $r_{ij} = ee$ to model the local self-transformation of *cause*, *tag* and *emotion*, respectively. As for inter-task interactions, first of all, regarding the scheme of *tag* task, there are two explicit

causal relations between *cause* and *tag* tasks: (1) if $y_i^c = 1$, then $y_i^{t,c} = C$, and vice versa; (2) if $y_i^c = 0$, then $y_i^{t,c} = O$, and vice versa. To model these causalities in MRG, we define $r_{ij} = ct$ and $r_{ij} = tc$ to achieve the mutual transformations between *cause* and *tag*. Besides, there are four kinds of causal relations between *tag* and *emotion* tasks. First, if $y_i^{t,c} = C$, there is at least one emotion clause among $x_{i-\gamma} \sim x_{i+\gamma}$: $y_{i-\gamma}^e = 1$ or $y_{i-\gamma+1}^e = 1$ or ... or $y_{i+\gamma-1}^e = 1$ or $y_{i+\gamma}^e = 1$. Second, if there is no emotion clause among $x_{i-\gamma} \sim x_{i+\gamma}$: $y_{i-\gamma}^e = 0$ and $y_{i-\gamma+1}^e = 0$ and ... and $y_{i+\gamma-1}^e = 0$ and $y_{i+\gamma}^e = 0$, then $y_i^{t,c} = O$ and $y_i^{t,d} = \perp$. Third, if $y_i^{t,d} = m$, then $y_{i+m}^e = 1$. Reversely, $y_{i+m}^e = 1$ cannot deduce $y_i^{t,d} = m$, but intuitively if $p(y_{i+m}^e = 1)$ increases, $p(y_i^{t,d} = m)$ should also increase. Forth, if $y_{i+m}^e = 0$, then $y_i^{t,d} \neq m$. Reversely, although $y_i^{t,d} \neq m$ cannot deduce $y_{i+m}^e = 0$, intuitively if $p(y_i^{t,d} = m)$ decreases, $p(y_{i+m}^e = 0)$ should increase. To model these *tag-emotion* causal relations in MRG, we define a set of relations represented by $r_{ij} = I_t(i)I_t(j) : rdis(ij)$, and some instances are shown in Fig. 2. For example, $r_{ij} = et : -2$ denotes the relation from node i (an *emotion* node) to node j (a *tag* node) and the relative distance between node i and node j .

In MRG, each inter-task relation corresponds to a fine-grained relative distance, consistent with the definition of the tagging scheme. Therefore, the inter-task graph transformations along these relations can achieve more sufficient and explicit multi-task reasoning.

3.2 CGR-Net

The overall architecture of our Co-evolving Graph Reasoning Network (CGR-Net) is shown in Fig. 3 (1). It consists of three components: Hierarchical Encoding, Initial Estimation, and Co-evolving Reasoning. Next, we depict the procedures of these three components.

Hierarchical Encoding.

Word-level Clause Encoding. The objective of clause encoding is to generate a representation containing the word-level dependencies for each clause. Following previous works, each clause is fed into BERT [8] encoder, then the last hidden state of [CLS] token is taken as the clause representation. Now we obtain the sequence of clause representation for \mathcal{D} : $H = (h_0, \dots, h_n)$.

Multi-task Clause-level Document Encoding. In this paper, we utilize BiLSTM [17] to generate the context-sensitive clause hidden states via modeling the inter-clause dependencies. To obtain task-specific clause hidden states for the three tasks, we separately apply three BiLSTMs over H to obtain the initial clause hidden states for *cause*, *tag* and *emotion*, respectively: $H_c^0 = (h_{c,1}^0, \dots, h_{c,n}^0)$, $H_t^0 = (h_{t,1}^0, \dots, h_{t,n}^0)$ and $H_e^0 = (h_{e,1}^0, \dots, h_{e,n}^0)$.

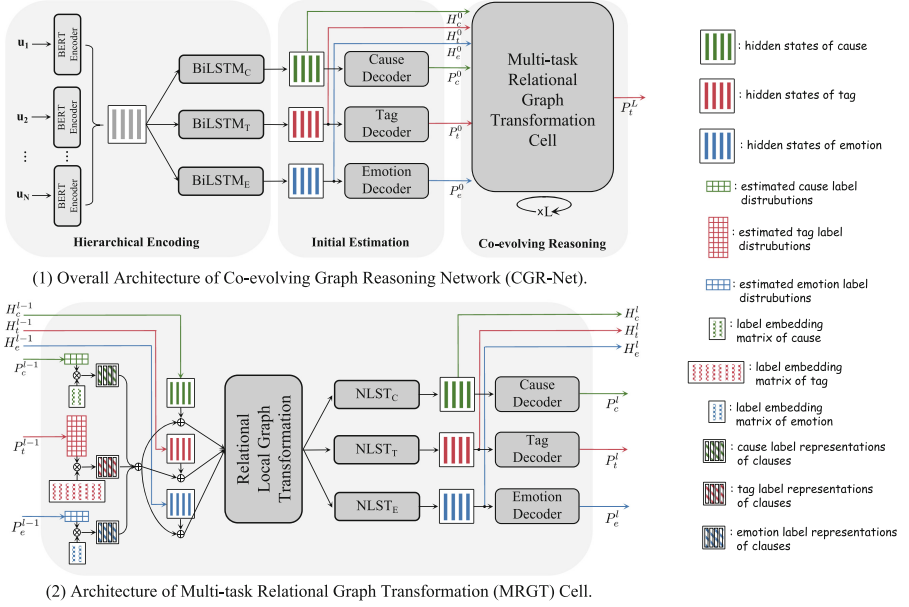


Fig. 3. The architectures of CGR-Net and MRGT cell. NLST denotes Non-Local Self-Transformation.

Initial Estimation. Since MRGT cell takes the three tasks' label distributions predicted in previous step as input, H_c^0 , H_t^0 and H_e^0 are separately fed into Cause Decoder, Tag Decoder and Emotion Decoder to produce the initial estimated label distributions:

$$\begin{aligned}
 P_c^0 &= \{P_{c,i}^0\}_{i=1}^n, P_t^0 = \{P_{t,i}^0\}_{i=1}^n, P_e^0 = \{P_{e,i}^0\}_{i=1}^n \\
 P_{c,i}^0 &= \text{softmax}(\text{MLP}_c(h_{c,i}^0)) = [p_{c,i}^0[1], p_{c,i}^0[2]] \\
 P_{t,i}^0 &= \text{softmax}(\text{MLP}_t(h_{t,i}^0)) = [p_{t,i}^0[1], \dots, p_{c,i}^0[|\mathcal{C}_t|]] \\
 P_{e,i}^0 &= \text{softmax}(\text{MLP}_e(h_{e,i}^0)) = [p_{e,i}^0[1], p_{e,i}^0[2]]
 \end{aligned} \tag{1}$$

Co-evolving Reasoning. Co-evolving Reasoning is achieved by the recurrent MRGT cell, whose details are shown in Fig. 3 (2). At step l , MRGT cell takes two streams of inputs: 1) hidden states of the three tasks: $H_c^{l-1} \in \mathbb{R}^{n \times d}$, $H_t^{l-1} \in \mathbb{R}^{n \times d}$ and $H_e^{l-1} \in \mathbb{R}^{n \times d}$; 2) label distributions of the three tasks: P_c^{l-1} , P_t^{l-1} and P_e^{l-1} . The procedure of an MRGT cell consists of three steps (1) projecting the input label distributions into vectors; (2) Relational Local Graph Transformation on MRG; (3) Non-local Self-Transformation.

Projection of Label Distribution. To achieve the prediction-level interactions, the input label distributions should be projected into vector form, and thus they can participate in representation learning. Accordingly, we use P_c^{l-1} , P_t^{l-1} and

P_e^{l-1} to respectively multiply the corresponding task-specific label embedding matrices $M_c^e \in \mathbb{R}^{|\mathcal{C}_c| \times d}$, $M_t^e \in \mathbb{R}^{|\mathcal{C}_t| \times d}$ and $M_e^e \in \mathbb{R}^{|\mathcal{C}_e| \times d}$, which are trained with the whole model. Specifically, x_i 's label representations for the three tasks are obtained as:

$$e_{t,i}^l = \sum_{k=1}^{|\mathcal{C}_t|} p_{t,i}^{l-1}[k] \cdot v_t^k; \quad e_{c,i}^l = \sum_{k'=1}^{|\mathcal{C}_c|} p_{c,i}^{l-1}[k'] \cdot v_c^{k'}; \quad e_{e,i}^l = \sum_{k''=1}^{|\mathcal{C}_e|} p_{e,i}^{l-1}[k''] \cdot v_e^{k''} \quad (2)$$

where v_t^k , $v_c^{k'}$ and $v_e^{k''}$ denotes the label embeddings of *tag*, *cau* and *emo*, respectively.

Relational Local Graph Transformation. Since MRG is based on local connections, we conduct relational local graph transformation inspired from [22, 28–31, 33–35] for multi-task reasoning. To achieve the self- and mutual-interactions between the semantics and predictions of the three tasks, for each node in MRG, we superimpose its corresponding clause's label representations of the three tasks on its hidden state:

$$\begin{aligned} e_i^l &= e_{c,i}^l + e_{t,i}^l + e_{e,i}^l, \\ \hat{h}_{e,i}^l &= h_{c,i}^{l-1} + e_i^l; \quad \hat{h}_{t,i}^l = h_{t,i}^{l-1} + e_i^l; \quad \hat{h}_{e,i}^l = h_{e,i}^{l-1} + e_i^l \end{aligned} \quad (3)$$

Thus each node representation contains the task-specific semantic features as well as the explicit correlative information conveyed by label representations, which are then integrated together into the relational local graph transformation to achieve semantics-level and prediction-level interactions.

Specifically, the relational local graph transformation updates the nodes on MRG as follows:

$$\bar{h}_i^l = W_1 \hat{h}_i^l + \sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} \frac{1}{|\mathcal{N}_i^r|} W_2^{rr} \hat{h}_j^l \quad (4)$$

where W_1 is the self-message matrix and W_2^r is the relation-specific matrix. \mathcal{N}_i^r denotes the neighbors set of node i along corresponding to the relation r . Now we obtain the updated hidden states: \bar{H}_c^l , \bar{H}_t^l and \bar{H}_e^l .

Non-Local Self-Transformation Despite the advantages of the relational local graph transformation, it has two potential issues: (1) due to the local self-transformation, some beneficial contextual dependencies between a node and its distant same-task nodes may be lost; (2) the information fusion weaken the task-specificity of the nodes to some extent, which is against predictions. To this end, inspired by [27, 32], we conduct non-local self-transformation (NLST) over the sequence of nodes of each task, and this is implemented by a task-specific BiLSTM which can capture long-range dependencies. The final hidden states of the three tasks at step l are obtained by:

$$H_c^l = \text{NLST}_C(\bar{H}_c^l); \quad H_t^l = \text{NLST}_T(\bar{H}_t^l); \quad H_e^l = \text{NLST}_E(\bar{H}_e^l) \quad (5)$$

Then H_c^l , H_t^l and H_e^l are fed to respective decoders to produce P_c^l , P_t^l and P_e^l .

Optimization with Logical Constraints. In CGR-Net, there are two vital logic rules. First, the label distributions estimated in the previous step should be relatively good to provide effective label representations for the current step. Otherwise, much incorrect and misleading explicit correlations would be introduced, harming multi-task reasoning. Second, ECPE and CE/EE are supposed to gradually promote each other via capturing more and more beneficial mutual knowledge and correlations in Co-evolving Reasoning. In other words, the estimated label distributions should be gradually improved along the steps. To satisfy these two rules, we propose a harness loss \mathcal{L}_{harn} that includes two terms: estimate loss \mathcal{L}_{est} and margin loss \mathcal{L}_{marg} , corresponding to the two rules, respectively.

Estimate Loss. Formally, \mathcal{L}_{est} is the cross-entropy loss. For ECPE task, $\mathcal{L}_{est}^{tag,l}$ is defined as:

$$\mathcal{L}_{est}^{tag,l} = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^{|\mathcal{C}_t|} y_{t,i}^k \log(p_{t,i}^l[k]), \quad (6)$$

Margin Loss. \mathcal{L}_{marg} works on the label distributions output in two adjacent steps, forcing CGR-Net to produce better predictions at step l than step $l-1$. For ECPE task, $\mathcal{L}_{marg}^{tag,l}$ is defined as:

$$\mathcal{L}_{marg}^{tag,(l,l-1)} = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^{|\mathcal{C}_t|} y_{t,i}^k \max(0, p_{t,i}^{l-1}[k] - p_{s,i}^l[k]) \quad (7)$$

Harness loss \mathcal{L}_{harn} is the weighted sum of \mathcal{L}_{est} and \mathcal{L}_{marg} . For ECPE task, \mathcal{L}_{harn}^{tag} is defined as:

$$\mathcal{L}_{harn}^{tag} = \sum_{l=0}^{L-1} \mathcal{L}_{est}^{tag,l} + \beta * \sum_{l=1}^L \mathcal{L}_{marg}^{tag,(l,l-1)} \quad (8)$$

where β is a hyper-parameter balancing the impact of the two kinds of punishments.

Final Training Objective The total loss for ECPE task (\mathcal{L}_{tag}) is the sum of \mathcal{L}_{harn}^{tag} and \mathcal{L}_{pred}^{tag} :

$$\mathcal{L}_{tag} = \mathcal{L}_{pred}^{tag} + \mathcal{L}_{harn}^{tag} \quad (9)$$

where \mathcal{L}_{pred}^{tag} is the cross-entropy loss of the produced tag label distributions at the final step L :

$$\mathcal{L}_{pred}^{tag} = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^{|\mathcal{C}_t|} y_{t,i}^k \log(p_{t,i}^L[k]) \quad (10)$$

The total losses of CE (\mathcal{L}_{cau}) and EE (\mathcal{L}_{emo}) can be derivated like Eqs. (6) to (10).

The final training objective of CGR-Net is the weighted sum of the total losses of the three tasks:

$$\mathcal{L} = \alpha * \mathcal{L}_{tag} + \frac{1 - \alpha}{2} \mathcal{L}_{cau} + \frac{1 - \alpha}{2} \mathcal{L}_{emo} \quad (11)$$

where α is a hyperparameter balancing the three tasks and it is intuitively set as 0.5 in this work.

4 Experiments

4.1 Datasets and Evaluation Metrics

The only benchmark dataset for ECPE task is released by [26] who construct it on an emotion-cause extraction corpus [16]. The dataset totally consists of 1,945 documents, among which 1,746 ones have one emotion-cause pair, 177 ones have two emotion-cause pairs, and 22 ones have more than two emotion-cause pairs. The average number of clauses per document is 14.77, and the max number is 73.

Following previous works, we adopt the 10-fold cross-validation for evaluations. And the averages of precision (P), recall (R), and F1-score over ten runs are adopted as metrics. Besides ECPE, we also report the results of EE and CE, which are evaluated based on the emotion clauses and cause clauses in the extracted emotion-cause pairs.

4.2 Implement Details

We adopt the BERT_{Chinese} implemented in PyTorch [24] as the clause encoder. And the three decoders are implemented as three 2-layer MLPs whose hidden size is set as 256. The AdamW optimizer [19] is used for model training, and the learning rate is $1e^{-5}$ for BERT and $1e^{-4}$ for other modules. The dimension d is 512, the max span γ is 3 and the margin loss coefficient β is $1e^{-3}$. The step number of Co-evolving Reasoning is 3. The dropout rate is 0.1, and the batch size is 4. The epoch number is 10, and the early stopping strategy is adopted. All experiments are conducted on a DGX A100 server.

4.3 Compared Baselines

We compare our CGR-Net with the following two groups of baselines.

Group 1: M₁: ECPE-2D (BERT) [10]; M₂: Hier-BiLSTM-BERT; M₃: PairGCN-BERT [3]; M₄: TransECPE [12]; M₅: RankCP+BERT [23]; M₆: UTOS+BERT [5]; M₇: ECPE-MLL(BERT) [11]; M₈: MGSAG(BERT) [1]; M₉: MaCa(BERT) [37];

Group 2: M₈: SLNT + BERT [38]; M₉: MTST + Refinement [13].

Our CGR-Net uses the same ECPE tagging scheme with the second group of baselines.

Table 2. Results comparison on ECPE task and the two subtasks. All scores are averages over 10 runs. All models adopt BERT for clause encoding. ^h denotes the results are reproduced by us. [†] and [‡] denote the results are retrieved from [13] and [5], respectively. * denotes our CGR-Net significantly overpasses M₁₀ and M₁₁ with $p < 0.05$ under t-test.

Models	Emotion-Cause Pair Ext.			Emotion Ext.			Cause Ext.		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
M ₁ : ECPE-2D(BERT)	72.92	65.44	68.89	86.27	92.21	89.10	73.36	69.34	71.23
M ₂ : Hier-BiLSTM-BERT	75.37	64.34	69.26	88.80	74.70	81.00	78.03	65.35	70.96
M ₃ : PairGCN-BERT	76.92	67.91	72.02	88.57	79.58	83.75	79.07	69.28	73.75
M ₄ : TransECPE [†]	77.08	65.32	70.72	88.79	83.15	85.88	78.74	66.89	72.33
M ₅ : RankCP+BERT [‡]	68.21	74.83	71.21	86.79	89.26	87.97	72.62	76.46	74.37
M ₆ : UTOS+BERT	73.89	70.62	72.03	88.15	83.21	85.56	76.71	73.20	74.71
M ₇ : ECPE-MLL(BERT)	77.00	72.35	74.52	86.08	91.91	88.86	73.82	79.12	76.30
M ₈ : MGSAG(BERT)	77.43	73.21	75.21	92.08	92.11	87.17	79.79	74.68	77.12
M ₉ : MaCa(BERT)	80.47	72.15	73.87	88.19	89.55	87.04	78.41	72.60	74.35
M ₁₀ : SLNT+BERT ^h	73.56	68.57	70.85	84.77	80.61	82.51	75.94	70.99	73.25
M ₁₁ : MTST+Refinement ^h	77.14	67.81	72.11	88.25	79.01	83.31	79.18	69.79	74.12
CGR-Net (ours)	77.62	75.49*	76.48*	89.65	86.23*	87.75*	79.68	77.84*	78.75*

4.4 Main Results

The overall results on ECPE and the two subtasks are shown in Table 2. We can observe that CE is much harder than EE, and CE determines the result of ECPE to a large extent. The reason is that CE plays a key role in identifying the causalities, which is difficult for machines. Our CGR-Net significantly outperforms the previous best-performing model M8 by 1.7%, 0.7%, and 2.1% in terms of F1 on ECPE, EE and CE, respectively. Using the same tagging scheme, our CGR-Net overpasses M₁₀ and M₁₁ by 6.1%, 5.3%, and 6.2% in F1 on the three tasks. In particular, we can find that the superior F1 of CGR-Net comes from the high recall and competitive precision. In contrast, previous models generally obtain low recalls because their single-round multi-task reasoning process and the one-way message only from CE/EE to ECPE are not competent enough to discover the emotion cause pairs sufficiently. Moreover, their implicit semantics interactions cannot effectively exploit the causal relations. We can find that our CGR-Net overpasses M₁₀ and M₁₁ by 10.1%, 7.0%, and 9.6% in terms of recall on the three tasks. This demonstrates that through Co-evolving Reasoning, CGR-Net can discover much more ground-truth emotion-cause pairs than baselines, while there are not many wrong-extracted pairs at the same time. CGR-Net’s satisfying results come from the advantages of MRG, the well-designed supervision signals, and the advanced architecture of MRGT.

4.5 Variants of MRG Structure

In this section, we investigate how the structure of MRG would affect our CGR-Net’s performance by applying different structures to MRG. Except for the orig-

Table 3. Results on different variants of MRG.

Variants	ECPE	EE	CE
	F1 (%)	F1 (%)	F1 (%)
MRG($\gamma = 3$)	76.48	87.75	78.75
MRG($\gamma = 1$)	75.15 (\downarrow 1.33)	86.31 (\downarrow 1.44)	77.32 (\downarrow 1.43)
MRG($\gamma = 2$)	75.49 (\downarrow 0.99)	87.55 (\downarrow 0.20)	77.79 (\downarrow 0.96)
MRG($\gamma = 4$)	74.87 (\downarrow 1.61)	86.78 (\downarrow 0.97)	77.21 (\downarrow 1.54)
OWM	74.71 (\downarrow 1.77)	86.30 (\downarrow 1.45)	77.22(\downarrow 1.53))
NoRel	74.63 (\downarrow 1.85)	85.33 (\downarrow 2.42)	77.10 (\downarrow 1.65)
FCG	73.65 (\downarrow 2.83)	84.89 (\downarrow 2.86)	76.56 (\downarrow 2.19)

inal MRG($\gamma = 3$), we design four variants: (1) MRG($\gamma = 1$), MRG($\gamma = 2$) and MRG($\gamma = 4$), which have different span limitation; (2) OWM (one-way message), in which the edges from *tag* nodes to *emotion* nodes and *cause* nodes are deleted, thus there is only one-way message from EE/CE to ECPE, like previous works; (3) NoRel, in which there is no relation on edges; (4) FCG (fully-connected graph), in which all *tag* nodes are fully connected with all *emotion* nodes and all *cause* nodes, while there is no relation. The results over these variants are listed in Table 3.

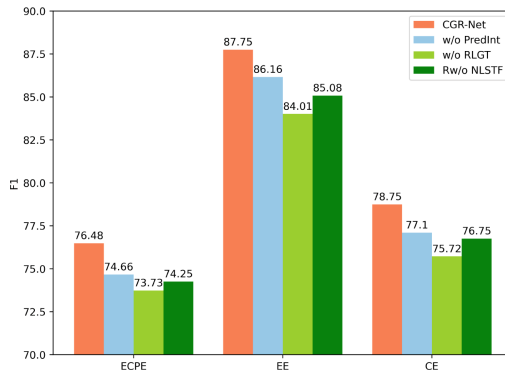
Several instructive observations can be made from the results. **Firstly**, with γ varying from 1 to 4, the results first increase and then decrease. The reason is that too small γ cannot capture all emotion-pairs, while too large γ makes it much harder to predict correctly because $|\mathcal{C}_t|$ is directly proportional to γ . More than 95% emotion-cause pairs’ spans in the dataset do not exceed 3, so intuitively $\gamma = 3$ performs best, and the results prove this. And this is consistent with the report of [13]. **Secondly**, compared with the original MRG, OWM’s performances on all tasks drop significantly. This proves that the reverse feedbacks from ECPE to EE/CE is crucial, while previous works ignore them. In this paper, we propose a new MTL framework based on the novel Co-evolving Reasoning mechanism to solve this issue. **Thirdly**, without relations, NoRel performs much worse than the original MRG. The distinct decrease of results proves that our designed relations are indispensable for capturing the casualties between ECPE and EE/CE and significantly improve the performances. **Finally**, FCG performs even worse than NoRel. This is because causalities often exist between close clauses. In FCG, useless information from distant nodes is integrated into the current node, making the crucial information diluted and discarded, resulting in poor results. And this proves the validity of the local connection rule in MRG.

4.6 Investigation of Supervision Signals

To investigate the necessities of different supervision signals, we remove different loss terms, and the results are listed in Table 4. Firstly, we can observe that removing \mathcal{L}_{emo} or \mathcal{L}_{cau} both lead to obvious result decreases on all tasks. This

Table 4. Results of removing different loss terms.

Variants	EE	CE	ECPE
	F1 (%)	F1 (%)	F1 (%)
CGR-Net	76.48	87.75	78.75
$-\mathcal{L}_{emo}$	74.51 (\downarrow 1.97)	86.21 (\downarrow 1.54)	77.03 (\downarrow 1.72)
$-\mathcal{L}_{cau}$	74.28 (\downarrow 2.20)	86.25 (\downarrow 1.50)	76.94 (\downarrow 1.81)
$-\mathcal{L}_{emo} - \mathcal{L}_{cau}$	73.76 (\downarrow 2.72)	85.05 (\downarrow 2.70)	76.13 (\downarrow 2.62)
$-\mathcal{L}_{est}$	75.43 (\downarrow 1.05)	87.54 (\downarrow 0.21)	77.97 (\downarrow 0.78)
$-\mathcal{L}_{marg}$	74.22 (\downarrow 2.26)	86.51 (\downarrow 1.24)	76.64 (\downarrow 2.11)
$-\mathcal{L}_{harn}$	74.89 (\downarrow 1.59)	85.61 (\downarrow 2.14)	77.60 (\downarrow 1.15)

**Fig. 4.** Ablation results of MRGT cell.

is because our model exploits the beneficial mutual correlations between ECPE and the two subtasks. So removing the supervision signal of a subtask harms not only the performance of itself but also the performances of ECPE and another subtask. If both subtasks do not have supervision signals, the performances of all tasks drop dramatically. Then we can find that the performances decrease remarkably if \mathcal{L}_{harn} or any of its two terms is removed. This is because without \mathcal{L}_{harn} CGR-Net is hard to achieve the virtuous cycle of Co-evolving Reasoning.

4.7 Ablation Study of MRGT Cell

We conduct ablation experiments to study the efficacies of the components in MRGT cell, and the results are listed in Fig. 4. When removing prediction-level interactions (PredInt), the performances drop significantly. This proves that only relying on semantics-level interactions is insufficient for multi-task reasoning. An essential advantage of our model is achieving prediction-level interactions that convey explicit correlations and provide feedback for semantics that can then rethink to improve. Without RLGT, Co-evolving Reasoning cannot be achieved,

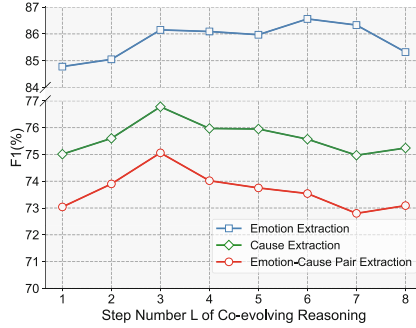


Fig. 5. CGR-Net’s performances on different L .

causing the worst results. Removing NLST also leads to sharp decreases in results. This is because without NLST the hidden states cannot obtain long-range crucial contextual information, and the three streams of hidden states output at each step are not task-specific enough, which both harm the predictions.

4.8 Step Number of Co-evolving Reasoning

We plot the performance (F1) trends of CGR-Net on the three tasks over different Co-evolving Reasoning step numbers, as presented in Fig. 5. The best overall performances are achieved when $L = 3$, which justifies the step number setting in Sec. 4.2. Generally, the performances of ECPE and the two subtasks steadily increase until $L = 3$, while then having dropping trends or fluctuate in a relatively narrow range when L continues increasing. This indicates that ECPE and CE/EE can gradually promote each other in the process of Co-evolving Reasoning, whose advantage is validated. However, after the performance reaches its peak, more steps lead to decreasing. We speculate the possible reason is that too many Co-evolving Reasoning steps may cause redundant information and over-fitting.

5 Conclusion and Prospect

In this paper, we improve ECPE on three aspects. First, we propose a new MTL based on Co-evolving Reasoning, allowing ECPE and its two subtasks to promote each other gradually. Besides, prediction-level interactions are integrated to model the explicit correlations. Second, we design a novel multi-task relational graph (MRG) to sufficiently exploit the causal relations. Finally, we propose a Co-evolving Graph Reasoning Network (CGR-Net) to implement our framework and conduct Co-evolving Reasoning on MRG. Experiment results demonstrate the superiority of our method, and detailed analyses further validate the advantages.

This work contributes a new paradigm not only for ECPE but also for a group of scenarios in which different tasks share the same input sequence. Future works include improving our method on ECPE and applying our paradigm to other MTL tasks.

Acknowledgments. This work was supported by Australian Research Council Grant DP200101328. Authors were also supported by A*STAR Centre for Frontier AI Research.

References

1. Bao, Y., Ma, Q., Wei, L., Zhou, W., Hu, S.: Multi-granularity semantic aware graph model for reducing position bias in emotion cause pair extraction. In: Findings of the Association for Computational Linguistics: ACL 2022, pp. 1203–1213. Association for Computational Linguistics, Dublin, Ireland (May 2022)
2. Chen, X., Li, Q., Wang, J.: A unified sequence labeling model for emotion cause pair extraction. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 208–218. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020)
3. Chen, Y., Hou, W., Li, S., Wu, C., Zhang, X.: End-to-end emotion-cause pair extraction with graph convolutional network. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 198–207. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020)
4. Chen, Y., Lee, S.Y.M., Li, S., Huang, C.R.: Emotion cause detection with linguistic constructions. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 179–187. Coling 2010 Organizing Committee, Beijing, China (Aug 2010)
5. Cheng, Z., Jiang, Z., Yin, Y., Li, N., Gu, Q.: A unified target-oriented sequence-to-sequence model for emotion-cause pair extraction. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 2779–2791 (2021)
6. Cheng, Z., Jiang, Z., Yin, Y., Yu, H., Gu, Q.: A symmetric local search network for emotion-cause pair extraction. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 139–149. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020)
7. De Beaugrande, R., Dressler, W.U.: *Introduction to Text Linguistics*. Longmans, London (1981)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019)
9. Ding, Z., He, H., Zhang, M., Xia, R.: From independent prediction to reordered prediction: integrating relative position and global label information to emotion cause identification. In: *AAAI*, pp. 6343–6350. AAAI Press (2019)
10. Ding, Z., Xia, R., Yu, J.: ECPE-2D: emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3161–3170. Association for Computational Linguistics (Online) (Jul 2020)

11. Ding, Z., Xia, R., Yu, J.: End-to-end emotion-cause pair extraction based on sliding window multi-label learning. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3574–3583. Association for Computational Linguistics (Online) (Nov 2020)
12. Fan, C., Yuan, C., Du, J., Gui, L., Yang, M., Xu, R.: Transition-based directed graph construction for emotion-cause pair extraction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3707–3717. Association for Computational Linguistics (Online) (Jul 2020)
13. Fan, C., Yuan, C., Gui, L., Zhang, Y., Xu, R.: Multi-task sequence tagging for emotion-cause pair extraction via tag distribution refinement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 2339–2350 (2021)
14. Ghazi, D., Inkpen, D., Szpakowicz, S.: Detecting emotion stimuli in emotion-bearing sentences. In: Gelbukh, A. (ed.) *CICLing 2015. LNCS*, vol. 9042, pp. 152–165. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18117-2_12
15. Gui, L., Hu, J., He, Y., Xu, R., Lu, Q., Du, J.: A question answering approach for emotion cause extraction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1593–1602. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017)
16. Gui, L., Wu, D., Xu, R., Lu, Q., Zhou, Y.: Event-driven emotion cause extraction with corpus construction. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1639–1649. Association for Computational Linguistics, Austin, Texas (Nov 2016)
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
18. Lee, S.Y.M., Chen, Y., Huang, C.R.: A text-driven rule-based system for emotion cause detection. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 45–53. Association for Computational Linguistics, Los Angeles, CA (Jun 2010)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019. OpenReview.net (2019)
20. Neviarouskaya, A., Aono, M.: Extracting causes of emotions from text. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 932–936. Asian Federation of Natural Language Processing, Nagoya, Japan (Oct 2013)
21. Russo, I., Caselli, T., Rubino, F., Boldrini, E., Martínez-Barco, P.: EMOCause: an easy-adaptable approach to extract emotion cause contexts. In: Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2011), pp. 153–160. Association for Computational Linguistics, Portland, Oregon (Jun 2011)
22. Schlichtkrull, M.S., Kipf, T.N., Bloem, P., van den Berg, R., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. In: *ESWC*, pp. 593–607 (2018)
23. Wei, P., Zhao, J., Mao, W.: Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3171–3181. Association for Computational Linguistics (Online) (Jul 2020)
24. Wolf, T., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics (Online) (2020)

25. Wu, S., Chen, F., Wu, F., Huang, Y., Li, X.: A multi-task learning neural network for emotion-cause pair extraction. In: ECAI. *Frontiers in Artificial Intelligence and Applications*, vol. 325, pp. 2212–2219. IOS Press (2020)
26. Xia, R., Ding, Z.: Emotion-cause pair extraction: A new task to emotion analysis in texts. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1003–1012. Association for Computational Linguistics, Florence, Italy (Jul 2019)
27. Xing, B., et al.: Earlier attention? aspect-aware lstm for aspect-based sentiment analysis. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, pp. 5313–5319. International Joint Conferences on Artificial Intelligence Organization (July 2019)
28. Xing, B., Tsang, I.: Co-guiding net: achieving mutual guidances between multiple intent detection and slot filling via heterogeneous semantics-label graphs. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 159–169. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022)
29. Xing, B., Tsang, I.: DARER: dual-task temporal relational recurrent reasoning network for joint dialog sentiment classification and act recognition. In: *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3611–3621. Association for Computational Linguistics, Dublin, Ireland (May 2022)
30. Xing, B., Tsang, I.: Dignet: digging clues from local-global interactive graph for aspect-level sentiment classification. *arXiv preprint [arXiv:2201.00989](https://arxiv.org/abs/2201.00989)* (2022)
31. Xing, B., Tsang, I.: Group is better than individual: exploiting label topologies and label relations for joint multiple intent detection and slot filling. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3964–3975. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022)
32. Xing, B., Tsang, I.: Neural subgraph explorer: reducing noisy information via target-oriented syntax graph pruning. In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022*, pp. 4425–4431. International Joint Conferences on Artificial Intelligence Organization (July 2022), main Track
33. Xing, B., Tsang, I.: Relational temporal graph reasoning for dual-task dialogue language understanding. *arXiv preprint [arXiv:2306.09114](https://arxiv.org/abs/2306.09114)* (2023)
34. Xing, B., Tsang, I.W.: Out of context: a new clue for context modeling of aspect-based sentiment analysis. *J. Artifi. Intell. Res.* **74**, 627–659 (2022)
35. Xing, B., Tsang, I.W.: Understand me, if you refer to aspect knowledge: Knowledge-aware gated recurrent memory network. *IEEE Trans. Emerging Topics Comput. Intell.* **6**(5), 1092–1102 (2022)
36. Yan, H., Gui, L., Pergola, G., He, Y.: Position bias mitigation: a knowledge-aware graph model for emotion cause extraction. 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)*, pp. 3364–3375. Association for Computational Linguistics (Online) (Aug 2021)

37. Yang, C., Zhang, Z., Ding, J., Zheng, W., Jing, Z., Li, Y.: A multi-granularity network for emotion-cause pair extraction via matrix capsule. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 4625–4629 (2022)
38. Yuan, C., Fan, C., Bao, J., Xu, R.: Emotion-cause pair extraction as sequence labeling based on a novel tagging scheme. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3568–3573. Association for Computational Linguistics, Online (Nov 2020)