Multilevel Position-Aware and Knowledge Enhancement for Causal Emotion Entailment

1023040823-herunjie

Nanjing University of Posts and Telecommunications

Abstract. The Dialogue Causal Emotion Entailment task aims to identify the utterances that elicit certain emotions in a dialogue. Positional information and implicit commonsense knowledge are critical for comprehensively interpreting the dialogue's meaning and perceiving its emotional content, thus identifying the cause utterances. Previous works only incorporated absolute positional information during encoding and simply added commonsense knowledge without screening out useful information. We provide position-aware information at multiple levels and propose a Multilevel Position-Aware and Commonsense Knowledge Enhancement (MPACKE) to address the shortcomings in the previous works. Specifically, we design a position-aware Commonsense Knowledge Selector (CKS) network and a Relative Position-Aware Graph (RPAG) network. We add Intermedia Information Compensate (IIC) between the candidate causal utterances and the target emotion utterance. Experimental results on public datasets show that our MPACKE model outperforms most existing emotion-cause extraction models.

Keywords: Emotion Cause Analysis · Graph Neural Network · Positional Embedding.

1 Introduction

Dialogue emotion analysis is a key part of perceptual artificial intelligence. Traditional emotion analysis research focuses on identifying emotion categories. The research on emotion generation mechanisms is insufficient. Accurate identification of dialogue emotional flows and the cause of emotions can effectively enhance the intelligence of dialogue agents and improve the human experience in using the agents[10]. Poria[13] proposed new tasks (RECCON- DD, RECCON- IE) to explore the emotional causes of dialogues, and defined the data set (RECCON) and subtasks: Causal Segment Extraction (CSE) and Causal Emotion Entailment (CEE). This paper focuses on CEE, which aims to identify cause utterances that trigger certain emotions from dialogue history.

Poria[13] treat CEE as a classification task of utterance pairs, which pairs target emotiona utterance with candidate cause utterances from the dialogue history, and then give the classification results. However, effective reasoning of emotion and a thorough comprehension of dialogue context are not fully taken into account. Challenges faced by CEE include:

(1) Timing of dialogue. Dialogue involves multiturn exchanges between two parties, leading to topic shifts and emotional changes. So the distance between cause utterances and emotional utterances in a dialogue is often close. Current methods neglect the effectiveness of sufficient location information encoding[3][19].

(2) Implicit information in dialogue. Understanding a dialogue heavily depends on context. The context contains a large amount of implicit information that is difficult for models to capture, which poses a challenge for CEE models to accurately infer the association between cause utterances and emotional utterances.

We present commonsense knowledge and propose a Multilevel Position-Aware and Knowledge Enhancement architecture in conjunction with rotated position encoding (ROPE)[16]. The architecture can determine the relative position between the candidate cause utterances and the target emotion utterance, as well as filter out task-relevant knowledge. Furthermore, we deploy a relative position-aware graph neural network (RPAG) to encode the turn structure of dialogue. We implemented intermediate information compensation (IIC), encoding the dialogue between cause utterances and emotion utterance into triples, to improve the detection of position distance and percept of the intermediate information.

Our contributions are summarized below:

1. Propose a Position-Aware Commonsense Knowledge Selector network (PACKS) to filter out knowledge clues useful for triggering emotions through location awareness.

2. We focus on the turn structure of the dialogue, proposing turn position-aware and Intermediate Information Compensation to perceive multiple levels of positional information.

3. Our model can outperform most of the baseline models on available data sets.

2 Related Works

2.1 Emotion Cause Extraction

Emotion cause extraction aims to identify the reasons that trigger a given emotion from a text. The task was first proposed by Lee, Chen, and Huang[9], who constructed a Chinese emotion cause annotation dataset and summarized two linguistic clues for detecting causes. Subsequent works employed rule-based methods for emotion cause extraction[2][6][7][12][14] to enhance the effectiveness. With the development of deep learning, Xia, Zhang, and Ding[20] used Transformer and integrated relative position information and global feature information into the encoder, achieving significant improvement in emotion-cause extraction.

2.2 Emotion Cause Pair Extraction

Emotion cause extraction task is limited by its dependence on manual emotion annotations, which does not align with the practical needs of large-scale text

processing. Xia and Ding[19] first proposed the Emotion Cause Pair Extraction task (ECPE), requiring models to simultaneously extract utterances with explicit emotion and the cause utterances that trigger that emotion, and defined a two stage framework. Wei, Zhao, and Mao[18] believed that the two stage pipeline method had issues with error propagation. Subsequently, Chen[1] proposed an end-to-end framework, constructing a PairGCN model to model the dependency relationships between candidate emotion cause pairs. Ding, Xia, and Yu[4] adopted a 2D matrix representation of all possible emotion cause pairs to model the interaction between different emotion cause pairs for end-to-end optimization. Ding, Xia, and Yu[5] extracted utterances and cause utterances separately through emotion-oriented sliding windows and cause-oriented sliding windows, then paired and filtered them. Zheng[24] proposed a Prompt-based method, adding a directional constraint mechanism and sequential learning mechanism to integrate emotion extraction, cause extraction, and emotion-cause pair extraction into a unified approach.

2.3 Dialogue Emotion Cause Analysis

Poria[13] first proposed the task of dialogue emotion cause recognition and defined Causal Fragment Extraction (CSE) and Causal Emotion Entailment (CEE), and annotated a dataset (RECCON) for emotion cause analysis. Zhang[22] proposed a dual stream model, encoding the emotional flow and speaker role flow separately, then fusing the information through an interaction module, achieving State Of The Art (SOTA) performance at the time. Li[10] expanded the directed acyclic graph with commonsense knowledge and proposed a knowledge-enhanced dialogue graph to assist in discovering causal clues. Gu[8] considered the positional relationships between utterances and designed a position-aware graph to model the relationship between emotions and causes in dialogues. Chi[21] proposed a token classification based BIO tagging model, achieving the current SOTA performance. Wang, Yu, and Xia[17] used a generative approach that pairs emotion sequence indices and causal sequence indices for emotion cause pairing.

The above methods did not consider the impact of positional information on encoding effects during knowledge selection or filtering, nor did they consider the turn structure of dialogues.

3 Methodology

3.1 Task Definition

We first define the task of dialogue causal emotion entailment(C2E2). Given a dialogue $D = \{u_1, u_2, \cdots, u_n\}$, the emotion of each utterance $E = \{e_1, e_2, \cdots, e_n\}$, and the role of each speaker $S = \{s_1, s_2, \cdots, s_n\} \varepsilon \{A, B\}$, the task is to identify the cause utterance u_i in the dialogue history that triggers the non-neutral emotion of the target utterance u_t(i<t).

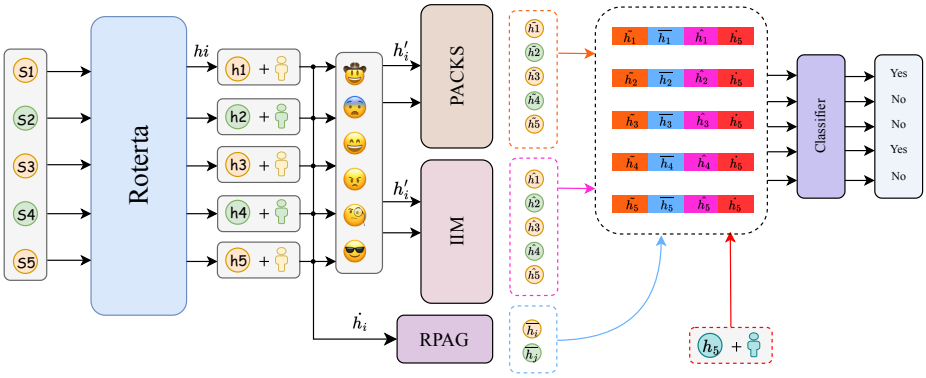

Fig. 1: MPACKE Structure Diagram

3.2 Model Architecture Overview

The model architecture mainly consists of three components: the utterance feature encoding module, the position-aware context understanding module, and the pairing prediction module, as shown in Fig. 1. In the utterance feature encoding module, the utterance is encoded to obtain representations that contain rich semantic information. Position-aware context understanding module focuses on modeling explicit and implicit clues of the dialogue context with position information. Specifically, we introduce a commonsense knowledge base to model the relationships between emotions and causes implied in the context, as well as Rotary Position Embedding (ROPE) to effectively select commonsense knowledge that is useful for the task. Considering the turn structure of the dialogue, we encode data with relational types through graph neural networks. Finally, in the pairing prediction process, we propose an intermedia information compensation mechanism, which encodes the utterances between the candidate cause utterance u_i and the target emotion utterance u_t to obtain compensatory information. The semantic representations obtained from the context understanding module, the compensatory information, and the representation of the target emotion utterance are paired and passed through a classification layer for binary prediction. If u_i contains the cause of the emotion expressed in u_t, then the output is Yes, otherwise it is No. The following sections will describe these three parts in detail.

3.3 Utterance Feature Encoding

We use a pretrained RoBERTa as the encoder. For each utterance in a dialogue, a special token [CLS] is prepended to the beginning of the utterance and [SEP] is appended at the end to indicate the end of the utterance, forming the input sequence $input = [CLS, w_1, w_2, \cdots, w_l, SEP]$. The words in the input are then converted into ID tokens for encoding. We perform MaxPooling on the representation of each word in the last layer to obtain the semantic representation of

each utterance at the sentence level, as shown in Equation 1:

$$h_i = MaxPooling\left(RoBERTa\left([CLS, w_1, w_2, \cdots, w_l, SEP]\right)\right) \tag{1}$$

where $h_i \varepsilon R^d$, d is the hidden dimension.

The utterance representation as speaker role features. Consistent with the speaker role encoding, we randomly initialize seven vectors to represent seven types of emotions, including neutrality. Ultimately, we obtain two basic utterance representations, as shown in Equation 2 and Equation 3:

$$\dot{h}_i = h_i + role_i \tag{2}$$

$$h_i^{'} = h_i + role_i + e_i \tag{3}$$

3.4 Commonsense Knowledge

We use the commonsense knowledge base ATOMIC-2020, which is a richly informative daily life commonsense knowledge graph that includes reasoning knowledge about various aspects of social, physical, and event contexts in life. It is widely used in dialogue emotion recognition and other dialogue scenarios for emotional computing tasks. ATOMIC-2020 contains a large number of triplets of knowledge composed of a head phrase, a relationship type, and a tail phrase, such as: “PersonX affords a car", xReact, and “proud". By providing a head phrase and a relationship type, it infers the tail phrase under the specified relationship based on the commonsense knowledge graph at the time of the current event. To conveniently obtain knowledge, we use a generative model COMET-ATOMIC2020 pre-trained on ATOMIC-2020 to obtain inferred knowledge representations. By concatenating each utterance and the relationship type as input, the model will generate beam size inferred knowledge, i.e., tail phrases. For example, given a dialogue “my car was broken." and the relationship“xReact", by concatenating“my car was broken." and “xReact [GEN]", the model will generate a textual description of the knowledge inferred from "my car was broken." under the relationship “xReact", such as “feel sad” or “feel frustrated".

Thus, we obtain knowledge under seven relationships, represented as K_i^r, $r \in [xWant, xReact, xIntent, xEffect, oWant, oReact, oEffect]$, and the knowledge under the relationship “isBefore" for the target emotion utterance, “x" denotes the impact on the speaker himself, “o" represents the impact on another speaker.

3.5 Position-Aware Commonsense Knowledge Screening (PACKS)

Based on the cause of the emotion in the target emotion utterance, we differentiate between Inter-speaker and Intra-speaker utterances. As shown in Fig. 2, first, the knowledge inferred by COMET is fused with the utterance representation to obtain a new representation containing inferred knowledge, denoted as $h_{ikx}^{'}$ and $h_{iko}^{'}$, k is a shorthand for knowledge. Since not all knowledge has a positive

Table 1: Examples of Different Social Interactive Relationship Types

head phrases	relation type	tail phrases
James votes for Nancy	xIntent	James wants to give support
	oEffect	Nancy receives praise
	oReact	Nancy feels confident
	oWant	Nancy wants to celebrate

effect, we propose a position-aware commonsense knowledge selection method. And we propose a position-aware commonsense knowledge selector to filter the knowledge that has negative effect. Specifically, we add position information to the utterance representations using Rotary Position Embedding to obtain $h_{ikx}^{'R}$ and $h_{iko}^{'R}$, using these two semantic representations as the Key and Value in the attention mechanism. We fuse the target emotion utterance representation with the knowledge ï£¡ï£¡isBeforeï£¡ï£¡ and emotional information as the Query. Additionally, we add the relative position information of the target emotion utterance using ROPE to obtain the final Query representation:

$$h_{ikx}^{'} = W_x \left(h_i^{'} + W_i ntra \left(xeffect; xintent; xreact; xwant \right) \right) \tag{4}$$

$$h_{iko}^{'} = W_o \left(h_i^{'} + W_{inter} \left(oeffect; oreact; owant \right) \right) \tag{5}$$

$$h_{ikx}^{'R} = ROPE \left(h_{ikx}^{'} \right) \quad h_{iko}^{'R} = ROPE \left(h_{iko}^{'} \right) \tag{6}$$

$$U_i = h_{ikx}^{'R} \quad E_i = h_{iko}^{'R} \tag{7}$$

$$U_t = ROPE \left(W_t \left(h_5^{'} + e_5 + isBefore \right) \right) \tag{8}$$

where $W_x \ \varepsilon \ \mathbb{R}^{(d \times d)}$, $W_i ntra \ \varepsilon \ \mathbb{R}^{(4d \times d)}$, $W_i nter \ \varepsilon \ \mathbb{R}^{(3d \times d)}$ and $W_t \ \varepsilon \ \mathbb{R}^{(d \times d)}$ are learnable parameters.

The attention weight representing the potential connection degree between each candidate cause utterance and the target emotion utterance can be calculated as follows:

$$s_{i,t}^{inter} = softmax \left(\frac{\left[E_i \odot U_t \right] mask^{inter}}{\sqrt{d}} \right) \tag{9}$$

$$s_{i,t}^{intra} = softmax \left(\frac{\left[U_i \odot U_t \right] mask^{intra}}{\sqrt{d}} \right) \tag{10}$$

where $\odot$ represents the dot product, $s_{i,t}^{inter}$ and $s_{i,t}^{intra}$ represent the attention weights of inter-interaction and intra-interaction respectively.

Finally, after obtaining the attention weights that represent the potential connection between candidate cause utterance and the target emotion utterance,

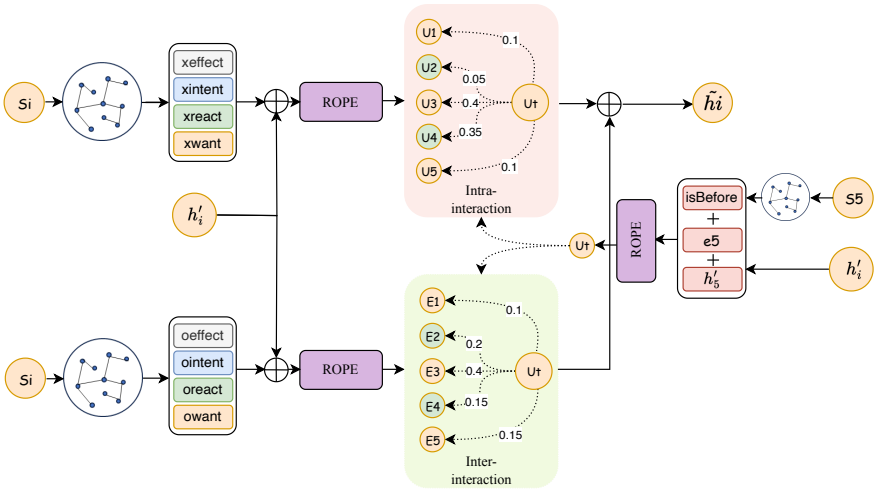

Fig. 2: Position-Aware Commonsense Knowledge Screening network

they are applied to the Values, obtaining the filtered utterance representations containing commonsense knowledge:

$$\tilde{h}_i = \begin{cases} s_{i,j}^{\text{intra}} * (h'_{ikx} + U_t) & \text{role}_{\ i} = \text{role}_{\ t} \\ s_{i,j}^{\text{intra}} * (h'_{iko} + U_t) & \text{role}_{\ i} \neq \text{role}_{\ t} \end{cases} \tag{11}$$

3.6 Relative position-aware graph (RPAG)

Each turn in a dialogue may have a different impact on the emotion of target emotion utterance. As shown in Fig. 3, we construct a directed heterogeneous relational graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to model utterances with different turns. The vertices $v_i \in \mathcal{V}$ in the graph are initialized by the utterance representation $\dot{h}_i$, and for each edge $e_{i,j} \in \mathcal{E}$ from v_{t-1} to v_{t-2} , the relationship is categorized according to the utterance turns. For instance, v_t denotes target emotion utterance, $e_{t,t-1}$ and $e_{t,t-2}$ denotes the first turn of the utterance closest to the target emotion utterance, the edges $e_{t,t-1}$ and $e_{t,t-2}$ are marked as the same type, the edges $e_{t,t-3}$ and $e_{t,t-4}$ are the second turn. The turn type encoding is shown in Equation 12:

$$r_{t,j} = \begin{cases} \frac{j-t-1}{2} - 1 & j < t \\ -\text{ win} & \text{if } \frac{j-t-1}{2} < -\text{ win} \end{cases} \tag{12}$$

For utterances that exceed the distance threshold "win", the edge types are marked as the same.

The self-circulating edge of the target utterance is also marked as the first round. We use RGCN (Schlichtkrull et al., 2018) for graph encoding, which is a

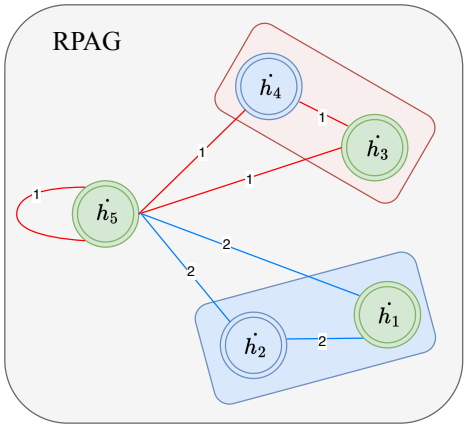

Fig. 3: Relative position-aware graph

relational graph convolutional network that aggregates nodes in the graph with different relationships to update node representations:

$$\overline{h_i} = \sigma\left(\sum_r \sum_{o \in \mathcal{N}_i^r} \frac{1}{c_{i,r}} W_r h_o + W_0 h_i\right) \tag{13}$$

where σ represents the activation function, N_i^r represents neighbor nodes of node v_i under the relationship r. W_r is the learnable parameter for aggregating neighbor nodes, W_0 is the learnable parameter used to retain its own information.

3.7 Intermediate information compensation (IIC)

We propose the intermedia information compensation mechanism to address the issue that pairwise combinations could not perceive the complete flow of dialogue information. We get a representation by encoding the utterances between the candidate cause utterances and the target emotion utterance and add it to the pairing to form a coherent dialogue information flow, which allows the model to better perceive the position of the candidate cause utterances. As shown in Fig. 4, assuming the dialogue $U = (u_1, u_2, u_3, u_4, u_5)$, u_5 is considered as the target emotion utterance and u_1 is the candidate cause utterance, previous works concatenate the semantic representations of u_1 and u_5 to form a pairwise representation $< u_1, u_5 >$. After introducing the IIC mechanism, the dialogue information between these two utterances is encoded as $u_{iic} = f\,(u_2, u_3, u_4)$, the pairing changes from a pairwise to a triplet $< u_1, u_{iic}, u_5 >$:

$$\widehat{h_i} = W(F(LSTM\,(h_i, h_{i+1}, \ldots, h_t))) \tag{14}$$

where $W \in R^{d \times d}$ is the learnable parameter, LSTM is the model for aggregating intermedia information, and F represents the addition function.

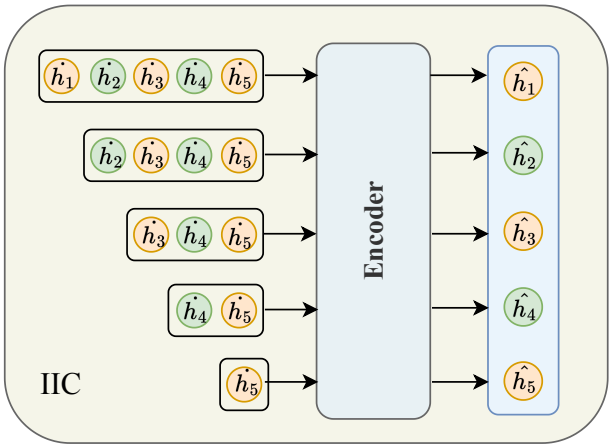

Fig. 4: Intermedia information compensate

3.8 Cause Predictor

After the above encoding, we obtain three different types of utterance representations: $\widehat{h}_i$, $\overline{h}_i$ and $\widetilde{h}_i$. The process of determining the candidate utterance u_i is the cause of the emotion in the target utterance u_t is shown in Equation 15:

$$p_{i,t} = sigmod(MLP1(MLP2(\widetilde{h}_i; \overline{h}_i); \widehat{h}_i; \dot{h}_t)) \tag{15}$$

where $(a; b)$ represents the concatenation of vector a and b. We use cross-entropy loss to calculate the loss between the prediction and the labels:

We use cross-entropy loss to calculate the difference between the predicted values and the true labels. The parameters are updated using the gradient descent method. The specific loss calculation is given by the Equation 16:

$$\mathcal{L}_{all} = \sum_{t \leq N} \sum_{i \leq t} (y_{i,t} \cdot \log p_{i,t} + (1 - y_{i,t}) \cdot (1 - p_{i,t})) \tag{16}$$

where $\mathcal{L}_{all}$ represents the total loss in a batch, N represents the total amount of data in a batch, and $y_{i,t} \in \{0, 1\}$ represents the label that whether u_i is the cause utterance of the emotion in u_t. Subsequent paragraphs, however, are indented.

4 Experimental Setup

4.1 Datasets

We select the RECCON-DD[13] as the benchmark dataset, which is extracted from the manually annotated dataset DailyDialog[11]. We further processed the original data so that when predicting the cause of the emotion utterances, the model can only focus on the dialogue history before the target emotion utterance. Some statistical data are shown in Table 2. Subsequent paragraphs, however, are indented.

Table 2: Statistics of the RECCON-DD dataset.

	Train	Test	Valid
Pos	7027	328	1767
Neg	20646	838	5330

4.2 Implementation Details

In the utterance-level encoding, we use the dimension of RoBERTa to be 768, and the dimensions for utterance encoding, role encoding, and knowledge encoding are set to 300. The model parameters are initialized using the xavier_uniform method. The batch size is 16, the learning rate and L2 regularization are respectively set to 0.00004 and 0.0003, the weight decay is 0.001, and the random seed is 42. When performing turn encoding, we set the context window to 8, and in the Intermedia Information Compensate module, the LSTM dropout is 0.5. We use an NVIDIA GeForce 3090 graphics card for training, save the model parameters that perform best on the validation set, and set the number of training epochs to 10.

4.3 Baselines and Comparison Models

We compare MPACKE with various baseline models on the Causal Emotion Entailment task. The selected baseline models are as follows:

(1) RoBERTa-Base: This is the baseline model introduced along with RECCON, which uses RoBERTa for utterance encoding. It concatenates the cause utterance, emotion utterance, and emotion vector to form the input and performs binary classification prediction.

(2) TSAM[22]: The model consists of an emotion attention network, a speaker attention network, and an interaction network between them.

(3) KEC[10]: The model constructs a directed acyclic graph, focusing particularly on the challenge of detecting causes in neutral utterances, proposing the use of social commonsense knowledge to alleviate the problem of limited clues brought by neutral utterances.

(4) ECPE-2D[5]: An end-to-end model that uses emotion cause interaction representation and a 2D Transformer to model different types of emotion cause pairs.

(5) ECPE-MLL[15]: The model uses a joint multilabel strategy, refining the cause of emotion by extracting causes and specifying emotions to find causes through a joint modeling framework of two frameworks.

(6) RankCP[18]: The model adopts a ranking based method to analyze the causes of emotions, models the correlation between utterances, then ranks them, and filters out the cause utterances based on ranking scores. This is an end-to-end method.

Table 3: Results of the model on RECCON-DD

Model	Pos. F1	Neg. F1	Macro F1
RoBERTa-Base	64.28	88.74	76.51
RoBERTa-Large	66.23	87.89	77.06
TSAM	68.59	89.75	79.17
KEC	66.76	95.74	81.25
ECPE-2D	55.50	94.96	75.23
ECPE-MLL	48.48	94.68	71.59
RankCP	33.00	97.30	65.15
KBCIN	68.59	89.65	79.12
Ours (MPACKE)	68.80	89.07	78.94

(7) KBCIN[23]: The model divides the introduced commonsense knowledge into event-centered semantic bridges and social interaction-centered emotional behavior bridges to construct a causal interaction network, further exploring hidden emotion cause clues in dialogue utterances.

Subsequent paragraphs, however, are indented.

4.4 Results and Analysis

The experimental results are shown in Table 3. Macro F1 represents an indicator considering the distribution of positive and negative examples; Pos. F1 represents an indicator of the correctness of predicting emotional cause utterances; and Neg. F1 represents the correctness of predicting which utterances are not caused utterances. Compared to the baseline models, our model achieved the highest Pos. F1 score, indi-cating that MPACKE can better identify the cause of emotion expressed in an utter-ance within a dialogue, demonstrating the effectiveness of MPACKE. Since the pur-pose of the Causal Emotion Entailment task is to find the cause that triggers emotion, we believe that making progress on Pos. F1 is more important than Neg. F1.

From Table 3, it can be noticed that ECPE-2D, ECPE-MLL, and RankCP have very low Pos. F1. The three models are originally designed to handle emotion-cause pair extraction tasks in document types. Therefore, they may not be suitable for dia-logue scenarios. Our model outperforms the two baseline models initially proposed for the dialogue emotion cause analysis task based on RoBERTa, and even surpasses the Large version of RoBERTa, reflecting the effectiveness of MPACKE. MPACKE, KEC, and KBCIN both enhance the ability to perceive hidden clues in the dialogue information flow by introducing commonsense knowledge. The difference is that KEC introduces too little knowledge, only considers knowledge in one aspect, and only focuses on the modeling of knowledge in the dialogue flow without explicitly modeling the relationship between knowledge and the triggered emotion. Compared to our model, KBCIN

Table 4: Results of ablation study.

Model	Pos. F1	Neg. F1	Macro F1
wo-rope	66.92	88.55	77.74
wo-PACKS	63.13	87.18	75.16
wo-RPAG	66.53	89.47	78.00
wo-IIC	66.07	85.94	76.01
Ours (MPACKE)	68.80	89.07	78.94

does not consider selecting knowledge, instead directly using all knowledge for the final classification semantic vector, which may bring knowledge noise.

4.5 Ablation Study

We conducted an ablation study to verify the effectiveness of MPACKE as shown in Table 4. During the knowledge selection phase, by removing the positional encoding, the selection process relies solely on the integrated knowledge of the target emotion utterances and the candidate cause utterances for attention distribution, neglecting the relative positional relationship between the Query and Key. This leads to a slight performance decline. However, benefit from the multi-level positional information mechanism, different sources of positional information can complement each other to some extent, mitigating the potential impact of the absence of a single positional information on performance.

Upon removing the Relative Position-Aware Graph (RPAG) module, which perceives turn information, there is a significant drop in the Pos. F1 score, while the Neg. F1 score slightly increases. After analyzing the dataset, we speculate that in some data, two utterances within a single turn are both the causes of a target emotion utterance. Therefore, when turn information perception is included, the model is likely to misjudge, i.e., the other utterance in turn is not the cause utterance. As a result, the removal of this feature leads to a slight improvement in Neg. F1. After removing the Intermedia information compensate (IIC) module, the model's performance drops significantly across all three metrics, with Neg. F1 is even lower than without commonsense reasoning. It indicates that even after refined dialogue level context modeling, the semantic features of the candidate cause utterances and the target emotion utterance cannot cover all intermediate dialogue information. Therefore, when predicting, relying solely on two semantic vectors is insufficient to fully perceive the entire emotional event. So the IIC module plays a crucial role in bridging the candidate cause utterances and the target emotion utterance, and its absence leads to a decline in model performance.

Furthermore, we attempted to remove the commonsense knowledge introduction module (PACKS), resulting in a significant degradation in Pos. F1 and Macro F1. It reaffirms that dialogue data contains a large amount of implicit information that requires inference based on daily experience and knowledge. Without commonsense reasoning, our model lacks a dialogue level information

Table 5: Results of ablation study.

Model	Pos. F1	Neg. F1	Macro F1
RNN	67.73	88.93	78.33
GRU	66.20	88.74	77.47
Ours - LSTM	68.80	89.07	78.94

interaction, with each utterance only perceiving itself or a small amount of surrounding information, falling into a ï£¡ï£¡local solutionï£¡ï£¡ and severely affecting the performance. Therefore, the introduction of commonsense knowledge is crucial for accurately understanding a dialogue.

In the Intermedia Information Compensate module, it is necessary to encode the dialogue information between the candidate cause utterances and the target emo-tional utterances at the dialogue level. We employed sequential models, experiment-ing with GRU, RNN, and LSTM, with the experimental results as shown in Table 5. LSTM demonstrates the best performance in encoding dialogue information. It is because LSTM effectively models contextual relationships and reduces information loss through its forget gate, input gate, output gate, and the hidden layer state and memory cell, and thus offers superior results compared to the unmodified RNN.

5 Conclusion

This paper presents the Multilevel Position-Aware and Commonsense Knowledge Enhancement (MPACKE) neural network for the task of Causal Emotion Entailment. To effectively perceive dialogue information, we first introduce commonsense knowledge to uncover implicit contextual clues between dialogues, then use the Posi-tion-Aware Commonsense Knowledge Selector (PACKS) to filter out knowledge that is useful for identifying cause utterance. Furthermore, we propose the Relative Posi-tion Aware Graph (RPAG) to encode the turn structure of the dialogue. Noticing that previous work simply concatenated candidate cause utterance and target utterances for prediction, we introduce an intermediate compensation mechanism to complete the information flow during prediction. The results show that our model achieves the best performance on the most critical Pos. F1 metric. We also validate the positive impact of each module of MPACKE.

References

1. 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 (2020)

2. 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 (2010)
3. 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: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6343–6350 (2019)
4. 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 (2020)
5. 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 (2020)
6. Gao, K., Xu, H., Wang, J.: Emotion cause detection for chinese micro-blogs based on ecocc model. In: Advances in Knowledge Discovery and Data Mining: 19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, May 19-22, 2015, Proceedings, Part II 19. pp. 3–14. Springer (2015)
7. Gao, K., Xu, H., Wang, J.: A rule-based approach to emotion cause detection for chinese micro-blogs. Expert Systems with Applications **42**(9), 4517–4528 (2015)
8. Gu, X., Lou, R., Sun, L., Li, S.: Page: A position-aware graph-based model for emotion cause entailment in conversation. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
9. 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 (2010)
10. Li, J., Meng, F., Lin, Z., Liu, R., Fu, P., Cao, Y., Wang, W., Zhou, J.: Neutral utterances are also causes: Enhancing conversational causal emotion entailment with social commonsense knowledge. arXiv preprint arXiv:2205.00759 (2022)
11. Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S.: Dailydialog: A manually labelled multi-turn dialogue dataset. arXiv preprint arXiv:1710.03957 (2017)
12. 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 (2013)
13. Poria, S., Majumder, N., Hazarika, D., Ghosal, D., Bhardwaj, R., Jian, S.Y.B., Hong, P., Ghosh, R., Roy, A., Chhaya, N., et al.: Recognizing emotion cause in conversations. Cognitive Computation **13**, 1317–1332 (2021)
14. Russo, I., Caselli, T., Rubino, F., Boldrini, E., Martínez-Barco, P., et al.: Emocause: an easy-adaptable approach to emotion cause contexts. Association for Computational Linguistics (ACL) (2011)
15. Song, H., Zhang, C., Li, Q., Song, D.: End-to-end emotion-cause pair extraction via learning to link. arXiv preprint arXiv:2002.10710 (2020)
16. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
17. Wang, F., Yu, J., Xia, R.: Generative emotion cause triplet extraction in conversations with commonsense knowledge. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023)

18. 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 (2020)
19. Xia, R., Ding, Z.: Emotion-cause pair extraction: A new task to emotion analysis in texts. arXiv preprint arXiv:1906.01267 (2019)
20. Xia, R., Zhang, M., Ding, Z.: Rthn: A rnn-transformer hierarchical network for emotion cause extraction. arXiv preprint arXiv:1906.01236 (2019)
21. Yoo, S., Jeong, O.: A token classification-based attention model for extracting multiple emotion–cause pairs in conversations. Sensors **23**(6), 2983 (2023)
22. Zhang, D., Yang, Z., Meng, F., Chen, X., Zhou, J.: Tsam: A two-stream attention model for causal emotion entailment. arXiv preprint arXiv:2203.00819 (2022)
23. Zhao, W., Zhao, Y., Li, Z., Qin, B.: Knowledge-bridged causal interaction network for causal emotion entailment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 14020–14028 (2023)
24. Zheng, X., Liu, Z., Zhang, Z., Wang, Z., Wang, J.: Ueca-prompt: Universal prompt for emotion cause analysis. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 7031–7041 (2022)