Multi-Granularity Semantic Aware Graph Model for Reducing Position Bias in Emotion-Cause Pair Extraction

Yinan Bao[1,2], **Qianwen Ma**[1,2], **Lingwei Wei**[1,2], **Wei Zhou**[1,*] and **Songlin Hu**[1,2]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
{baoyinan, maqianwen, weilingwei, zhouwei, husonglin}@iie.ac.cn

Abstract

The Emotion-Cause Pair Extraction (ECPE) task aims to extract emotions and causes as pairs from documents. We observe that the relative distance distribution of emotions and causes is extremely imbalanced in the typical ECPE dataset. Existing methods have set a fixed size window to capture relations between neighboring clauses. However, they neglect the effective semantic connections between distant clauses, leading to poor generalization ability towards position-insensitive data. To alleviate the problem, we propose a novel **M**ulti-**G**ranularity **S**emantic **A**ware **G**raph model (MGSAG) to incorporate fine-grained and coarse-grained semantic features jointly, without regard to distance limitation. In particular, we first explore semantic dependencies between clauses and keywords extracted from the document that convey fine-grained semantic features, obtaining keywords enhanced clause representations. Besides, a clause graph is also established to model coarse-grained semantic relations between clauses. Experimental results indicate that MGSAG surpasses the existing state-of-the-art ECPE models. Especially, MGSAG outperforms other models significantly in the condition of position-insensitive data.

1 Introduction

Emotion Cause Analysis (ECA) has attracted increasing research interest in recent years (Wei et al., 2020; Sun et al., 2021; Singh et al., 2021; Yu et al., 2021), because of the great potential of applying in consumer review mining, public opinion monitoring, and online empathetic chatbot building. Its goal is to detect causes or stimuli for a certain emotion expressed in text.

Emotion Cause Pair Extraction (ECPE) (Xia and Ding, 2019) is a new task related to ECA, which is concerned with causal relationships be-
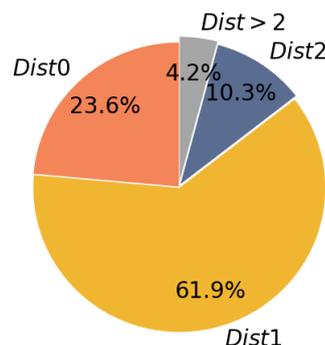
---
* Corresponding author.

Figure 1: The distribution of the relative distance of an emotion clause and a cause clause that comprise a pair in the ECPE dataset (Xia and Ding, 2019). $Dist0$, $Dist1$, and $Dist2$ mean the relative distances between the two clauses are 0, 1, and 2 respectively. $Dist > 2$ means the relative distances are larger than 2.

tween emotions and causes. It's a much more challenging task. Because we need a comprehensive understanding of document content and structure to perform emotion-cause co-extraction and discriminate emotion-cause clause pairs from negative ones (Wei et al., 2020). As shown in the following example, an emotion clause c_7 and a cause clause c_2 construct an emotion-cause pair (c_7, c_2) which is needed to be extracted by an ECPE model.

Example. *When the driver was about to start the bus to leave the station* (c_1), *an old lady ran to the front of the bus with a fast speed and sat down on the ground* (c_2). *Passengers standing in the front of the bus can see this scene clearly* (c_3). *Seeing this scene* (c_4), *the passengers in the car immediately became restless* (c_5), *and had a heated debate* (c_6). *Some of the passengers were **angry*** (c_7), *and told the driver he shouldn't be meddlesome* (c_8).

In general, the number of candidate emotion-cause pairs is the square of the number of clauses in a document. However, most documents contain only one emotion-cause pair. Due to the problem of the tremendous search space, most existing

1203

methods have fully exploited relative position features to decrease the number of candidate pairs. For instance, ECPE-MLL (Ding et al., 2020b) and SLSN (Cheng et al., 2020) set a fixed size window around a certain clause, and the central clause and other clauses inside the window comprise candidate pairs. However, models heavily relying on the relative position features ignore the distant semantic cues, resulting in poor generalization ability towards position-insensitive data in which the cause clause is not in proximity to the emotion clause.

According to Figure 1, we can observe that there is a position bias problem in ECPE. For the most 85% emotion-cause pairs, the relative distances between its emotion clauses and corresponding cause clauses are less than 2. It means that most cause clauses either appear immediately preceding/following their corresponding emotion clauses or are the emotion clauses themselves. Existing methods mainly focus on the position-sensitive data (majority) and neglect the position-insensitive data (minority). How to improve the performance on the two parts of data instead of only focusing on one of them, has become an intractable challenge.

Some proposed methods (Xia and Ding, 2019; Chen et al., 2020a) without relative position information seem to be position-insensitive, but overlook the effective semantic connections between distant clauses which convey causal cues. Thus, they can not alleviate the position bias problem.

To alleviate this problem, we propose a multi-granularity semantic aware graph model (MGSAG). We assume that fine-grained semantic features conveyed by global keywords in a document are conducive to exploring causal cues, especially cues implied in distant clauses. Besides, coarse-grained semantics between clauses is also important to find causal relations implied in the context. From the two perspectives, we realize multi-granularity semantic enhanced clause relationships modeling based on two graphs: clause-keyword bipartite graph and fully connected clause graph, utilize fine-grained and coarse-grained semantic features jointly. Experimental results show that MGSAG outperforms all of the state-of-the-art baselines. Especially, it achieves a significant improvement on position-insensitive test data. In summary, our contributions are three-fold:

- To alleviate the position bias problem in ECPE, we propose MGSAG to achieve fine-grained and coarse-grained semantic enhanced clause representation learning.

- To value model performance on emotion-cause clause pairs consisting of distant clauses, we split the original test set into two parts according to the relative distances of emotion clauses and cause clauses, and evaluate models on them.

- Experimental results prove that our model achieves remarkable improvement over best-performing approaches on the original test set. Especially, it outperforms other methods in the condition of position-insensitive data.

2 Related Work

According to whether the relative position information is used explicitly or not, existing ECPE works can be divided into two categories: position-sensitive approaches and position-insensitive approaches.

Position-Sensitive Approaches. Most methods (Ding et al., 2020a; Cheng et al., 2020; Ding et al., 2020b) have set a fixed size window to reduce the number of candidate pairs according to the inherent position bias in the dataset, because of the sparsity of true emotion-cause pairs compared with candidate emotion-cause pairs. Besides, Chen et al. (2020b) leveraged the relative position information explicitly in the process of pair representation learning. The ECPE-MLL model proposed by Ding et al. (2020b) is the state-of-the-art method in the ECPE task. An over-reliance on relative position information makes these methods have poor generalization ability towards position-insensitive data.

Position-Insensitive Approaches. Some sequence-based methods without relative position information (Xia and Ding, 2019; Chen et al., 2020a; Fan et al., 2020) seem to be position-insensitive. Xia and Ding (2019) proposed a RNN-based framework and generate candidate pairs by applying the Cartesian product. Chen et al. (2020a) reformulated the ECPE task as a unified sequence labeling problem. Fan et al. (2020) modeled the extraction of emotion-cause pairs as performing a sequence of transitions and actions. However, these methods have shown poor performance on position-insensitive data due to the neglect of effective semantic connections between distant clauses.

Different from the above methods, our model incorporates fine-grained and coarse-grained semantic features jointly, which can alleviate the position bias problem well.
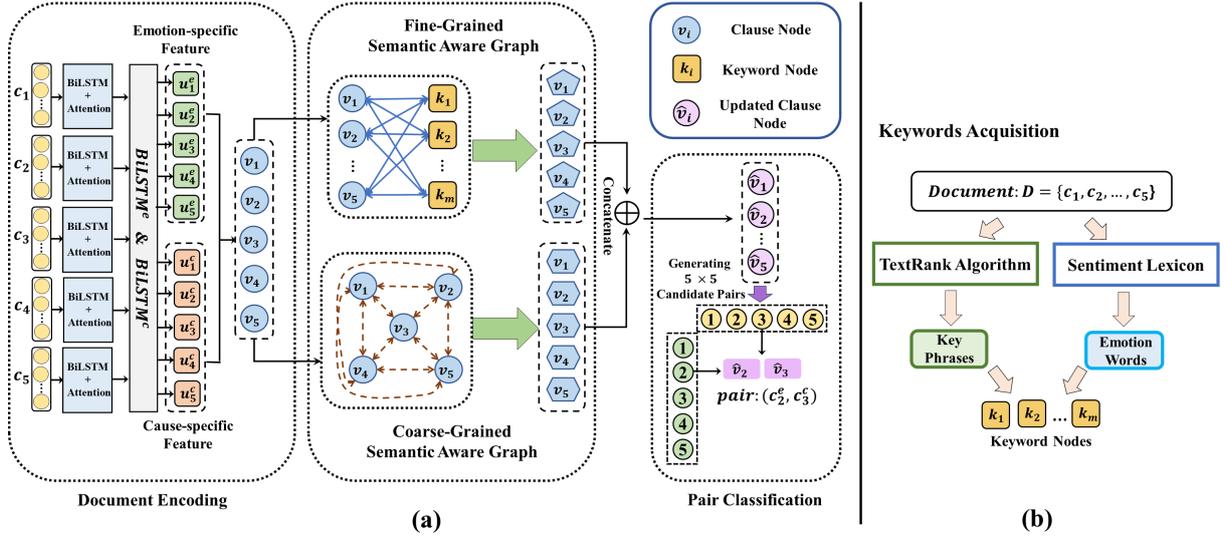
Figure 2: (a) shows an overview of MGSAG. (b) shows the process of keywords acquisition.

3 Problem Formulation

Given a document $D = \{c_1, c_2, ..., c_{|D|}\}$ where $|D|$ is the number of clauses, the clauses are formed into $|D| \times |D|$ candidate emotion-cause pairs using Cartesian product: $P = \{..., (c_i^e, c_j^c), ...\}$, where c_i^e is clause c_i serving as a candidate emotion clause, c_j^c is clause c_j serving as a candidate cause clause. The ECPE task is to assign a binary label to each candidate pair (c_i^e, c_j^c), where "1" means that clause c_i is an emotion clause and clause c_j provides the cause of it, otherwise "0".

4 Methodology

We propose a multi-granularity semantic aware graph model to alleviate the position bias problem in ECPE. More concretely, we obtain fine-grained semantic aware clause representations based on a clause-keyword bipartite graph. Simultaneously, coarse-grained semantic aware clause representations are generated based on a fully connected clause graph. As shown in Figure 2, the model consists of four components: 1) document encoding, 2) fine-grained semantic aware graph (FGSAG), 3) coarse-grained semantic aware graph (CGSAG), 4) pair classification.

4.1 Document Encoding

Given a document $D = \{c_1, c_2, ..., c_{|D|}\}$ consisted of $|D|$ clauses, we adopt a hierarchical recurrent neural network to encode context information and generate emotion-specific and cause-specific clause representations for each clause in the document.

Word-Level Encoder. For each clause $c_i = \{w_1^i, w_2^i, ..., w_{|c_i|}^i\}$, we first adopt a word-level BiLSTM network to encode the context by passing words' information along the clauses forwards and backwards, and then obtain the clause's hidden state sequence $(h_1^i, h_2^i, ..., h_{|c_i|}^i)$. An attention layer is adopted to combine them and return a state vector $\mathbf{h}_i = \sum_{j=1}^{|c_i|} \alpha_j h_j^i$ for clause c_i, where $\alpha_j = \mathbf{softmax}(\mathbf{W}_a h_j^i)$ is the attention weight of the j-th word in clause c_i, $\mathbf{W}_a$ is a trainable weight matrix for attention score calculation.

Clause-Level Encoder. In order to extract the emotion features and the cause features respectively, the clause-level encoder consists of two BiLSTM networks. The document D's clause state sequence $(\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{|D|})$ is fed into two clause-level BiLSTM networks to produce emotion-specific and cause-specific clause representations, respectively:

$$
\begin{aligned}
\mathbf{u}_i^e &= \mathbf{BiLSTM}^e(\mathbf{h}_i)\,, \\
\mathbf{u}_i^c &= \mathbf{BiLSTM}^c(\mathbf{h}_i)\,,
\end{aligned} \tag{1}
$$

where $\mathbf{BiLSTM}^e$ and $\mathbf{BiLSTM}^c$ generate the emotion-specific and cause-specific clause representation $\mathbf{u}_i^e, \mathbf{u}_i^c \in \mathbb{R}^{2d_h \times 1}$ of clause c_i, respectively. d_h means the number of hidden units in BiLSTM.

Afterwards, we use a gate mechanism to fuse the emotion feature $\mathbf{u}_i^e$ and the cause feature $\mathbf{u}_i^c$ to obtain clause representation $\mathbf{v}_i \in \mathbb{R}^{2d_h \times 1}$:

$$
\begin{aligned}
\mathbf{g}_i &= \sigma(\mathbf{W}_g \mathbf{u}_i^e + \mathbf{b}_g)\,, \\
\mathbf{v}_i &= \mathbf{g}_i \mathbf{u}_i^c + (1 - \mathbf{g}_i)\mathbf{u}_i^e\,,
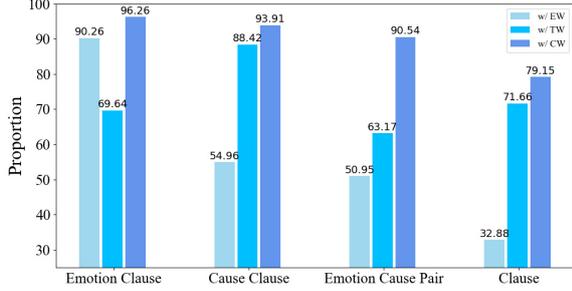\end{aligned} \tag{2}
$$

Figure 3: The influence of the two types of keywords from an intuitive aspect. It shows the proportion of emotion clauses, cause clauses, emotion-cause pairs, and clauses that are covered by the extracted key phrases or emotion words or both of them. "w/ EW", "w/ TW", and "w/ CW" means using emotion words, key phrases obtained by TextRank or both of them, respectively.

where $\mathbf{W}_g \in R^{1 \times 2d_h}$ and $\mathbf{b}_g$ are parameters; σ is the sigmoid function.

In the training process, we leverage the emotion labels and cause labels as auxiliary supervision signals to facilitate the clause representation learning in the clause-level encoder:

$$\begin{aligned} \hat{\mathbf{y}}_i^e &= \mathbf{softmax}(\mathbf{W}_e \mathbf{u}_i^e + \mathbf{b}_e), \\ \hat{\mathbf{y}}_i^c &= \mathbf{softmax}(\mathbf{W}_c \mathbf{u}_i^c + \mathbf{b}_c), \end{aligned} \quad (3)$$

where $\mathbf{W}_e, \mathbf{W}_c \in \mathbb{R}^{1 \times 2d_h}$ are trainable parameters and $\mathbf{b}_e, \mathbf{b}_c$ are bias terms.

4.2 Fine-Grained Semantic Aware Graph

To obtain fine-grained semantic enhanced clause representations, we leverage external knowledge to extract keywords in the document first. Then, we build a clause-keyword bipartite graph to model the relations between clauses. In this way, the keywords which convey fine-grained semantic features can help highlight the potential causal features contained in the clause representations.

Keywords Acquisition. We use the TextRank algorithm (Mihalcea and Tarau, 2004) to extract key phrases and a sentiment lexicon (Xu et al., 2008)[1] to obtain emotion words in a document. We take the union of the two sets as the final keyword set.

To measure the influence of the two types of keywords from an intuitive view, we count the proportions of emotion clauses, cause clauses, emotion-cause pairs, and clauses that are covered by the emotion words or key phrases or both of them. Noted that if emotion clause and cause clause that

comprise a pair both contain any keyword, we think that the pair is covered by the keywords.

From Figure 3 we observe that if we use the key phrases extracted by TextRank alone, only about 69% of emotion clauses can be found; if we use the emotion words alone, only about 54% of cause clauses can be identified. With the use of emotion words or key phrases, only about 50% or 63% of emotion-cause pairs can be figured out. Consequently, we take the union of the two sets as the final keyword set. However, given the complete keyword set, clauses that contain keywords account for a large proportion (79%), which means that the imported keywords may introduce noise as well. To this end, it's necessary to measure the importance of different keywords when modeling the interaction between clauses and keywords.

Clause-Keyword Bipartite Graph Construction. Given a document D, we denote the clause-keyword bipartite graph as $\mathcal{G}_b = (\mathcal{V}, \mathcal{E}_b)$, where $\mathcal{V} = \mathcal{V}_c \cup \mathcal{V}_k$ represents a node set composing of clause nodes and keyword nodes and $\mathcal{E}_b$ denotes edges between nodes. $\mathcal{V}_k = \{k_1, k_2, ..., k_m\}$ and $\mathcal{V}_c = \{c_1, c_2, ..., c_{|D|}\}$ mean there are m keywords and $|D|$ clauses in the document D. We establish edges between each node in $\mathcal{V}_c$ and each node in $\mathcal{V}_k$, which means every element e_{ij} in $\mathcal{E}_b \in \mathbb{R}^{|D| \times m}$ is 1. It is because the average length of clauses is too short, many keywords only appear once in one clause. Thus, an adjacency matrix based on keyword-clause co-occurrence is extremely sparse.

For keywords in $\mathcal{V}_k$, their feature vectors are initialized by the word embedding vectors released by Xia and Ding (2019). As for clause nodes $c_i \in \mathcal{V}_c$, they are initialized with the corresponding context-aware clause representation $\mathbf{v}_i$ generated from the clause-level encoder. We denote the feature matrices of keyword and clause nodes as $\mathbf{X}_k = \{\mathbf{k}_1, ..., \mathbf{k}_m\} \in \mathbb{R}^{m \times d_w}$ and $\mathbf{X}_c = \{\mathbf{v}_1, ..., \mathbf{v}_{|D|}\} \in \mathbb{R}^{|D| \times 2d_h}$ respectively, where d_w is the dimension of the word embedding and is equal to $2d_h$ in our setting.

Attention Guided Clause Representations Update. We propose a graph attention module to model the semantic interaction between clauses and keywords, aiming to utilize the fine-grained semantic features implied in keywords to facilitate clause representation learning.

Intuitively, the clause-keyword bipartite graph realizes fine-grained semantic connections between distant clauses, which is helpful to extract emotion-

---

cause pairs composed of distant clauses. Nevertheless, for a specific clause, the importance of various keywords is different. Therefore, we use the graph attention mechanism (Velickovic et al., 2018) to measure the document-level keyword preference-degree of each clause, where the attention weight is computed as the edge weight between the clause node c_i and the keyword node k_j in a document:

$$\alpha_{ij} = \frac{\exp(w^\top[\mathbf{W}_1\mathbf{v}_i; \mathbf{W}_2\mathbf{k}_j])}{\sum_{t=1}^{|D|} \exp(w^\top[\mathbf{W}_1\mathbf{v}_t; \mathbf{W}_2\mathbf{k}_j])}, \quad (4)$$

where $\mathbf{v}_i$ and $\mathbf{k}_j$ are features of clause c_i and keyword k_j respectively; $[\cdot; \cdot]$ is the concatenation operation; $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d_w \times d_w}$ and $w \in \mathbb{R}^{2d_w \times 1}$ are trainable parameters.

Then, clause c_i is encoded as the fine-grained semantic enhanced representation $\mathbf{v}_i^b$ as follows:

$$\mathbf{v}_i^b = \tanh\left(\left(\mathbf{v}_i + \sum_{j=1}^{m}\left(\alpha_{ij}\left(\sum_{t=1}^{|D|}\alpha_{tj}\mathbf{W}_3\mathbf{v}_t\right)\right)\right) + \mathbf{b}\right), \quad (5)$$

where $\sum_{t=1}^{|D|}\alpha_{tj}\mathbf{W}_3\mathbf{v}_t$ means the representation of the keyword $\mathbf{k}_j$, and $\sum_{j=1}^{m}(\alpha_{ij}(\sum_{t=1}^{|D|}\alpha_{tj}\mathbf{W}_3\mathbf{v}_t))$ is the weighted added of keyword representations for generating fine-grained semantic enhanced clause representation. $\mathbf{W}_3 \in \mathbb{R}^{d_w \times d_w}$ is a trainable parameter and $\mathbf{b}$ is a bias term.

4.3 Coarse-Grained Semantic Aware Graph

Coarse-grained semantic relationships between clauses are useful for finding causal cues implied in the context. We establish a fully connected clause graph and leverage graph attention mechanism to model the coarse-grained semantic relationships between clauses.

Given a document D, we define the clause graph as $\mathcal{G}_c = (\mathcal{V}_c, \mathcal{E}_c)$, where $\mathcal{V}_c$ represents a node set and $\mathcal{E}_c$ denotes an edge set. Each node in the fully connected graph is a clause in D, and every two nodes have an edge. Self-loop edge is added to every node because a clause can be an emotion clause and a cause clause simultaneously. We use clause representation $\mathbf{v}_i$ generated from the clause-level encoder for node feature initialization. Based on the self-attention mechanism (Vaswani et al., 2017) which aggregated neighboring clauses' information, the graph attention network propagates information among clauses by stacking multiple graph attention layers. The representation of clause

c_i in the t-th layer is updated as follows:

$$\mathbf{v}_i^{(t)} = \mathbf{ReLU}\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(t)}\mathbf{W}_1^{(t)}\mathbf{v}_j^{(t-1)} + \mathbf{b}^{(t)}\right), \quad (6)$$

where $\mathbf{W}_1^{(t)} \in \mathbb{R}^{d_w \times d_w}$ is a transform matrix and $\mathbf{b}^{(t)}$ is a bias term; $\mathcal{N}(i)$ represents the neighbouring clauses of c_i; $\mathbf{v}_i^{(0)} = \mathbf{v}_i$. The attention weight $\alpha_{ij}^{(t)}$ is learned as follows:

$$e_{ij}^{(t)} = w^{(t)^\top}\tanh([\mathbf{W}_2^{(t)}\mathbf{v}_i^{(t-1)}; \mathbf{W}_3^{(t)}\mathbf{v}_j^{(t-1)}]),$$

$$\alpha_{ij}^{(t)} = \frac{\exp(\mathbf{LeakyReLU}(e_{ij}^{(t)}))}{\sum_{k \in \mathcal{N}(i)} \exp(\mathbf{LeakyReLU}(e_{ik}^{(t)}))}, \quad (7)$$

We stack two graph attention layers and obtain $\mathbf{v}_i^c = \mathbf{v}_i^{(2)}$ as the updated representation for c_i.

4.4 Pair Classification

We concatenate the two types of clause representations and obtain $\hat{\mathbf{v}}_i = [\mathbf{v}_i^b; \mathbf{v}_i^c]$ as the final representation of clause c_i.

Emotion Cause Pair Extraction. For a candidate pair $(c_i^e, c_j^c) \in P$, we pass its representation $\mathbf{v}_{ij}^p = [\hat{\mathbf{v}}_i; \hat{\mathbf{v}}_j]$ to a fully-connected layer with softmax activation function to predict the label of it:

$$\hat{\mathbf{p}}_{ij} = \mathbf{softmax}(\mathbf{W}_p^\top\mathbf{v}_{ij}^p + \mathbf{b}_p), \quad (8)$$

where $\mathbf{W}_p \in \mathbb{R}^{4d_w \times 2}$ and $\mathbf{b}_p \in \mathbb{R}^{2 \times 1}$ are trainable parameters. We obtain the predicted label $\hat{\mathbf{EC}}_{ij}$ for the candidate pair (c_i^e, c_j^c) according to the probability distribution $\hat{\mathbf{p}}_{ij}$.

During model training, we use two cross-entropy loss functions $\mathcal{L}_{emo}$ and $\mathcal{L}_{cau}$ to supervise the clause representation learning in the clause-level encoder and a cross-entropy loss function $\mathcal{L}_{pair}$ to supervise the final emotion-cause pair prediction. The loss function $\mathcal{L}$ is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{pair} + \mathcal{L}_{emo} + \mathcal{L}_{cau}. \quad (9)$$

Emotion Extraction and Cause Extraction. Following Chen et al. (2020b), we implement emotion extraction and cause extraction based on the predictions of all candidate pairs. For emotion extraction, the predicted label $\hat{\mathbf{E}}_i$ for clause c_i can be obtained as follows:

$$\hat{\mathbf{E}}_i = \begin{cases} 1, & if \sum_{j=1}^{|D|}(\hat{\mathbf{EC}}_{ij}) > 0 \\ 0, & otherwise \end{cases}. \quad (10)$$

For cause extraction, the predicted label $\hat{\mathbf{C}}_i$ for clause c_i can be obtained similarly.

1207

5 Experiments

We conduct a series of experiments to verify the effectiveness of MGSAG.

5.1 Experimental Setup

5.1.1 Dataset and Evaluation Metrics

We use the benchmark dataset released by Xia and Ding (2019) for experiments. This typical and widely used dataset is constructed based on an emotion cause extraction corpus (Gui et al., 2016) that contains 1,945 Chinese documents from SINA city news[2]. To obtain statistically credible results, we adopt the same data split setting (10-fold cross-validation) used by Xia and Ding (2019), repeat the experiments 10 times, and report the average results of precision (P), recall (R), and F_1-score (F_1) on the main task: emotion-cause pair extraction (**ECPE**), and two sub-tasks: emotion extraction (**EE**) and cause extraction (**CE**), following existing works (Xia and Ding, 2019; Ding et al., 2020b,a; Chen et al., 2020a,b; Cheng et al., 2020).

5.1.2 Redistricting of Original Test Set

As ECPE is a newly proposed task, there is only one typical and widely used dataset. Because of the inherent position bias in ECPE, how to improve the performance on both position-sensitive (majority) and position-insensitive data (minority), has become one of the challenges. Therefore, it is essential to measure the reliance of existing methods on the relative position information.

To this end, we split the original test set ($Test_{all}$) of each fold into two parts according to the relative distance between emotions and causes. The first part ($Test_{Bias}$) contains documents with only one pair and the relative distance between the two clauses is less than 2. The second part ($Test_{NoBias}$) is the complement of the first part, which means $Test_{all} = Test_{Bias} \cup Test_{NoBias}$ and $Test_{Bias} \cap Test_{NoBias} = \emptyset$. We conduct experiments on the original test set first, and then use $Test_{Bias}$ and $Test_{NoBias}$ to evaluate various methods respectively. To ensure fairness, we use the same model parameters which produce results on $Test_{all}$ to obtain the results on the two subsets: $Test_{Bias}$ and $Test_{NoBias}$.

5.1.3 Comparative Approaches

We compare MGSAG with the following methods, which can be divided into two types: position-

---

[2]http://news.sina.com.cn/society/

insensitive and position-sensitive methods.

Position-insensitive Methods. Following methods haven't utilized the relative position information explicitly. **Indep / Inter-CE / Inter-EC** (Xia and Ding, 2019): these two-step approaches first extracted emotions and causes separately to form candidate emotion-cause pairs and then trained a classifier to recognize true pairs. **IE-CNN** (Chen et al., 2020a) reformulated the ECPE task as a sequence labeling task and extracted pairs in an end-to-end fashion.

Position-sensitive Methods. Following methods take relative position information as a crucial feature to recognize pairs. **PairGCN** (Chen et al., 2020b) is a method highly dependent on position information when modeling relations between pairs. **ECPE-2D** (Ding et al., 2020a) extracted pairs through 2D representation, interaction, and prediction. The window-constrained 2D Transformer achieved the best performance. **SLSN-U** (Cheng et al., 2020) extracted pairs through a process of local search which was defined by the setting of the local context window. **RankCP** (Wei et al., 2020) utilized kernel-based relative position embedding to enhance the clause representations obtained from inter-clause modeling module. **ECPE-MLL** (Ding et al., 2020b) used a multi-label learning method inside each sliding window which was defined manually.

5.1.4 Implementation Details

To conduct a fair comparison with the baselines, we utilize the same word embeddings followed Xia and Ding (2019). The dimension of word embedding is 200. The numbers of hidden units of BiLSTM in the word-level and clause-level encoder are set to 200 and 100, respectively. We stack two graph attention layers to build a graph attention network and add dropout (Srivastava et al., 2014) with the rate of 0.1 for each layer to reduce overfitting. During the training process, we use the Adam (Kingma and Ba, 2015) optimizer to update all parameters. We report the results of BERT (Devlin et al., 2019) in the appendix.

5.2 Experimental Results

5.2.1 Results on Original Test Set

Table 1 reports the comparative results on emotion cause pair extraction and two sub-tasks. We can observe that position-sensitive models perform better than position-insensitive models on average, indicating the effectiveness of using relative position

Category	Model	Emotion Ext.			Cause Ext.			EC Pair Ext.		
		P	R	F_1	P	R	F_1	P	R	F_1
Position-insensitive Baselines	Indep	0.8375	0.8071	0.8210	0.6902	0.5673	0.6205	0.6832	0.5082	0.5818
	Inter-CE	0.8494	0.8122	0.8300	0.6809	0.5634	0.6151	0.6902	0.5135	0.5901
	Inter-EC	0.8364	0.8107	0.8230	0.7041	0.6083	0.6507	0.6721	0.5705	0.6128
	IE-CNN	0.8614	0.7811	0.8188	0.7348	0.5841	0.6496	0.7149	0.6279	0.6686
Position-sensitive Baselines	PairGCN	0.8587	0.7208	0.7829	0.7283	0.5953	0.6541	0.6999	0.5779	0.6321
	ECPE-2D	0.8512	0.8220	0.8358	0.7272	0.6298	0.6738	0.6960	0.6118	0.6496
	SLSN-U	0.8406	0.7980	0.8181	0.6992	0.6588	0.6778	0.6836	0.6291	0.6545
	RankCP	0.8703	0.8406	**0.8548**	0.6927	**0.6743**	0.6824	0.6698	**0.6546**	0.6610
	ECPE-MLL	0.8582	**0.8429**	0.8500	0.7248	0.6702	0.6950	0.7090	0.6441	0.6740
Our Model	**MGSAG**	**0.8721**	0.7911	0.8287	**0.7510**	0.6713	**0.7080**	**0.7243**	0.6507	**0.6846**

Table 1: Comparison of varying approaches on the original test set ($Test_{all}$).

Model	$Test_{Bias}$	$Test_{NoBias}$
Inter-EC	0.6783	0.3318
IE-CNN	0.7666	0.3484
PairGCN	0.7246	0.3355
ECPE-2D	0.7590	0.3830
SLSN-U	0.7456	0.3978
RankCP	0.7467	0.3857
ECPE-MLL	0.7673	0.3988
MGSAG	**0.7730**	**0.4301**

Table 2: F_1 results of varying approaches on $Test_{Bias}$ and $Test_{NoBias}$, focusing on EC Pair Ext.

Model	$Test_{Bias}$	$Test_{NoBias}$	$Test_{all}$
w/o FGSAG	0.7594	0.3894	0.6519
w/o CGSAG	0.7654	0.4027	0.6529
w/o FGSAG+CGSAG	0.7264	0.3269	0.6242
MGSAG	**0.7730**	**0.4301**	**0.6846**

Table 3: F_1 results of ablation study on $Test_{Bias}$, $Test_{NoBias}$, and $Test_{all}$, focusing on EC Pair Ext.

information. However, our method MGSAG hasn't utilized relative position information, aiming to alleviate the position bias problem in ECPE. In spite of this, MGSAG still outperforms the existing state-of-the-art methods. Especially, MGSAG achieves the best F_1 on the main task: emotion-cause pair extraction. The F_1 score of MGSAG on ECPE is **1.06%** higher than that of ECPE-MLL, which indicates the efficiency of capturing multi-granularity semantic relations between clauses.

For the two sub-tasks, MGSAG outperforms other baselines in terms of cause extraction compared with emotion extraction. This indicates that the effective clause representation learning based on MGSAG is beneficial to extract cause clauses and further facilitate the extraction of emotion-cause pairs.

5.2.2 Results on $Test_{Bias}$ and $Test_{NoBias}$

To evaluate if MGSAG is vulnerable when the causes are not in proximity to the emotion, we evaluate it on the two subsets as shown in 5.1.2. Table 2 shows the results on $Test_{Bias}$ and $Test_{NoBias}$. Noted that when we get the best results on the original test set as shown in Table 1, we use the same parameters to evaluate models on the two subsets ($Test_{Bias}$ and $Test_{NoBias}$).

From Table 2 we observe that there is a significant gap (34∼41%) between the results on $Test_{Bias}$ and $Test_{NoBias}$, for all of the methods. One of the reasons should be the imbalanced data of $Test_{Bias}$ and $Test_{NoBias}$, which means the proportion of position-insensitive data is very small. More importantly, most of the methods exploit the relative position information explicitly or implicitly, leading to poor performance on $Test_{NoBias}$.

However, MGSAG outperforms existing state-of-the-art baselines on both of the two subsets ($Test_{Bias}$ and $Test_{NoBias}$), proving its generalization ability towards position-sensitive and position-insensitive data. Specially, the F_1 score of MGSAG on $Test_{NoBias}$ is **3.13%** higher than that of ECPE-MLL. The results verify the effectiveness of capturing causal relations between clauses via multi-granularity semantics encoding.

5.3 Discussions

We conduct ablation studies to analyze the effects of different components and settings in our method MGSAG.

5.3.1 Influence of Different Components

As shown in Table 3, we remove FGSAG, CGSAG, and both of them respectively to verify the effectiveness of the proposed two graphs with the semantics of different granularity.
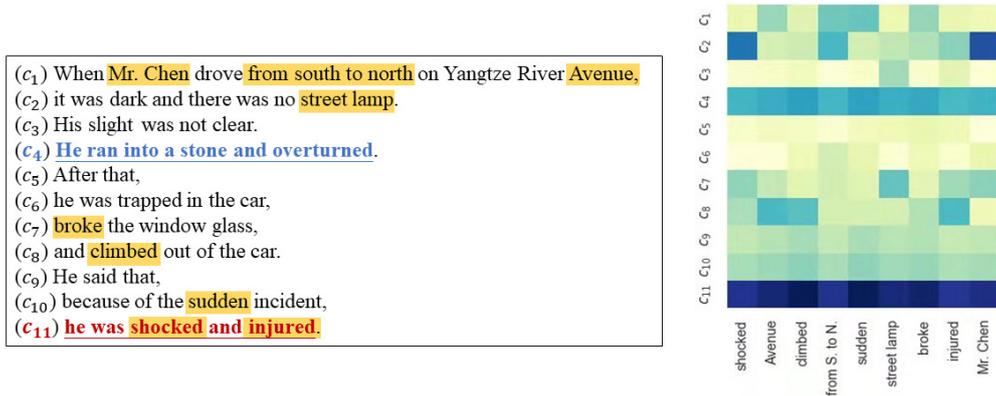
Figure 4: An example that MGSAG extracts the emotion cause pair (c_{11}, c_4) correctly, while ECPE-MLL fails. Words shaded in yellow are keywords. The heatmap presents attention scores in the clause-keyword bipartite graph. Rows of c_{11} and c_4 are the top-two darkest rows, means that keywords pay more attention to them and facilitate MGSAG to extract pair (c_{11}, c_4) correctly.

Effect of Fine-Grained Semantic Aware Graph. We remove the FGSAG to verify the effect of fine-grained semantic enhanced relations. Table 3 shows that removing FGSAG results in significant performance degradation, indicating that it is indeed useful for pair prediction. Especially, the result of F_1 on $Test_{NoBias}$ decreases **4.07%** without the FGSAG, proving its efficiency of alleviating position bias.

Effect of Coarse-Grained Semantic Aware Graph. We remove CGSAG which is used for coarse-grained semantic enhanced relations to verify its effect. Table 3 reports that model without CGSAG results in a clear drop (**2.74%/3.17%**) on $Test_{NoBias}$ and $Test_{all}$, but a limited drop (**0.76%**) on $Test_{Bias}$. It shows that modeling the coarse-grained semantic relations between clauses can alleviate position bias as well.

Effect of Semantic Aware Graph Model. We further evaluate the effect of dual graph-based modules by removing FGSAG and CGSAG simultaneously. As shown in Table 3, the model without the two graphs performs worse than without any one of them. The significant performance decline of the F_1 score on all of the test sets verifies that the fine-grained semantics and coarse-grained semantics are complementary to each other. Thus, it's necessary to take both of them into account.

5.3.2 Influence of Two-Level Supervision

We use the two-level supervised signals to train MGSAG. A low-level signal $\mathcal{L}_{emo} + \mathcal{L}_{cau}$ supervises the clause representation learning at the clause-level encoder and a high-level signal $\mathcal{L}_{pair}$ supervises the pair representation learning at the

Loss Function	P	R	F_1
$\mathcal{L}_{pair}$	0.6940	**0.6533**	0.6720
$\mathcal{L}_{pair} + \mathcal{L}_{emo} + \mathcal{L}_{cau}$	**0.7243**	0.6507	**0.6846**

Table 4: Comparison of different supervised signals for our method.

Model	$Test_{Bias}$	$Test_{NoBias}$	$Test_{all}$
w/ RW	0.7596	0.4078	0.6674
w/o EW	0.7669	0.3920	0.6686
w/o TW	0.7658	0.4271	0.6771
MGSAG	**0.7730**	**0.4301**	**0.6846**

Table 5: Comparative F_1 results on $Test_{Bias}$, $Test_{NoBias}$, and $Test_{all}$ of our variant models, focusing on EC Pair Ext. "w/ RW" means using random embeddings for keyword feature initialization. "w/o EW" and "w/o TW" means removing emotion words and key phrases obtained by TextRank, respectively.

classification stage. To evaluate the effectiveness of low-level supervision, we only use $\mathcal{L}_{pair}$ to train the model, and the results are shown in Table 4. It shows that training with low-level supervision brings an improvement mainly on precision, which indicates that the low-level supervision is helpful to learn more accurate emotion-specific and cause-specific features and eventually facilitates the performance on emotion-cause pair extraction.

5.3.3 Influence of Different Keyword Settings

As shown in Table 5, we use different keyword settings to verify the effectiveness of our proposed keywords, which is the union of emotion words obtained from a sentiment lexicon (Xu et al., 2008)

and key phrases obtained by TextRank (Mihalcea and Tarau, 2004). Removing any one of them results in a performance decline on all of the test sets. It proves that it's necessary to take both of them into account. Moreover, we replace the keyword features with randomly initialized embeddings, showing a significant drop on $Test_{NoBias}$. It indicates that the fine-grained semantics implied in keywords does help to alleviate the position bias problem.

5.3.4 Case Study

As shown in Figure 4, the distance between the emotion clause c_{11} and the cause clause c_4 is 7. Although the cause clause c_4 doesn't contain any keywords, global keywords in the document convey crucial fine-grained semantics, helping MGSAG extracts (c_{11}, c_4) correctly.

6 Conclusion and Future Work

In this paper, we propose MGSAG to alleviate the position bias problem in the ECPE task. Our approach implements clause representation learning via fine-grained semantics introduced by keywords and coarse-grained semantics among clauses. Experimental results show that MGSAG surpasses the state-of-the-art baselines, and outperforms other methods significantly on the position-insensitive data. In the future, we would like to tackle the problem of imbalanced data by reducing non-emotion-cause pairs, based on a position-insensitive approach.

References

Xinhong Chen, Qing Li, and Jianping Wang. 2020a. A unified sequence labeling model for emotion cause pair extraction. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 208–218.

Ying Chen, Wenjun Hou, Shoushan Li, Caicong Wu, and Xiaoqiang Zhang. 2020b. End-to-end emotion-cause pair extraction with graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 198–207.

Zifeng Cheng, Zhiwei Jiang, Yafeng Yin, Hua Yu, and Qing Gu. 2020. A symmetric local search network for emotion-cause pair extraction. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 139–149.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020a. 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, ACL 2020, Online, July 5-10, 2020*, pages 3161–3170.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020b. 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 2020, Online, November 16-20, 2020*, pages 3574–3583.

Chuang Fan, Chaofa Yuan, Jiachen Du, Lin Gui, Min Yang, and Ruifeng Xu. 2020. Transition-based directed graph construction for emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3707–3717.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1639–1649.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 404–411.

Aaditya Singh, Shreeshail Hingane, Saim Wani, and Ashutosh Modi. 2021. An end-to-end network for emotion-cause pair extraction. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EACL 2021, Online, April 19, 2021*, pages 84–91.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Qixuan Sun, Yaqi Yin, and Hong Yu. 2021. A dual-questioning attention network for emotion-cause pair extraction with context awareness. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, pages 1–8.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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

Penghui Wei, Jiahao Zhao, and Wenji Mao. 2020. 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, ACL 2020, Online, July 5-10, 2020*, pages 3171–3181.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1003–1012.

Linhong Xu, Hongfei Lin, Yu Pan, Ren Hui, and Jianmei Chen. 2008. Constructing the affective lexicon ontology. *Journal of the China Society For Scientific and Technical Information*, 27(2):180–185.

Jiaxin Yu, Wenyuan Liu, Yongjun He, and Chunyue Zhang. 2021. A mutually auxiliary multitask model with self-distillation for emotion-cause pair extraction. *IEEE Access*, 9:26811–26821.

A Experimental Results with BERT

Model	$Test_{Bias}$	$Test_{NoBias}$
PairGCN	0.7246	0.3355
MGSAG	**0.7730**	**0.4301**
PairGCN (BERT)	**0.8219**	0.4005
MGSAG (BERT)	0.8214	**0.5004**

Table 6: F_1 results of varying approaches with and without BERT on $Test_{Bias}$ and $Test_{NoBias}$, focusing on emotion cause pair extraction.

We implement MGSAG with the pre-trained BERT (Devlin et al., 2019) to explore the effect of pre-trained language model, where we use the

Model	Emotion Ext.		
	P	R	F_1
ECPE-2D	0.8512	0.8220	0.8358
PairGCN	0.8587	0.7208	0.7829
RankCP	0.8703	0.8406	**0.8548**
ECPE-MLL	0.8582	**0.8429**	0.8500
MGSAG	**0.8721**	0.7911	0.8287
ECPE-2D (BERT)	0.8627	**0.9221**	0.8910
PairGCN (BERT)	0.8857	0.7958	0.8375
RankCP (BERT)	0.9123	0.8999	**0.9054**
ECPE-MLL (BERT)	0.8608	0.9191	0.8886
MGSAG (BERT)	**0.9208**	0.9211	0.8717

Model	Cause Ext.		
	P	R	F_1
ECPE-2D	0.7272	0.6298	0.6738
PairGCN	0.7283	0.5953	0.6541
RankCP	0.6927	**0.6743**	0.6824
ECPE-MLL	0.7248	0.6702	0.6950
MGSAG	**0.7510**	0.6713	**0.7080**
ECPE-2D (BERT)	0.7336	0.6934	0.7123
PairGCN (BERT)	0.7907	0.6928	0.7375
RankCP (BERT)	0.7461	0.7788	0.7615
ECPE-MLL (BERT)	0.7382	**0.7912**	0.7630
MGSAG (BERT)	**0.7979**	0.7468	**0.7712**

Model	Emotion Cause Pair Ext.		
	P	R	F_1
ECPE-2D	0.6960	0.6118	0.6496
PairGCN	0.6999	0.5779	0.6321
RankCP	0.6698	**0.6546**	0.6610
ECPE-MLL	0.7090	0.6441	0.6740
MGSAG	**0.7243**	0.6507	**0.6846**
ECPE-2D (BERT)	0.7292	0.6544	0.6889
PairGCN (BERT)	0.7692	0.6791	0.7202
RankCP (BERT)	0.7119	**0.7630**	0.7360
ECPE-MLL (BERT)	0.7700	0.7235	0.7452
MGSAG (BERT)	**0.7743**	0.7321	**0.7521**

Table 7: Comparison of varying approaches with and without BERT on the original test set ($Test_{all}$).

base Chinese model[3]. We replace the word-level encoder with the [CLS] embeddings of a clause which is obtained by BERT. Results on $Test_{Bias}$ and $Test_{NoBias}$ with and without BERT are shown in Table 6. Results on the original test set with and without BERT are shown in Table 7.

During the training process, we use the Adam (Kingma and Ba, 2015) optimizer to update all parameters. The mini-batch size with BERT is set to 2. The learning rate with BERT is set to 1e-5.

As shown in Table 7, methods with BERT perform better than those without BERT on the origi-

---

[3]We download the pre-trained model from this link: https://s3.amazonaws.com/models.huggingface.co/bert/bert-base-chinese.tar.gz

nal test set, which shows the effectiveness of utilizing the pre-trained BERT. As shown in Table 6, results of models with BERT on $Test_{Bias}$ and $Test_{NoBias}$ indicate that using BERT as the encoder cannot make up for the deficiency caused by position bias. MGSAG still outperforms other methods on $Test_{all}$ and $Test_{NoBias}$. The results verify the effectiveness of capturing the causal semantic relations between clauses via fine-grained and coarse-grained semantics encoding.