

# Bidirectional Hierarchical Attention Networks based on Document-level Context for Emotion Cause Extraction

Guimin Hu, Guangming Lu, Yi Zhao\*

Harbin Institute of Technology (Shenzhen), China

rice.hu.x@gmail.com, {luguangm, zhao.yi}@hit.edu.cn

## Abstract

Emotion cause extraction (ECE) aims to extract the causes behind the certain emotion in text. Some works related to the ECE task have been published and attracted lots of attention in recent years. However, these methods neglect two major issues: 1) pay few attentions to the effect of document-level context information on ECE, and 2) lack of sufficient exploration for how to effectively use the annotated emotion clause. For the first issue, we propose a bidirectional hierarchical attention network (BHA) corresponding to the specified candidate cause clause to capture the document-level context in a structured and dynamic manner. For the second issue, we design an emotional filtering module (EF) for each layer of the graph attention network, which calculates a gate score based on the emotion clause to filter the irrelevant information. Combining the BHA and EF, the EF-BHA can dynamically aggregate the contextual information from two directions and filters irrelevant information. The experimental results demonstrate that EF-BHA achieves the competitive performances on two public datasets in different languages (Chinese and English). Moreover, we quantify the effect of context on emotion cause extraction and provide the visualization of the interactions between candidate cause clauses and contexts.

## 1 Introduction

In recent years, emotion cause extraction (ECE), which aims to identify the causes with respect to certain emotion in the text (Li et al., 2018b; Ding et al., 2019; Xia et al., 2019), has obtained increasing attention in academics and industry. Mining the causes of certain emotion has a wide range of applications, such as public opinion monitoring and product review mining (Wang et al., 2016; Tang et al., 2016; Ma et al., 2017; Xia et al., 2019; Phan

and Ogunbona, 2020; Ding et al., 2020). The goal of ECE is to find out the cause clause (e.g.,  $c_3$ ) that contains the emotion cause for the given emotion clause (e.g.,  $c_4$ ), as presented in Example 1.

**Example 1** ( $c_1$ ) *Wu was diagnosed with advanced liver cancer at the beginning of 2014* ( $c_2$ ) *since he began to update his health condition in Microblog* ( $c_3$ ) *If Wu didn't update his microblog for a long time* ( $c_4$ ) *people worried that he may have passed away*

We divide the existing works related to ECE into rule-based methods, traditional machine learning algorithms and deep learning methods. The rule-based methods are dependent on linguistic rules and common-sense knowledge, which requires plenty of manual operations (Lee et al., 2010; Russo et al., 2011). The traditional machine learning algorithms generally rely on feature engineering to manually select the features as the inputs of model (Chen et al., 2010; Gui et al., 2016). Recently, a number of works adopted deep neural networks like self-attention (Xia et al., 2019), co-attention (Li et al., 2018b), hierarchical attention (Gui et al., 2017; Ding et al., 2019) to capture the relations among clauses. Some works utilized the multi-task learning (Xia and Ding, 2019; Hu et al., 2020a; Chen et al., 2018) to extract emotion cause clauses.

For Example 1, if the contextual clauses (e.g.,  $c_1$  and  $c_2$ ) are ignored, there may be no direct causal relationship between the emotion clause (e.g.,  $c_4$ ) and the cause clause (e.g.,  $c_3$ ), since not updating one's social media account will not cause worried. In fact, context has been utilized in many causal relation tasks to provide semantic information and improve the model performance (Kruengkrai et al., 2017; Li and Mao, 2019; Sridhar and Getoor, 2019; Kayesh et al., 2019). Chen et al. (2020) has mentioned that the causal relationship between the emotion and cause clauses may only be valid in a specific context. However, few works related to emo-

\*Corresponding author.

tion cause extraction consider the document-level context information. To determine whether a clause is the cause of a certain emotion clause, it actually requires to understand the entire document. In this paper, we take each clause in the document as the candidate cause clause of the given emotion, and propose a bidirectional hierarchical attention networks (BHA) to capture the document-level context for a specific candidate cause clause in a structured and dynamic manner. The “bidirectional” denotes the document-level context is divided into forward and backward context based on the position of current candidate cause clause in document, and “hierarchical” denotes that we use hierarchical attention networks to selectively focus on the context information related to the current candidate cause clause at word and clause levels. In contrast to the hierarchical attention networks (Werlen et al., 2018) and self attention model (Vaswani et al., 2017), the BHA allows dynamically access to the context from two directions and distinguishes the effects of forward and backward context on candidate cause clause. Moreover, the previous works are generally limited to calculate the relative position of candidate cause clause to the emotion clause for leveraging emotion clause (Ding et al., 2019), **ignoring the role of emotion clause in filtering the irrelevant information**. So we design an emotional filtering module (EF), which computes a gate score for each layer of graph attention networks (GAT) based on the emotion clause to filter the irrelevant information. Combining the BHA and EF, we propose a bidirectional hierarchical attention networks (EF-BHA) with emotional filtering to appropriately encode contextual features into the clause representation for emotion cause extraction.

The main contributions of this paper are summarized as follows:

1. Different from the hierarchical attention networks and self-attention mechanism, the proposed bidirectional hierarchical attention (BHA) dynamically integrates the forward and backward contexts related to the specified candidate cause clause into the clause representation in a multi-granularity way.
2. We design an emotional filtering module (EF) for each layer of graph attention networks, which calculates a gate score based on the emotion clause to filter the irrelevant information.

3. Experimental results on two public datasets in different languages (Chinese and English) demonstrate that the effectiveness of EF-BHA and further provide the visualization of the interactions between candidate cause clauses and contexts.

## 2 Related Work

Lee et al. (2010) firstly gave the definition of emotion cause extraction (ECE), and manually constructed a corpus from the Academia Sinica Balanced Chinese Corpus. Based on this corpus, a multi-label approach was proposed with the basic of linguistic features as cues (Chen et al., 2010). Russo et al. (2011) proposed an approach that automatically identified linguistic contexts. However, taking the word as labeling granularity of ECE brings some drawbacks including incompleteness in meaning and analysis difficulties. To overcome these shortcomings, Gui et al. (2016) released a clause-level Chinese emotion cause corpus and proposed an event-driven multi-kernel SVM model for this corpus.

Recently, a number of works adopted deep learning networks to solve ECE task. Based on the original memory network (Sukhbaatar et al., 2015), Gui et al. (2017) proposed to store local context of each word in different memory slots to extract emotion cause clause. Chen et al. (2018) captured the interactions between emotion classification and cause detection in an end-to-end fashion. Li et al. (2018b) built the co-attention relationship between emotion clause and each candidate clause with emotional context. Ding et al. (2019) viewed ECE as a reordered prediction problem and incorporated the dynamic global labels into the model. Xia et al. (2019) firstly employed Transformer (Vaswani et al., 2017) to encode the global level information on all the clauses rather than relying solely on the hidden state of one clause. Hu et al. (2020b) proposed a graph convolutional structure with fusion of semantics and structural constraints (FSS-GCNs) to automatically learned how to selectively attend the relevant clauses useful for emotion cause extraction.

However, we notice that these methods related to ECE task ignore two major issues: 1) pay few attentions to the effects of document-level context on ECE task. 2) lack of sufficient exploration about how to effectively use the annotated emotion clause. In this paper, we try to incorporate the document-level context into the ECE task and use the emotion

clause to filter the irrelevant contextual information simultaneously.

### 3 Methodology

#### 3.1 Task Definition

Given a document  $d = \{c_1, c_2, \dots, c_n\}$ ,  $c_i = \{w_{i1}, w_{i2}, \dots, w_{im}\}$  contains  $m$  words, where  $w_{ij}$  is the  $j$ -th word in the clause  $c_i$ . Each document contains an emotion clause and one or more corresponding emotion cause clauses, and each clause is annotated with label  $\in \{0, 1\}$ , where label “1” denotes the clause is a cause clause. We formalize the ECE task as a binary classification problem, and our goal is to determine which clauses contain the emotion cause according to the given emotion clause.

#### 3.2 Overall Architecture

The overall architecture of EF-BHA is shown in Figure 1. It contains BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), word-level context attention module (WCA), GATs with emotional filtering (EF-GATs), clause-level context attention module (CCA) and context aggregation module (CAM). BERT is used to encode hidden states of word and clause (see Section 3.3). The word-level context attention module is used to extract those words that are related to candidate cause clauses from forward and backward contexts and then integrate them to generate high-level context information (see Section 3.4). The GATs with emotional filtering is the modified version of graph attention networks, which aims to filter the irrelevant information to the emotion in each layer of GATs and capture the inter-clause dependency simultaneously (see Section 3.5). Following the GATs, the clause-level context attention module is used to capture those clauses that related to the candidate cause clause and then summarize them into two vectors to represent the forward context and backward context, respectively (Section 3.6). After obtaining the context representation in both direction, we use the context aggregation module to encode contextual information into clause representation (Section 3.7).

#### 3.3 BERT Encoder

Following the idea of the RANKCP (Wei et al., 2020), each clause in the document is processed into the sequence that takes  $[CLS]$  as the start token and  $[SEP]$  as the end token, and we concate-

nate these sequences as the input of BERT encoder, where  $[CLS]$  is a special token that aggregates the sequence features as the hidden state of clause and  $[SEP]$  is a dummy token not used for this task. We take the hidden state  $h_{ij}$  in the BERT model as the representation of  $w_{ij}$  and the hidden state of  $[CLS]$  as the vector of the clause.

#### 3.4 Word-level Context Attention

For the candidate cause clause  $c_i$ , its document-level context can be divided into forward context  $c_i^a = \{c_1, c_2, \dots, c_{i-1}\}$  and backward context  $c_i^b = \{c_{i+1}, c_{i+2}, \dots, c_n\}$ . The forward context and the backward context do not include the current candidate cause clause as the context aggregation module will take it as one of the features. The word-level context attention summarizes information for each clause of the bidirectional context, e.g., clauses  $c_j \in c_i^a$  and  $c_r \in c_i^b$ , which uses the *Multi\_Attention* proposed by Vaswani et al. (2017) to capture different types of relations between the words in  $\{c_j, c_r\}$  and  $c_i$ . The key and value are both the word-level hidden states of BERT encoder and then aggregate those words to hidden states  $x_j$  and  $x_r$ , respectively:

$$q_i^w = f_w(h_{iCLS}) \quad (1)$$

$$x_j = Multi\_Attention(q_i^w, h_{jk}) \quad (2)$$

$$x_r = Multi\_Attention(q_i^w, h_{rs}) \quad (3)$$

where  $q_i^w$  is the word-level query representation and  $f_w$  is a linear transformation with the parameters randomly initialized.  $h_{jk}$  and  $h_{rs}$  denote the hidden states of  $w_{jk}$  and  $w_{rs}$  respectively. Especially,  $h_{iCLS}$  is the hidden state of  $i$ -th token  $[CLS]$ , which is the representation of  $c_i$  encoded by the BERT encoder. Similarly, we can obtain the representation  $x_e$  for emotion clause after the word-level context attention module.

#### 3.5 Graph Attention Network with Emotional Filtering

We employ the graph attention networks (GATs) (Velickovic et al., 2018) to capture the inter-clause dependency, in which each clause can be viewed as a node in the graph and every two nodes have an edge to represent the relation between nodes. To retain self-information, each node adds a self-loop edge. 1-layer GATs encodes only information about immediate neighbors, while GATs with stacking  $L$  graph attention layers aggregates the  $L$ -order

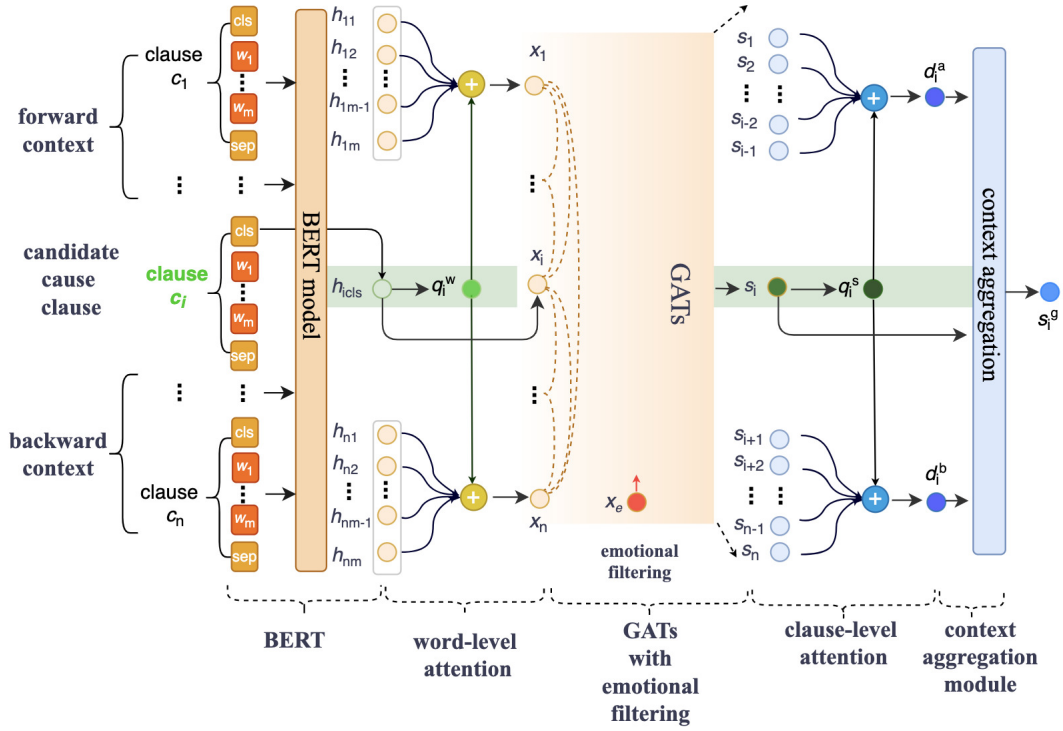


Figure 1: The architecture of EF-BHA for the candidate emotion clause  $c_i$ .

neighbor nodes.  $h_i^l$  denotes the representation of clause  $c_i$  after  $l$ -layer GATs:

$$h_i^l = \text{Relu}\left(\sum_{j \in N_i} \alpha_{ij}^l W^l h_j^{l-1} + b^l\right) \quad (4)$$

where  $\alpha_{ij}^l = f(v_l^T \tanh([W_a^l h_i^{l-1}, W_b^l h_j^{l-1}]))$  is the attention weight of clauses  $c_j$  to  $c_i$  with an MLP parameterized by  $\{v_l, W_a^l, W_b^l\}$ , and  $f$  is the normalization function *Softmax*.  $N_i$  denotes the neighbor clauses of  $c_i$ ,  $\{W^l, b^l\}$  are learnable parameters. The node  $c_j$ 's initial hidden state  $h_j^1 = x_j$ , which is encoded by the word-level context attention.

To ensure that the information transmitted between nodes is related to the annotated emotion, we modify the original GAT by adding an emotional filtering module in each layer of GATs to filter the irrelevant information.  $g^{l-1} \in \mathbb{R}^d$  denotes the gate score of  $(l-1)$ -th layer of GATs, where  $d$  is the dimension of node representation. We apply  $g^{l-1}$  over the hidden vector  $h_j^{l-1}$  via the element-wise multiplication operation  $\circ$ . The aggregation of node representation  $h_i^l$  after emotional filtering is given by:

$$g^{l-1} = \sigma(W_g^{l-1} x_e) \quad (5)$$

$$\hat{h}_i^l = \text{Relu}\left(\sum_{j \in N_i} \alpha_{ij} W^l (g^{l-1} \circ h_j^{l-1}) + b^l\right) \quad (6)$$

where  $x_e$  is the representation of emotion clause encoded by the word-level context attention module,  $\sigma(\cdot)$  is the sigmoid function and  $W_g^{l-1} \in \mathbb{R}^{d \times d}$  is a learnable matrix.

Considering that each emotional filtering module may retain different aspects of node representation, we concatenate these node representations generated in the previous  $L$  layers of GATs as  $m_i^L = [\hat{h}_i^1, \hat{h}_i^2, \dots, \hat{h}_i^L]$ . We transform  $m_i^L$  into  $s_i \in \mathbb{R}^d$  through a non-linear transformation  $s_i = \sigma(W m_i^L)$  parameterized by  $W \in \mathbb{R}^{d \times (L \times d)}$ .

### 3.6 Clause-level Context Attention

Similar to the word-level context attention module, the clause-level context attention summarizes the forward context and backward context into  $d_i^a$  and  $d_i^b$  based on the interactions between the candidate cause clause  $c_i$  and contexts, respectively:

$$q_i^s = f_s(s_i) \quad (7)$$

$$d_i^a = \text{Multi\_Attention}(q_i^s, s_j) \quad (8)$$

$$d_i^b = \text{Multi\_Attention}(q_i^s, s_r) \quad (9)$$

where  $f_s$  is a linear transformation to obtain clause-level query representation  $q_i^s$ . The key and value are both the expressions of clauses based on WCA. In particular,  $s_j$  and  $s_r$  denote the representations of  $c_j \in c_i^a$  and  $c_r \in c_i^b$ , respectively.



### 3.7 Context Aggregation Module

$d_i^a$  and  $d_i^b$  represent the aggregation of forward context and backward context, respectively. In order to selectively incorporate context information into the representation of candidate cause clause  $c_i$ , we use two scores less than 1, i.e.  $\lambda_a$  and  $\lambda_b$ , to control which information of forward context and backward context can flow to the final clause representation. Following this idea, we can have the following aggregation:

$$s_i^g = (1 - \lambda_a - \lambda_b) * s_i + \lambda_a * d_i^a + \lambda_b * d_i^b \quad (10)$$

where  $\lambda_a = \sigma(W_s^a s_i + W_d^a d_i^a)$  and  $\lambda_b = \sigma(W_s^b s_i + W_d^b d_i^b)$  denote the gate scores for  $d_i^a$  and  $d_i^b$ , respectively.  $\{W_s^a, W_s^b, W_d^a, W_d^b\}$  are learnable weight matrices. With this aggregation module, the model can handle forward context and backward context to capture the specific contextual information required by the candidate cause clause. We take  $s_i^g$  as the final feature for emotion cause prediction:

$$\hat{p}_i = f(W_p s_i^g + b_p) \quad (11)$$

where  $f$  is the *Softmax* function, and  $\{W_p, b_p\}$  are the learnable parameters. Specifically, the model is trained by using standard gradient descent algorithm with the cross-entropy loss, which is given as:

$$J(\Theta) = \frac{1}{N} \sum_k \sum_i^n -[p_i \log \hat{p}_i + (1 - p_i)(1 - \log \hat{p}_i)] \quad (12)$$

where  $N$  is the number of training instances and  $p_i$  is the real distribution.

## 4 Experiments

### 4.1 Dataset and Metrics

The proposed model is evaluated on two public datasets: a Chinese public benchmark dataset (Chi dataset) (Gui et al., 2016) and an English Dataset (Eng dataset) (Gao et al., 2017). Chi dataset is collected from SINA city news and Eng dataset is collected from an English novel. Note that each document of both datasets contains only one emotion clause and one or more emotion cause clauses corresponding to it. We adopt the same experimental setting as RTHN (Xia et al., 2019), that is, we use 10-fold cross validation to conduct experiments with 9 folds as training data and remaining 1 fold as testing data. Table 1 gives the details about the two datasets. We repeat the experiments 20 times

to report the average result and perform one sample t-test on the experimental results. We adopt the precision (P), recall (R) and F1 score (F1) as the metrics for evaluation, which are defined as:

$$P = \frac{n_{cc}}{n_{pc}}, R = \frac{n_{cc}}{n_{gc}}, F1 = \frac{2 \times P \times R}{P + R}, \quad (13)$$

where  $n_{cc}$ ,  $n_{pc}$  and  $n_{gc}$  denote the correctly predicted causes, predicted causes and the ground-truth causes respectively.

Range	Number	Percentage(%)
Chi dataset:		
one cause	2046	97.2
two and more causes	59	2.8
All	2105	100
Eng dataset:		
one cause	1949	90.4
two and more causes	196	9.1
All	2156	100

Table 1: The proportion of documents with different number of emotion causes.

### 4.2 Implementation Details

For the Chi dataset, we implement the EF-BHA based on the BERT that is initialized using BERT-Base, Chinese<sup>1</sup>. We use AdamW optimizer (Loshchilov and Hutter, 2019) and 20 epochs with early stopping to optimize the model. We set the batch size and learning rate to 4 and  $1e - 5$  respectively, and apply a scheduler to adjust the learning rate, that is, the first 10% of all training steps is linear warmup phrase, and then linear decay phrase. Furthermore, we set the weight decay for BERT model and downstream model to 0.01 and  $2e - 5$ , respectively. The proposed EF-BHA achieves the best performance when GATs adopts 2-layer emotional filtering, where the first layer has one attention head and the second layer has four attention heads. For Eng dataset, we implement the proposed EF-BHA based on the BERT that is initialized using BERT-Base, English<sup>2</sup>. The model using  $2e - 5$  learning rate with 15 epochs achieves the best performance on the Eng dataset when the GATs adopts single-layer emotional filtering.

### 4.3 Baseline Methods

We compare the proposed EF-BHA with the existing methods. We summarize the baseline methods

<sup>1</sup><https://github.com/google-research/bert>

<sup>2</sup>[github.com/huggingface/transformers](https://github.com/huggingface/transformers)

	P	R	F1
RB	0.6747	0.4287	0.5243
CB	0.2672	0.7130	0.3887
Word2vec	0.4301	0.4233	0.4136
SVM	0.4200	0.4375	0.4285
Multi-kernel	0.6588	0.6972	0.6752
Memnet	0.5922	0.6354	0.6134
ConvMS-Memnet	0.7076	0.6838	0.6953
CANN	0.7721	0.6891	0.7266
PAE-DGL	0.7619	0.6908	0.7242
RTHN	0.7697	0.7662	0.7677
MANN	0.7843	0.7587	0.7706
FSS-GCN	0.7861	0.7572	0.7714
RHNN	0.8112	0.7725	<b>0.7914</b>
EF-BHA	0.7938	<b>0.7808</b>	0.7868

Table 2: Comparison of the proposed EF-BHA with existing methods using precision, recall, and F1 score as metrics on Chi dataset. These results are reprinted from the corresponding publications.

into the following groups:

- Rule-based and commonsense-based methods: **RB** is a traditional rule-based method based on linguistic rules (Lee et al., 2010). **CB** is a commonsense-based method proposed by Russo et al. (2011).
- Machine learning methods: **SVM** uses common facts and unigrams, bigrams and trigrams as the features to train an SVM (Cortes and Vapnik, 1995) classifier (Chen et al., 2010). **Multi-kernel** is a method using multi-kernel convolution to learn the relations between emotion cause clause and events (Gui et al., 2016). **Word2vec** uses the word representations pre-trained by Word2vec (Mikolov et al., 2013) to train SVM as classifier.
- Deep learning method: **Memnet** uses the memory networks (Sukhbaatar et al., 2015) to capture the mutual impacts between emotion word and emotion causes (Gui et al., 2017). **ConvMS-Memnet** stores relevant contexts in different memory slots with convolution operation (Gui et al., 2017). **CANN** builds the co-attention interaction between emotion and each candidate clause (Li et al., 2018b). **PAE-DGL** take ECPE task as a reordered prediction problem (Ding et al., 2019) to extract emotion causes. **RTHN** uses Transformer (Vaswani et al., 2017) to capture the relation

	P	R	F1
Word2vec	0.1651	0.8673	0.2774
SVM	0.2757	0.6416	0.3856
ConvMS-Memnet	0.4605	0.4177	0.4381
MANN	0.7933	0.4081	0.5328
FSS-GCN	0.6743	0.5303	0.5948
RHNN	0.6901	0.5267	0.5975
EF-BHA	0.7277	0.5305	<b>0.6137</b>

Table 3: Comparison of the proposed EF-BHA with existing results that are implemented in Li et al. (2019b) on Eng dataset.

among the clauses (Xia et al., 2019). **MANN** uses the multi-attention mechanism and CNN layer to extract critical features from the text (Li et al., 2019b). **FSS-GCN** learns how to selectively focus on the relevant clauses by fusing the semantics and structural information (Hu et al., 2020b). **RHNN** adopts the hierarchical attention and knowledge-based regularization to extract emotion cause (Fan et al., 2019).

#### 4.4 Main Results

The experimental results on Chi and Eng datasets are shown in Table 2 and Table 3, respectively. We firstly focus on the Table 2. It can be observed that EF-BHA is better than most competitive baselines. RB and CB seem difficult to achieve a balance between precision and recall. For traditional machine learning, Word2vec and SVM cannot achieve better results in accuracy and recall. The recent works, such as CANN, RTHN and FSS-GCN, model the relations among clauses to incorporate more information, and obtain significant improvements. The F1 score of EF-BHA is at least 1.6% higher than those of these works, which is close to the state-of-the-art method RHNN. This improvement is significant with p-value less than 0.01 in one sample t-test. The reason for the improvement is that EF-BHA dynamically incorporates the document-level context information according to the candidate cause clause for emotion cause extraction. Next, we focus on the Table 3. The performances of the existing methods on Eng datasets are generally low. We can find that the proposed EF-BHA achieves the best performance on Eng dataset, outperforming the state-of-the-art (RHNN) method by 1.62% in F1 measure. These results illustrate that the proposed EF-BHA can better encode clause representation by effectively attending to those words or

	P	R	F1
WCA	0.7843	0.7587	0.7701
CCA+CAM	0.7224	0.7298	0.7218
WCA+CCA+CAM	0.7171	0.7135	0.7144
EF	0.7637	0.7458	0.7530
EF-GATs	0.7367	0.7213	0.7314
EF-BHA	0.7938	0.7808	0.7868

Table 4: An ablation study for EF-BHA model.

clauses related to current candidate cause clause in the bidirectional hierarchical texts, so as to obtain improvements.

## 4.5 Analysis and Discussion

### 4.5.1 Ablation Study

We conduct an ablation study by removing each module or the combination of module separately to verify the effect of each module in EF-BHA. Experimental results are shown in Table 4. Removing the WCA bring degradation on F1 scores by 1.67%. If we only remove the EF, that is, the original GATs is used to model the inter-clause dependency, the performance will be impaired (over 3% drops in F1 score), which illustrates that emotional filtering module can help encode the clause representation better. When EF and GATs are removed from the model, the F1 score drops to 0.7314, which may be caused by the insufficiency of modeling the relation of inter-clause. Additionally, we remove the combination of modules, such as CCA+CAM, and then the F1 score declines to 72.18%. We also eliminate WCA+CCA+CAM, which leads to a poor F1 score of 71.44%. These experimental results show that these modules and their mining features are significant in the emotion cause extraction.

### 4.5.2 Re-evaluating ECE Models

Ding and Kejriwal (2020) pointed out that some existing deep neural networks make use of the bias in the benchmark to achieve better performance, where the bias denotes the location imbalance distribution phenomenon of emotional cause location as most of the cause clauses appear near the emotion clause. To verify the ability of EF-BHA to understand the actual context, we conduct the experiments on “de-bias” dataset (Ding and Kejriwal, 2020) and the results are shown in Table 5. The previous works like RTHN and PAE-GDL usually leverage the position bias in the benchmark dataset to obtain improvement and they are position-aware

	P	R	F1
PAE	0.5511	0.3078	0.3851
PAEDGL	0.5525	0.3279	0.4096
RTHN	0.5467	0.5466	0.5445
EF-BHA	0.5640	0.6549	0.6061

Table 5: Comparison of the proposed methods with existing results that are implemented in Ding and Kejriwal (2020) on ‘de-bias’ dataset.

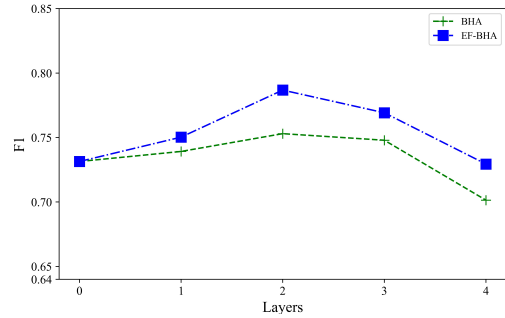


Figure 2: Comparison of BHA and EF-BHA with different graph attention layers.

models. Different from RTHN, PAE-GDL, EF-BHA is position-insensitive model, which attends to understand actual context instead of depending on the dataset bias. Compared with the typical methods, the F1 score of EF-BHA is still higher than PAE, PAEDGL and RTHN although the F1 scores about this dataset is relatively low. We speculate that due to the small size of “de-bias” dataset, the problem of over fitting may occur in the parametric deep networks based on BERT model.

### 4.5.3 Performance against GAT layers

We vary the number of graph attention layer (ranging from 0 to 4) to verify its effect on EF-BHA and BHA, and the results are shown in Figure 2. When removing the graph attention layer, the performance of EF-BHA drops a lot (0.7314 in F1 score). In this case, the F1 scores of EF-BHA is same as that of BHA as the emotional filtering performs on each graph attention layer. When using 2-layer graph attention networks, the proposed EF-BHA achieves the best performance, and more layers result in the deterioration of model performance. This finding just confirms the conclusive results that stacking more layers in a graph neural network (GNN) could lead to over smoothing and finally the features of graph vertices converge to the same value (Li et al., 2018a, 2019a).

$c_1$  to save the woman as soon as possible  
 $c_2$  the commander worked out a rescue plan  
 $c_3$  the first group laid life-saving air cushion  
 $c_4$  and evacuated irrelevant people around  
 $c_5$  another group climbed up to the sixth floor  
 $c_6$  persuading women in the building  
 $c_7$  in the process of persuasion  
 $c_8$  fire officers and soldiers understand  
 $c_9$  due to the other party's project arrears  
 $c_{10}$  her family is badly in need of money  
 $c_{11}$  she lives a stressful life  
 $c_{12}$  she helplessly intends to jump off a building to commit suicide

Table 6: Visualization results of word-level context attention module for the candidate cause clause  $c_4$ .

#### 4.6 Case Study

**Example 2** ( $c_1$ ) 为尽快将女子救下 ( $c_2$ ) 指挥员制订了救援方案 ( $c_3$ ) 第一组在楼下铺设救生气垫 ( $c_4$ ) 并对周围无关人员进行疏散 ( $c_5$ ) 另一组队员爬上6楼 ( $c_6$ ) 在楼内对女子进行劝说 ( $c_7$ ) 劝说过程中 ( $c_8$ ) 消防官兵了解到 ( $c_9$ ) 该女子是由于对方拖欠工程款 ( $c_{10}$ ) 家中又急需用钱 ( $c_{11}$ ) 生活压力大 ( $c_{12}$ ) 无奈才选择跳楼轻生

*Translation:* ( $c_1$ ) to save the woman as soon as possible ( $c_2$ ) the commander worked out a rescue plan ( $c_3$ ) the first group laid life-saving air cushion ( $c_4$ ) and evacuated irrelevant people around ( $c_5$ ) another group climbed up to the sixth floor ( $c_6$ ) persuading women in the building ( $c_7$ ) in the process of persuasion ( $c_8$ ) fire officers and soldiers understand ( $c_9$ ) due to the other party's project arrears ( $c_{10}$ ) her family is badly in need of money ( $c_{11}$ ) she lives a stressful life ( $c_{12}$ ) she helplessly intends to jump off a building to commit suicide

To deeper understand the bidirectional context attention networks, we choose one document (Example 2) from the Chi dataset to visualize the attention weights extracted from the word-level attention module using sequence labeling toolkit (Yang and Zhang, 2018). For Example 2,  $c_{12}$  is the emotion clause containing the emotion “helplessly”, and  $\{c_9, c_{10}, c_{11}\}$  are the corresponding cause clauses.

Here, the attention distribution refers to the importance of words to the current candidate cause clause. Note that the intensity of color is proportional to the weight value (that is, dark color means large weight). Table 6 and Table 7 show the visualizations about the attention weights of bidirectional context under the candidate cause clauses  $c_4$

$c_1$  to save the woman as soon as possible  
 $c_2$  the commander worked out a rescue plan  
 $c_3$  the first group laid life-saving air cushion  
 $c_4$  and evacuated irrelevant people around  
 $c_5$  another group climbed up to the sixth floor  
 $c_6$  persuading women in the building  
 $c_7$  in the process of persuasion  
 $c_8$  fire officers and soldiers understand  
 $c_9$  due to the other party's project arrears  
 $c_{10}$  her family is badly in need of money  
 $c_{11}$  she lives a stressful life  
 $c_{12}$  she helplessly intends to jump off a building to commit suicide

Table 7: Visualization results of word-level context attention module for the candidate cause clause  $c_9$ .

and  $c_9$ , respectively. In the comparison of Table 6 and Table 7, it can be obviously observed that the larger attention weights (the red color blocks) focus on the different parts of text. Specifically, when candidate cause clause is  $c_4$ , the words with larger attention weights are mainly concentrated in the contents above  $c_4$ , i.e., clauses  $\{c_1, c_2, c_3\}$ , which indicates the words in these clauses is more related to  $c_4$ . When candidate cause clause is  $c_9$ , the words with larger attention weights are mainly concentrated in the contents below  $c_9$ , i.e., clauses  $\{c_{10}, c_{11}, c_{12}\}$ , which indicates the words below  $c_9$  is more related to  $c_9$ . Moreover, those words related to clause  $c_9$  are usually negative, such as “stressful”, “badly” and “suicide”. This finding may reflect that the emotion triggered by  $c_9$  is negative if  $c_9$  is the cause clause, which corresponds to the emotion “helplessly” in  $c_{12}$ . These results illustrate the proposed BHA can dynamically capture the contextual information from two directions to distinguish the effects of bidirectional context on the current candidate cause clause. Through the visualization of word-level attention weights, we quantify the effect of the interaction between the clause and the contexts at the word level.

## 5 Conclusions and Future Work

In this work, we propose the EF-BHA to model the relations between candidate cause clauses and the contexts. Especially, EF-BHA extracts the relevant contextual information according to the candidate cause clause, and then summarize contextual information with different granularity into clause representation. Moreover, we propose to add an



emotional filtering module for each layer of GATs to filter the irrelevant information. The experimental results on two public datasets demonstrate that the proposed EF-BHA achieves competitive performance in comparison with the existing methods, thereby validating the effectiveness. We further visualize the attention weight extracted by the bidirectional hierarchical context attention module, aiming to provide the visualization of the interactions between candidate cause clauses and contexts. EF-BHA may introduce some irrelevant information when integrating context information. **In future work, we will only use the previous  $|w|$  clauses and the following  $|w|$  clauses of the candidate cause clause as the context to alleviate this problem.**

## 6 Acknowledgement

This work was supported by the Innovative Research Project of Shenzhen under Grant No. KQJSCX20180328165509766 and the Natural Science Foundation of Guangdong, China under Grant No. 2020A1515010812 and 2021A1515011594.

## References

- Xinhong Chen, Qing Li, and Jianping Wang. 2020. [Conditional causal relationships between emotions and causes in texts](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3111–3121.
- Ying Chen, Wenjun Hou, Xiyao Cheng, and Shoushan Li. 2018. [Joint learning for emotion classification and emotion cause detection](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 646–651.
- Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Churen Huang. 2010. [Emotion cause detection with linguistic constructions](#). In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 179–187.
- Corinna Cortes and Vladimir Vapnik. 1995. [Support-vector networks](#). *Mach. Learn.*, 20(3):273–297.
- 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.
- Jiayuan Ding and Mayank Kejriwal. 2020. [An experimental study of the effects of position bias on emotion cause extraction](#). *CoRR*, abs/2007.15066.
- Zixiang Ding, Huihui He, Mengran Zhang, and Rui Xia. 2019. [From independent prediction to re-ordered prediction: Integrating relative position and global label information to emotion cause identification](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6343–6350.
- Zixiang Ding, Rui Xia, and Jianfei Yu. 2020. [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.
- Chuang Fan, Hongyu Yan, Jiachen Du, Lin Gui, Lidong Bing, Min Yang, Ruifeng Xu, and Ruibin Mao. 2019. [A knowledge regularized hierarchical approach for emotion cause analysis](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5613–5623.
- Qinghong Gao, H Jiannan, X Ruifeng, G Lin, Y He, KF Wong, and Q Lu. 2017. [Overview of ntcir-13 eca task](#). In *Proceedings of the NTCIR-13 Conference*.
- Lin Gui, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du. 2017. [A question answering approach to emotion cause extraction](#). *CoRR*, abs/1708.05482.
- 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.
- Guimin Hu, Guangming Lu, and Yi Zhao. 2020a. [Emotion-cause joint detection: A unified network with dual interaction for emotion cause analysis](#). In *Natural Language Processing and Chinese Computing - 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14-18, 2020, Proceedings, Part I*, pages 568–579.
- Guimin Hu, Guangming Lu, and Yi Zhao. 2020b. [Fss-gcn: A graph convolutional networks with fusion of semantic and structure for emotion cause analysis](#). *Knowledge-Based Systems*, page 106584.
- Humayun Kayesh, Md. Saiful Islam, and Junhu Wang. 2019. [On event causality detection in tweets](#). *CoRR*, abs/1901.03526.

- Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, Julien Kloetzer, Jong-Hoon Oh, and Masahiro Tanaka. 2017. [Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks](#). In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3466–3473.
- Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. [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*, pages 45–53, Los Angeles, CA. Association for Computational Linguistics.
- Guohao Li, Matthias Müller, Ali K. Thabet, and Bernard Ghanem. 2019a. [Deepgcns: Can gcns go as deep as cnns?](#) In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9266–9275.
- Pengfei Li and Kezhi Mao. 2019. [Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts](#). *Expert Syst. Appl.*, 115:512–523.
- Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018a. [Deeper insights into graph convolutional networks for semi-supervised learning](#). In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3538–3545.
- Xiangju Li, Shi Feng, Daling Wang, and Yifei Zhang. 2019b. [Context-aware emotion cause analysis with multi-attention-based neural network](#). *Knowl. Based Syst.*, 174:205–218.
- Xiangju Li, Kaisong Song, Shi Feng, Daling Wang, and Yifei Zhang. 2018b. [A co-attention neural network model for emotion cause analysis with emotional context awareness](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4752–4757.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.
- Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. [Interactive attention networks for aspect-level sentiment classification](#). In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4068–4074.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. [Distributed representations of words and phrases and their compositionality](#). In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.
- Minh Hieu Phan and Philip O. Ogunbona. 2020. [Modelling context and syntactical features for aspect-based sentiment analysis](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3211–3220.
- Irene Russo, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco. 2011. [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@ACL 2011, Portland, OR, USA, June 24, 2011*, pages 153–160.
- Dhanya Sridhar and Lise Getoor. 2019. [Estimating causal effects of tone in online debates](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 1872–1878.
- Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. [End-to-end memory networks](#). In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2440–2448.
- Duyu Tang, Bing Qin, and Ting Liu. 2016. [Aspect level sentiment classification with deep memory network](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 214–224.
- 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, 4-9 December 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*.
- Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. [Attention-based LSTM for aspect-level sentiment classification](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural*

*Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 606–615.

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.

Lesly Miculicich Werlen, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. [Document-level neural machine translation with hierarchical attention networks](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2947–2954.

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.

Rui Xia, Mengran Zhang, and Zixiang Ding. 2019. [RTHN: A rnn-transformer hierarchical network for emotion cause extraction](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5285–5291.

Jie Yang and Yue Zhang. 2018. [NCRF++: an open-source neural sequence labeling toolkit](#). In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations*, pages 74–79.