IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 30, 2021, accepted February 3, 2021, date of publication February 8, 2021, date of current version February 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3057880

A Mutually Auxiliary Multitask Model With Self-Distillation for Emotion-Cause Pair Extraction

JIAXIN YU[1], **WENYUAN LIU**[1,2], **(Member, IEEE), YONGJUN HE**[3], **AND CHUNYUE ZHANG**[4]

[1]School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
[2]Engineering Research Center for Network Perception and Big Data of Hebei Province, Qinhuangdao 066004, China
[3]School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China
[4]Department of Computer Science, Harbin Finance University, Harbin 150036, China

Corresponding authors: Wenyuan Liu (wyliu_ysu@aliyun.com) and Yongjun He (holywit@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672448 and Grant 61673142, in part by the Key Research and Development Project of Hebei Province under Grant 18270307D, and in part by the Natural Science Foundation of Heilongjiang Province of China under Grant JJ2019JQ0013.

ABSTRACT Emotion-cause pair extraction (ECPE), which aims to extract emotions and the corresponding causes in documents, has a wide range of applications in network public opinion analysis. Current two-stage methods first extract emotion and cause clauses, and then pair them. However, there are two problems in these methods: 1) the unidirectional enhancement between emotion and cause extraction fails to make full use of the correlation between them; 2) the errors from the first stage directly degrade the performance of the second stage. To address these problems, we firstly propose a mutually auxiliary multitask model to promote the extraction of emotion and cause clauses by adding two auxiliary tasks which are identical to the original tasks. The proposed model uses the predicted results generated by the two auxiliary tasks as extra features of each other's main tasks, so as to establish the bidirectional correlation between emotion and cause extraction. Secondly, to reduce the influence of error propagation on the second stage, we design a self-distillation method for pairwise tasks to train the proposed model, which further improve the accuracy of emotion and cause extraction. Experimental results on the ECPE benchmark dataset show that the proposed model has achieved good performance on emotion-cause pair extraction, outperforming the baseline models by 1.92% in F1 score.

INDEX TERMS Emotion cause extraction, multitask learning, neural network, self-distillation.

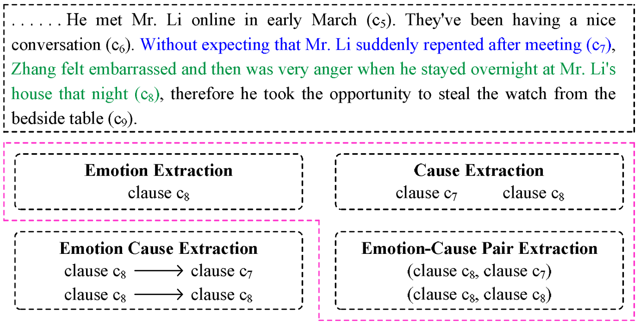
I. INTRODUCTION

Emotion detection has been widely concerned in the field of natural language processing (NLP) and computer vision [1]–[3]. Textual emotion detection which aims at detecting whether a text contains emotion or recognizing the emotion category, plays an important and fundamental role in NLP [4]–[14]. Compared with emotion detection, emotion cause extraction (ECE) has more important application value. This task was first defined as a word-level sequence labeling problem by Lee *et al.* [15]. To make good use of contextual information, the ECE task was redefined as a clause-level classification problem [16], [17]. However, the ECE task needs to annotate emotions before extracting causes, which is very labor-consuming. To overcome this limitation, Xia and Ding [18] put forward a new task named emotion-cause pair extraction (ECPE), and proposed two-stage methods which extract emotion clauses coupled with their cause clauses. As shown in Fig. 1, c_8 is an emotion clause reflecting ‘‘anger''. Since c_7 induces the emotion of c_8, it is identified as a cause clause. The clause c_8, which itself includes the causality, is labeled as both an emotion and a cause clause. Hence, the document contains two emotion-cause pairs, labeled (c_8, c_7) and (c_8, c_8).

However, these two-stage ECPE methods either non-interactively extract emotions and causes or use one of the two tasks to promote another. In most cases, the predicted results of emotion extraction are utilized to facilitate cause extraction. Hence, it is failed for these methods to achieve the mutual promotion between the two tasks. In fact, cause and emotion are interrelated and inseparable. Moreover, the two-stage methods depend heavily on the first-stage model. For example, when c_8 is not extracted correctly in the first stage, extracting (c_8, c_7) and (c_8, c_8) in the following stage is

The associate editor coordinating the review of this manuscript and approving it for publication was Huiling Chen.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/



. . . . . . He met Mr. Li online in early March (c_5). They've been having a nice conversation (c_6). Without expecting that Mr. Li suddenly repented after meeting (c_7), Zhang felt embarrassed and then was very anger when he stayed overnight at Mr. Li's house that night (c_8), therefore he took the opportunity to steal the watch from the bedside table (c_9).

Emotion Extraction
clause c_8
Cause Extraction
clause c_7 clause c_8
Emotion Cause Extraction
clause c_8 $\longrightarrow$ clause c_7
clause c_8 $\longrightarrow$ clause c_8
Emotion-Cause Pair Extraction
(clause c_8, clause c_7)
(clause c_8, clause c_8)

FIGURE 1. An intuitive example of the difference between the emotion extraction, cause extraction, ECE, and ECPE tasks.

necessarily failed. Hence, the errors from the first stage directly degrade the performance of the second stage.

The purpose of this paper is to improve the performance of ECPE by constructing the bidirectional correlation of emotion and cause extraction and training the proposed model by self-distillation method. Specifically, we attempt to solve the above problems with: 1) We design a multitask model to jointly extract emotion and cause clauses. We add two auxiliary tasks which are identical to the original tasks, and then treat the predicted results of the auxiliary tasks as extra clause features of the main tasks. This manner can make emotion and cause extraction benefit from each other, so as to improve the predicted accuracy of them. 2) We design a self-distillation method for pairwise tasks to train our multitask model. In the training process, each pair of tasks in the student model can take extra supervision from the corresponding main task of the teacher model. Through several generations of teacher-student training, the accuracy of emotion and cause extraction is further improved. The impact of error propagation is alleviated.

Our main contributions can be summarized as follows:

- We propose a mutually auxiliary multitask model (MAM) which establishes the bidirectional correlation between emotion and cause extraction and realizes their mutual promotion.
- We design a self-distillation method for pairwise tasks and apply it to train our multitask model, which further improve the accuracy of emotion and cause extraction.
- We evaluate our models by comparative experiments on the benchmark ECPE corpus to demonstrate the improvements of our model.

The rest of this paper is organized as follows. The related work about ECE, ECPE, and knowledge distillation is introduced in Section 2. Our proposed models for ECPE is presented in details in Section 3. Experimental settings and results analysis are provided in Section 4. Finally, the conclusion is made in section 5.

II. RELATED WORK

Emotion analysis is one of the most active research topics in NLP. Here, we focus on two challenging tasks which are ECE and ECPE. Knowledge Distillation (KD) can improve the model performance, however a comprehensive survey of the KD is beyond the scope of this paper. Therefore, we only briefly review some most relevant work to our research.

A. ECE & ECPE

The ECE task is first defined as a word-level sequence labeling problem by Lee *et al.* [15]. Considering the correlation between emotions and cause events, they proposed an emotion cause detection method based on linguistic rules (RB). Also based on RB, Chen *et al.* [19] transformed emotion cause detection into a multi-label classification problem. Russo *et al.* [20] introduced common-sense knowledge into the RB for emotion cause recognition. Gao *et al.* [21], [22] did further research work around the rule-based approach. Since rules cannot cover all language phenomena, some researchers tried to apply machine learning approaches to ECE. Gui *et al.* [23] used the Support Vector Machine (SVM) and Conditional Random Field (CRF) to extract emotional causes. In addition, they also constructed a corpus of emotion causes based on Chinese microblogs. Ghazi *et al.* [24] used the CRF to extract emotional causes, but their method was only applicable to that the emotion and its causes are contained in the same sentence. To utilize discourse information well, Gui *et al.* [16] constructed a new clause-level corpus and adopted multi-kernel SVM to extract emotional causes.

Since its emergence, deep learning has shown persuasive ability in representation learning, so it is used in the ECE task widely. Gui *et al.* [25] treated the ECE as an answer retrieval task and solved it by designing a method based on the Convolutional Neural Network (CNN) and Memory Network. Cheng *et al.* [26] improved the performance of emotion cause extraction by modeling context with the help of long and short-term memory networks (LSTM). A hierarchical CNN model for emotion cause detection was designed in another study by their team [27]. Subsequently, a variety of explorations on hierarchical design were made by researchers. Li *et al.* [28] explored the relevance of word context and used a co-attention neural network to model the representation of clauses. Later, Li *et al.* [29] transformed the computational granularity of attention from words to clauses, and proposed a neural network model built on multi-attention. Yu *et al.* [30] deepened hierarchical representation by introducing phrase-level representation. Xia *et al.* [31] used the Transformer to encode clauses, which achieved excellent performance. In order to overcome the lack of training data, Fan *et al.* [32] and Hu *et al.* [33] coincidentally introduced external emotion knowledge on the basis of hierarchical design to improve the accuracy of the model. In addition, inspired by methods on other NLP tasks, researchers have made a lot of new attempts. For example, Ding *et al.* [34] and Xu *et al.* [35] both transformed the ECE task into a clause ordering problem in the perspective of information retrieval. Learning from machine reading comprehension, Diao *et al.* [36] designed a multi-granularity attention network. Xiao *et al.* [37] regarded the ECE task as a sequence labeling problem, and used multiple attention to obtain multi-view clause representations.

The above ECE methods are all based on the premise of known emotion. Inspired by multitask learning, Chen *et al.* [38] studied jointly learning for emotion classification and emotion cause extraction, and proved the correlation between the two sub-tasks. Recently, Xia *et al.* [18] further redefined this problem and proposed the ECPE task, aiming to extract the emotion and its causes in pairs. Tang *et al.* [39] designed a joint model of emotion detection and emotion-cause pair extraction. In addition, some researchers attempted to transform the ECPE from the sequence classification to other NLP tasks. Song *et al.* [40] regarded pair extraction as a link prediction task and presented an end-to-end multitask model. Wu *et al.* [41] solved the relationship classification task together with emotion and cause extraction in a unified model. Wei *et al.* [42] proposed a one-step approach to emphasize inter-clause modeling from a ranking perspective. Ding *et al.* [43] designed a 2D Transformer and its two variants to model the interaction between emotion-cause pairs. Fan *et al.* [44] proposed a novel method, transforming the relationship classification into the process of constructing directed graphs. These methods all solved multiple tasks in a framework, which achieved good performance. Hence, most recent studies focused on extracting emotion-cause pairs in an end-to-end manner [45]–[53].

B. KNOWLEDGE DISTILLATION

In 2015, Hinton *et al.* [54] first proposed the concept of KD and applied it successfully in deep neural network. They utilized the class probabilities produced by the large model (teacher) as "soft targets" to train the small model (student), realizing the knowledge transfer between them, so that the performance of the student can be as close as possible to or beyond that of the teacher. Furlanello *et al.* [55] proposed a self-distillation method called the Born-Again Network (BAN), which aims not to compress the model, but to train students with the same structure as the teacher. By this way, the students perform significantly better than the teacher in language modeling tasks. Hence, Yang *et al.* [56] utilized the self-distillation to accurately detect the text in the image and optimized the teacher-student training process. Clark *et al.* [57] applied the BAN to multitask learning and validated its effectiveness in other NLP tasks such as textual similarity, textual entailment, and so on.

In summary, without depending on the given emotion annotations, ECPE can extract emotion and cause clauses simultaneously, so it is more preferable. However, the two-stage ECPE can lead to cross-stage error propagation. Since KD has the potential to improve the performance of multitask model, we design a mutually auxiliary multitask model with self-distillation to further improve the model performance of the first stage, and then alleviate the impact of error propagation to the second stage.

III. METHODOLOGY

Similar to the original ECPE methods [18], our work is still a two-stage based method: emotion and cause clauses are extracted in Stage 1, and then paired in Stage 2.

A. PROBLEM DEFINITION

Given a document $d = [c_1, \cdots c_i \cdots, c_m]$ composed of a sequence of clauses, and each clause can be further decomposed into a sequence of words, represented as $c_i = [term_{i,1}, \cdots term_{i,j} \cdots, term_{i,n}]$, where m indicates the number of clauses in the document, and n denotes the length of the word sequence contained in the clause. ECPE aims to extract all emotion-cause pairs $C^{pair} = \{(c_l^{emo}, c_l^{cau})\}_{l=1}^{|C^{pair}|}$. Here, c_l^{emo} and c_l^{cau} represent the emotion clause and cause clause in the l-th emotion-cause pair, respectively.

B. OVERALL ARCHITECTURE

An overview of our proposed model is shown in Fig. 2. Inspired by the multitask learning based on auxiliary task [58], our model contains four sub-tasks which are a pair of tasks for emotion extraction and another pair for cause extraction. These tasks are trained by self-distillation in a framework. In order to model different granularity of language representation, a hierarchical architecture is designed. The bottom is the word encoding layer, which aims to obtain the word representations. The middle layers are the inner-clause and inter-clause encoding layers. They transform the word representations into the contextual clause representations. The top level is the classification layer that predicts whether a clause is an emotion/cause clause or not.

C. EMBEDDING & WORD ENCODING LAYER

In the embedding layer, each word is transformed into a v-dimensional vector. The vectors of all the words in the clause line up an embedding matrix. On the formal definition, the i-th clause in the document can be denoted by the embedding matrix $c_i = [w_{i,1}, \cdots w_{i,j} \cdots, w_{i,n}]$.

The purpose of the word encoding layer is to establish contextualized word representations. Here, the BiLSTM is adopted to encode words. To capture the specific features for emotion and cause, two word-level Bi-LSTMs are performed to generate the emotion-specific and cause-specific clause representations. The hidden states of the two BiLSTMs are respectively as follows:

$$[h_{i,1}^e, \ldots h_{i,j}^e \ldots, h_{i,n}^e] = \mathrm{BiLSTM}_{word}^e([w_{i,1}, \ldots w_{i,j} \ldots, w_{i,n}]) \tag{1}$$

$$[h_{i,1}^c, \ldots h_{i,j}^c \ldots, h_{i,n}^c] = \mathrm{BiLSTM}_{word}^c([w_{i,1}, \ldots w_{i,j} \ldots, w_{i,n}]) \tag{2}$$

where $w_{i,j}$ represents the word vector of the j-th word in the i-th clause. $h_{i,j}^e = [\rightarrow LSTM^e(w_{i,j}); \leftarrow LSTM^e(w_{i,j})]$ is the emotion-specific word representation of $w_{i,j}$, and $h_{i,j}^c$ denotes the cause-specific word representation.

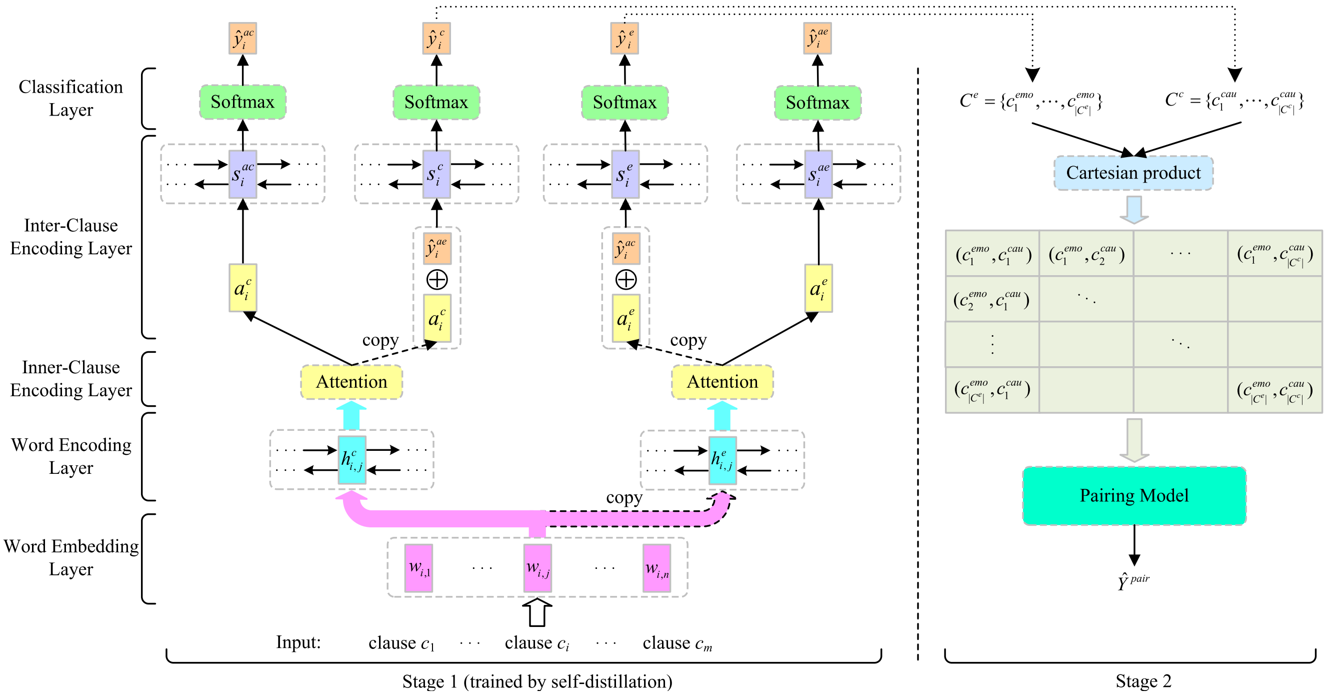

Classification Layer
Inter-Clause Encoding Layer
Inner-Clause Encoding Layer
Word Encoding Layer
Word Embedding Layer
Softmax
Attention
copy
Input: clause c_1 · · · clause c_i · · · clause c_m
$C^e = \{c_1^{emo}, \cdots, c_{|C^e|}^{emo}\}$
$C^c = \{c_1^{cau}, \cdots, c_{|C^c|}^{cau}\}$
Cartesian product
Pairing Model
$\hat{Y}^{pair}$
Stage 1 (trained by self-distillation)
Stage 2

FIGURE 2. An overview of our two-stage method for ECPE. The left half illustrates the proposed mutually auxiliary multitask model with self-distillation.

D. CLAUSE ENCODING LAYER

1) INNER-CLAUSE ENCODING LAYER

To perform aggregation operations on the word language representations and encode the inner-clause contextual information, the attention layer proposed in [36] is adopted, which enables the model to focus on more informative words. Because the words that function as emotional keywords may be different from those express causes in the same clause, two attention layers are used to generate different language representations for one clause. The procedure of obtaining the inner-clause representation for emotion extraction is as follows:

$$u_{i,j}^e = \tanh(W_1^e \cdot h_{i,j}^e + b_1^e) \tag{3}$$

$$score_{i,j}^e = \frac{\exp((u_{i,j}^e)^\top \cdot W_2^e)}{\sum_t \exp((u_{i,t}^e)^\top \cdot W_2^e)} \tag{4}$$

$$a_i^e = \sum_j score_{i,j}^e \cdot h_{i,j}^e \tag{5}$$

where W_1^e and W_2^e are trainable weight matrices, and b_1^e is the bias parameter. As the contextual representation of the j-th word, $h_{i,j}^e$ is transformed by a fully-connected neural network to obtain the relevance to the specific target. $score_{i,j}^e$ denotes the weight of the j-th word in the i-th clause, which is obtained by a softmax operation. $\top$ represents the transpose of matrix. a_i^e is the inner-clause representation of the i-th clause for emotion extraction, which is the weighted sum of $[h_{i,1}^e, \ldots h_{i,j}^e \ldots, h_{i,n}^e]$. Except for different parameters, a_i^c can be calculated by $[h_{i,1}^c, \ldots h_{i,j}^c \ldots, h_{i,n}^c]$ in the same way.

2) INTER-CLAUSE ENCODING LAYER

The fourth layer of our model is the inter-clause encoding layer. In order to capture a specific representation for each task, four clause-level BiLSTMs are constructed to generate task-specific inter-clause representations. In the two auxiliary tasks, $\text{BiLSTM}_{clause}^{ae}$ and $\text{BiLSTM}_{clause}^{ac}$ take the inner-clause representation sequences $[a_1^e, \ldots a_i^e \ldots, a_m^e]$ and $[a_1^c, \ldots a_i^c \ldots, a_m^c]$ as their inputs respectively. Different from the auxiliary tasks, the inputs received by the inter-clause encoding layers of the two main tasks are $a_i^e \oplus \hat{y}_i^{ac}$ and $a_i^c \oplus \hat{y}_i^{ae}$ in the i-th time-step respectively, where $\oplus$ represents the concatenation operation. The inter-clause representations of four tasks are modeled by:

$$s_i^{ac} = \text{BiLSTM}_{clause}^{ac}(a_i^c) \tag{6}$$

$$s_i^{ae} = \text{BiLSTM}_{clause}^{ae}(a_i^e) \tag{7}$$

$$s_i^{c} = \text{BiLSTM}_{clause}^{c}(a_i^c \oplus \hat{y}_i^{ae}) \tag{8}$$

$$s_i^{e} = \text{BiLSTM}_{clause}^{e}(a_i^e \oplus \hat{y}_i^{ac}) \tag{9}$$

where s_i^{ac}, s_i^{ae}, s_i^c and s_i^e respectively represent the hidden states output by four clause-level BiLSTMs in the i-th time-step. The superscript ac denotes the auxiliary task targeting cause extraction, and ae represents the auxiliary task that aims at extracting emotion clauses. The superscript c and e correspond to the two main tasks of cause extraction and emotion extraction respectively. $\hat{y}_i^{ae}$ represents the emotion distribution predicted by one auxiliary task, and $\hat{y}_i^{ac}$ is the cause distribution predicted by another auxiliary task.

E. CLASSIFICATION LAYER

The final language representation of each clause is fed to the classification layer for determining whether a clause is an emotion/cause clause or not. Here, a full-connected layer is used to map the inter-clause representation into targeted categories. Then, the probability distribution of clause on all categories is computed by the softmax operation. Due to fact that the classification layers for different tasks have the same structure, only the process of obtaining $\hat{y}_i^e$ is described in the following:

$$o_i^e = W_3^e \cdot s_i^e + b_3^e \tag{10}$$

$$\hat{y}_i^e = \frac{\exp(o_i^e)}{\sum_{q\in\{0,1\}} \exp(o_{i,q}^e)} \tag{11}$$

where W_3^e and b_3^e are the weight matrix and bias, respectively. $\hat{Y}^e = [\hat{y}_1^e, \ldots \hat{y}_i^e \ldots, \hat{y}_m^e]$ denotes the emotion distributions of all clauses in the document which is output by the main task that aims to emotion extraction. The label with the highest probability in $\hat{y}_i^e$ is the predicted category of i-th clause.

F. SELF-DISTILLATION TRAINING

Since good performance in extraction of emotion and cause is helpful for pairing, we try to employ a self-distillation training to further improve our model. The characteristic of our model is that the tasks appear in pairs, and each pair of tasks corresponds to the main and auxiliary tasks for the same target (such as emotion extraction) respectively. Hence, in each generation of teacher-student training, we determine that only the predicted results of the main tasks in the teacher model are regarded as the knowledge to be transferred to the corresponding pair of tasks in the student model. Through several generations of teacher-student training, students can surpass the teacher. The process of knowledge transfer between the teacher and student models is illustrated in Fig. 3.

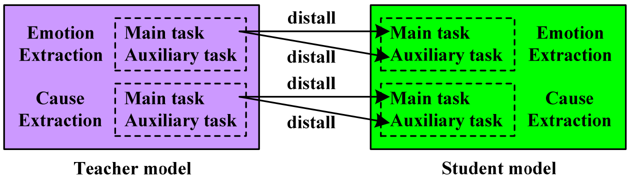


FIGURE 3. **The knowledge distillation from teacher to student.**

1) LOSS FUNCTION

When training our mutually auxiliary multitask model, the cross-entropy loss function (denoted by *CEL*) is utilized. Because our model contains four sub-tasks, the loss needs to be calculated for each task. We also take the main task e as an example in the following:

$$CEL(Y^e, \hat{Y}^e, \theta^e) = -\sum_{i=1}^{m} (y_i^e \cdot \log(\hat{y}_i^e)) + \frac{\lambda}{2} \cdot \left\|\theta^e\right\|^2 \tag{12}$$

where $Y^e = [y_1^e, \cdots y_i^e \cdots, y_m^e]$ is the ground truth. θ^e stands for the parameters that need to be optimized for the task e. λ denotes a coefficient for L_2-norm regularization.

In the mode of teacher-student knowledge distillation, the data labels represented by the one-hot vectors can lead to the loss of similarity information between categories. Here, based on the cross-entropy loss *CEL*, another loss term is added to measure the Kullback-Leibler Divergence between the teacher and the student, which is as follows:

$$KL(\tilde{Y}^e, \hat{Y}^e) = -\sum_{i=1}^{m} (\tilde{y}_i^e \cdot \log(\frac{\tilde{y}_i^e}{\hat{y}_i^e})) \tag{13}$$

where $\tilde{Y}^e = [\tilde{y}_1^e, \ldots \tilde{y}_i^e \ldots, \tilde{y}_m^e]$ denotes the emotion distribution output by the trained teacher model in non-training mode. Combining the above two terms, the loss function becomes:

$$L^e = \alpha \cdot CEL(Y^e, \hat{Y}^e, \theta^e) + \beta \cdot KL(\tilde{Y}^e, \hat{Y}^e) \tag{14}$$

where the weight of the two loss terms is adjusted by adding hyper-parameters α and β with $\alpha + \beta = 1$. It should be noted that the KL term of L^{ae} also depends on $\tilde{Y}^e$ rather than $\tilde{Y}^{ae}$, because the outputs of the auxiliary tasks are only used inside the model, but not used as the teacher signal to supervise the training of students. L^{ae} is denoted as follows:

$$L^{ae} = \alpha \cdot CEL(Y^e, \hat{Y}^{ae}, \theta^{ae}) + \beta \cdot KL(\tilde{Y}^e, \hat{Y}^{ae}) \tag{15}$$

Similar to the process of computing L^e and L^{ae}, L^c and L^{ac} can also be obtained. For a document, the total loss of the four sub-tasks is

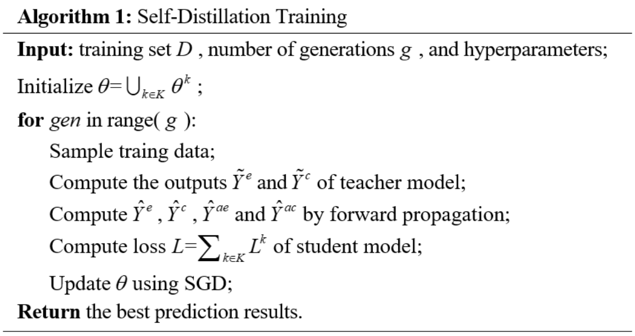
$$L = L^e + L^{ae} + L^c + L^{ac}. \tag{16}$$

2) TRAINING

The entire process of self-distillation training is partitioned into g generations. Therefore, there are $g - 1$ teachers, i.e., the best iteration in prior generations is taken as the teacher in this generation. The training process is shown in Fig. 4. Given a training set D, the number of generations g and all hyperparameters, the model parameters θ needs to be initialized. θ^k denotes those parameters related to the subtask k. $K = \{e, c, ae, ac\}$ represents a set of tasks, where the meanings of e, c, ae and ac are the same as those described in Section III.D. In every generation, firstly, $\tilde{Y}^e$ and $\tilde{Y}^c$ are computed by the teacher model in the non-training mode. Secondly, the student model computes the predicted results of four sub-tasks by forward propagation. Then, the total loss of the four sub-tasks is calculated according to $\tilde{Y}^e$ and $\tilde{Y}^c$. Next, θ can achieve effective updating and optimizing through stochastic gradient descent (SGD). Finally, the best predicted results in the g generations of training are returned.

G. PAIRING EMOTION AND CAUSE

After extracting emotion and cause clauses, the second-stage method in [18] is employed for pairing. Briefly speaking,



```
Algorithm 1: Self-Distillation Training
Input: training set $D$ , number of generations $g$ , and hyperparameters;
Initialize $\theta = \bigcup_{k \in K} \theta^k$ ;
for $gen$ in range( $g$ ):
    Sample traing data;
    Compute the outputs $\tilde{Y}^e$ and $\tilde{Y}^c$ of teacher model;
    Compute $\hat{Y}^e$ , $\hat{Y}^c$ , $\hat{Y}^{ae}$ and $\hat{Y}^{ac}$ by forward propagation;
    Compute loss $L = \sum_{k \in K} L^k$ of student model;
    Update $\theta$ using SGD;
Return the best prediction results.
```

FIGURE 4. The training process of our model with self-distillation.

according to the predicted results in the first stage, the emotion clause set C^e and cause clause set C^c are constructed. Then, the Cartesian product of C^e and C^c is applied to obtain the set C^{pair} of all possible emotion-cause pairs. C^{pair} is used as the dataset for the second stage. As shown in the right half of Fig. 2, (c_1^{emo}, c_1^{cau}) represents an element in C^{pair}. Next, a neural network model is adopted to determine whether each pair of clauses has a causal relationship. In the pairing model, all layers are the same as those of the first-stage model in single-task mode, except for the inter-clause layer not adopted. Finally, the predicted distribution $\hat{Y}^{pair}$ is output.

IV. EXPERIMENTS

A. BENCHMARK DATASET

We utilized the benchmark ECPE dataset released by Xia and Ding [18] which consists of 1945 news documents. Each document is artificially divided into multiple clauses. This dataset is built on an ECE corpus, which assumes that there is at least one emotion in each document, and each emotion can correspond more than one causes. Moreover, a pair of emotion and cause may be in the same clause or in different clauses. Table 1 shows the summary statistics of the dataset. In order to adopt the verification method proposed by Xia and Ding [18], the dataset was randomly divided into 10 equal subsets. Nine of them were used as training data and the remaining as test data.

TABLE 1. Statistics about the dataset. Offset indicates the absolute distance between a pair of emotion and cause clauses. Doc. is short for Document.

Item	Number	Item	Number
doc.	1945	pairs	2167
doc. with 0 pair	0	pairs with offset 0	511
doc. with 1 pair	1746	pairs with offset 1	1342
doc. with 2 pairs	177	pairs with offset 2	224
doc. with ≥ 3 pairs	22	pairs with offset ≥3	90

B. EXPERIMENTAL SETTINGS

The word vectors pre-trained by Xia and Ding [18] are used to initialize the word embedding layer. The dimensions of word embedding and relative position embedding are set to 200 and 50, respectively. In all tasks, the number of hidden units of all BiLSTM is set to 100, and all weight matrix and bias are randomly initialized by the continuous uniform distribution $U(-0.01, 0.01)$. In order to relieve the overfitting problem, we apply dropout to the word embeddings and set 0.2 as the probability with which the elements of the feature vector are randomly zeroed. The Adam optimizer is used, and the mini-batch size is set to 16. The learning rate and the coefficient of L_2-norm regularization are set to 0.005 and 0.00001, respectively. In the self-distillation stage, the max generation of students is set to 3.

To obtain credible results, we repeated the experiments 20 times and averaged the results. The precision P, recall R, and F1 score were selected as the assessment metrics of performance. Since the pairing task depends on the results of emotion and cause extraction in the two-stage method, the performance of the three tasks needs to be evaluated.

C. COMPARED METHODS

In order to evaluate the performance on the ECPE task, three two-stage approaches proposed by Xia and Ding [18] are selected as the baselines.

- **Indep** independently extracts emotion and cause clauses with multitask learning.
- **Inter-CE** uses the predicted results of cause extraction to improve emotion extraction.
- **Inter-EC** utilizes the predicted results of emotion extraction to enhance cause extraction.

To further verify the effectiveness of our proposed models, they are also compared with the following approaches.

- **Inter-ECNC** [46] is a variant of Inter-EC, replacing the LSTM by a Transformer to model the cause clauses.
- **E2EECPE-asw** [40] is an end-to-end link prediction model using CNN and biaffine attention.
- **E2EECPE-gtw** [40] employs Ground Truth Weight Matrix instead of Asymmetric Position Weight Matrix.
- **PairGCN** [53] adopts a graph convolutional network to model the dependent relations between clauses.
- **PairGCN-BERT** [53] is a PairGCN enhanced by pre-trained BERT [59].

D. RESULT ANALYSIS

1) OVERALL PERFORMANCE

Firstly, the comparison between our model and the two-stage methods is analyzed, and the results are shown in Table 2. Compared with Inter-CE, MAM not only improves the F1 score of emotion extraction by 0.19%, but also achieves better performance on cause extraction (precision, recall and F1 are increased by 2.26%, 4.59% and 3.65%, respectively). Without reducing the performance of cause extraction, MAM outperforms Inter-EC by 1.22% and 0.89% in the precision and F1 score of emotion extraction, respectively. And, the precision of MAM on the ECPE task is improved by 2.92% compared with the best-performing two-stage method (Inter-ECNC), and the F1 score is increased by 0.27%.

TABLE 2. Comparison of experimental results on the emotion extraction, cause extraction, and ECPE.

Method	Emotion extraction			Cause extraction			Emotion-cause pair extraction		
	P	R	F1	P	R	F1	P	R	F1
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
Inter-ECNC	-	-	-	0.6863	0.6254	0.6544	0.6601	0.5734	0.6138
E2EECPE-asw	0.8595	0.7915	0.8238	0.7062	0.6030	0.6503	0.6478	0.6105	0.6280
E2EECPE-gtw	0.8552	0.8024	0.8275	0.7048	0.6159	0.6571	0.6491	0.6195	0.6315
PairGCN	0.8587	0.7208	0.7829	0.7283	0.5953	0.6541	0.6999	0.5779	0.6321
PairGCN-BERT	0.8857	0.7958	0.8375	0.7907	0.6928	0.7375	0.7692	0.6791	0.7202
Indep-SD	0.8420	0.8194	0.8299	0.6854	0.5930	0.6348	0.6755	0.5329	0.5939
Inter-CE-SD	0.8461	0.8228	0.8336	0.6834	0.5885	0.6305	0.6802	0.5417	0.6024
Inter-EC-SD	0.8404	0.8212	0.8303	0.7217	0.6223	0.6677	0.6753	0.5836	0.6221
MAM	**0.8486**	**0.8123**	**0.8319**	**0.7035**	**0.6093**	**0.6516**	**0.6893**	**0.5604**	**0.6165**
MAM-SD	**0.8554**	**0.8141**	**0.8339**	**0.7202**	**0.6375**	**0.6751**	**0.6963**	**0.5799**	**0.6320**

Since Indep non-interactively extracts emotion and cause clauses and ignores the mutual indication between emotions and causes, its performance is the worst. It can be seen that the performance of Inter-CE is better than that of Inter-EC on the emotion extraction task but worse on cause extraction. This verifies the conclusion that the unidirectional correlation between emotions and causes extraction can only improve the performance of one task. Since Transformer rather than the LSTM is used to model the cause clauses, Inter-ECNC outperforms Inter-EC on emotion extraction and ECPE. It is worth noting that Inter-EC, Inter-CE and Inter-ECNC all promote one task while ignoring another. Different from them, our MAM establishes the bidirectional interaction between emotion and cause extraction, so as to achieve the mutual promotion of the two tasks.

Secondly, our proposed MAM with self-distillation (MAM-SD) achieves further improvements over MAM on all tasks. In particular, the F1 score of MAM-SD is increased by 2.35% and 1.55% on cause extraction and ECPE, respectively. The self-distillation training is also applied to the three baselines to further verify its effectiveness. These methods (denoted by Indep-SD, Inter-CE-SD and Inter-EC-SD respectively) outperform their versions without self-distillation, respectively. However, the performance of MAM-SD is still better than that of Inter-EC-SD (F1 increased by 0.99%). This not only indicates that our self-distillation method is effective in the joint extraction of emotion and cause clauses, but also that the performance improvement of the first stage is very important for the two-stage ECPE method.

Thirdly, the comparison between our model and some end-to-end approaches is made. As shown in Table 2, the best F1 score of ECPE is achieved by PairGCN-BERT. However, the performance of MAM-SD is almost the same as that of PairGCN, which is another version of PairGCN-BERT. The main difference between these two approaches is whether BERT is employed. Therefore, the performance of our model can also be significantly improved after adopting pre-trained BERT. In addition, the F1 score of MAM-SD is 0.4% and 0.05% higher than those of E2EECPE-asw and E2EECPE-gtw, respectively. The reason why E2EECPE-asw and E2EECPE-gtw are slightly worse than MAM-SD is due to the data imbalance in the set of emotion-cause pairs. In our method, this set comes from the predicted emotion and cause clauses rather than all clauses.

2) ABLATION STUDY

As shown in Table 3, we also conduct ablation experiments on the proposed basic model MAM to verify each component. The ablation models are listed as follows.

- **-individual:** shares an inner-clause encoding layer.
- **-attention:** removes the inner-clause encoding layer.
- **-word:** removes the word encoding layer.
- **-clause:** removes the inter-clause encoding layer.
- **-hierarchy:** removes the word encoding layer and inner-clause encoding layer.

a: INDIVIDUAL

Although the proportion of clauses labeled as both emotion and cause is small in the dataset, sharing an inner-clause encoding layer still decreased F1 score by about 0.8% and 0.41% on emotion extraction and cause extraction, respectively. Hence, it is necessary to get different language representations of the same clause in two perspectives.

b: ATTENTION

Instead of the output of inner-clause encoding layer, the hidden state of clause-level BiLSTM at the last moment is used as the inner-clause contextual representation of clause. As a result, the clause representation becomes worse, which lead to the decrease of the F1 score by 1.69%, 1.74% and 2.57% on emotion extraction, cause extraction and ECPE, respectively.

TABLE 3. Experimental results of structural ablation.

Method	Emotion extraction			Cause extraction			Emotion-cause pair extraction		
	P	R	F1	P	R	F1	P	R	F1
MAM	**0.8486**	**0.8123**	**0.8319**	0.7035	**0.6093**	**0.6516**	**0.6893**	**0.5604**	**0.6165**
-individual	0.8471	0.8028	0.8239	**0.7056**	0.6002	0.6475	0.6891	0.5538	0.6133
-attention	0.8378	0.7955	0.8150	0.6872	0.5923	0.6342	0.6560	0.5386	0.5908
-word	0.8153	0.7501	0.7805	0.6083	0.4549	0.5185	0.6042	0.4141	0.4885
-clause	0.7855	0.7372	0.7590	0.4676	0.2809	0.3467	0.5719	0.2293	0.3256
-hierarchy	0.7743	0.6261	0.6917	0.6519	0.4324	0.5187	0.6232	0.3711	0.4639

c: WORD

Through the observation on the corpus, we can see that emotions are mostly presented as independent emotion-words in clauses, while the causes are often distributed in the context of emotion keywords. If the word encoding layer is not employed to model the temporal relationship of words, the performance of the model will be degraded. Especially, the F1 of cause extraction and pairing drops by around 13%.

d: CLAUSE

When the inter-clause encoding layer is removed from MAM, the F1 on cause extraction significantly drops by 30.49%. Since the error transmission between the two phases, the performance of the model in the second stage is also reduced substantially (F1 dropped by 29.09%). Because emotion clauses usually contain explicit emotion words, the extraction of emotion clauses are not completely dependent on their context. However, it is difficult to extract keywords which represent causes, so the extraction of cause clauses depends on their context heavily. The experimental results illustrate the importance of modeling the context of clauses.

e: HIERARCHY

The mean value of all word vectors in a clause is directly used as the clause vector. The F1 degradation on ECPE is approximately equivalent to the sum of the performance loss of -attention and -word (drops 15.26%). The experimental results show that the model without hierarchical design cannot effectively encode the clause representations. Because there are natural hierarchical relationships among grammatical units of human languages, capturing these grammatical features is necessary.

3) EVALUATION ON EMOTION CAUSE EXTRACTION

In order to obtain a wider comparison, we also evaluate the performance of our model on the ECE task. The descriptions of all baselines are omitted to save space. These baselines can be divided into three types: rule-based approaches (RB and CB), feature-based approaches (RB + CB + ML and Multi-kernel), and neural network-based approaches.

As shown in Table 4, the F1 of MAM-SD is higher 15.08% than that of the best-performing rule-based approach (RB). Compared with the feature-based approaches, the performance of MAM-SD in the F1 score is almost the same as that of Multi-kernel, which is improved by 11.54% over the RB + CB + ML. Furthermore, our method achieves higher F1 value than CNN and Memnet (F1 is increased by 6.75% and 6.17%, respectively), but slightly lower F1 than the ConvMS-Memnet. Although CANN and RTHN outperform our method, their modified versions (CANN_E and RTHN-APE) are outperformed by our method (F1 dropped by 29.54% and 10.57%).

TABLE 4. Comparison of experimental results on ECE.

Method	P	R	F1
RB [15]	0.6747	0.4287	0.5243
CB [20]	0.2672	0.7130	0.3887
RB+CB+ML [19]	0.5921	0.5307	0.5597
Multi-Kernel [16]	0.6588	0.6927	0.6752
CNN [41]	0.6215	0.5944	0.6076
Memnet [25]	0.5922	0.6354	0.6134
ConvMS-Memnet [25]	0.7076	0.6838	0.6955
CANN [28]	0.7721	0.6891	0.7266
RTHN [31]	0.7677	0.7697	0.7662
CANN_E [18]	0.4826	0.3160	0.3797
RTHN-APE [42]	0.5800	0.5618	0.5694
Inter-EC [18]	0.7041	0.6083	0.6507
MAM-SD	**0.7202**	**0.6375**	**0.6751**

It is worth noting that the methods listed in the top half of Table 4 all utilize known emotion clauses as input. Even so, our model still achieves comparable performance with many baselines dependent on emotion annotations. Compared with the methods listed in the bottom half of Table 4, our model achieves the better performance. The comparison results show that the rule-based approaches have the worst performance. The reason for this is that they suffer from the insufficient coverage of manual rule and pattern detection methods. Compared with the feature-based approaches, our method can learn different granularity features and deep semantic representation by using neural networks. Different from the neural network-based approaches, our method can still utilize emotion extraction to promote cause extraction without known emotion annotations. The experimental results confirm the effectiveness and advantages of our method on the ECE task.

TABLE 5. Setting of weight threshold.

β	P	R	F1
0.2	0.6714	0.5753	0.6191
0.3	0.6853	0.5689	0.6204
0.4	0.6925	**0.5804**	0.6306
0.5	0.6935	0.5747	0.6272
0.6	**0.6963**	0.5799	**0.6320**
0.7	0.6554	0.5738	0.6111
0.8	0.6750	0.5733	0.6193

4) EFFECT OF WEIGHT FOR DISTILLATION LOSS

The optimal weight of distillation loss is determined through experiments. We let β change from 0 to 1 with a step of 0.1. The experimental results show that the F1 score of MAM-SD reaches the peak when the weight $\beta = 0.6$. Taking 0.6 at the center, the more the value of β deviates from the center, the more obvious the performance degradation is. However, the recall is the highest when β is 0.4, which can be due to the presence of some emotion-cause pairs that were not extracted. When β drops to 0, the current generation of training ignores the teacher's supervision signal, and degenerates to the original process. In this case, the method which only relies on the one-hot form of real labels without self-distillation may lose the similarity information between categories.

On the other hand, with the increase of β from 0.6 to 1, the F1 score decreased more obviously. When $\beta = 1$, the training process is carried out in a pure distillation mode, and only depends on the performance of teacher. This indicates that it is very difficult for the students to surpass the teacher without the ground truth. We also observed that the optimal value is not 0.5. One reason is that a pair of main and auxiliary tasks both depend on the prediction results of the same main task in the previous generation model to calculate their Kullback-Leibler losses. Therefore, the weight of these losses needs to be slightly large.

5) CASE STUDY

To analyze our model in more detail, we present two cases, as shown in Fig. 1 and Fig. 5. In Case 1, clause c_7 serves as one cause for the emotion clause c_8, and clause c_8 also constitutes a causality by itself. In this situation, our model can correctly extract two emotion-cause pairs (c_8, c_7) and (c_8, c_8), while Inter-EC fails to identify the emotion-cause pair (c_8, c_8). As shown in the results, with the help of auxiliary tasks, our model can extract the clauses that contain a causality better than Inter-EC.

In Case 2, neither model can make an accurate prediction for all pairs. Inter-EC only identifies the emotion clause c_{12}, but predicts clause c_4 as the cause clause by mistake. Our model can correctly extract the emotion-cause pair (c_{12}, c_{10}). These illustrate that our model has better precision than Inter-EC. Moreover, the predicted results of both Inter-EC and our model miss the emotion-cause pairs (c_{12}, c_9) and (c_{12}, c_{11}). Significantly, there are three emotion-cause pairs with the same emotion clause in the document. However, only 22 documents with three or more emotion-cause pairs are in the dataset. Such a small number of samples are not enough to learn the features of complex emotion-cause relationship well. Therefore, further research needs to be conducted to deal with more complex emotion-cause relationship in the future.

Case 1: The detailed description is shown in Fig. 1.

True: (c_8, c_7) (c_8, c_8) Inter-EC: (c_8, c_7) Ours: (c_8, c_7) (c_8, c_8)

Case 2: ...... Since the other party is in arrears with the project payment (c_9), her family is in urgent need of money (c_{10}), and the pressure of life is great (c_{11}), so the woman was helpless but to jump off a building and commit suicide (c_{12}).

True: (c_{12}, c_9) (c_{12}, c_{10}) (c_{12}, c_{11}) Inter-EC: (c_{12}, c_4) Ours: (c_{12}, c_{10})

FIGURE 5. Case study.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a mutually auxiliary multitask model which aims at jointly extracting the emotions and their causes. By adding two auxiliary tasks which are identical to the original tasks, the model establishes the bidirectional correlation between emotion and cause extraction, and improves the performance of both the two tasks. To further enhance the accuracy of emotion and cause extraction, we design a self-distillation method to train our multitask model. Experimental results show that the proposed method achieves better performance than the baseline methods on ECPE.

In our two-stage ECPE method, cross-stage error propagation can be relieved, but still have not yet been solved completely. Hence, in the future work, we will attempt to extract emotion-cause pairs via an end-to-end model. In addition, how to better capture the implicit features of complex causality is also worth focusing on.

REFERENCES

[1] L. H. Son, A. Kumar, S. R. Sangwan, A. Arora, A. Nayyar, and M. Abdel-Basset, "Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network," *IEEE Access*, vol. 7, pp. 23319–23328, 2019.

[2] M. Mukhopadhyay, S. Pal, A. Nayyar, P. K. D. Pramanik, N. Dasgupta, and P. Choudhury, "Facial emotion detection to assess learner's state of mind in an online learning system," in *Proc. 5th Int. Conf. Intell. Inf. Technol.*, Feb. 2020, pp. 107–115.

[3] A. Kumar, S. R. Sangwan, and A. Nayyar, "Multimedia social big data: Mining," in *Multimedia Big Data Computing for IoT Applications*. Singapore: Springer, 2020, pp. 289–321.

[4] A. Seyeditabari, N. Tabari, and W. Zadrozny, "Emotion detection in text: A review," 2018, *arXiv:1806.00674*. [Online]. Available: http://arxiv.org/abs/1806.00674

[5] M. Abdul-Mageed and L. Ungar, "EmoNet: Fine-grained emotion detection with gated recurrent neural networks," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 718–728.

[6] T. Al-Moslmi, N. Omar, S. Abdullah, and M. Albared, "Approaches to cross-domain sentiment analysis: A systematic literature review," *IEEE Access*, vol. 5, pp. 16173–16192, 2017.

[7] A. Onan, S. Korukoğlu, and H. Bulut, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification," *Expert Syst. Appl.*, vol. 62, pp. 1–16, Nov. 2016.

[8] A. Onan and M. A. Tocoglu, "A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification," *IEEE Access*, vol. 9, pp. 7701–7722, 2021.

[9] A. Onan, "Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks," *Concurrency Comput., Pract. Exper.*, Jun. 2020, Art. no. e5909.

[10] A. Onan, "Deep learning based sentiment analysis on product reviews on Twitter," in *Big Data Innovations and Applications*. Cham, Switzerland: Springer, Aug. 2019, pp. 80–91.

[11] A. Onan and M. A. Toçoğlu, "Weighted word embeddings and clustering-based identification of question topics in MOOC discussion forum posts," *Comput. Appl. Eng. Educ.*, pp. 1–15, May 2020.

[12] A. Onan, "Mining opinions from instructor evaluation reviews: A deep learning approach," *Comput. Appl. Eng. Edu.*, vol. 28, no. 1, pp. 117–138, Jan. 2020.

[13] A. Onan, "Sentiment analysis in turkish based on weighted word embeddings," in *Proc. 28th Signal Process. Commun. Appl. Conf. (SIU)*, Oct. 2020, pp. 1–4.

[14] A. Onan, "Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering," *IEEE Access*, vol. 7, pp. 145614–145633, 2019.

[15] S. Y. M. Lee, Y. Chen, and C.-R. Huang, "A text-driven rule-based system for emotion cause detection," in *Proc. NAACL HLT Workshop Comput. Approaches Anal. Gener. Emotion Text*, Jun. 2010, pp. 45–53.

[16] L. Gui, D. Wu, R. Xu, Q. Lu, and Y. Zhou, "Event-driven emotion cause extraction with corpus construction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1639–1649.

[17] R. Xu, J. Hu, Q. Lu, D. Wu, and L. Gui, "An ensemble approach for emotion cause detection with event extraction and multi-kernel SVMs," *Tsinghua Sci. Technol.*, vol. 22, no. 6, pp. 646–659, Dec. 2017.

[18] R. Xia and Z. Ding, "Emotion-cause pair extraction: A new task to emotion analysis in texts," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, Jul. 2019, pp. 1003–1012.

[19] Y. Chen, S. Lee, S. Li, and C. Huang, "Emotion cause detection with linguistic constructions," in *Proc. Conf. Comput. Linguist.*, 2010, pp. 179–187.

[20] I. Russo, T. Caselli, F. Rubino, E. Boldrini, and P. Martínez-Barco, "EMOCause: An easy-adaptable approach to emotion cause contexts," in *Proc. 2nd Workshop Comput. Approaches Subjectivity Sentiment Anal.*, Jun. 2011, pp. 153–160.

[21] K. Gao, H. Xu, and J. Wang, "Emotion cause detection for Chinese micro-blogs based on ECOCC model," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, 2015, pp. 3–14.

[22] K. Gao, H. Xu, and J. Wang, "A rule-based approach to emotion cause detection for Chinese micro-blogs," *Expert Syst. Appl.*, vol. 42, no. 9, pp. 4517–4528, Jun. 2015.

[23] L. Gui, L. Yuan, R. Xu, B. Liu, Q. Lu, and Y. Zhou, "Emotion cause detection with linguistic construction in Chinese Weibo text," in *Natural Lang. Process. Chin. Comput.* Berlin, Germany: Springer, pp. 457–464, 2014.

[24] D. Ghazi, D. Inkpen, and S. Szpakowicz, "Detecting emotion stimuli in emotion-bearing sentences," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*, 2015, pp. 152–165.

[25] L. Gui, J. Hu, Y. He, R. Xu, L. Qin, and J. Du, "A question answering approach for emotion cause extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sep. 2017, pp. 1593–1602.

[26] X. Cheng, Y. Chen, B. Cheng, S. Li, and G. Zhou, "An emotion cause corpus for chinese microblogs with multiple-user structures," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 17, no. 1, pp. 1–19, Nov. 2017.

[27] Y. Chen, W. Hou, and X. Cheng, "Hierarchical convolution neural network for emotion cause detection on microblogs," in *Proc. Artif. Neural Netw. Mach. Learn.*, 2018, pp. 115–122.

[28] X. Li, K. Song, S. Feng, D. Wang, and Y. Zhang, "A co-attention neural network model for emotion cause analysis with emotional context awareness," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, Oct. 2018, pp. 4752–4757.

[29] X. Li, S. Feng, D. Wang, and Y. Zhang, "Context-aware emotion cause analysis with multi-attention-based neural network," *Knowl.-Based Syst.*, vol. 174, pp. 205–218, Jun. 2019.

[30] X. Yu, W. Rong, Z. Zhang, Y. Ouyang, and Z. Xiong, "Multiple level hierarchical network-based clause selection for emotion cause extraction," *IEEE Access*, vol. 7, pp. 9071–9079, 2019.

[31] R. Xia, M. Zhang, and Z. Ding, "RTHN: A RNN-transformer hierarchical network for emotion cause extraction," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5285–5291.

[32] C. Fan, H. Yan, J. Du, L. Gui, L. Bing, M. Yang, R. Xu, and R. Mao, "A knowledge regularized hierarchical approach for emotion cause analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, Nov. 2019, pp. 5614–5624.

[33] J. Hu, S. Shi, and H. Huang, "Combining external sentiment knowledge for emotion cause detection," in *Natural Language Processing and Chinese Computing*. Cham, Switzerland: Springer, 2019, pp. 711–722.

[34] Z. Ding, H. He, M. Zhang, and R. Xia, "From independent prediction to reordered prediction: Integrating relative position and global label information to emotion cause identification," in *Proc. Conf. Assoc. Advancement Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 6343–6350.

[35] B. Xu, H. Lin, Y. Lin, Y. Diao, L. Yang, and K. Xu, "Extracting emotion causes using learning to rank methods from an information retrieval perspective," *IEEE Access*, vol. 7, pp. 15573–15583, 2019.

[36] Y. Diao, H. Lin, L. Yang, X. Fan, Y. Chu, D. Wu, K. Xu, and B. Xu, "Multi-granularity bidirectional attention stream machine comprehension method for emotion cause extraction," *Neural Comput. Appl.*, vol. 32, no. 12, pp. 8401–8413, Jun. 2020.

[37] X. Xiao, P. Wei, W. Mao, and L. Wang, "Context-aware multi-view attention networks for emotion cause extraction," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2019, pp. 128–133.

[38] Y. Chen, W. Hou, X. Cheng, and S. Li, "Joint learning for emotion classification and emotion cause detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Oct. 2018, pp. 646–651.

[39] H. Tang, D. Ji, and Q. Zhou, "Joint multi-level attentional model for emotion detection and emotion-cause pair extraction," *Neurocomputing*, vol. 409, pp. 329–340, Oct. 2020.

[40] H. Song, C. Zhang, Q. Li, and D. Song, "An end-to-end multi-task learning to link framework for emotion-cause pair extraction," 2020, *arXiv:2002.10710*. [Online]. Available: http://arxiv.org/abs/2002.10710

[41] S. Wu, F. Chen, F. Wu, Y. Huang, and X. Li, "A multi-task learning neural network for emotion-cause pair extraction," in *Proc. 24th Euro. Conf. Artif. Intell. (ECAI)*, Sep. 2020, pp. 1–8.

[42] P. Wei, J. Zhao, and W. Mao, "Effective inter-clause modeling for End-to-End emotion-cause pair extraction," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 3171–3181.

[43] Z. Ding, R. Xia, and J. Yu, "ECPE-2D: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 3161–3170.

[44] C. Fan, C. Yuan, J. Du, L. Gui, M. Yang, and R. Xu, "Transition-based directed graph construction for emotion-cause pair extraction," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 3707–3717.

[45] C. Yuan, C. Fan, J. Bao, and R. Xu, "Emotion-cause pair extraction as sequence labeling based on a novel tagging scheme," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Nov. 2020, pp. 3568–3573.

[46] J. Shan and M. Zhu, "A new component of interactive multi-task network model for emotion-cause pair extraction," *J. Phys., Conf. Ser.*, vol. 1693, no. 1, Dec. 2020, Art. no. 012022.

[47] G. Hu, G. Lu, and Y. Zhao, "Emotion-cause joint detection: A unified network with dual interaction for emotion cause analysis," in *Natural Language Processing and Chinese Computing*. Cham, Switzerland: Springer, 2020, pp. 568–579.

[48] R. Fan, Y. Wang, and T. He, "An end-to-end multi-task learning network with scope controller for emotion-cause pair extraction," in *Natural Language Processing and Chinese Computing*. Cham, Switzerland: Springer, 2020, pp. 764–776.

[49] Z. Ding, R. Xia, and J. Yu, "End-to-end emotion-cause pair extraction based on sliding window multi-label learning," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Nov. 2020, pp. 3574–3583.

[50] H. Bi and P. Liu, "ECSP: A new task for emotion-cause span-pair extraction and classification," 2020, *arXiv:2003.03507*. [Online]. Available: http://arxiv.org/abs/2003.03507

[51] Z. Cheng, Z. Jiang, Y. Yin, H. Yu, and Q. Gu, "A symmetric local search network for emotion-cause pair extraction," in *Proc. 28th Int. Conf. Comput. Linguistics*, Madrid, Spain, Dec. 2020, pp. 139–149.

[52] X. Chen, Q. Li, and J. Wang, "A unified sequence labeling model for emotion cause pair extraction," in *Proc. 28th Int. Conf. Comput. Linguistics*, Madrid, Spain, Dec. 2020, pp. 208–218.

[53] Y. Chen, W. Hou, S. Li, C. Wu, and X. Zhang, "End-to-end emotion-cause pair extraction with graph convolutional network," in *Proc. 28th Int. Conf. Comput. Linguistics*, Madrid, Spain, Dec. 2020, pp. 198–207.

[54] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: http://arxiv.org/abs/1503.02531

[55] T. Furlanello, Z. C. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," 2018, *arXiv:1805.04770*. [Online]. Available: http://arxiv.org/abs/1805.04770

[56] P. Yang, G. Yang, X. Gong, P. Wu, X. Han, J. Wu, and C. Chen, "Instance segmentation network with self-distillation for scene text detection," *IEEE Access*, vol. 8, pp. 45825–45836, 2020.

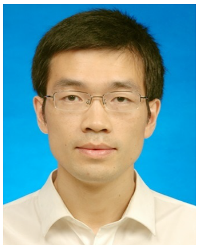
[57] K. Clark, M.-T. Luong, U. Khandelwal, C. D. Manning, and Q. V. Le, "BAM! born-again multi-task networks for natural language understanding," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, Jul. 2019, pp. 5931–5937.

[58] H. Martínez Alonso and B. Plank, "When is multitask learning effective? Semantic sequence prediction under varying data conditions," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 44–53.

[59] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL HLT*, Jun. 2019, pp. 4171–4186.

JIAXIN YU received the B.S. degree in computer science and technology from Harbin Engineering University, China, in 2004, and the M.S. degree in computer science and technology from the Harbin Institute of Technology, China, in 2006. He is currently pursuing the Ph.D. degree in computer science with Yanshan University, Qinhuangdao, China. He is also a Lecturer with the School of Information Science and Engineering, Yanshan University. His research interests include natural language processing, text mining, and sentiment analysis.

WENYUAN LIU (Member, IEEE) received the B.S. and M.S. degrees in computer science from the Northeast Heavy Machinery Institute, Heilongjiang, China, in 1991 and 1993, respectively, and the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2000. Since 2000, he has been with the School of Information Science and Engineering, Yanshan University, Qinhuangdao, China, where he is currently a Chair Professor. He is also the Director of the Engineering Research Center for Network Perception and Big Data of Hebei Province, Qinhuangdao. His research interests include artificial intelligence and mobile networks.



YONGJUN HE received the B.S. degree in computer science and technology from the Harbin University of Science and Technology, Harbin, China in 2003, and the M.S. and Ph.D. degrees with the School of Computer Science, Harbin Institute of Technology, Harbin, in 2006 and 2008, respectively. He is currently a Professor with the School of Computer Science and Technology, Harbin University of Science and Technology. His research interests include speaker recognition, speech recognition, and machine learning.

CHUNYUE ZHANG received the B.S. and M.S. degrees in computer science and technology from the Harbin Institute of Technology, Harbin, China, in 2007 and 2010, respectively. He is currently a Teaching Assistant with the Department of Computer Science, Harbin Finance University. His research interests include natural language processing, machine translation, and text mining.