Multi-level Emotion Cause Analysis by Multi-head Attention Based Multi-task Learning

Xiangju Li, Shi Feng, Yifei Zhang, and Daling Wang(✉)

School of Computer Science and Engineering, Northeastern University,
No. 195, Chuangxin Road, Hunnan District, Shenyang 110207, China
1610543@stu.neu.edu.cn, {fengshi,zhangyifei,wangdaling}@cse.neu.edu.cn

Abstract. Emotion cause analysis (ECA) aims to identify the potential causes behind certain emotions in text. Lots of ECA models have been designed to extract the emotion cause at the clause level. However, in many scenarios, only extracting the cause clause is ambiguous. To ease the problem, in this paper, we introduce multi-level emotion cause analysis, which focuses on identifying emotion cause clause (ECC) and emotion cause keywords (ECK) simultaneously. ECK is a more challenging task since it not only requires capturing the specific understanding of the role of each word in the clause but also the relation between each word and emotion expression. We observe that ECK task can incorporate the contextual information from the ECC task, while ECC task can be improved by learning the correlation between emotion cause keywords and emotion from the ECK task. To fulfill the goal of joint learning, we propose a multi-head attention based multi-task learning method which utilizes a series of mechanisms including shared and private feature extractor, multi-head attention, emotion attention and label embedding to capture features and correlations between the two tasks. Experimental results show that the proposed method consistently outperforms the state-of-the-art methods on a benchmark emotion cause dataset.

Keywords: Emotion cause analysis · Emotion cause clause · Emotion cause keywords · Multi-task learning

1 Introduction

Emotion cause analysis (ECA), a new field in emotion analysis, attempts to comprehend a given text, and then extracts potential causes that lead to emotion expressions in the text. There has been an increasing interest in the research community on ECA more recently since it is widely used in many scenarios. For example, restaurants are eager to find out why people like or dislike their

Supported by National Key R&D Program of China (2018YFB1004700), National Natural Science Foundation of China (61772122, 61872074).

S. Li et al. (Eds.): CCL 2021, LNAI 12869, pp. 77–93, 2021.
https://doi.org/10.1007/978-3-030-84186-7_6

food or services from users' comments or reviews. Similarly, instead of gauging public opinions towards policies or political issues just using frequency counts, governments would like to further know the triggering factors of certain attitudes expressed online.

ECA is a challenging emotion analysis task since it requires a comprehensive understanding of natural languages and the ability to do further inference. Restricted by the lack of annotated corpora, early studies used rule-based methods [1,17] and crowd-sourcing methods [24] to tackle this task. Until recently, Gui et al. [11] released a reasonable ECA corpus, based on which they developed the first deep learning model for the task [10], and by following that, various other ECA approaches were proposed and achieved superior results [5,7,20,21,27].

Example 1. $[\mathbf{x_1}]$: Entertainment reporter Jucy interviewed that LAM Raymond and Xinyue Zhang to get married in 2020. $[\mathbf{x_2}]$: Also, $[\mathbf{x_3}]$: she interviewed a piece of explosive news that **the wedding ceremony of Tina Tang will be held**. $[\mathbf{x_4}]$: The Tina Tang's fans were very happy and congratulated her[1].
*(*Original text: $[\mathbf{x_1}]$: 娱乐记者Jucy采访到林峰和张馨月在2020年结婚. $[\mathbf{x_2}]$: 同时, $[\mathbf{x_3}]$: 她采访到一条唐艺昕**将要举办婚礼**的爆炸性新闻. $[\mathbf{x_4}]$: 唐艺昕的粉丝非常开心并且庆祝她.*)*

Most of the existing studies identify which clause contains the emotion cause. Example 1 shows a piece of text from Sina Weibo, in which the emotion word is "*happy*" and the exact emotion cause of "*happy*" is "*the wedding ceremony of Tina Tang will be held*". We call all the words in the exact emotion cause as the emotion cause keywords and the clause which contains the emotion word as the emotion clause. For instance, in this example, the emotion clause, emotion cause clause and the emotion cause keywords are clause $[\mathbf{x_4}]$, clause $[\mathbf{x_3}]$ and {"*the*", "*wedding*", "*ceremony*", "*of*", "*Tina*", "*Tang*", "*will*", "*be*", "*held*"}, respectively. With the existing methods, the emotion cause clause $[\mathbf{x_3}]$ is expected to be extracted because the cause of "*happy*" is "*the wedding ceremony of Tina Tang will be held*" that is a part of $[\mathbf{x_3}]$.

However, only identifying which clause contains the emotion cause is flawed and ambiguous. In Example 1, the content "*she interviewed a piece of explosive news*" in $[\mathbf{x_3}]$ is not the cause of "Tina Tang's fans happiness. If $[\mathbf{x_4}]$ becomes "*The reporter felt very happy and immediately won the boss's praise*", the content "*she interviewed a piece of explosive news*" is the cause and "*the wedding ceremony of Tina Tang will be held*" is not the cause in $[\mathbf{x_3}]$. Therefore, with only an emotion cause clause extracted, it is common that one cannot exactly tell the real stimulus of a given emotion.

Extracting the exact emotion cause is very challenging. It needs not only deep text understanding including the role of each word in the emotion expression, but also requires specific semantic inference based on what is understood. Meanwhile, it is difficult to precisely determine the boundary of the cause segment, which

[1] Each instance in the ECA corpus contains presumably a unique emotion and at least one emotion cause clause. A clause is typically a text segment separated by punctuation marks (e.g., ',', '.', '?', '!', etc.) in the given document.

differs from the traditional Question Answering (QA) task for why questions. Because the emotion clause expressions in ECA triggering the cause finding are typically much more diverse and ambiguous, and the real cause to be extracted is generally much finer-grained. We argue that rather than only locating the coarse-grained emotion cause clause or precisely finding the exact cause segment(s), it would be more practical to adopt a hybrid extraction strategy considering clause level and word level to help us get the emotion cause.

In this paper, we attempt to extract the Emotion Cause Clause (ECC) and Emotion Cause Keywords (ECK) simultaneously. Given an emotion event, the goal of ECC task is to identify which clause contains the stimulants of emotion. ECK is a finer-grained emotion cause analysis task, which aims to identify which word(s) in the clause contribute to stimulate the emotion expression. Basically, ECK is more difficult to identify than ECC but more light-weighted than the exact emotion cause identification. The ECK task requires not only capturing the relationship between the words and emotion expression but also understanding the role of each word in the clause. However, it does not only need to identify the complete and precise cause content but also the keywords that help us better understand the emotion cause from the clause. For example, we can find that the specific cause of "*happy*" emotion in Example 1 can be better conveyed if both the emotion cause clause $[\mathbf{x_3}]$ and emotion cause keywords, e.g., "*wedding*", "*ceremony*", "*Tina*" and "*Tang*" are identified.

To this end, we propose a Multi-head Attention based Multi-task learning network for Multi-level Emotion Cause Analysis (MamMeca). In the MamMeca, both ECC and ECK tasks make use of the semantic information of the text and the emotion expression to infer the cause of the emotion, for which the ECK and ECC mutually enhance each other in the unified framework. The proposed model consists of a shared feature extractor and a private feature extractor, where multi-head attention and label embedding mechanisms are designed to facilitate capturing the relationship between the two tasks. The contribution of our paper is three-fold:

- We present a multi-level ECA problem, based on the hypothesis that ECC and ECK tasks together can help us better identify the specific emotion cause and both tasks can benefit each other by mutual enhancement. To the best of our knowledge, this work is the first attempt to incorporate the two sub-tasks into a unified framework for ECA.
- We propose an extensible and effective multi-head attention based multi-task neural network for multi-level ECA. The model utilizes a shared private feature extractor to get effective representations of the keywords and clause. Meanwhile, multi-head attention and label embedding mechanisms are designed to further capture the inter-task correlations.
- Our results on a dominating benchmark dataset validate the feasibility and effectiveness of our proposed MamMeca model.

2 Related Work

Various learning methods have been applied to emotion cause analysis, which are mainly categorized as rule-based models, feature-driven models and feature-learning models.

Rule-Based Models. Lee et al. [17] first gave the formal definition of emotion cause analysis task and constructed a small-scale corpus from the Academia Sinica Balanced Chinese Corpus. Based on the corpus, Lee et al. [18] developed a rule-based system for emotion cause detection based on various linguistic rules. Some studies then extended rule-based approaches to informal texts such as Gao et al. [9]. Li et al. [19] also constructed an automatic rule-based system to detect the cause event of emotional post on Chinese microblog posts.

Feature-Driven Models. Chen et al. [1] developed two sets of linguistic features based on linguistic cues and a multi-label approach, and utilized SVM to detect emotion causes. Similarly, Gui et al. [12] extended the linguistic rules as features and used SVM model for emotion cause extraction. More recently, Gui et al. [11] released a Chinese emotion cause corpus based on public city news, which has inspired a large-scale ECA research campaign. Meanwhile, they presented a multi-kernel SVM approach for emotion cause extraction. Xu et al. [28] used LambdaMART algorithm incorporating both emotion-independent features and emotion-dependent features to identify emotion cause clause. The above models have achieved highly competitive results for ECA task, but the models heavily depend on the design of effective features.

Feature-Learning Models. Inspired by deep learning, Gui et al. [10] utilized the deep memory network model to capture the relationship between the clause and the emotion word, and then identified the emotion cause clause. Yu et al. [31] presented a hierarchical network-based clause selection framework for ECA, which considered three levels (word-phrase-clause) of information. Li et al. [21] proposed a co-attention mechanism to capture the relationship between the emotion expression and the candidate clause, and then extracted the emotion cause clause. Li et al. [20] took advantage of clues provided by the context of the emotion word and proposed a multi-attention-based neural network to identify which clause contained emotion cause. Ding et al. [5] proposed a neural network architecture to incorporate the relative position of the clause and the prediction label of previous clauses information for emotion cause clause extraction. Xia et al. [27] proposed a hierarchical network architecture based on RNN and Transformer to capture the different levels features for emotion cause clause identification. Fan et al. [7] designed a regularized hierarchical neural network (RHNN) which utilized the discourse context information and the relative position information for emotion cause clause extraction. Hu et al. [14] proposed a graph convolutional network to fuse the semantics and structural information, which automatically learned how to selectively attend the relevant clauses useful for emotion cause analysis. Recently, Xia et al. [26] proposed a new task: emotion-cause pair extraction, which aims to extract all potential pairs of emo-

Table 1. An example of illustrating the ECC task.

Clause	Content	y^c
x_1	Entertainment . . . get married in 2020. (娱乐. . . 在2020年结婚.)	0
x_2	Also, (同时，)	0
x_3	she interviewed . . . held. (她采访到. . . 新闻.)	1
x_4	The Tina Tang's fans . . . congratulated her. (唐艺昕的粉丝. . . 庆祝她.)	0

tion clause and corresponding cause clause in a text. Following this, many deep learning models [6,8,14,25,30] were designed for this task.

Discussion. Most of the previous studies attempt to extract which clause contains the emotion causes for a given emotion cause event. It is not enough to identify which clause contains the emotion cause in many application scenario, and Example 1 has illustrated this situation clearly. Only Gui et al. [10] utilized the emotion cause keyword to identify which clause contains the emotion cause, however, they still extract the emotion cause at clause level. That is, the clause is identified as emotion cause clause if it contains the emotion cause keyword in their model. Different from the previous studies, we propose to extract both the emotion cause clause and the indicative emotion cause keywords in one shot which is the first of such effort.

3 Methodology

3.1 Task Definition

Given a document d, which is a passage about an emotion cause event, it contains an emotion expression and the cause of the emotion. The document usually consists of multiple clauses $\{x_1, x_2, \cdots, x_m\}$, and each $x_i = \{w_{i1}, w_{i2}, \cdots, w_{in_i}\}$ is a clause where w_{ij} is the j-th word of x_i. Each document is assumed to have *a unique emotion* and at least one corresponding **emotion cause clause**. Let $x^e = \{w_1^e, \cdots, w_{l_e}^e, \cdots, w_{n_e}^e\}$ be the **emotion clause** containing the concerned emotion word $w_{l_e}^e$ which is the l_e-th word of x^e. In our work, both ECC and ECK tasks are seen as a binary classification problem. The expected labels of the clause or word obtained by the model is either 1 (yes) or 0 (no).

ECC Task. The goal of ECC task is to identify which clause stimulates the emotion expression. Then, the task can be formulated as

$$p_i^{y^c} = f_{ECC}(x_i, x^e) \tag{1}$$

where the function f_{ECC} identifies whether the clause x_i stimulates the emotion expressed in the emotion clause x^e, and $p_i^{y^c}$ is the predicted probability of x_i ($y^c = 1$ if x_i stimulates the emotion expressed in the x^e, or $y^c = 0$ otherwise). Table 1 illustrates ECC task clearly.

Table 2. An example of illustrating the ECK task. (Entertainment: 娱乐; reporter: 记者; the wedding ceremony: 婚礼; her: 她)

w	Entertainment	reporter	...	that	the	wedding	ceremony	...	her	.
y^w	0	0	...	0	1	1	1	...	0	0

ECK Task. ECK task aims to identify which word participates to stimulate the emotion $w^e_{l_e}$, which is formulated as

$$p^{y^w}_{ij} = f_{ECK}\big(w_{ij}, x_i, x^e\big) \tag{2}$$

where x_i and x^e are the i-th clause and the emotion clause of document, respectively. w_{ij} is the j-th word of x_i. The function f_{ECK} outputs the probability that the word w_{ij} stimulates the emotion expression or not, $p^{y^w}_{ij}$ is the predicted probability for $w_{ij} \in x_i$, and $y^w \in \{1, 0\}$. To illustrate this definition, we show the labels of words in Example 1 in Table 2.

3.2 Model Description

In this section, we introduce our proposed MamMeca model that will learn task-shared feature (Sect. 3.2) and the task-private feature (Sect. 3.2). The architecture of MamMeca is given in Fig. 1, which mainly consists of three components: (1) task-shared feature extracting layer; (2) task-private feature extracting layer; and (3) classification layer. The task-shared feature extracting layer aims to capture the common features of the ECC and ECK tasks, which mainly contains two parts: shared Bi-GRU and emotion attention mechanism. After this layer, we can obtain the emotion weighted word representations, which will be further fed into the private feature extracting layer. Task-private feature extracting layer mainly contains three parts: private Bi-GRU, multi-head attention mechanism, and label embedding mechanism. Private Bi-GRUs are used for ECC task and ECK tasks to get the word level and clause level representations respectively. The emotion cause keywords must appeared in emotion cause clause which can be seen the definitions of two tasks in Sect. 3.1. Hence, the labeling embedding and multi-head attention mechanisms are designed to enhance the performance of the ECC task and ECK task by using the predicted word labels in ECK task and the clause presentation obtained in ECC task. The classification layer aims to get the class distribution of the clauses and words for ECC and ECK tasks respectively.

Task-Shared Feature Extracting Layer. This layer extracts common features shared between the two tasks, which contains two parts: (1) shared Bi-GRU encoder; (2) emotion attention mechanism.

Shared Bi-GRU Encoder. Bi-directional gated recurrent units (Bi-GRU) leverages gates to control the information flow from previous and future words,

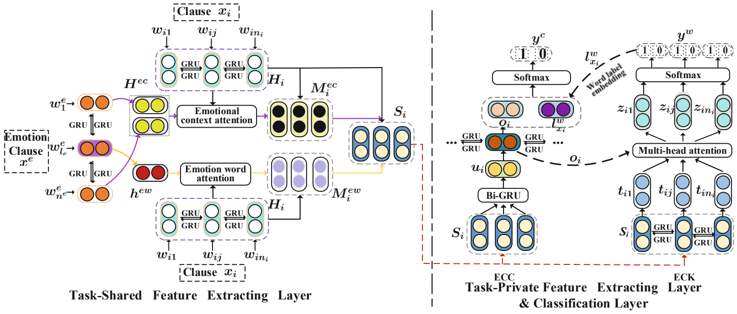

Fig. 1. The architecture of the MamMeca model. The model contains three main parts: Task-Shared feature Extracting Layer, Task-Private Feature Extracting Layer and Classification Layer. Task-Shared Feature Extracting Layer contains shared Bi-GRU encoder and emotion attention mechanism. This layer aims to capture the shared features for the ECC and ECK tasks. Task-Private Feature Extracting Layer includes private Bi-GRU encoder for specific task extraction, multi-head attention mechanism for enhancing the word representation by using the clause representation obtained by ECC task, Label embedding mechanism for enhancing the clause representation by utilizing the word label obtained in ECK task. Classification Layer is able to get the class distribution of the words and clauses, respectively.

which can better capture long term dependencies than basic RNNs, and are often chosen in practice [2]. Thus, we adopt Bi-GRU to incorporate information from both the forward and the backward directions of input sequence. In this work, we first map each word into a low dimensional embedding space and then feed the whole document into a Bi-GRU word encoder to extract word sequence features.

$$\overrightarrow{h}_{ij} = \overrightarrow{GRU}(\overline{w}_{ij}), \quad \overleftarrow{h}_{ij} = \overleftarrow{GRU}(\overline{w}_{ij}), \quad j \in \{1, \cdots, n_i\} \tag{3}$$

where $\overline{w}_{ij} \in \mathbb{R}^{d_w}$ is the embedding vector for the word w_{ij} in clause x_i at time step j and n_i is the length of clause x_i. The j-th word representation in the clause x_i can be expressed as $h_{ij} = [\overrightarrow{h}_{ij} \oplus \overleftarrow{h}_{ij}]$, where $\oplus$ denotes concatenation, $h_{ij} \in \mathbb{R}^{2d_h}$, and d_h is the size of Bi-GRU hidden vector. Therefore, we can obtain the representation matrix $H_i = [h_{i1}; h_{i2}; \cdots; h_{in_i}]$ $(H_i \in \mathbb{R}^{n_i \times 2d_h})$ of clause x_i. Symmetrically, we can obtain the emotion word $(w_{l_e}^e)$ representation vector $h^{ew} \in \mathbb{R}^{2d_h}$ and the emotion context word (w_i^e) representation $h_i^{ec} \in \mathbb{R}^{2d_h}$ $(i \in \{1, \ldots, l_e - 1, l_e + 1, \ldots, n_e\})$.

Emotion Attention Mechanism. The relationship between the candidate cause clause and the emotion clause plays an important role in emotion cause identification, which has been verified in [21]. We introduce an emotion attention mechanism to extract such words that are important to the emotion expression of the clause and aggregate the representation of these informative words to construct the clause vector. Specifically, we differentiate emotion word and emotion context which usually express different types of information. The emotion word "*happy*" in Example 1 aims to convey the emotion polarity directly while the emotion context "*The Tina Tang's fans were very - and congratulated her*" provides the related event information about the emotion, such as *"Tina Tang's fans congratulated her"* (dubbed as *emotion event*). These two types information play different roles in emotion cause identification. Hence, we get separate clause representations based on emotion word attention and emotion context attention.

(1) *Emotion word attention.* Emotion word attention is applied over the words embedding to allow the model to focus on words that contribute highly to the emotion category expression of the clause:

$$m_{ij}^{ew} = \alpha_{ij} * h_{ij}; \qquad \alpha_{ij} = \frac{\exp(h_{ij}^{\top} h^{ew})}{\sum_{j'=1}^{n_i} \exp(h_{ij'}^{\top} h^{ew})} \tag{4}$$

where h^{ew} is the emotion word vector obtained by Bi-GRU encoder, α_{ij} is the attention weight indicating the importance of word w_{ij}, and m_{ij}^{ew} is the emotion word attention-based representation of w_{ij}. We then obtain the emotion weighted representation of x_i as $M_i^{ew} = [m_{i1}^{ew}; \dots; m_{in_i}^{ew}]$ where $M_i^{ew} \in \mathbb{R}^{n_i \times 2d_h}$.

(2) *Emotion context attention.* Emotion context attention allows the model to focus on words that contribute to the emotion event of the clause. The relation matrix between the clause x_i and the emotion context is constructed as $A = (H_i W_1) * (H^{ec} W_2)^{\top}$, where $H^{ec} = [h_1^{ec}; \cdots; h_{l_e-1}^{ec}; h_{l_e+1}^{ec}; \dots; h_{n_e}^{ec}]$, $H^{ec} \in \mathbb{R}^{(n_e-1) \times 2d_h}$, $W_1, W_2 \in \mathbb{R}^{2d_h \times 2d_h}$ are trainable parameters. Each element a_{jk} ($j \in \{1, \dots, n_i\}$, $k \in \{1, \dots, l_e - 1, l_e + 1, \dots, n^e\}$) of A represents the relationship between the j-th word of clause x_i and the k-th word of emotion context of x^e. The importance of the j-th word of x_i to the emotion event expression can be obtained as follows:

$$\beta_{ij} = \frac{\exp(\theta_{ij})}{\sum_{j'=1}^{n_i} \exp(\theta_{ij'})}; \quad \theta_{ij} = max(a_{j1}, a_{j2}, \dots, a_{jn_i}) \tag{5}$$

θ_{ij} represents the most influential values for the emotion context obtained by x_i. Then we can obtain the new representation of x_i considering the emotion context as: $M_i^{ec} = [m_{i1}^{ec}; \dots; m_{in_i}^{ec}]$ where $m_{ij}^{ec} = \beta_{ij} * h_{ij}$, $M^{ec} \in \mathbb{R}^{n_i \times 2d_h}$.

Finally, the high-level representation of the clause x_i can be obtained by combining the original clause representation, the emotion word attention weighted clause representation and the emotion context attention weighted clause representation:

$$S_i = Relu((M_i^{ew} \oplus M_i^{ec}) * W_3) \oplus H_i \tag{6}$$

where $W_3 \in \mathbb{R}^{4d_h \times 2d_h}$ is the trainable parameter.

Task-Private Feature Extracting Layer. This layer extracts private features that are specific to each task being updated exclusively, which contains three parts: (1) private Bi-GRU encoder; (2) multi-head attention mechanism; (3) label embedding mechanism.

Private Bi-GRU Encoder. For the ECC task, two private Bi-GRUs are utilized, one applied at word level and the other at clause level.

To capture the task-specific information, a private Bi-GRU is used at word level to get the representation of x_i as $u_i = [\overrightarrow{GRU}(s_{in_i}) \oplus \overleftarrow{GRU}(s_{i1})] \quad i \in \{1, 2, \ldots, m\}$, where s_{i1} and s_{in_i} are the first and the n_i-th word vectors of S_i (see Eq. (6)). The semantic expression of a clause is usually impacted by its context. Hence, we utilize another Bi-GRU applied at clause level to model the latent relation among different clauses on top of u_i. The clause-level representation of x_i can be obtained as $o_i = [\overrightarrow{GRU}(u_i) \oplus \overleftarrow{GRU}(u_i)]$, where $o_i \in \mathbb{R}^{2d_h}$.

For the ECK task, we utilize a single Bi-GRU to obtain the specific word representation for each word w_{ij} as $t_{ij} = [\overrightarrow{GRU}(s_{ij}) \oplus \overleftarrow{GRU}(s_{ij})] \quad j \in \{1, 2, \ldots, n_i\}$, where $s_{ij} \in \mathbb{R}^{2d_h}$ is the word vector of w_{ij} in S_i.

Multi-head Attention Mechanism. ECC and ECK tasks are closely related as the emotion cause keywords must appear in emotion cause clause. Our core idea is to utilize the cause clause representation generated by the ECC task to enhance the learning of cause keyword representation in the ECK task. We exploit multi-head attention mechanism to capture word correlation in each clause, based on which the high-level word representation is obtained for further classification.

Let τ denote the number of heads in the multi-head attention. We first linearly project the queries, keys and values by using different linear projections: $q_{ij} = t'_{ij}W^q$, $k_{ij} = t'_{ij}W^k$, $\quad v_{ij} = t_{ij}W^v$. Where $t'_{ij} = t_{ij} \oplus o_i$ and $t'_{ij} \in \mathbb{R}^{4d_h}$, $W^q \in \mathbb{R}^{d_k \times d_k}$, $W^k \in \mathbb{R}^{d_k \times d_k}$ and $W^v \in \mathbb{R}^{d_k/2 \times d_k}$ are trainable parameters, and $d_k = 2d_h/\tau$. Then the attention value of the j-th word to the k-th word of clause x_i can be computed below:

$$\eta_{jk} = \frac{\exp(q_{ij} * k_{ik}^{\top})}{\Sigma_{k'=1}^{n_i} \exp(q_{ij} * k_{ik'}^{\top})} \tag{7}$$

The final representation of the j-th word is obtained by fusing the attention weighted vector and the query (q_{ij}): $z'_{ij} = \eta_{jk}v_{ik} + q_{ij}$, where z'_{ij} is the word representation taking into account word correlations in the clause.

Label Embedding Mechanism. The emotion cause keywords can provide important signals for locating the emotion cause clause. Therefore, we can enhance the ECC representation learning using the cause keyword labels obtained by the ECK task.

Let $l_{y^w} \in R^{d_w}$ be the embedding vector of keyword label y^w. Note that the clause, which contains emotion cause keywords, is the emotion cause clause. Therefore, the keyword label in the clause x_i also plays an important role in emotion cause clause identification. Let $\{y^w_{i1}, y^w_{i2}, \cdots, y^w_{in_i}\}$ represent the keywords labels predicted by ECK task (see Sect. 3.2). Then, the predicted keywords label embedding vector of x_i can be presented as: $l^w_{x_i} = [l_{y^w_{i1}} \oplus l_{y^w_{i2}} \oplus \cdots \oplus l_{y^w_{in_i}}] * W_l$. Finally, we obtain the new clause vector by concatenating the label embedding vector and the original clause representation vector as $o'_i = [o_i \oplus l^w_{x_i}]$.

Classification Layer. In the classification layer, the class distribution of a keyword w is computed using softmax as $p_{ij}^{y^w} = softmax(W_w z_{ij} + b_w)$, where z_{ij} is the combination of τ representation vectors (z'_{ij}), and W_w and b_w are learnable parameters. Similarly, the class distribution of clause x_i is computed as $p_i^{y^c} = softmax(W_c o'_i + b_c)$, where W_c and b_c are training parameters.

3.3 Training and Parameter Learning

Given a document d, the loss functions of ECC task and ECK task can be defined as follows:

$$\mathcal{L}_{ECC} = -\sum_{x_i \in d} \mathcal{G}(x_i) \log(p_i^{\mathcal{G}(x_i)}) \qquad \mathcal{L}_{ECK} = -\sum_{x_i \in d} \sum_{w_{ij} \in x_i} \mathcal{Y}(w_{ij}) \log(p_{ij}^{\mathcal{Y}(w_{ij})}) \tag{8}$$

where $\mathcal{G}(x_i)$ and $\mathcal{Y}(w_{ij})$ denote the ground-truth label of x_i and w_{ij}, respectively, and $p_i^{\mathcal{G}(x_i)}$ and $p_{ij}^{\mathcal{Y}(w_{ij})}$ are the corresponding class probability predicted. The final loss function of the proposed model is given as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ECC} + \lambda_2 \mathcal{L}_{ECK} \tag{9}$$

where λ_1 and λ_2 are hyper-parameters.

In the training phrase, we use Adam [16] to optimize the final loss function. After learning the parameters, we feed the test instance into the model and take the label with the highest probability as the predicted category.

4 Experiments and Results

4.1 Dataset and Settings

Dataset. Our experiments are conducted on a Chinese emotion cause analysis dataset publicly available and widely used for ECA evaluation which was collected from Sina News[2] by Gui et al. [11]. The dataset is manually annotated with the clause labels and keyword labels which contains 2,105 documents, 11,799 clauses and 2,167 emotion cause clauses. Most of the documents contain one emotion cause clause. Each clause is word segmented by Jieba[3] and the average number of words in the clause is 7.

Experimental Settings. We follow the settings of previous works to split the datasets for train/test [10,27]. We apply fine-tuning for the word vectors, which can help us improve the performance. The word vectors are initialized by word embeddings that are pre-trained on the emotion cause dataset with CBOW [22], where the dimension is 100. The trainable model parameters are given initial values by sampling from uniform distribution $\mathcal{U}(-0.01, +0.01)$. The learning rate is initialized as 0.001. Dropout [13] is taken to prevent overfitting, and the

[2] http://hlt.hitsz.edu.cn/?page%20id=694.

[3] https://github.com/fxsjy/jieba.

dropout rate is 0.5. The size of Bi-GRU hidden states d_h is set as 50. λ_1 and λ_2 are set as 1.0 and 0.75, respectively. Both the batch size and epochs are set to 20. The metrics of both tasks we use in evaluation include precision (P), recall (R) and F1 score $(F1)$, which are the most commonly used evaluation metrics for emotion cause analysis [10,27].

4.2 Comparison of Different Methods

For the ECC task we compare our proposed model with the following three groups models. (1) **Group I (Rule-based and knowledge-based models):** *RB* extracts the emotion cause by utilizing two sets of linguistic rules proposed by Lee et al. [17]. *KB* is a knowledge-based method [24] that uses the Chinese Emotion Cognition Lexicon [29] as the common-sense knowledge base. (2) **Group II (Feature-driven models):** *SVM (RB+KB)*, *SVM (Word2vec)* and *SVM(n-grams)* use linguistic rules [17] plus Emotion Cognition Lexicon [29], Word2vec embeddings [23], and n-grams as features, respectively, to train a SVM classifier. *SVM (MK)* uses the multi-kernel SVMs based on structured representation of events to extract emotion cause [11]. *LambdaMART* utilizes LambdaMART algorithm incorporating emotion independent and dependent features to identify emotion cause [28]. (3) **Group III (Feature-learning models):** *ConvMS-Memnet* is a convolutional multiple-slot deep memory network for the ECC task [10]. *CANN* [21] and *MANN* [20] takes advantage of the emotion context information and designed different attention model to capture the relationship between the emotion clause and clause for ECC task. *PAE-DGL* is a reordered prediction model, which incorporates relative position information and dynamic global label for emotion cause extraction [5]. *RTHN* is a transformer hierarchical network which utilizes RNN to encode multiple words in each clause and transforms to learn the correlation between multiple clauses in a document [27]. *RHNN* is a regularized hierarchical neural network [7]. *FSS-GCN* is a graph convolutional networks with fusion of semantic and structure for emotion cause clause identification [14].

Among these methods, only **RB** and **ConvMS-Memnet** are able to identify emotion cause keywords. To test the performance on ECK task, we compare the proposed model with the rule-based model (**RB**), feature-driven model (**SVM**), and Feature-learning models (**ConvMS-Memnet**, **Bi-GRU**, **Bi-LSTM**). Furthermore, we compare the proposed model with question answering which is relevant to the ECA problem. In our experiment, we adopt **BERT** ($BERT_{BASE}$ version[4]) [4], a pre-trained bidirectional Transformer-based language model which achieves a good performance on various public question answering datasets recently [3,15].

[4] https://storage.googleapis.com/bert_models/2018_11_03/chinese_L\discretionary-12_H\discretionary-768_A\discretionary-12.zip.

Results and Analysis. Table 3 shows the results of our proposed MamMeca model and baselines on ECC task. We can observe that: (1) MamMeca outperforms state-of-art baselines for ECC task on all the evaluation metrics, which clearly confirms the effectiveness of joint identification of emotion cause clause and keywords with our multi-task learning framework. (2) The $F1$ value obtained by MamMeca model outperforms the strongest baseline RHNN by 3.1%, which verifies the effectiveness of incorporating the label embedding and emotion attention mechanisms. (3) MamMeca outperforms the BERT-based QA model, which further verifies advantage of our proposed model. This is because standard QA task assumes that the question is a complete question expression while in our case the emotion clause is most likely incomplete or ambiguous rendering a more challenging problem. MamMeca can better deal with it since the complex relationship between the emotion clause and the cause clause can be captured with the joint learning.

Table 3. Results on ECC task. The results with superscript ◇ are reported in Gui et al. [10], and the rest are reprinted from the corresponding publications.

Compared with Group I and Group II				Compared with Group III			
Method	P	R	$F1$	**Method**	P	R	$F1$
RB◇	0.675	0.429	0.524	ConvMS-Memnet◇	0.708	0.689	0.696
KB◇	0.267	0.713	0.389	CANN	0.772	0.689	0.727
RB+KB◇	0.544	0.531	0.537	MANN	0.784	0.759	0.771
SVM (RB+KB)◇	0.592	0.531	0.560	PAE-DGL	0.762	0.691	0.742
SVM (n-grams)◇	0.420	0.4375	0.429	RTHN	0.770	0.766	0.768
SVM (Word2vec)◇	0.430	0.423	0.414	RHNN	0.811	0.773	0.791
SVM (MK)◇	0.659	0.693	0.675	FSS-GCN	0.786	0.757	0.771
LambdaMART	0.772	0.750	0.761	BERT	0.782	0.757	0.769
MamMeca	**0.849**	**0.798**	**0.822**	MamMeca	**0.849**	**0.798**	**0.822**

Table 4. Results on ECK task.

Method	P	R	$F1$
RB	0.228	0.643	0.337
SVM (Word2vec)	0.024	0.006	0.010
Bi-LSTM	0.150	0.332	0.207
Bi-GRU	0.149	0.311	0.202
ConvMS-Memnet	0.625	0.614	0.620
BERT	0.710	0.749	0.729
MamMeca	**0.714**	**0.774**	**0.742**

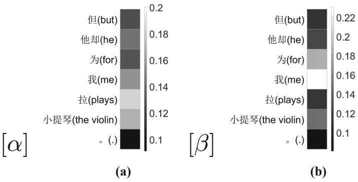

Fig. 2. Visualization of attention. Darker color represents lower attention weight.

Table 4 shows the results of the emotion cause keyword extraction. From this table, we find that our MamMeca model outperforms all the baselines including the state-of-the-art model ConvMS-Memnet [10] and the strong QA model

BERT. It gains improvement more than 12% in $F1$ compared to ConvMS-Memnet, which indicates that the proposed model's strong ability to capture the relationships between the emotion expression and the candidate cause words expressions. BERT achieves a good performance on many QA datasets, however performs worse than MamMeca on the ECK task as well. It further confirms that the QA models is not a better choice for tackling the ECA problem. In general, the emotion cause extraction is concerned about the cause of the given emotion expression instead of the relevance or similarity between the question and text.

4.3 Ablation Study

To understand the effect of different components, we compare several sub-networks of our model.

Full is the full MamMeca model. We use **Full-X** to represent the model without component **X**, where **X** can be ECK, ECC, EA, MA and LE corresponding to ECK private parameters, ECC private parameters, Emotion Attention, Multi-head Attention, and Labeling Embedding mechanisms, respectively.

The performance of above models are shown in Tables 5 and 6. As expected, the results in F1-score of the sub-networks all drop. This clearly demonstrates the usefulness of these components. Both Full-ECK and Full-ECC are worse which confirms that joint training of two tasks is helpful for learning the effective features. On the one hand, the word label predicted by ECK task is able to provide the important emotion cause signal which help inferring that whether the clause is the emotion cause clause. For example, if there are some words are predicted as emotion cause keywords, the model will increase the probability of the current clause being predicted as an emotion cause clause. On the other hand, the clause representation obtained by ECC task is able to give a positive impact for emotion cause keyword prediction. That is, if the current clause is predicted as emotion cause clause, the words in this clause more likely be the emotion cause keywords. Full gains 1.6% improvement in $F1$ over Full-EA, which indicates that the emotion attention can provide important information for emotion cause keywords extraction. In Table 5, when removing the word label embedding mechanism, the $F1$ score of Full-LE decreases 2.9%, which indicates the word label embedding from ECK task is conducive to ECC task. Also, Full gains 10.5% improvement in $F1$ over Full-MA indicating that the ECC task can enhance the performance of the ECK task by multi-head attention mechanism in Table 6. We also find that Full-ECK outperforms the strong baseline RHNN, which maybe due to the case that considering the emotion word and context differently is effective.

Table 5. Ablation test results of ECC task.

Model	P	R	$F1$
Full	**0.849**	**0.798**	**0.822**
Full-ECK	0.807	0.786	0.796
Full-EA	0.818	0.821	0.819
Full-LE	0.830	0.761	0.793
Full-MA	0.816	0.779	0.796

Table 6. Ablation test results of ECK task.

Model	P	R	$F1$
Full	**0.714**	**0.774**	**0.742**
Full-ECC	0.662	0.690	0.674
Full-EA	0.689	0.771	0.726
Full-LE	0.696	0.745	0.718
Full-MA	0.621	0.655	0.637

4.4 Case Study

To show how emotion attention and self-attention mechanisms work, we visualize the attention weights α_{ij} (in Eq. (4)) and β_{ij} (in Eq. (5)) with heatmap. Example 2 illustrates the detail with a training example.

Example 2. $[x_1]$: 后士凤心中充满感激。$[x_2]$: 她说: $[x_3]$: 虽然我们并不熟悉, $[x_4]$: 但他却**为我拉小提琴**, $[x_5]$: 我十分开心。*(***In English:** $[x_1]$: Shifeng Hou's heart is full of gratitude. $[x_2]$: She said: $[x_3]$: we are not familiar, $[x_4]$: but he **plays the violin for me**, $[x_5]$: and I'm very happy.*)*

Figure 2(a) and (b) represent the attention distribution of emotion word and emotion context to the each word of x_4. In Fig. 2(a), "*but*", and "." have low attention score as they are indeed irrelevant with respect to the emotion cause expression. Figure 2(b) shows that the words "*for*" and "*me*" in clause x_4 are paid more attention by the emotion context, which means that the emotion cause has a close relation with these two words. From Fig. 2(a) and (b), we can easily find the words "*me*", "*plays*", "*the violin*" in the clause x_4 have higher attention weights than "*but*" and punctuation ".", implying that the words, which help express the cause, are more important and thus captured by the emotion attention mechanism. These again verify the effectiveness of our proposed emotion attention mechanism on emotion cause analysis.

4.5 Error Analysis

We notice that for some passages which have the long distance between the emotion word and the cause, our model may have a difficulty in detecting the correct emotion cause keywords. We show an example to illustrate this situation (see Example 3). From the example, we can find the emotion cause of the emotion "angry" is "the old lady who was helped up ran to the front of the bus and sat down on the ground". However, the emotion cause keywords obtained by our model is "Seeing this scene". It is a challenging task to properly model the words which have long-distance with the emotional expression. In the feature, we will explore different network architecture with consideration of the various relationship between the words and emotion expression.

Example 3. 没想到, 徐连林刚准备发动汽车离开车站，那位**被扶起的老太以迅雷不及掩耳之势跑到了公交车前一屁股坐在了地上**. 站在公交车前部的乘客都将这一幕看得一清二楚，看到这一幕，车上的乘客立刻炸开了锅，激烈争论起来。其中一部分乘客很气愤，一边数落徐连林“不该多事”一边给他“上课”：“我叫你们别去管这事吧。”
In English: Unexpectedly, when Lianlin Xu was just about to activate the car and leave the station, **the old lady who was helped up ran to the front of the bus and sat down on the ground**. The passengers standing in the front of the bus saw the scene clearly. Seeing this scene, the passengers on the bus immediately burst into a boiling pot and argued fiercely. Some of the passengers were very angry. They accused Lianlin Xu of “not being too busy” while giving him a lesson: “I told you not to take care of this.”

5 Conclusions

In this paper, we study the multi-task learning approach to identify emotion cause at clause level and word level simultaneously. We propose an effective multi-head attention based multi-task learning network, which utilizes shared-private feature extractor, multi-head attention mechanism and label embedding mechanism to enable two tasks to interact with each other for better learning the task-oriented representations. Results on benchmark dataset for ECA task demonstrate that our model can effectively extract multi-level emotion causes, and outperform the strong QA-based system and other strong ECA baselines by large margins. In the future, we plan to focus on extracting the specific cause(s) in a more accurate granularity for improving emotion cause analysis.

References

1. Chen, Y., Lee, S.Y.M., Li, S., Huang, C.R.: Emotion cause detection with linguistic constructions. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 179–187 (2010)
2. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) EMNLP 2014, pp. 1724–1734 (2014)
3. Cui, Y., et al.: A span-extraction dataset for Chinese machine reading comprehension. arXiv preprint arXiv:1810.07366 (2018)
4. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT 2019, pp. 4171–4186 (2019)
5. Ding, Z., He, H., Zhang, M., Xia, R.: 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 (2019)
6. Ding, Z., Xia, R., Yu, J.: ECPE-2D: emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) ACL, pp. 3161–3170 (2020)

7. Fan, C., et al.: A knowledge regularized hierarchical approach for emotion cause analysis. In: EMNLP-IJCNLP 2019, pp. 5618–5628 (2019)
8. Fan, C., Yuan, C., Du, J., Gui, L., Yang, M., Xu, R.: Transition-based directed graph construction for emotion-cause pair extraction. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) ACL 2020, pp. 3707–3717 (2020)
9. Gao, K., Xu, H., Wang, J.: Emotion cause detection for Chinese micro-blogs based on ECOCC model. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 3–14 (2015)
10. Gui, L., Hu, J., He, Y., Xu, R., Lu, Q., Du, J.: A question answering approach to emotion cause extraction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1953–1602 (2017)
11. Gui, L., Wu, D., Xu, R., Lu, Q., Zhou, Y.: Event-driven emotion cause extraction with corpus construction. In: EMNLP, pp. 1639–1649 (2016)
12. Gui, L., Yuan, L., Xu, R., Liu, B., Lu, Q., Zhou, Y.: Emotion cause detection with linguistic construction in Chinese Weibo text. In: Natural Language Processing and Chinese Computing, pp. 457–464 (2014)
13. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
14. Hu, G., Lu, G., Zhao, Y.: FSS-GCN: a graph convolutional networks with fusion of semantic and structure for emotion cause analysis. Knowl. Based Syst. **212**, 106584 (2021)
15. Hu, M., Peng, Y., Huang, Z., Li, D.: A multi-type multi-span network for reading comprehension that requires discrete reasoning. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) EMNLP-IJCNLP 2019, pp. 1596–1606 (2019)
16. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations, pp. 1–15 (2015)
17. Lee, S.Y.M., Chen, Y., Huang, C.R.: A text-driven rule-based system for emotion cause detection. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 45–53 (2010)
18. Lee, S.Y.M., Ying, C., Huang, C.R., Li, S.: Detecting emotion causes with a linguistic rule-based approach. Comput. Intell. **29**(3), 390–416 (2013)
19. Li, W., Hua, X.: Text-based emotion classification using emotion cause extraction. Expert Syst. Appl. **41**(4), 1742–1749 (2014)
20. Li, X., Feng, S., Wang, D., Zhang, Y.: Context-aware emotion cause analysis with multi-attention-based neural network. Knowl.-Based Syst. **174**, 205–218 (2019)
21. Li, X., Song, K., Feng, S., Wang, D., Zhang, Y.: 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, pp. 4752–4757 (2018)
22. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Bengio, Y., LeCun, Y. (eds.) ICLR 2013 (2013)
23. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
24. Russo, I., Caselli, T., Rubino, F., Boldrini, E., Martínez-Barco, P.: EMOCause: an easy-adaptable approach to emotion cause contexts. In: Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pp. 153–160 (2011)

25. Tang, H., Ji, D., Zhou, Q.: Joint multi-level attentional model for emotion detection and emotion-cause pair extraction. Neurocomputing **409**, 329–340 (2020)
26. Xia, R., Ding, Z.: Emotion-cause pair extraction: a new task to emotion analysis in texts. In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) ACL 2019, pp. 1003–1012 (2019)
27. Xia, R., Zhang, M., Ding, Z.: RTHN: A RNN-transformer hierarchical network for emotion cause extraction. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019 (2019)
28. Xu, B., Lin, H., Lin, Y., Diao, Y., Yang, L., Xu, K.: Extracting emotion causes using learning to rank methods from an information retrieval perspective. IEEE Access **7**, 15573–15583 (2019)
29. Xu, R., et al.: A new emotion dictionary based on the distinguish of emotion expression and emotion cognition. J. Chinese Inf. Process. **27**(6), 82–90 (2013)
30. Yu, J., Liu, W., He, Y., Zhang, C.: A mutually auxiliary multitask model with self-distillation for emotion-cause pair extraction. IEEE Access **9**, 26811–26821 (2021)
31. Yu, X., Rong, W., Zhang, Z., Ouyang, Y., Xiong, Z.: Multiple level hierarchical network-based clause selection for emotion cause extraction. IEEE Access **7**, 9071–9079 (2019)