# A Unified Sequence Labeling Model for Emotion Cause Pair Extraction

**Xinhong Chen[1], Qing Li[2], Jianping Wang[1]**
[1] Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
[2] Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong
xinhchen2-c@my.cityu.edu.hk, jianwang@cityu.edu.hk
qing-prof.li@polyu.edu.hk

## Abstract

Emotion-cause pair extraction (ECPE) aims at extracting emotions and causes as pairs from documents, where each pair contains an emotion clause and a set of cause clauses. Existing approaches address the task by first extracting emotion and cause clauses via two binary classifiers separately, and then training another binary classifier to pair them up. However, the extracted emotion-cause pairs of different emotion types cannot be distinguished from each other through simple binary classifiers, which limits the applicability of the existing approaches. Moreover, such two-step approaches may suffer from possible cascading errors. In this paper, to address the first problem, we assign emotion type labels to emotion and cause clauses so that emotion-cause pairs of different emotion types can be easily distinguished. As for the second problem, we reformulate the ECPE task as a unified sequence labeling task, which can extract multiple emotion-cause pairs in an end-to-end fashion. We propose an approach composed of a convolution neural network for encoding neighboring information and two Bidirectional Long-Short Term Memory networks for two auxiliary tasks. Experiment results demonstrate the feasibility and effectiveness of our approaches.

## 1 Introduction

Recently, causal relationships between human emotions and their corresponding causes have received a lot of attention. Recognizing the causes of a specific emotion in a document is considered as more useful than only identifying the emotions, due to the great potential of helping people make reasonable decisions and avoid unnecessary loss (Gui et al., 2017; Li et al., 2018; Xia et al., 2019; Ding et al., 2019). The first task concerning causal relationships between emotions and causes is the Emotion Cause Extraction (ECE) task proposed by Gui et al. (2016), which aims to extract the causes for the annotated emotion in a document. To avoid the limitation that emotions must be annotated before cause extraction in the ECE task, Xia and Ding (2019) proposed the Emotion-Cause Pair Extraction (ECPE) task to extract emotions and causes as pairs from documents without the emotion annotations, where the extracted pairs are called "emotion-cause pairs (ECPs)". To address the ECPE task, they proposed a two-step framework, where emotions and causes are first extracted separately by two binary classifiers, and then another binary classifier is trained to pair them up. However, such two-step approaches may suffer from two major shortcomings as follows.

Firstly, binary classifiers cannot distinguish ECPs of differernt emotion types, which limits the applicability of the proposed approaches since the extracted ECPs may not be directly usable in real-world applications, such as discovery and analysis of the possible causes for a specific emotion type. To further see this, let us consider a document with two ECPs shown in Table 1, where the red clauses (i.e., Clause 3 and 7) are emotion clauses and the blue clauses (i.e., Clause 2 and 8) are cause clauses. Following the two-step approaches proposed by Xia and Ding (2019), the extracted ECPs are: (3, 2) and (7, 8). However, the model itself does not know what type of emotion is contained in these two ECPs. It would be desirable to directly distinguish the extracted emotion and cause clauses by their related emotion types,

Table 1: An example document in the ECPE dataset

| No. | Document Content | Clause Label | |
|---|---|---|---|
| | | Emotion | Cause |
| 1 | After 10 years of working and struggling, | 0 | 0 |
| 2 | Wu finally got his legal citizenship in Shenzhen | 0 | 1 |
| 3 | and his family was happy about this news. | 1 | 0 |
| 4 | According to the laws, | 0 | 0 |
| 5 | his wife could also get her legal citizenship after two years. | 0 | 0 |
| 6 | However, half year later, | 0 | 0 |
| 7 | Wu was diagnosed with lung cancer and he was very depressed, | 1 | 0 |
| 8 | since his wife may not be able to get her citizenship if he died soon. | 0 | 1 |

so as to only pair up those emotion and cause clauses of the same emotion type and extract ECPs with more useful information. To this end, we propose in this paper to assign emotion type labels (e.g., Happiness, Sadness, etc.) to emotion and cause clauses, so that ECPs of different emotion types can be easily distinguished.

Secondly, a two-step framework such as the one proposed by Xia and Ding (2019) may produce possible cascading errors, which reduces the training accuracy. In the first step, their approaches extract emotion and cause clauses via two binary classifiers, and generate candidate ECPs for the second step by pairing up the extracted emotion clauses with the extracted cause clauses one by one. However, these candidate ECPs may already contain incorrect emotion or cause clauses. Therefore, to avoid such cascading errors, we propose to extract ECPs from documents by an end-to-end approach.

Taking these two problems into consideration, the main challenge is to extract multiple ECPs of different emotion types in an end-to-end fashion. In this paper, we reformulate the ECPE task as a sequence labeling task, which can assign different labels to different clauses and extract multiple ECPs of different emotion types simultaneously by simply pairing up clauses with matching labels. To pair up emotion clauses and cause clauses, we design a unified label for each clause, which is composed of two types of labels: a causal identity label denoting the clause being a cause or emotion clause (e.g., `Cause(C)`, `Emotion(E)`), and an emotion type label denoting the related emotion type of the clause (e.g., `Happiness(H)`, `Sadness(Sa)`).

With such a formulation, we devise our end-to-end sequence labeling approach based on a stacked neural network framework, namely **IE-CNN**, where the upper network predicts the unified labels and the lower networks perform auxiliary prediction tasks. Specifically, since our unified labels consist of causal identity labels and emotion type labels, two Bidirectional Long-Short Term Memory (BiLSTM) networks are employed in the lower level to separately predict these two types of labels for each clause, and obtain feature vectors with causal identity information and emotion type information for the subsequent unified label prediction. Then, based on the observations made by Ding et al. (2019) that most emotion clauses and their corresponding cause clauses appear near each other, the resultant feature vectors of the two BiLSTMs are concatenated and fed to a convolutional neural network (CNN) in the upper level to encode the neighboring information. Experiment results show that our approach significantly outperforms the existing baseline models. The main contributions of this work can be summarized as follows:

- We assign fine-grained emotion type labels to emotion and cause clauses so that emotion-cause pairs of different emotion types can be easily distinguished.

- We reformulate the ECPE task as a sequence labeling task with specially designed unified labels, so as to extract multiple emotion-cause pairs of different emotion types in an end-to-end fashion.

- We propose a sequence labeling approach based on a stacked neural network framework, incorporating two Bidirectional Long-Short Term Memory networks and one convolution neural network.

## 2 Problem formulation

In this section, we formally redefine the ECPE task. In the benchmark ECPE dataset, the following observations form the basis of our task reformulation. Firstly, there are possibly multiple ECPs in a document, and each emotion clause can be paired up with multiple cause clauses. Secondly, a clause can be both a cause clause and an emotion clause, and each clause will only be associated with at most one emotion type, which means that there will not be multiple emotion clauses of the same emotion type, or a single clause with multiple emotion types in one document. Therefore, formulating the ECPE task as a sequence labeling task is a suitable choice since ECPs of different emotion types can be extracted simultaneously from the final label sequence.

We design a set of unified labels to pair up emotion and cause clauses. Specifically, except "O" which denotes that a clause is neither a cause clause nor an emotion clause, each unified label contains a causal identity label and an emotion type label. The set of causal identity labels is denoted by: $\mathcal{Y}^I = \{$Outside(O), Cause(C), Emotion(E), Both-Cause&Emotion(B)$\}$, and the set of emotion type labels is denoted by: $\mathcal{Y}^E = \{$O, Happiness(H), Sadness(Sa), Anger(A), Disgust(D), Surprise(Su), Fear(F)$\}$.

With these two sets of labels, the unified label set is constructed as: $\mathcal{Y}^U = \{$O, C-H, C-Sa, C-A, C-D, C-Su, C-F, E-H, E-Sa, E-A, E-D, E-Su, E-F, B-H, B-Sa, B-A, B-D, B-Su, B-F$\}$. Using these unified labels, emotion and cause clauses of the same emotion type can be easily paired up. For example, C-H denotes the clause being a cause clause of emotion Happiness, E-Sa denotes the clause being an emotion clause of Sadness, while B-A denotes the clause being both cause and emotion clauses of Anger. These specially-designed unified labels enable the proposed model to handle documents with multiple ECPs of different emotion types as desired.

Thus, ECPE can be redefined as: given a clause sequence $X = \{x_1, ..., x_T\}$ with $T$ clauses, predict a label sequence $Y^U = \{y_1^U, ..., y_T^U\}$, where $y_t^U \in \mathcal{Y}^U$.

## 3 Approach

As shown in Figure 1, our proposed **IE-CNN** is based on a stacked neural network framework, where the upper network can utilize the useful information delivered from the lower networks. Since our unified labels contain two parts, namely, causal identity labels and emotion type labels, we first employ two BiLSTM networks, BiLSTM$^I$ and BiLSTM$^E$, in the lower level to predict these two types of labels for each clause, respectively. Then, the hidden representations of two BiLSTM networks are concatenated to construct feature vectors with both causal identity and emotion type information, which are further fed to the CNN layer in the upper level to encode the neighboring information. In the following, we introduce our proposed framework in detail.

### 3.1 Clause Embedding

To obtain an embedding vector for each word, we use the word embedding vectors released by Xia and Ding (2019), which are trained using the word2vec algorithm (Mikolov et al., 2013). As for encoding word embedding vectors into a clause embedding vector, we adopt a word-level BiLSTM network with an attention module, which is capable of generating an informative vector for each clause by passing words' information along the clause forwards and backwards. Specifically, for the $t$-th clause, its clause embedding vector $x_t$ is calculated as follows:

$$\begin{aligned} x_t &= \sum_{i=1}^{n} \alpha_i \cdot BiLSTM(v_{w_i}) \\ \alpha_i &= \frac{e^{W_c \cdot v_{w_i}}}{\sum_{i=1}^{n} e^{W_c \cdot v_{w_i}}} \end{aligned} \tag{1}$$

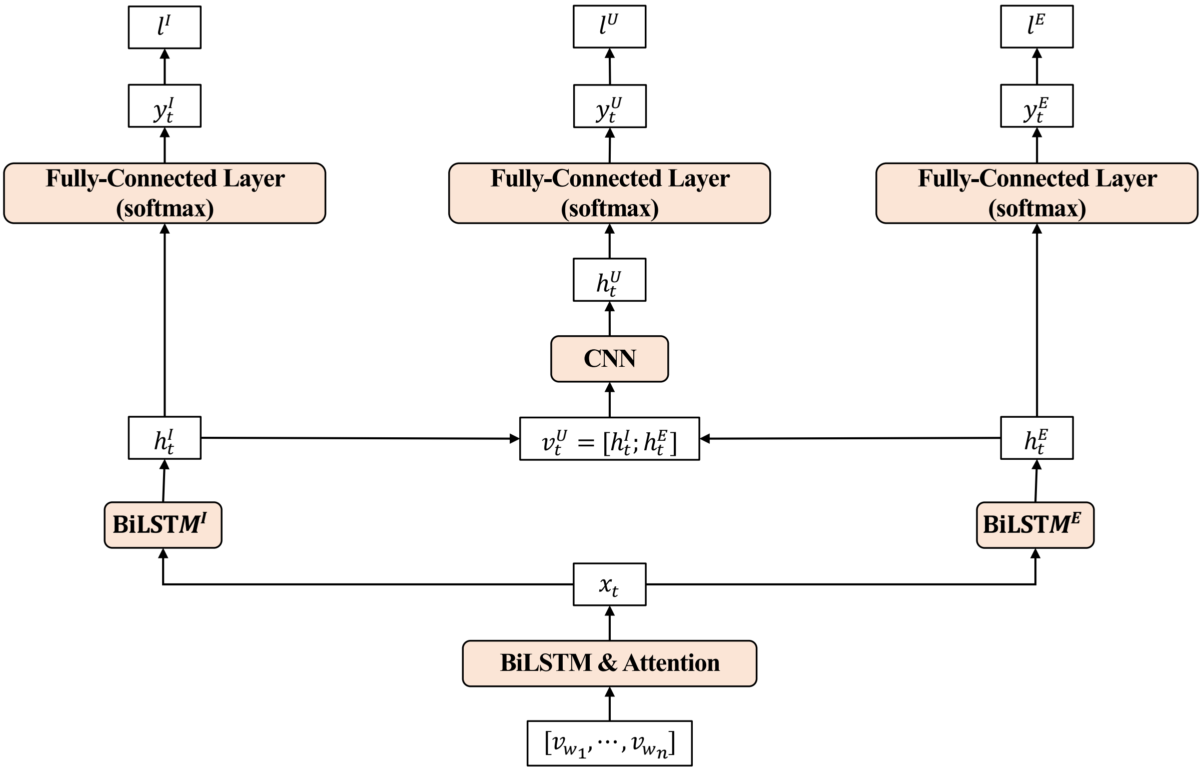

Figure 1: Architecture of our **IE-CNN** model

where $n$ is the length of the clause, $v_{w_i}$ is the word embedding vector of the $i$-th word, and $W_c$ is a trainable weight matrix for attention score calculation.

### 3.2 Cause Identity Label and Emotion Type Label Prediction

In order to obtain an embedding vector with causal identity and emotion type information, we employ two BiLSTM networks to predict a causal identity label and an emotion type label for each clause. Specifically, for the $t$-th clause, we pass its embedding vector $x_t$ to the two BiLSTM networks, $BiLSTM^I$ and $BiLSTM^E$, in parallel. Then, the resultant embedding vectors $h_t^I$ and $h_t^E$ are fed to fully-connected layers with softmax activation function to predict a causal identity label $y_t^I$ and an emotion type label $y_t^E$, respectively.

$$\begin{aligned} h_t^I &= \text{BiLSTM}^I(x_t) \\ h_t^E &= \text{BiLSTM}^E(x_t) \\ y_t^I &= \text{Softmax}(W^I h_t^I) \\ y_t^E &= \text{Softmax}(W^E h_t^E) \end{aligned} \tag{2}$$

where $y_t^I \in \mathcal{Y}^I$, $y_t^E \in \mathcal{Y}^E$, $W^I$ and $W^E$ are trainable weight matrices.

### 3.3 Unified Label Prediction

According to an observation made by Ding et al. (2019), more than 97% of the emotions and their corresponding causes appear near each other. Therefore, we employ a CNN layer in the upper level to encode the neighboring information and further enhance each clause's embedding vector. Specifically, for the $t$-th clause, two hidden embedding vectors from the two BiLSTM networks are first concatenated and passed to the CNN layer to generate the neighborhood-encoded embedding vector $h_t^U$. Then, we pass $h_t^U$ to a fully-connected layer with softmax activation function or a Conditional Random Field

Table 2: An example document with our specially-designed unified labels

| No. | Document Content | Clause Label |
|---|---|---|
| 1 | After 10 years of working and struggling, | O |
| 2 | Wu finally got his legal citizenship in Shenzhen | C-H |
| 3 | and his family was happy about this news. | E-H |
| 4 | According to the laws, | O |
| 5 | his wife could also get her legal citizenship after two years. | O |
| 6 | However, half year later, | O |
| 7 | Wu was diagnosed with lung cancer and he was very depressed, | E-Sa |
| 8 | since his wife may not be able to get her citizenship if he died soon. | C-Sa |

(CRF) decoder layer to predict a unified label $y_t^U$ for the $t$-th clause.

$$\begin{aligned} v_t^U &= [h_t^I, h_t^E] \\ h_t^U &= \mathrm{CNN}(v_t^U) \\ y_t^U &= \mathrm{Softmax}(W^U h_t^U) \text{ or } y_t^U = \mathrm{CRF}(W^U h_t^U) \end{aligned} \tag{3}$$

where $W^U$ is a trainable weight matrix. The performance of these two label prediction layers will be discussed in Section 4.3.

### 3.4 Model training

All the components we have proposed above are differentiable, therefore we use gradient-based optimization methods to train the model. For the loss function, we apply the cross-entropy loss function for all subtasks, and the aggregated loss function is given as:

$$\mathcal{L}^* = -\frac{1}{\sum_{d=1}^{D} T_d} \sum_{d=1}^{D} \sum_{t=1}^{T_d} l_t^* = -\frac{1}{\sum_{d=1}^{D} T_d} \sum_{d=1}^{D} \sum_{t=1}^{T_d} \mathbb{I}(y_t^{*,g}) \circ y_t^*, * \in \{I, E, U\} \tag{4}$$

$$\mathcal{J}(\theta) = \mathcal{L}^I + \mathcal{L}^E + \mathcal{L}^U + \lambda ||\theta||^2 \tag{5}$$

where $D$ denotes the number of documents, $T_d$ denotes the number of clauses in the $d$-th document, $\mathbb{I}(y)$ denotes the one-hot vector with only the dimension of the ground-truth label being 1, $y_t^{*,g}$ denotes the ground truth label, $\circ$ denotes the element-wise production, $\theta$ denotes all model parameters, and $\lambda$ denotes the weight of the L2-regularization term. More training details are given in Section 4.1.

## 4 Experiments

In this section, we present our experimental studies to evaluate the feasibility and effectiveness of our proposed approach.

### 4.1 Dataset Description, Metrics and Experiment Setting

As mentioned earlier, existing two-step approaches such as the one proposed by Xia and Ding (2019) treat the extraction of emotion and cause clauses as two binary classification tasks, which cannot distinguish ECPs of different emotion types. In out work, to address this problem, we assign different emotion type labels to emotion and cause clauses. Since these emotion type labels are already contained in the original ECPE dataset, we directly utilize them to construct a unified label for every clause in every document. Also, it's notable that there are some documents where a single clause is related to multiple emotion types in the ECPE dataset. To assure that each clause has a unique label, we randomly select one emotion type

Table 3: Dataset details of the ECPE dataset

| | Number |
|---|---|
| # of documents | 1945 |
| # of documents with one ECP | 1746 |
| # of documents with two ECPs | 177 |
| # of documents with three ECPs | 22 |

for these clauses. An example document with our specially-designed unified labels is shown in Table 2, and the dataset details are shown in Table 3.

For the evaluation, we follow the same process of (Xia and Ding, 2019) to evaluate the performance of emotion extraction, cause extraction, and pair extraction. We randomly divide the whole dataset into 10 folds, where 9 folds of them are selected for training and the remaining fold is used for testing. We conduct each experiment 10 times for each of our proposed model and each time we use one fold of the dataset as the testing set, in which way 10 folds can all be used for testing once. Following such an experiment design, we report the average scores of ten experiments to reduce the impact of randomness. The metrics used for all tasks are Precision (P), recall (R) and F1 scores.

$$\begin{aligned} P &= \frac{\sum correct\_ECPs}{\sum predicted\_ECPs} \\ R &= \frac{\sum correct\_ECPs}{\sum annotated\_ECPs} \\ F1 &= \frac{2 * P * R}{P + R} \end{aligned} \tag{6}$$

To conduct fair comparison with the baselines, we follow the same experiment setting used in (Xia and Ding, 2019), utilizing the existing word vectors that were pre-trained with word2vec toolkit. Words outside of vocabulary are assigned with randomized vectors. The dimension of word embedding is 200 and the number of hidden units of all BiLSTM modules is set to 100. The filter size of the CNN model is 3, while the number of filters is 128. The number of heads in Self-Attention module is set to 2. All weights and biases are randomly initialized by a uniform distribution $\mathcal{U}(-0.01, 0.01)$.

As for the training, we apply Adam (Kingma and Ba, 2015) to optimize our loss function, with batch size set to 32 and learning rate set to 0.005. Also, for regularization, dropout is applied with dropout rate set to 0.2, and a L2-norm regularization term is added to constraint the softmax parameters, where the weight of the regularization term $\lambda$ is set to $1e^{-5}$.

### 4.2 Baseline Models

We compare our proposed approach with the following methods (i.e., baselines):

- **Indep / Inter-CE / Inter-EC**: these two-step approaches proposed by Xia and Ding (2019) first extract emotions and causes separately to form emotion-cause pair candidates, and then train a binary classifier to filter out pairs without causal relations.

- **Softmax-Joint**: this is a sequence labeling baseline model that jointly trains two separate BiLSTMs for predicting causal identity labels and emotion type labels, using softmax decoding layers to get these two types of labels.

- **CRF-Joint**: this model by Mitchell et al. (2013) is almost the same as Softmax-Joint, except for using CRF as the decoding layer.

- **BiLSTM-Softmax**: this is another sequence labeling baseline model that trains a single standard BiLSTM model to directly predict a unified label for each clause by using a softmax decoding layer.

- **BiLSTM-CRF**: proposed by Lample et al. (2016), this is also very similar to BiLSTM-Softmax, except for using a CRF decoding layer instead of a softmax decoding layer.

- **Self-Attention**: this is a baseline model with 1-layer-multi-head self-attention module proposed by Vaswani et al. (2017). This model first obtains each clause's embedding vector by the BiLSTM&Attention module in our framework, and then passes these embedding vectors to the self-attention module to get context-encoded embedding vectors for the final prediction.
- **Multi-Self-Attention**: this is almost the same as the Self-Attention baseline, but with multi-layer-multi-head self-attention module.[1]
- **IE-BiLSTM / IE-Self-Attention**: these are two variant models of our proposed approach, which uses BiLSTM and Self-Attention module, respectively, instead of CNN in the unified label prediction component.
- **RANKCP**: proposed by Wei et al. (2020), they made use of the graph attention network to propagates information among clauses and ranked the candidate ECPs with the learned pair representations to get the prediction results.
- **E2EECPE**: proposed by Song et al. (2020), they treated the ECPE task as a link prediction problem and proposed to make use of a biaffine attention module to model the relationships between any two clauses.
- **LAE-Joint-MANN-BERT**: proposed by Tang et al. (2020), they enhanced the encoder with BERT and proposed a joint model for the ECPE task.
- **ECPE-2D (Inter-EC)**: proposed by Ding et al. (2020), they deployed the self-attention module to calculate the biaffine attention matrix among all clauses and applied a 2D transformer to achieve the interaction between ECPs.
- **ECPE-2D-BERT (Inter-EC)**: this is almost the same baseline model as ECPE-2D, except for using BERT to obtain word embedding vectors and clause embedding vectors.
- **TDGC-LSTM**: proposed by Fan et al. (2020) recently, they formulated the recognition of the ECPs as a set of actions and transitions and built a directed graph to model the transition process. Compared with the below baseline, this one uses LSTM to encode words' embedding vectors into clause embedding vectors.
- **TDGC-BERT**: this is almost the same model as TDGC-LSTM, except for using BERT to obtain word embedding vectors and clause embedding vectors.

### 4.3 Overall performance

We conduct the performance evaluation of all the approaches on three tasks, i.e., emotion extraction, cause extraction and pair extraction. As shown in Table 4, all the approaches perform competitively in the emotion extraction and cause extraction tasks, yet the best F1 score of emotion extraction is achieved by Ding et al. (2020). As our main focus here is on the pair extraction task, we shall compare the approaches based on the performance of pair extraction in the following.

Firstly, compared with the two-step approaches proposed by Xia and Ding (2019) (i.e., Indep, Inter-CE, and Inter-EC), the Joint baselines and the Unified baselines of sequence labeling already achieve better F1 scores on pair extraction, which demonstrates the effectiveness of formulating the ECPE task as a sequence labeling problem. Specifically, unified baselines perform generally better than the joint baselines. However, it is noticeable that the performance of "Self-Attention" is unexpectedly low. Also, with the increase of the layers in self-attention module, we observe a decrease of the performance, as shown by "Multi-Self-Attention" in Table 4. We think such observations are due to insufficient training examples for deeper neural networks and uninformative words' embedding vectors.

[1] Here the number of layer is set to 2 for demonstration, since a larger number of layers causes a decrease in performance according to our preliminary experiments.

Table 4: Performance comparison

| Catogory | Model | emotion extraction | | | cause extraction | | | pair extraction | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | P(%) | R(%) | F1(%) | P(%) | R(%) | F1(%) | P(%) | R(%) | F1(%) |
| Baselines without BERT | Indep | 84.14 | 80.27 | 82.10 | 70.01 | 56.99 | 62.67 | 67.99 | 50.65 | $57.82^{\ddagger}$ |
| | Inter-CE | 85.17 | 80.60 | 82.76 | 68.28 | 58.68 | 62.89 | 68.26 | 53.04 | $59.47^{\ddagger}$ |
| | Inter-EC | 82.63 | 81.32 | 81.92 | 69.24 | 61.04 | 64.75 | 66.11 | 56.22 | $60.64^{\ddagger}$ |
| | RANKCP | 87.03 | 84.06 | 85.48 | 69.27 | 67.43 | 68.24 | 66.98 | **65.46** | 66.10 |
| | E2EECPE | 85.95 | 79.15 | 82.38 | 70.62 | 60.30 | 65.03 | 64.78 | 61.05 | 62.80 |
| | ECPE-2D (Inter-EC) | 85.12 | 82.20 | 83.58 | 72.72 | 62.98 | 67.38 | 69.60 | 61.18 | 64.96 |
| | TDGC-LSTM | 80.80 | 84.39 | 82.56 | 67.42 | 65.34 | 66.36 | 65.15 | 63.54 | 64.34 |
| Baselines with BERT | LAE-Joint-MANN-BERT | **89.90** | 80.00 | 84.70 | - | - | - | 71.10 | 60.70 | 65.50 |
| | ECPE-2D-BERT (Inter-EC) | 86.27 | **92.21** | **89.10** | 73.36 | **69.34** | **71.23** | 72.92 | 65.44 | 68.89 |
| | TDGC-BERT | 87.16 | 82.44 | 84.74 | **75.62** | 64.71 | 69.74 | **73.74** | 63.07 | 67.99 |
| Joint sequence labeling baselines | Softmax-joint | 86.09 | 78.39 | 82.00 | 72.19 | 57.53 | 63.78 | 69.23 | 59.89 | $64.07^{\dagger}$ |
| | CRF-joint | 84.95 | 79.41 | 82.02 | 70.93 | 58.33 | 63.95 | 66.45 | 62.27 | $64.29^{\dagger}$ |
| Unified sequence labeling baselines | BiLSTM-Softmax | 85.35 | 77.69 | 81.29 | 73.86 | 53.17 | 61.76 | 69.61 | 60.69 | $64.63^{\dagger}$ |
| | BiLSTM-CRF | 86.48 | 77.79 | 81.88 | 72.50 | 57.24 | 63.92 | 67.85 | 61.43 | $64.48^{\dagger}$ |
| | Self-Attention | 84.11 | 78.15 | 80.98 | 72.99 | 44.45 | 55.08 | 66.02 | 59.22 | $62.33^{\dagger}$ |
| | Multi-Self-Attention | 83.35 | 78.24 | 80.68 | 70.60 | 48.44 | 57.20 | 64.03 | 59.63 | $61.61^{\dagger}$ |
| Ours | IE-BiLSTM+Softmax | 85.95 | 78.05 | 81.73 | 73.53 | 54.80 | 62.54 | 68.43 | 62.74 | 65.46 |
| | IE-BiLSTM+CRF | 85.76 | 78.32 | 81.80 | 73.14 | 58.92 | 65.17 | 69.33 | 61.43 | 65.04 |
| | IE-Self-Attention+Softmax | 84.74 | 77.29 | 80.82 | 73.76 | 51.82 | 60.24 | 70.41 | 60.06 | 64.66 |
| | IE-Self-Attention+CRF | 85.67 | 77.81 | 81.49 | 72.31 | 58.56 | 64.51 | 68.79 | 61.09 | 64.59 |
| | IE-CNN+Softmax | 86.19 | 78.05 | 81.80 | 72.14 | 56.54 | 63.25 | 67.10 | 64.07 | 65.55 |
| | IE-CNN+CRF | 86.14 | 78.11 | 81.88 | 73.48 | 58.41 | 64.96 | 71.49 | 62.79 | **66.86** |

[*] F1 scores of pair extraction with superscripts of † and ‡ denote that our proposed approaches perform significantly better than these baseline models with paired t-test where $p$ is smaller than 0.05 and 0.001, respectively.

[*] “-” denotes that the evaluation metric is not reported in the original paper.

[*] The bold values indicate the best performance, where our proposed model “IE-CNN+CRF” gives the highest F1 score in pair extraction compared with all baselines without BERT.

[*] Red underlined values indicate better performance compared with our proposed model, but our model outperforms these two baselines when without BERT.

Secondly, compared with all existing baseline without BERT, our proposed models generally achieve better performance, among which “IE-CNN+CRF” gives the highest F1 score (i.e., 0.6686) for the pair extraction task. We think that the main reason of such performance lies in the usage of CNN module in our model, which has the ability to maintain the emotion information within a small sliding window and enable the neighboring clauses to share the same emotion labels. Also, as highlighted by the superscripts and p-value attached in the table, our “IE-CNN+CRF” model significantly outperforms these baseline models in terms of F1 scores of the pair extraction task.

When compared with two baseline models using BERT, our proposed models achieve a lower F1 score in terms of pair extraction (see the red underline values in the table), which we think is mainly due to the usage of BERT in these two baselines since our model outperforms them when without BERT. We believe that if we replace the BiLSTM in the clause embedding module of our framework with BERT, we can further improve the prediction performance of our model and outperform these two baselines.

Thirdly, we further replace CNN in the unified label prediction component by a BiLSTM network and a Self-Attention module to create two variants of our model (i.e., IE-BiLSTM and IE-Self-Attention), where IE-BiLSTM achieves a slightly lower F1 score in pair extraction and IE-Self-Attention performs the worst. This is possibly due to that Self-Attention module may affect the training of the two BiLSTMs used for auxiliary tasks, which results in insufficient training of the whole model. Comparing the proposed models containing a softmax decoding layer with those containing a CRF decoding layer, using

Table 5: Ablation study of our approach

| Catogory | Model | emotion extraction | | | cause extraction | | | pair extraction | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | P(%) | R(%) | F1(%) | P(%) | R(%) | F1(%) | P(%) | R(%) | F1(%) |
| Proposed approches | IE-CNN+Softmax | 86.19 | 78.05 | 81.80 | 72.14 | 56.54 | 63.25 | 67.10 | **64.07** | 65.55 |
| | IE-CNN+CRF | 86.14 | 78.11 | 81.88 | 73.48 | 58.41 | 64.96 | **71.49** | 62.79 | **66.86** |
| Variant approches | IE+Softmax | 85.23 | 78.73 | 81.83 | 71.51 | 55.85 | 62.65 | 68.31 | 61.68 | 64.81 |
| | IE+CRF | 83.76 | 78.39 | 80.95 | 70.72 | 54.60 | 61.55 | 68.63 | 62.19 | 65.14 |
| | I-CNN+Softmax | 86.07 | 78.01 | 81.80 | **75.19** | 57.21 | 64.90 | 67.27 | 63.25 | 65.20 |
| | I-CNN+CRF | 85.44 | 78.46 | 81.74 | 73.00 | 59.61 | 65.54 | 67.15 | 63.30 | 65.17 |
| | E-CNN+Softmax | **87.01** | 75.87 | 81.05 | 73.66 | 54.68 | 62.63 | 67.09 | 59.60 | 63.07 |
| | E-CNN+CRF | 85.53 | **78.87** | **82.05** | 67.95 | 59.70 | 63.56 | 66.25 | 61.73 | 63.91 |
| | CNN+Softmax | 85.68 | 78.73 | 82.03 | 73.46 | 59.12 | 65.30 | 66.70 | 63.56 | 65.09 |
| | CNN+CRF | 85.46 | 78.31 | 81.69 | 71.72 | **60.31** | **65.42** | 68.20 | 62.62 | 65.29 |

CRF with CNN to get the final predicted label leads to a better performance.

### 4.4 Ablation Study

To further demonstrate the effectiveness of our proposed model, we conduct ablation study and report the performance of multiple variants of our approach in Table 5.

Specifically, the first type of variant, "IE", is to remove the CNN layer and directly use the concatenation of the two resultant feature vectors from two BiLSTMs to predict the unified labels. The results show that the removal of the CNN layer leads to a decrease of the F1 score in pair extraction. In other words, the CNN layer is important to encode the neighboring information for better extraction of the emotion-cause pairs.

As for the second type of variant, instead of using two BiLSTMs to predict causal identity labels and emotion type labels, we only use 1 BiLSTM to predict causal identity labels or emotion type labels, so as to see if the two BiLSTM networks can help capture more useful information for the unified label prediction. As shown by the results, "I-CNN" and "E-CNN" both perform worse than the original proposed model, where BiLSTM$^I$ seems to be more important since the performance drops more significantly without it.

The last type of variant is to remove the two BiLSTM networks entirely and only use CNN layer to get the embedding vectors for each clause, the result of which also indicates that these two BiLSTM networks are indispensable to better extract the emotion-cause pairs.

## 5 Related Work

Recently, there have been several works focusing on the ECPE task, aiming to propose end-to-end approaches to avoid the possible cascading errors brought by the two-step manner in the three approaches proposed by Xia and Ding (2019). Wei et al. (2020) proposed a RANKCP model which makes use of the graph attention network to propagates information among clauses and ranks the candidate ECPs based on the learned pair representations to get the prediction results. Tang et al. (2020) proposed a BERT-base model called LAE-Joint-MANN-BERT to deal with the emotion detection (ED) and the ECPE task jointly. Specifically, they calculated the biaffine attention values among all clauses to predict the probability of each pair being an emotion-cause pair. Similarly, Song et al. (2020) treated the ECPE task as a link prediction problem and proposed E2EECPE to address the ECPE task, which utilizes a biaffine attention module to model the relationships between any two clauses. The ECPE-2D model proposed by Ding et al. (2020) deployed a hierarchical self-attention module to calculate the biaffine attention matrix among all clauses and a 2D transformer to model the interaction among ECPs.

All of the above works are constructed based on calculating the attention value between any two clauses of the given document. Fan et al. (2020) proposed to address the ECPE task from a different angle by modeling the extraction of the ECP as performing a sequence of transitions and actions. Specifically, they constructed a Transitional Directed Graph for each given document and transformed the original dataset into sequences of transitions and actions, based on which they trained a model to predict the next state of the sequence given the current state and the predicted transition.

Different from the above works, we reformulate the ECPE task as a unified sequence labeling problem and design our unified labels based on two sets of labels: causal identity labels and emotion type labels. Such a formulation can help us address the two problems analyzed in Section 1 (i.e., uninformative extraction results and possible cascading error of the two-step approaches), and enable our model to extract multiple ECPs of different emotion types simultaneously.

A conventional solution for sequence labeling with two sets of labels is to train, jointly, two models to predict these two sets of labels separately (Mitchell et al., 2013; Zhang et al., 2015). However, using such approaches, the two sets of labels may not match correctly, resulting in meaningless labels. For example, a clause can be assigned with “Cause” as its causal identity label, and “Outside” as its emotion type label which means the clause is not related to any emotion. In this case, the predicted causal identity label and emotion type label cannot be combined to construct a meaningful unified label.

Therefore, a better solution is directly predicting unified labels so as to avoid the matching error (Lample et al., 2016; Li et al., 2019). In this regard, many existing frameworks (e.g., a stacked recurrent neural network framework proposed by Li et al. (2019)) have been proposed and are applicable to address the ECPE task. In our approach introduced in Section 3, we adopt a stacked neural network framework containing two BiLSTMs and a CNN, given that our unified labels contain two parts and encoding neighboring information may improve the task performance.

## 6 Conclusion

In this paper, we have investigated the ECPE task which aims to extract emotion-cause pairs from documents. To tackle the limitation that the existing two-step approaches cannot distinguish emotion-cause pairs of different emotion types, our proposed approach assigns emotion type labels to both emotion and cause clauses. Besides, we have reformulated the ECPE task as a sequence labeling task, and proposed a unified sequence labeling approach to extract multiple emotion-cause pairs in an end-to-end fashion. This allows us to handle documents with multiple types of emotions and avoid cascading errors compared to the existing baselines. Based on stacked neural networks, our proposed approach consists of two BiLSTM networks in the lower level to separately predict causal identity label and emotion type label, and a CNN layer in the upper level to encode the neighboring information. Experiment results show that our approach achieves better performance in pair extraction compared with the existing baseline without BERT, and we believe that adding BERT in our proposed model can further improve the prediction performance. Furthermore, ablation study well reveals the effectiveness of our devised framework and individual components.

## Acknowledgements

The research described in this paper has been supported by the Hong Kong Research Grants Council through a Collaborative Research Fund (Project No. C1031-18G) and a Research Impact Fund (Project No. R5060-19).

## References

Zixiang Ding, Huihui He, Mengran Zhang, and Rui Xia. 2019. From independent prediction to reordered prediction: Integrating relative position and global label information to emotion cause identification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 6343–6350.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020. Ecpe-2d: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3161–3170.

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

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

Lin Gui, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du. 2017. A question answering approach for emotion cause extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1593–1602, September.

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

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, June.

Xiangju Li, Kaisong Song, Shi Feng, Daling Wang, and Yifei Zhang. 2018. 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*, pages 4752–4757, October-November.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 6714–6721.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654, October.

Haolin Song, Chen Zhang, Qiuchi Li, and Dawei Song. 2020. End-to-end emotion-cause pair extraction via learning to link. *arXiv preprint arXiv:2002.10710.*

Hao Tang, Donghong Ji, and Qiji Zhou. 2020. Joint multi-level attentional model for emotion detection and emotion-cause pair extraction. *Neurocomputing*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Penghui Wei, Jiahao Zhao, and Wenji Mao. 2020. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3171–3181.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, July.

Rui Xia, Mengran Zhang, and Zixiang Ding. 2019. Rthn: A rnn-transformer hierarchical network for emotion cause extraction. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, pages 5285–5291.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2015. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 612–621, September.