Conditional Causal Relationships between Emotions and Causes in Texts

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

The causal relationships between emotions and causes in text have recently received a lot of attention. Most of the existing works focus on the extraction of the causally related clauses from documents. However, none of these works has considered the possibility that the causal relationships among the extracted emotion and cause clauses may only be valid under a specific context, without which the extracted clauses may not be causally related. To address such an issue, we propose a new task of determining whether or not an input pair of emotion and cause has a valid causal relationship under different contexts, and construct a corresponding dataset via manual annotation and negative sampling based on an existing benchmark dataset. Furthermore, we propose a prediction aggregation module with low computational overhead to fine-tune the prediction results based on the characteristics of the input clauses. Experiments demonstrate the effectiveness and generality of our aggregation module.

1 Introduction

Recently, the research on the causal relationships between human emotions and their corresponding causes has received much attention. Recognizing the causes of a specific emotion in a document is considered as more useful than only identifying the emotion, 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).

There are currently two well-designed tasks concerning the causal relationships between emotions and their causes, the Emotion Cause Extraction (ECE) task (Lee et al., 2010; Chen et al., 2010; Gui et al., 2016a,b) and the Emotion-Cause Pair Extraction (ECPE) task (Xia and Ding, 2019; Chen et al., 2018). Specifically, the ECE task focuses on

extracting the causes for a given emotion, while the ECPE task focuses on extracting emotions and the corresponding causes as pairs.

Despite their increasing popularity, both tasks only aim to extract the clauses containing causal relationships, and have neglected the possibility that the context clauses may be indispensable for the extracted clauses to have a valid causal relationship. Let us consider the following example:

- *Wu was diagnosed with advanced liver cancer at the beginning of 2014,*
- *since when he began to update his health condition in Microblog and has attracted much attention from many users.*
- *If Wu didn't update his microblog for a long time,*
- *people worried that he may have passed away.*
- *There was one time that Wu hadn't updated his microblog for about two months,*
- *and he had received a lot of messages concerned about his health conditions.*
- *...*

Figure 1: Example document, where the yellow-green color clauses are the context clauses with important information, the red one is the corresponding cause clause, and the blue one is the targeted emotion clause.

In the above example, the cause and emotion clauses may not have a causal relationship if we ignore the context clauses, since the reasons of not updating one's social media account can be more than the owner passing away (e.g., forgetting his/her password, using a new account, etc.). Only when extra information contained in the context clauses is available, can these two clauses have a valid causal relationship.

Therefore, it is essential to consider the context clauses as seriously as the targeted emotion and cause clauses when determining their causal relationships. With these context-dependent causal relations figured out, more complete and meaningful information about these causal relationships

can be extracted. Specifically, we can learn that some types of events may evoke different types of emotions under different circumstances, while some will always evoke one type of emotion under any circumstance. Such useful information about causal relationships can be beneficial in many emotion-related applications, such as accurately predicting one's emotions with context taken into consideration when a specific event occurs.

Despite the ubiquitousness of the context-dependent causal relationships described above, few works have paid attention to them. In this work, we articulate the importance of context in the problem of causal relationship recognition, and make the first step of studying such special causal relationships in text data.

We propose a task to determine whether or not the input pair of emotion and cause clauses has a causal relationship given some specific context clauses. As our task is new without any existing dataset available, we manually label the documents in the ECPE dataset constructed by Xia and Ding (2019) and follow the procedure of negative sampling (Mikolov et al., 2013) to build our own dataset. The constructed dataset can be further processed and used in some other important tasks that we aim to focus on in the future, such as quantifying the effect of context on the causal relationships.

Furthermore, we propose a prediction aggregation module with low computational overhead to fine-tune the prediction results according to the characteristics of the input clauses. The experiments on our constructed dataset demonstrate the effectiveness and generality of our aggregation module.

The contributions of this work can be summarized as follows.

- To address the issue that context can be indispensable for some causal relationships to be valid, we define a new task to determine whether or not an input pair of emotion and cause has a causal relationship under different contexts.

- Based on the ECPE dataset, we construct a dataset for our proposed task via manual annotation and negative sampling, which can also be used in some other important tasks, such as quantifying the effect of context on the causal relationships.

- We propose a general prediction aggregation module with low computational overhead, which can be used together with most existing models and significantly improve the prediction results on our proposed task.

2 Related Works

As known to all, context has been utilized in many text-related applications to provide semantic information and improve task performance (Kruengkrai et al., 2017; Kayesh et al., 2019; Zhou et al., 2016; Li and Mao, 2019).

In traditional causal reasoning, the term "context" is mostly discussed in the task of *causal effect estimation*, which is to estimate the influence of the cause variable on the effect variable (Guo et al., 2018). In some cases, the change of the effect variable relies not only on the cause variable, but also on some other relevant variables which can be viewed as the context variables and called *confounders*. Therefore, these confounders should be discovered and "removed" by some specially designed algorithms in order to accurately estimate the causal effect, such as propensity score method (Gu and Rosenbaum, 1993; Austin, 2011; Imbens, 2004; Lunceford and Davidian, 2004), front-door criterion (Pearl, 1995), instrumental variable estimator combined with structural causal models or potential outcome framework (Guo et al., 2018), etc.

Unfortunately, most of these existing works are not designed for text data of unstructured data format. Also, before the noisy causal effect from the confounders can be removed, these confounders need to be discovered and represented in an explicit form, which is another challenge for text data to fit in the existing causal effect estimation models. A potential solution proposed by Sridhar and Getoor (2019) is to represent the confounders in text data as the topic distribution vectors achieved by Latent Dirichlet Allocation model. However, such a representation is not theoretically explainable to carry enough information to represent these confounders. Also, before estimating the causal effect, we need to first discover these causal relationships which are determined together by cause and context in text data.

The tasks concerning causality extraction in text can be mainly divided into two categories, causal phrase extraction and causal clause extraction, where the former focuses on extracting word

Table 1: Examples of documents containing conditional emotion-cause pairs and non-conditional ones

Type	Document content
Document with a conditional pair	Wu was diagnosed with advanced liver cancer at the beginning of 2014, since when he began to update his health condition in Microblog. If Wu didn't update his microblog for a long time, people worried that he may have passed away. ...
Document with a non-conditional pair	The convenience store was at the corner of the street. Recalling the bloody murder in the early morning, the store owner still felt terrified. She was tallying the goods when she heard a scream from the outside. ...

phrases that have causal relationship in one sentence (Hashimoto et al., 2014; Zhao et al., 2017), and the later focuses on extracting multiple clauses from a document (Gui et al., 2016a,b; Xia and Ding, 2019; Chen et al., 2018). Existing models for causal phrase extraction are mostly based on combinations of syntactic patterns and machine learning techniques, which first extract candidate phrases based on predefined templates, and then train a classifier to classify the candidate causal pairs. On the other hand, existing models for causal clause extraction are mostly based on deep learning models to extract abstract features for each clause, in order to accurately classify whether or not some clauses are causally related. Although the context of the input document is always involved in providing more semantic information to enhance the embedding vectors of clauses, none of these works has paid attention to the possible effect of context on the causal relationship itself.

Moreover, as we focus on emotion causal relationships in this paper, for some emotions (e.g., shame, envy, guilt, etc.) to arise in the first place, a particular social setting may be necessary. Therefore, taking the social contexts into consideration may be an essential step to study whether a specific event can cause an emotion (Wilutzky, 2015; Marsella et al., 2010; Jurafsky, 2004). In our work, we articulate the importance of context in the problem of causal relationship recognition, in view that the context can be essential in order for a pair of emotion and cause to have a valid causal relationship.

3 Task Definition

In this section, we first formally define the term "conditional" based on the concept of emotion-cause pair used in the ECE and ECPE tasks, and then formulate our proposed task based on such

conditional emotion-cause pairs.

As defined by Xia and Ding (2019), an emotion-cause pair (ECP) contains an emotion clause indicating an emotion (e.g., Happiness) and a set of corresponding cause clauses. We define "conditional ECP" as follows.

Definition 1 (Conditional Emotion-Cause Pair)
If an emotion-cause pair is considered to have causal relationship only when a specific context is given, it is called a conditional emotion-cause pair.

Examples of documents with a conditional pair and a non-conditional pair can be found in Table 1. Specifically, for the document with a conditional pair, in general, most people would not worry about whether a social media user updates his/her account or not, but with specific context like the one in the document, Wu had already gained much attention and hence people cared about his life. As for the document with a non-conditional pair, one shall feel frightened whenever he/she witnesses a bloody murder, which is unlikely to change with different contexts.

Definition 1 indicates that the conditional pairs should not be judged to have causal relationship when an irrelevant context or no context is given. Considering such a property, our task is formulated as follows.

The proposed task Given a specific context con_i and an emotion-cause pair $x_i = (C_i, e_i)$ containing a set of cause clauses C_i and an emotion clause e_i, determine a binary label y_i to indicate whether or not the input pair x_i has causal relationship under the context con_i.

As defined above, our proposed task is not to directly distinguish the conditional pairs from the non-conditional ones. The reason is that the recognition of a conditional pair is based on its different labels under different contexts instead of the text

Table 2: Details of the manually labeled dataset

	Conditional (N_{con})	Non-conditional (N_{noncon})
# of documents	561	1524

itself. Therefore, such a task formulation spares the models from worrying about how to transform the labels of causal relationship to the labels of conditional pairs, and simplifies the process of training.

4 Dataset Construction

Different from the ECPE task (Xia and Ding, 2019), the ECPs are directly given in our task and the goal of our task is to judge whether or not they are causally related under a specific context. Therefore, our proposed task is new without any existing annotation available, so we construct our own dataset based on the ECPE dataset (Xia and Ding, 2019) by the following two steps: manual annotation and negative sampling.

4.1 Manual Annotation

In the ECPE dataset, each document contains one ECP composed of an emotion clause and a set of cause clauses. These documents are mainly snippets of news articles or social media documents. To manually annotate the documents with the labels of conditional pairs, we have recruited three human experts who are required to give a binary label to each document: 1 indicates the ECP in this document is conditional, and 0 indicates it is not. Specifically, these three experts are experienced academic partners in the area of emotion cause extraction.

To label an ECP as a conditional one, the cause events and the effect emotions should be less or not relevant under normal circumstances. For example, in general one shall not reject the care of nurses when he/she is ill, but someone with racial prejudice may feel disgusted with foreign nurses. Such context information contained outside of the cause and emotion clauses is what the three experts are required to find and judge whether these context information is essential for the targeted ECP to have a causal relationship.

We have inspected the labels provided by the three experts and the average kappa value among them is **0.8675**, which indicates the fidelity of these manual labels. With three labels from the three human experts, we adopt the majority vot-

ing scheme to determine the final label for each document. For example, given a document, if two experts agree on labeling the ECP in this document as a conditional pair, then the final label for this document is 1. The details of the annotated dataset are shown in Table 2, where N_{noncon} denotes the number of documents with non-conditional pairs, and N_{con} denotes the number of documents with conditional pairs.

4.2 Negative Sampling

Although through manual annotation we have obtained the labels of conditional pairs, all ECPs in the resultant dataset are supposed to have valid causal relationships, since the conditional ones are all given their correct contexts and the non-conditional ones do not depend on any context. In other words, the current dataset only has "positive" instances, but for our proposed task we also need "negative" instances to train a classification model.

To generate such "negative" samples, we follow the procedure of negative sampling (Mikolov et al., 2013). Specifically, we define the following two types of "negative" samples:

- **Context-type**: The context-type negative sample of a document is generated by replacing its original context with a randomly sampled context from the other documents, while keeping the ECP unchanged.

- **Emotion-type** : The emotion-type negative sample of a document is generated by replacing its emotion clause with a randomly sampled emotion clause (indicating a different emotion) from the other documents, while keeping the other clauses unchanged.

Table 3 shows examples of the two types of generated "negative" samples. Specifically, compared with the original document, the generated context-type document has a totally different set of context clauses (i.e., the italic clauses enclosed by "⟨" and "⟩"), which does not provide the information that Wu was diagnosed with cancer and has been updated his health condition in Microblog. Therefore, the generated context-type document is expected to have no causal relationship due to the irrelevant context. As for the generated emotion-type document, the emotion clause is replaced with a clause of Happiness, which makes no sense since the cause clause and the new emotion clause are now irrelevant.

Table 3: Examples of two types of generated "negative" documents

Type	Document content
Original	Wu was diagnosed with advanced liver cancer at the beginning of 2014 and began to update his health condition in Microblog. If Wu didn't update his microblog for a long time, people worried that he may have passed away. ...
Context-type	⟨ *When Bai was notified that his advice was adopted by the National public security bureau, he was cooking dinner for his children.* ⟩ If Wu didn't update his microblog for a long time, people worried that he may have passed away. ⟨ ... ⟩
Emotion-type	Wu was diagnosed with advanced liver cancer at the beginning of 2014 and began to update his health condition in Microblog. If Wu didn't update his microblog for a long time, ⟨ *he was really happy that he could finally afford his dream house.* ⟩ ...

[1] The italic clauses enclosed by the angle brackets are the replaced clauses in the generated documents.

Table 4: Details of the constructed dataset with $n = 2$

	Positive	Negative
# of documents	5133	5292

To summarize the labels of causal relationships of the generated documents, for the generated context-type documents, those with a conditional pair will not have causal relationships due to their irrelevant contexts, while those with a non-conditional pair will still have causal relationships. As for the generated emotion-type documents, all of them will not have causal relationships, since the original cause clauses should only lead to the original emotion given the original context.

Therefore, suppose that n denotes the number of each type of "negative" documents generated for each original document, we can calculate the number of documents with and without causal relationships in the constructed dataset as follows:

$$N_{pos} = N_{noncon} * n + N_{noncon} + N_{con}$$
$$N_{neg} = N_{noncon} * n + N_{con} * 2 * n \qquad (1)$$

where N_{pos} denotes the number of documents with causal relationships, N_{neg} denotes the number of documents without causal relationships.

In order to generate a balanced dataset for our proposed task, we need to make sure that:

$$\frac{N_{pos}}{N_{neg}} \approx 1 \qquad (2)$$

from which n should be around 1.858. Since n should be an integer, the possible choices are $n = 1$

and $n = 2$, while a larger n may create an imbalanced dataset with too many negative samples and cause the models biased towards the negative labels. Based on our preliminary experiments conducted to validate the setting of n, the results show that the constructed dataset with $n = 1$ is not reasonable[1]. Therefore, we set n to 2 to construct our dataset and the details of the constructed dataset are shown in Table 4.

5 Architecture

In this section, we introduce our architecture for our task, and propose a simple, general and effective prediction aggregation module based on the characteristics of conditional ECPs.

5.1 The Framework

As shown in Figure 2, our framework contains three main modules: a clause embedding module, a context encoding module, and a newly proposed prediction aggregation module. First, in the clause embedding module, the word embedding vectors of the input clauses are passed into a Bi-directional Long Short-Term Memory (BiLSTM) model to obtain informative clause embedding vectors for each clause. Then, in the context encoding module, context information is encoded into the clause embedding vectors of the input ECPs by using one of the three classic methods: explicit concatenation, implicit encoding and attention-based method. Finally, the context-encoded embedding vectors and the original ones of the input ECPs are passed to the proposed prediction aggregation module to

---

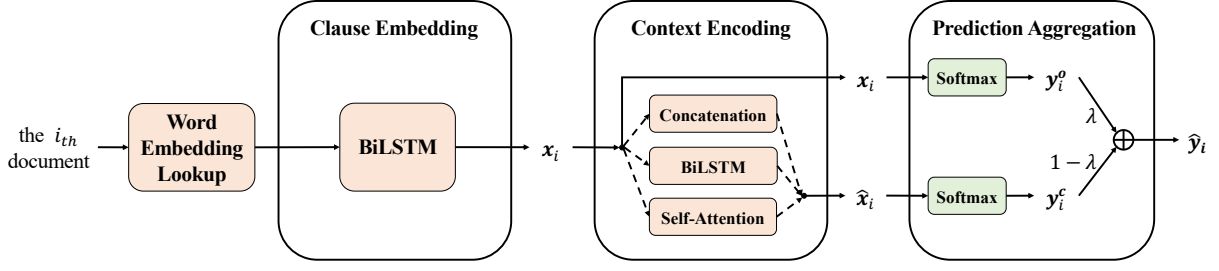[1]More details and discussion on the preliminary experiment results are given in Section 6.3.

Figure 2: Architecture of our framework, where the dashed arrows indicate that we have three different ways to encode context information.

generate the final prediction.

Below, we introduce these three modules in detail. For simplicity, the formulas in subsequent discussions only consider the case where there is only one cause clause. Note that it can be easily extended to multiple cause clauses by concatenating their embedding vectors together.

5.2 Clause Embedding Module

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). To encode words' embedding vectors into a clause embedding vector, we adopt BiLSTM model, which is capable of generating an informative vector for each clause by passing words' information along the clause forwards and backwards. Specifically, for the i-th document, the input of this module includes three parts: the cause clause c_i, the emotion clause e_i, and the context clauses con_i. Note that the cause clauses and the emotion clause are already annotated in the dataset, so the remaining clauses of the input document are denoted as the context clauses.

$$
\begin{aligned}
\mathbf{c}_i &= BiLSTM(c_i) \\
\mathbf{e}_i &= BiLSTM(e_i) \\
\mathbf{con}_i &= BiLSTM(con_i) \\
\mathbf{x}_i &= [\mathbf{c}_i; \mathbf{e}_i]
\end{aligned} \tag{3}
$$

5.3 Context Encoding Module

After we retrieve an embedding vector for each clause, in order to determine the causal relationship of the input ECP under a specific context, we need to encode context information into the embedding vectors of the input ECP for the subsequent prediction. In this aspect, we consider three classic methods used most frequently in the area of text processing: explicit concatenation, implicit encoding, and attention-based method. The performance

of these methods will be shown and discussed in Section 6.

Explicit concatenation As indicated by the name, this method directly concatenates the embedding vectors of the context clauses to those of the input pair, and passes them to the next module so that the final prediction is based on all clauses, i.e.,

$$
\hat{\mathbf{x}}_i = [\mathbf{c}_i; \mathbf{e}_i; \mathbf{con}_i] \tag{4}
$$

Implicit encoding The second method aims to encode context information implicitly into the embedding vectors of the input ECP via an extra layer, such as BiLSTM or Convolutional Neural Network (CNN). Considering that the relevant information may locate anywhere of the context clauses, CNN may not be a good choice due to its fixed neighborhood size. Therefore, we adopt BiLSTM model for implicit encoding of context information.

$$
\begin{aligned}
\hat{\mathbf{c}}_i &= BiLSTM(\mathbf{c}_i) \\
\hat{\mathbf{e}}_i &= BiLSTM(\mathbf{e}_i) \\
\hat{\mathbf{x}}_i &= [\hat{\mathbf{c}}_i; \hat{\mathbf{e}}_i]
\end{aligned} \tag{5}
$$

Attention-based method The third method is based on the self-attention module proposed by Vaswani et al. (2017), which has achieved great success recently in translation work. We adopt the 1-layer-multi-head self-attention module to encode context information. Specifically, instead of calculating the attention scores among all sentences in the original self-attention module, we only calculate the attention scores between the input pair and the context clauses, which reduces unnecessary attention weights, and targets at generating the context-encoded embedding vectors of the input

ECP for the subsequent prediction.

$$\hat{\mathbf{c}}_i = \mathbf{c}_i + \sum_{j \in con_i} \alpha_{c,j} \cdot \mathbf{con}_{i,j}$$

$$\hat{\mathbf{e}}_i = \mathbf{e}_i + \sum_{j \in con_i} \alpha_{e,j} \cdot \mathbf{con}_{i,j}$$

$$\alpha_{c,j} = \frac{\exp(\mathbf{c}_i \cdot \mathbf{con}_{i,j})}{\sum_{k \in con_i} \exp(\mathbf{c}_i \cdot \mathbf{con}_{i,k})} \quad (6)$$

$$\alpha_{e,j} = \frac{\exp(\mathbf{e}_i \cdot \mathbf{con}_{i,j})}{\sum_{k \in con_i} \exp(\mathbf{e}_i \cdot \mathbf{con}_{i,k})}$$

$$\hat{\mathbf{x}}_i = [\hat{\mathbf{c}}_i; \hat{\mathbf{e}}_i]$$

5.4 Prediction Aggregation Module (PAM)

As defined in Section 3, the conditional pairs will no longer have causal relationships if an irrelevant context or no context is given, whereas the non-conditional pairs will always have valid causal relationships. Taking such a difference into consideration, here we propose a simple, general and effective prediction aggregation module.

First, to get the prediction with context, we pass the context-encoded embedding vectors of the input pair, $\hat{\mathbf{x}}_i$, to a fully-connected layer with a softmax activation function:

$$P(y_i^c) = softmax(W_c \hat{\mathbf{x}}_i)) \quad (7)$$

where W_c is a trainable weight matrix.

Next, we add an extra step of predicting the labels of causal relationship directly based on the original embedding vectors of the input pair, without encoding the context information. Specifically, we pass the original embedding vectors achieved in the clause embedding module, $\mathbf{x}_i$, to a fully-connected layer with a softmax activation function:

$$P(y_i^o) = softmax(W_o \mathbf{x}_i) \quad (8)$$

where W_o is a trainable weight matrix.

The proposed module works as follows. If $P(y_i^o)$ has already shown that the input pair has a valid causal relationship (i.e., $P(y_i^o = 1) > P(y_i^o = 0)$), then this pair is more likely to still have a causal relationship under any specific context, and the final result should depend more on the prediction without encoding context information. On the other hand, if the input pair is predicted to have no causal relationship without context, the final result should give more weight to the prediction taking context information into consideration. Following this logic, we can have the following aggregation formula:

$$P(y_i) = \lambda * P(y_i^o) + (1 - \lambda) * P(y_i^c),$$
$$\lambda = P(y_i^o = 1) \quad (9)$$

With this aggregation module, the model can handle both conditional and non-conditional pairs, and give a better prediction on the causal relationship of an input pair under a specific context.

6 Experiment

In Section 4, we have described the process of dataset construction and the details of the constructed dataset[2]. In this section, we conduct experimental studies to evaluate our approach, and analyze the experiment results pragmatically.

6.1 Baseline Models

As mentioned in Section 5.3, there are three options for the context encoding module. Therefore, we consider three baseline models without PAM, each of which contains one of the three context encoding methods we described.

- **BiLSTM + Concatenation**: this baseline model uses BiLSTM at word level to get the clause embedding vectors and directly concatenates the context clauses' vectors to those of the input pair.

- **BiLSTM + BiLSTM**: this model uses BiLSTM at both word level and clause level to get the context-encoded embedding vectors of the input pair for the final prediction.

- **BiLSTM + Self-Attention**: this model uses BiLSTM at word level and uses Self-Attention at clause level to encode the context information.

6.2 Experiment Settings

We randomly select 90% of the data for training and the remaining 10% for testing. To avoid the effect of randomness, we divide the whole dataset into 10 folds and repeat the experiments 10 times with each fold being testing data. The average experiment results are reported in the following sections. Since our proposed task is a binary classification task, we adopt the traditional precision, recall and F1 scores to evaluate the prediction performance.

---

[2]The constructed dataset and our programs can be found in: https://github.com/mark-xhchen/Conditional-ECPE.

Table 5: Preliminary experiment results of baselines

Models	BiLSTM + Concatenation		BiLSTM + BiLSTM		BiLSTM + Self-Attention	
Parameter	$n = 1$	$n = 2$	$n = 1$	$n = 2$	$n = 1$	$n = 2$
Precision (%)	57.34	54.12	57.93	66.06	57.81	57.66
Recall (%)	98.29	71.19	99.64	74.00	99.58	77.70
F1 (%)	72.77	61.27	73.26	69.76	73.16	66.05

[1] The values in gray cells are too high to be reasonable.

Table 6: Performance of models with and without PAM

Models	Containing PAM (✓ or ×)	P(%)	R(%)	F1(%)
BiLSTM + Concatenation	×	54.12	71.19	61.27
	✓	60.24	75.91	**67.10**[‡] (↑ 5.83)
BiLSTM + BiLSTM	×	66.06	74.00	69.76
	✓	65.11	78.30	**71.10**[†] (↑ 1.34)
BiLSTM + Self-Attention	×	57.66	77.70	66.05
	✓	61.95	78.65	**69.29**[‡] (↑ 3.24)

- [†] and [‡] denote the statistical significance for $p < 0.01$ and $p < 0.001$, respectively.

As for the detailed design of the models, the hidden units in BiLSTM is set to 100, and the heads in Self-Attention module is set to 2. All weight matrices are randomly initialized with uniform distribution. For training, we use the stochastic gradient descent algorithm and Adam optimizer, 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 is set to $1e^{-5}$.

6.3 Experiment Results

In this section, we report the experiment results in Table 5 to validate our setting of n. We conduct experiments on the constructed datasets by setting n to 1 and 2. We notice that all our baseline models achieve unbelievably high recall values (i.e., $0.98 \sim 0.99$, see gray cells in Table 5) when $n = 1$. After looking into the detailed predictions, we find that when $n = 1$, the models unreasonably predict all test samples to have positive labels, which reveals that the models are heavily biased towards the positive labels due to insufficient negative samples. In contrast, the performance of the baseline models becomes more reasonable when $n = 2$. Therefore, we conduct our following experiments on the constructed dataset using $n = 2$.

6.4 Effect of PAM

To validate the effectiveness of PAM, we conduct experiments on the three baselines and report the results in Table 6. As shown in the table, before we add PAM, "BiLSTM + BiLSTM" achieves the highest F1 score compared with the other two models, possibly due to that the Self-Attention module needs a larger-scale dataset to train well, while simple concatenation cannot get semantic embedding vectors. After adding PAM, the F1 scores of the three baseline models are improved on average by **3.47%**, and the results with p-value attached indicate the models containing PAM significantly outperform those without PAM.

Specifically, the effect of adding PAM from high to low is "BiLSTM + Concatenation", "BiLSTM + Self-Attention", and "BiLSTM + BiLSTM". This seems to imply that the improvement of PAM should be small when the model without PAM can already encode contexts well. The above results demonstrate the generality of PAM that it can be easily used together with existing classic models, and for our proposed task, PAM can improve their prediction performance significantly.

6.5 Case Study

To further illustrate how our aggregation module (i.e., PAM) improves the performance, we inspect the predictions of four examples given by the "BiLSTM + Concatenation + PAM" model. As shown in Table 7, Documents #1 and #2 share the same conditional ECP, and Documents #3 and #4 share the same non-conditional pair.

For Documents #1 and #2, since people would not care about Wu's health condition if he did

Table 7: Examples of predictions from the model "BiLSTM + Concatenation + PAM"

Index	Document	$P(y_i^o)$	$P(y_i^c)$	$P(y_i)$	True Label
#1	Wu was diagnosed with advanced liver cancer at the beginning of 2014 and began to update his health condition in Microblog. If Wu didn't update his microblog for a long time, people worried that he may have passed away. ...	$[0.657, 0.343]$	$[0.265, 0.735]$	$[0.3995, \mathbf{0.6005}]$	1
#2	When Bai was notified that his advice was adopted by the National public security bureau, he was cooking dinner for his children. If Wu didn't update his microblog for a long time, people worried that he may have passed away. ...	$[0.657, 0.343]$	$[0.712, 0.288]$	$[\mathbf{0.6931}, 0.3069]$	0
#3	The convenience store was at the corner of the street. Recalling the bloody murder in the early morning, the store owner still felt terrified. She was tallying the goods when she heard a scream from the outside. ...	$[0.211, 0.789]$	$[0.182, 0.818]$	$[0.2049, \mathbf{0.7951}]$	1
#4	Yu came to visit her relatives in ChuanCang Village early in the morning of April 5. Recalling the bloody murder in the early morning, the store owner still felt terrified. She didn't notice rushing river from upstream due to her poor hearing. ...	$[0.211, 0.789]$	$[0.254, 0.746]$	$[0.2201, \mathbf{0.7799}]$	1

[1] $P(y_i^o)$ is the prediction based on only the input pair, $P(y_i^c)$ is the prediction with context encoded, and $P(\hat{y})$ is the final predicted probability.
[2] Red clause is the cause clause and blue clause is the emotion clause

not begin to update his information in his social media account, we can judge that Document #1 should have a causal relationship while document #2 should not, corresponding to their true labels being 1 and 0, respectively. The prediction without context $P(y_i^o)$ indicates that both documents have no causal relationship since the pair is a conditional pair. Taking context into consideration, the prediction with context $P(y_i^c)$ indicates that document #1 has a causal relationship, while document#2 still has no causal relationship due to its irrelevant context. The difference among these predictions corresponds to the characteristics of the conditional ECPs, which is to depend more on $P(y_i^c)$ when $P(y_i^o)$ indicates no causal relationship.

As for Documents #3 and #4, one shall feel frightened whenever he/she witnesses a bloody murder around him/her, which is unlikely to change with different contexts. Therefore, both documents should have causal relationships. As shown in the table, $P(y_i^o)$ already indicates that the pair is causally related regardless of context and hence the final prediction indicates the same result.

The above cases illustrate that our simple aggregation module enables the model to simultaneously deal with documents containing conditional and non-conditional ECPs, and to fine-tune the final predictions accordingly.

7 Conclusion and Future Work

In this paper, we articulate the importance of context in determining the causal relationships between emotions and their causes. To address this problem, we define a new task of determining whether or not an input emotion-cause pair has a causal relationship under a specific context. We construct a dataset for our task through manual annotation and negative sampling based on the ECPE dataset. Furthermore, we propose a prediction aggregation module (PAM) with low computational complexity, to enable the models to dynamically adjust the final prediction according to the type of emotion-cause pair contained in a document. Experiments demonstrate the effectiveness and generality of our proposed PAM.

In view of the importance of context in the conditional causal relationships we define in this work, what we have done is only the first step. There remain many important and interesting problems ahead of us. For example, how to quantify the effect of context on the targeted causal relationship is another important task to study this type of causal relationship. Besides, how to enable the existing emotion-cause pair extraction models to consider the effect of context is also a meaningful task.

8 acknowledgement

References

Peter C. Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424. PMID: 21818162.

Ying Chen, Wenjun Hou, Xiyao Cheng, and Shoushan Li. 2018. Joint learning for emotion classification and emotion cause detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 646–651, Brussels, Belgium. Association for Computational Linguistics.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 179–187.

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

Xing Sam Gu and Paul R. Rosenbaum. 1993. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420.

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.

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

Lin Gui, Ruifeng Xu, Qin Lu, Dongyin Wu, and Yu Zhou. 2016b. Emotion cause extraction, a challenging task with corpus construction. In *Social Media Processing*, pages 98–109.

Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. 2018. A survey of learning causality with data: Problems and methods. *Computing Research Repository*, arXiv:1809.09337. Version 3.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 987–997, Baltimore, Maryland. Association for Computational Linguistics.

Guido W. Imbens. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29.

Daniel Jurafsky. 2004. 26 pragmatics and computational linguistics. *The handbook of pragmatics*, page 578.

Humayun Kayesh, Md. Saiful Islam, and Junhu Wang. 2019. On event causality detection in tweets. *Computing Research Repository*, arXiv:1901.03526. Version 1.

Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, Julien Kloetzer, Jong-Hoon Oh, and Masahiro Tanaka. 2017. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3466–3473. AAAI Press.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53.

Pengfei Li and Kezhi Mao. 2019. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Systems with Applications*, 115:512 – 523.

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.

Jared K. Lunceford and Marie Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960.

Stacy Marsella, Jonathan Gratch, Paolo Petta, et al. 2010. Computational models of emotion. *A Blueprint for Affective Computing-A sourcebook and manual*, 11(1):21–46.

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.

Judea Pearl. 1995. Causal diagrams for empirical research: Discussion of 'Causal diagrams for empirical research' by J. Pearl. *Biometrika*, 82(4):689–690.

Dhanya Sridhar and Lise Getoor. 2019. Estimating causal effects of tone in online debates. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1872–1878. International Joint Conferences on Artificial Intelligence Organization.

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.

Wendy Wilutzky. 2015. Emotions as pragmatic and epistemic actions. *Frontiers in Psychology*, 6:1593.

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.

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.

Sendong Zhao, Quan Wang, Sean Massung, Bing Qin, Ting Liu, Bin Wang, and ChengXiang Zhai. 2017. Constructing and embedding abstract event causality networks from text snippets. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, page 335–344, New York, NY, USA. Association for Computing Machinery.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.