Emotion Cause Extraction on Social Media without Human Annotation

Debin Xiao, Rui Xia,[*] and Jianfei Yu
School of Computer Science and Engineering,
Nanjing University of Science and Technology, China
{debinxiao, rxia, jfyu}@njust.edu.cn

Abstract

In social media, there is a vast amount of information pertaining to people's emotions and the corresponding causes. The emotion cause extraction (ECE) from social media data is an important research area that has not been thoroughly explored due to the lack of fine-grained annotations. Early studies referred to either unsupervised rule-based methods or supervised machine learning methods using a number of manually annotated data in specific domains. However, the former suffers from limitations in extraction performance, while the latter is constrained by the availability of fine-grained annotations and struggles to generalize to diverse domains. To address these issues, this paper proposes a new ECE framework on Chinese social media that achieves high extraction performance and generalizability without relying on human annotation. Specifically, we design a more dedicated rule-based system based on constituency parsing tree to discover causal patterns in social media. This system enables us to acquire large amounts of fine-grained annotated data. Next, we train a neural model on the rule-annotated dataset with a specific training strategy to further improve the model's generalizability. Extensive experiments demonstrate the superiority of our approach over other methods in unsupervised and weakly-supervised settings.

1 Introduction

The Emotion Cause Extraction (ECE) task was firstly introduced by Lee et al. (2010b), which aims to identify the underlying causes of a given emotion expression in textual data. Previous studies mainly focused on extracting emotion causes from news articles (Gui et al., 2016a; Xu et al., 2019; Li et al., 2018; Xia et al., 2019; Fan et al., 2019; Yan et al., 2021). One representative study among them is Gui et al. (2016a), which constructed a new corpus based on SINA City News. The corpus has attracted much attention in subsequent studies and become a benchmark dataset for the ECE task. In addition to news articles, microblog has nowadays become an important platform for Internet users to publish instant posts and share their personal opinions about hot events or topics, which contains a huge amount of subjective emotional expressions. Tracing the potential causes behind these subjective emotions is helpful to obtain a deep insight into the public emotions, discover the essential causes of the public opinion, and provide an important basis for governments to promptly adjust their political strategies.

However, the ECE task faces significant challenges due to the wide range of topics, diverse domains, and the prevalence of informal expressions in social media. Early studies aimed to address these challenges by approaching the task from a linguistic perspective and employing rule-based methods to detect emotion cause expressions on social media (Gui et al., 2014; Li and Xu, 2014; Gao et al., 2015a,b; Yada et al., 2017). Although these rule-based methods are generally designed and can be applied to different domains or topics, their performance remains limited. Some recent studies further employed statistical machine learning or deep learning models to extract emotion causes in social media. However, most of these studies primarily focus on training their models on small-scale manually annotated corpora in several specific domains (Cheng et al., 2017; Chen et al., 2018a,b; Liu et al., 2021). Despite obtaining better extraction performance, these studies heavily rely on fine-grained cause annotations and are solely suitable for specific domains. Due to the huge amount of data and diverse domains in social media, it is impossible to manually construct an annotated corpus for each domain when we build a machine learning-based ECE system, which greatly limits the large-scale applications in real-world social media scenarios.

[*] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 1455–1468
July 9-14, 2023 ©2023 Association for Computational Linguistics

To address the aforementioned problems, in this work, we propose a new approach to extract emotion causes on social media without human annotation. Our framework is centered around a rule-based method, bolstered by a specialized training strategy. Firstly, a Constituent-Based Rule (CBR) method is proposed to extract the emotion causes by utilizing the syntactic patterns in emotion and cause expressions, and obtain a large rule-annotated dataset without relying on human annotation. Secondly, a Rule-Guided Pseudo Supervised Learning (RGPS) framework is introduced to develop a general system for emotion cause extraction on social media. This method involves training a model on the rule-annotated dataset by masking the cue words and includes a label refinement module for iterative learning.

In this work, CBR is a rule-based method that relies on the constituent syntactic structure in the Chinese language. Unlike previous methods that mostly utilized word-level patterns of emotion cause expressions, CBR employs carefully designed rules based on the constituency parsing tree. This approach effectively improves the performance of span-level ECE and achieves high precision in extraction. We employ CBR to a large-scale unannotated corpus to automatically obtain rule-annotated data. With such a rule-annotated dataset, we then train a neural model for extracting emotion cause spans based on a pre-trained language model. We propose to alleviate the problem of overfitting inherent rule patterns by masking significant rule features such as causal cue words. Additionally, we propose a label refinement module to enhance the diversity and accuracy of data labels through iterative training, which enables the model to further improve its extraction performance and generalization ability.

To evaluate the effectiveness of our approach, we construct a new emotion cause dataset, COVID19-ECE, which focuses on the topic of the COVID-19 pandemic. We conduct experiments on COVID19-ECE and another social media ECE dataset named CoEmoCause (Liu et al., 2021). Our experimental findings are as follows: 1) compared to previous rule-based methods, our proposed constituent-based rules demonstrate a significant enhancement in span-level emotion cause extraction performance. Moreover, this approach achieves high precision, which is highly beneficial for large-scale practical applications in social media. 2) The pseudo-supervised ECE model significantly improves the recall of emotion cause extraction based on CBR, which leads to a noteworthy improvement in the F-score. Based on the rule-based pseudo annotation of 25,600 instances, our RGPS approach achieves comparable span-level extraction performance to standard supervised learning methods that uses hundreds of human-annotated instances. 3) By leveraging our RGPS approach and incorporating a small amount of human-annotated data, we achieve further improvements. E.g. with the help of 200 instances of human-annotated data, our approach yields results that are comparable to those obtained using full human annotations on that dataset.

2 Approach

Traditional ECE aims at extracting emotion causes at the clause level (Gui et al., 2016a). Several recent studies extend the task to extract the fine-grained span-level causes (Oberländer and Klinger, 2020; Li et al., 2021a,b). Due to the short and informal nature of social media posts, span-level is more suitable for identifying the emotion causes in social media texts. Therefore, this work focuses on the span-level ECE task (Li et al., 2021b), which is formalized as follows: Given a post S containing a sequence of N tokens $S = [w_1, w_2, ..., w_N]$ and an annotated emotion expression $E = [e_1, e_2, ..., e_K]$ in S , the span-level ECE task aims to detect the boundaries of the emotion cause span from S, which stimulates the emotion expressions.

In the upcoming sections, we introduce our emotion cause extraction approach, which includes two main stages: Constituent-Based Rule (CBR) and Rule-Guided Pseudo-Supervised Learning (RGPS).

2.1 Constituent-Based Rule for Emotion Cause Extraction

Previous works found that some specific words are indicative of emotion causes, and summarized these cue words into seven categories (Lee et al., 2010b). Earlier rule-based methods for the ECE task typically designed different word-level rules and constraints for these cue words to extract verb-centered cause triples, i.e., (Noun, Verb, Noun) (Li and Xu, 2014; Gui et al., 2014; Chen et al., 2010), which leads to limited performance in span-level ECE. To this end, we propose constituent-based rules to extract continuous emotion cause spans.

In order to introduce CBR more clearly and pro-

Error Tolerance	0 token	3 token	5 token
Coverage	62.07%	79.41%	85.89%

Table 1: Statistics of cause constituent coverage on the constituency parsing tree of the COVID19-ECE dataset.

vide a comprehensive understanding of their implementation, we first conduct preliminary analysis and establish certain assumptions as a foundation.

2.1.1 Observation and Assumption

As shown in Table 1, our statistics based on the human-annotated dataset (COVID19-ECE) show that 62.07% of the gold emotion cause spans overlap entirely with the text span corresponding to an individual constituent on the constituency parsing tree. The coverage improves to 85.89% when we allow for a 5-token error between the boundary of the constituent and the gold annotation. This indicates that most of the emotion cause span is a relatively complete and independent constituent and can be covered by a single constituent of the constituency parsing tree. Moreover, the statistics on the annotated corpus show that more than 90% of the cause constituent types in the constituency parsing tree belong to IP (Simple Clause), VP (Verb Phrase) and NP (Noun Phrase). Based on this observation, we propose our basic assumption that the span-level emotion cause extraction problem can be converted to a cause constituent recognition problem with specific constituent types on the constituency parsing tree.

Specifically, we define some key constituents and causal syntactic relation on a constituency parsing tree for the ECE problem as follows.

- **Emotion Constituent**: the constituent that completely covers the emotion word and has the deepest depth.
- **Cue Word Constituent**: the constituent that completely covers the cue word and has the deepest depth.
- **Cause Constituent**: the constituent with maximum coverage of cause span and has the deepest depth.
- **Causal Syntactic Relation**: the connections between the cause constituent and the cue word constituent.

Pattern	Cue Words	Examples	Number
A	Prepositions	for	9
	Conjunctions	because	4
B	Reported Verbs	think/talk	22
	Epistemic Verbs	see/hear/know	34
	Copula	is	1
C	Light Verbs	let/make	8
	Causal Verbs	cause/lead to	9
D	Emotion Verbs	fear/hate	-

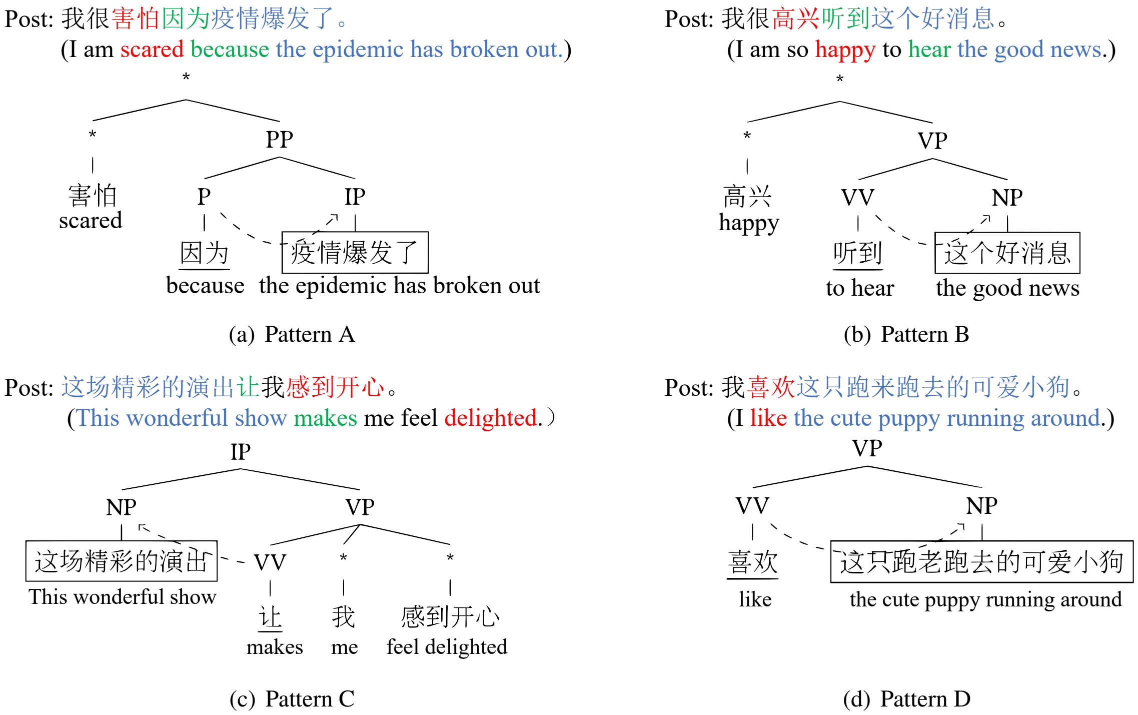
Table 2: Cue words for proposed constituent-based rules.

2.1.2 Patterns, Cue Words, and Rule Details

Based on these preliminary analysis and assumptions, we describe the details of CBR. In constituency parsing trees, there are relatively fixed syntactic patterns between specific types of cue words and the corresponding cause constituents they indicate. Based on this observation, we summarize four representative and general causal syntactic patterns. By employing pattern matching, we can identify the causal patterns present in the text and extract the corresponding cause constituents. Table 2 presents the categories and quantities of cue words corresponding to each pattern, and Figure 1 illustrates templates and examples for the four causal syntactic patterns. Due to space limitation, in the following we only introduce the first causal pattern in detail. The matching procedure for the other types of causal patterns remains similar.

In Pattern A, as illustrated in Figure 1(a), the corresponding cue word category is preposition or conjunction. In constituency parsing, the category of the cue word is represented by ‘P’ to indicate a preposition, and its parent node type is ‘PP’ (prepositional phrase). When the right child node of it belongs to ‘IP/VP/NP’, the right child node is recognized as a potential cause constituent. Specifically, in the input text with the explicit emotion word ‘scared’, we first match the preposition cue word ‘because’. Then, we check if the pattern of the subtree matches the template. If the validation is successful, we can locate the position of the cause constituent in the tree and map the corresponding node to the cause span ‘the epidemic has broken out’.

Note that it is possible to match multiple candidate constituents during the matching process. Following the practice in previous related works (Li and Xu, 2014; Gui et al., 2014; Chen et al., 2010), we only regard the candidate constituent closest to the emotion constituent or cue word constituent as


Post: 我很害怕因为疫情爆发了。
(I am scared because the epidemic has broken out.)
*
*
PP
害怕
scared
P
IP
因为
疫情爆发了
because
the epidemic has broken out
Post: 我很高兴听到这个好消息。
(I am so happy to hear the good news.)
*
*
VP
高兴
happy
VV
NP
听到
这个好消息
to hear
the good news
Post: 这场精彩的演出让我感到开心。
(This wonderful show makes me feel delighted.)
IP
NP
VP
这场精彩的演出
This wonderful show
VV
*
*
让
我
感到开心
makes
me
feel delighted
Post: 我喜欢这只跑来跑去的可爱小狗。
(I like the cute puppy running around.)
VP
VV
NP
喜欢
这只跑老跑去的可爱小狗
like
the cute puppy running around

(a) Pattern A

(b) Pattern B

(c) Pattern C

(d) Pattern D

Figure 1: Illustrations of our proposed Constituent-Based Rule. It contains four types of constituent-based patterns and eight types of cue words. We use the colors red, green and blue to indicate emotion expressions, cue words and cause spans in the post text. In the constituency parsing tree, we use boxes to indicate cause constituent, underlines to indicate cue word constituent, and $*$ to indicate that any types are satisfied. We provide the corresponding English translations below the Chinese examples.

the cause constituent.

Finally, we would like to state that although this work was carried out based on Chinese microblogs, its key idea can be applied in English as well. According to our observation, the syntactic patterns of emotion cause expression in Chinese and English are quite similar, and the vast majority of the CBR rules are mutually compatible, except for some minor differences.

2.2 Rule-Guided Pseudo-Supervised Learning

The rule-annotated dataset reflects specific patterns but does not cover all causal patterns. This will affect the generalization performance of the model. Therefore, we propose the Rule-Guided Pseudo-Supervised learning method to alleviate this limitation. We use CBR for automatic data annotation to obtain a large-scale rule-annotated dataset. Next, we use RGPS to train a model based on the rule-annotated dataset, as shown in Figure 2.

The span-level ECE task can be formalized as a sequence labeling task (Li et al., 2021b). Specifically, a post $S = [w_1, w_2, ..., w_N]$ and a given emotion expression $E = [e_1, e_2, ..., e_K]$ in S are concatenated to form a combined sequence X as the input fed into a pre-trained model like BERT: [CLS], $w_1, ..., w_N$,[SEP], $e_1, ..., e_K$,[SEP], where [CLS] and [SEP] are special tokens. The output of the model is the contextualized representation of each token in the combined sequence, and then we use a Conditional Random Field (CRF) layer to predict the labels of the input post. We use $\{B, I, O\}$ as the label set. Here, B, I, and O represent the beginning, inside, and outside of a cause span, respectively, indicating the span ranges.

In order to train the model to learn causal relations beyond the inherent patterns from the rule-annotated dataset, we propose a method called 'rule masking'. It involves randomly masking a specific proportion of the causal cue words and emotion cue words and then guiding the model to predict the emotion cause spans without seeing these explicit cues. This procedure breaks the inherent causal patterns in the rule-annotated dataset and prevents the model from relying on the cue words features during the encoding process, allowing the model to focus more on other effective and meaningful information. As shown in Figure 2, the [MASK] tokens in green and red represent the masked causal cue words and emotion cue words, respectively.

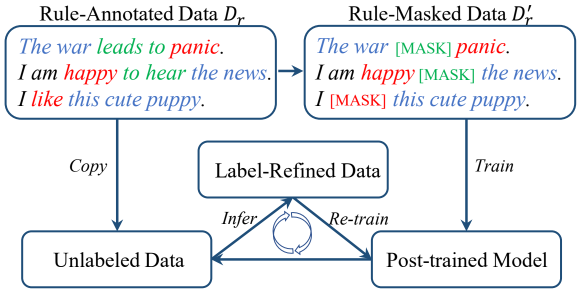


Figure 2: The Rule-Guided Pseudo-Supervised Learning for span-level ECE.

Note that, we mask the cue words in the attention mask of the model's input layer. During training, the model is trained to learn the original sequence labels, facilitating the capturing of connections between the emotion expression and the emotion cause span. We use Dr' to denote the modified dataset.

The labels generated by CBR are limited and often inaccurate. After preliminary experiments, we found that the model obtained by initial training with the pseudo labels already significantly outperforms the rule-based method. Therefore, we propose label refinement to update the original labels during the training process iteratively. Specifically, we train an initial model $\theta^{(0)}$ over the masked rule-annotated dataset D_r' with the initial rule labels. In the subsequent iterative rounds, we employ the predicted labels of the previous round's model on the training set as the supervision labels for training the current round's model. Consequently, in the t-th iteration, the model's output is $\hat{y}^{(t)} = \mathrm{BERT\text{-}CRF}(\mathrm{x}; \theta^{(\mathrm{t})})$. In each iteration, we use the model that has already converged in the previous iteration to initialize the model for the current round. The information in the original rule-annotated dataset is propagated through the iterative training procedure. The subsequent models are trained on a new refined dataset with more accurate and diverse labels, which helps the model learn more efficiently.

3 Experiments

3.1 Evaluation Setup

3.1.1 Evaluation Datasets and Metrics

In this work, there are two COVID-19 related datasets for evaluation: a new dataset COVID19-ECE constructed by us and an open-source emotion cause dataset CoEmoCause. We describe the two manually annotated datasets as follows:

Item	COVID19-ECE	CoEmoCause
Number of posts	1,793	1,997
Number of instances	2,016	2,610
Number of cause spans	2,797	2,969
Avg. length of post	148.2	78.0
Avg. length of cause	14.4	8.0

Table 3: Statistic of two evaluation datasets.

COVID19-ECE. We selected a portion of data from the crawled large-scale corpus for human annotation of the emotion causes. We hired three annotators (all native Chinese speakers) to manually annotate 5,500 Chinese microblog posts. Two annotators work independently during the annotation process. They are required to annotate the cause span corresponding to the emotion words on the tweets that have been pre-matched with the emotion lexicon (Gui et al., 2016a). When the annotators have different opinions on the annotations, we involve the third annotator as the arbitrator. Finally, we end up with 1,793 labeled posts and we name it COVID19-ECE.

CoEmoCause. This dataset is constructed by Liu et al. (2021) and originally comes from the epidemic dataset of the SMP2020 microblog emotion classification competition.[1] There are 5,195 posts in the dataset with span-level annotations. There are nine emotion categories: respect, support, anger, happiness, surprise, disgust, sadness, fear, and anticipation. To ensure the emotion categories of this dataset conform to the rule-annotated dataset, we remove the samples belonging to support, anticipation and respect. Finally, we get 1,997 posts. The details of these two datasets are listed in Table 3.

We split COVID19-ECE and CoEmoCause into a training set, validation set and a testing set with [75%, 10%, 15%] and [50%, 25%, 25%], respectively. Our primary approach is evaluated solely on the testing set. The training and validation sets will be used for other comparative methods and different settings. Due to space constraints, the main evaluation metrics are span-level Precision, Recall, and F1-score based on exact matching. If using span-level relaxed matching, our approach still achieves satisfactory results.

3.1.2 Pusedu-Supervised Training Dataset

Since the Evaluation Datasets are related to COVID-19, so we use CBR to obtain a large-

[1] https://smp2020.aconf.cn/smp.html#4.

Item	Number
Number of posts	48,014
Number of instances	51,200
Number of cause spans	58,466
Avg. length of post	142.3
Avg. length of cause	8.0

Table 4: Statistic of the pusedu-supervised training dataset.

scale rule-annotated dataset focusing on the topic of COVID-19. We collect data from the Chinese SINA MicroBlog, under the COVID-19 epidemic topic, from February 2020 to June 2020 as the raw corpus. We use the dictionary of causal cue words proposed by Lee et al. (2010b) and the dictionary of emotion words in Gui et al. (2016a) for keyword matching. Note that we made slight modifications to the aforementioned two dictionaries. We use Berkeley Neural Parser (Kitaev and Klein, 2018) to perform CBR for the automatic annotating in the pre-processed corpus. We finally obtained a rule-annotated dataset containing about 400K posts.

We randomly sample 51,200 instances to form our training set, denoted as Dr. The statistics of the rule-annotated training set are shown in Table 4. We partition the training set into several training subsets with increasing data sizes, consisting of 400, 1,600, 3,200, 6,400, 12,800, 25,600, 38,400, and 51,200 instances, respectively. We select training subsets for model training based on different experimental setups.

3.1.3 Implementation Details

We implemented our models with the PyTorch version of the Huggingface Transformers (Wolf et al., 2020). We use a learning rate of 1e-5 for BERT/RoBERTa and 1e-2 for CRF layer. We performed grid search for batch size in [16, 32, 64, 128, 256] and set it to 16. Warmup is applied on the initial 10% steps. The dropout rates between transformer layers are set to 0.1. AdamW is used as the optimizer. In fully-supervised setting, we use the validation set to preserve the checkpoints for final testing. In unsupervised setting, we train the model until it converges and use the average evaluation result on the test set of the last 10 steps as the final result. All results are averaged over 4 randomized replicate experiments.

3.2 Compared Methods

We compare our proposed CBR and RGPS approach with other methods in two settings, depending on whether a small amount of human annotation is used.

w/o human annotations. WBR (Word-Based Rules): Li and Xu (2014) adopts the word-level rules for ECE on social media. We take the span that is covered by the cause triples as the prediction. **CBR**: This refers to our proposed constituent-based rules. We use it to extract the span-level emotion causes directly. **CBR+RGPS**: This is our proposed approach to train the model on 25,600 rule-annotated instances. Note that we chose the proportion of the causal word mask to be 0.6 and the emotion cue mask to be 0.8 based on our empirical experiments. In addition, we conduct a single round of label refinement.

w/ few human annotations. In this setting, we use additional 200 human-annotated instances as a supplement. **Supervised Training**: We use the pre-trained language model BERT$_{base}$ as the backbone. We add a CRF layer on top of the model and fine-tune it on the human-annotated data directly. The performance of the supervised method serves as a baseline for other methods. **Self-Training**: This method utilizes a model obtained through supervised learning on 200 manually annotated instances and conducts self-training on 25,600 unannotated instances (Du et al., 2021). **CBR+BERT-CRF+FT**: This method uses the BERT-CRF model to train on 25,600 rule-annotated instances, and then fine-tunes it on 200 human-annotated instances. **CBR+RGPS+FT**: This method firstly trains the model with RGPS on 25,600 rule-annotated instances, and then fine-tunes it on 200 human-annotated instances.

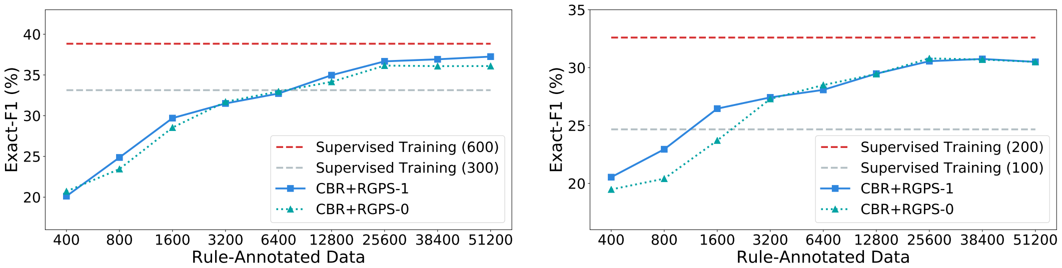
w/ full human annotations. We directly fine-tune the BERT-CRF model using full human-annotated data (Li et al., 2021b). The performance serves as an upper bound for other baseline models.

3.3 Main Results

Based on the two experimental settings mentioned above, we compare these methods on two human-annotated datasets, Covid19-ECE and CoEmo-Cause, as shown in Table 5. Without the need for any human annotations, our proposed CBR outperforms WBR in all metrics and achieves high precision scores, which is highly beneficial for large-scale practical applications in social media. We believe that the constituent parsing provides more accurate cause span candidates, combined with the causal syntactic patterns we proposed,

	Method	COVID19-ECE			CoEmoCause		
		P	R	F1	P	R	F1
w/o human annotations	WBR (Li and Xu, 2014)	13.85	8.99	10.90	16.93	8.99	11.74
	CBR (Ours)	**48.84**	14.89	22.83	**69.41**	18.71	29.47
	w/ rule-annotated data						
	CBR+RGPS (Ours)	46.79	**31.06**	**37.21**	32.82	**28.59**	**30.56**
w/ few human annotations	Supervised Training (Li et al., 2021b)	23.16	26.31	24.54	31.20	34.70	32.60
	Self-Training (Du et al., 2021)	33.91	34.53	34.21	30.83	45.99	36.91
	w/ rule-annotated data						
	CBR+BERT-CRF+FT	45.00	41.56	43.14	36.41	40.86	38.50
	CBR+RGPS+FT (Ours)	54.23	40.19	46.16	36.07	45.31	40.15
w/ full human annotations	Supervised Training (Li et al., 2021b)	45.94	45.93	45.90	40.57	43.94	42.05

Table 5: Main Results on COVID19-ECE and CoEmoCause. We highlight the best results for the *w/o human annotations* setting using boldface. For the *w/ few human annotations* setting, we underline the best results.



(a) COVID19-ECE

(b) CoEmoCause

Figure 3: Size of rule-annotated data vs. Performance. Supervised Training (x) denotes the BERT-CRF model trained on x human-annotated instances. CBR+RGPS-n denotes RGPS trained on 25600 rule-annotated instances with n rounds of label refinement.

Method	COVID19-ECE			CoEmoCause		
	P	R	F1	P	R	F1
RGPS	46.8	**31.1**	**37.2**	32.8	**28.6**	30.6
w/o LR	**50.1**	28.5	36.2	36.7	28.2	**31.6**
w/o RM	40.1	23.1	29.2	36.5	22.8	27.7
w/o LR/RM	41.8	21.7	28.6	**39.4**	21.5	27.5

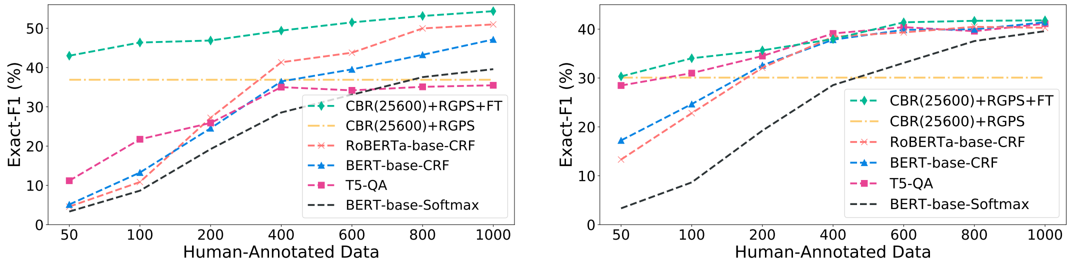
Table 6: Ablation study on RGPS.

resulting in better overall performance. Based on 25600 rule-annotated instances, the proposed method CBR+RGPS shows a significant improvement in recall and F1 score when compared to the rule-based methods. This indicates that RGPS can improve the low coverage issues that exist with rule-based methods. However, we observe an obvious drop in precision on the CoEmoCause dataset, which may be attributed to differences in the dataset distribution.

By incorporating an additional 200 instances of human-annotated data, our proposed CBR+RGPS+FT model demonstrates further improvements over the CBR+RGPS baseline. It achieves comparable or even better results compared to the fully-supervised baseline that requires 1,300 human-annotated instances. CBR+RGPS+FT also outperforms semi-supervised and CBR+BERT-CRF+FT with the highest F1 score. These findings suggest that the rule-annotated data can provide valuable knowledge to the model, and our proposed RGPS module can assist in efficiently utilizing such rule-annotated data.

3.4 Ablation Study on RGPS

We study the effectiveness of each component of RGPS. Specifically, we use the following abbreviation to denote each component of RGPS: Rule Masking (RM) and Label Refinement (LR). As we can see from Table 6, Rule Masking plays a crucial role in RGPS by providing a solid foundation for further improvements. By employing Rule Masking, the model becomes capable of understanding the deep connections between emotions and causes, which facilitates enhanced performance. And we find that RGPS achieves the best results when cue words are masked at a ratio of 60%-80%. Furthermore, Label Refinement contributes to improving the generalization performance of the model and enhancing recall rates, albeit at the cost of sacrificing a certain level of precision.



(a) COVID19-ECE (b) CoEmoCause

Figure 4: Size of human-annotated data vs. Performance of different methods.

3.5 The Size of Rule-Annotated Data

We explore the performance change when having different sizes of rule-annotated data. We uniformly sample data from the rule-annotated dataset and train our neural model on it. The curves can be found in Figure 3. The X-axis is the number of rule-annotated data, while the Y-axis is the F1 score at the span level on the testing set. There is a significant growing trend when the number of rule-annotated instances is below 10K. Thereafter, the growing speed slow down and converges when the amount of data reaches approximately 25K. Furthermore, by further expanding the training data to around 50K instances on the COVID19-ECE dataset, the performance of our method CBR+RGPS approaches that of the model trained using 600 manually annotated instances. Additionally, the model subjected to one round of label refinement, CBR+RGPS-1, demonstrates superior performance compared to CBR+RGPS-0, providing evidence for the effectiveness of our approach.

3.6 Discussion on Human-Annotated Data

We explore the dependence of different models on different sizes of human-annotated data on two datasets, COVID19-ECE and CoEmoCause, as shown in Figure 4. The X-axis indicates the number of human-annotated data, while the Y-axis indicates the F1 score on the testing set. CBR(25600)+RGPS indicates our proposed training framework, and we use BERT$_{base}$-CRF as the backbone for post-training based on 25600 rule-annotated instances, which serves as a baseline. CBR(25600)+RGPS+FT indicates that we post-train the model based on CBR(25600)+RGPS, and then fine-tune it on human-annotated data. We use a pre-trained Chinese T5-small (Wang et al., 2022) as another strong baseline, and formalize the ECE task as a Question Answering (QA) task.

In the COVID19-ECE dataset, CBR(25600)+ RGPS+FT can achieve an F1 score of 47.5% by fine-tuning with only 100 human-annotated instances. The other models require 600 and more data to train from scratch to achieve similar results. In the CoEmoCause dataset, CBR(25600)+RGPS+FT also performs better than other methods, which indicates that our method can better alleviate the reliance of the model on human-annotated data.

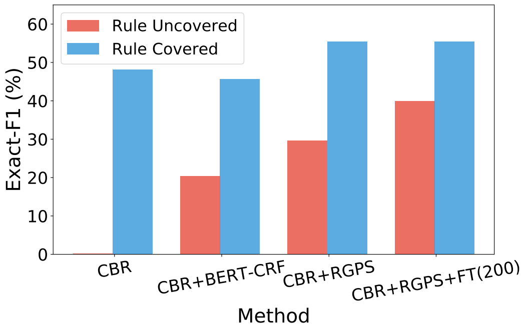


Figure 5: Generalizability test on COVID19-ECE.

3.7 Generalizability Beyond the Rules

This section further investigates the model's generalization ability on testing set within and outside the scope of rule coverage. Firstly, we divide the testing set into two subsets based on whether an instance can be successfully matched by the CBR. As shown in Figure 5, the red color denotes data outside the rule coverage and the blue color denotes the data within the rule coverage. All the models are trained on 25600 rule-annotated instances. We observe that CBR+RGPS outperforms CBR and CBR+BERT-CRF notably on the data instances outside the rule coverage. This indicates that our proposed pseudo-supervised framework assists the model in generalizing to data beyond the scope of rule coverage. Additionally, it also enhances the extraction performance of data within the rule coverage. With the incorporation of a small amount of human-annotated data, CBR+RGPS+FT(25600) achieves further improvements on data outside the rule coverage, but the performance gains on data within the rule coverage are limited. This suggests that the model's performance on data outside the rule coverage is one of the bottlenecks limiting the overall model performance.

4 Related Work

Early works on emotion cause extraction (ECE) mainly focused on rule-based method. Lee et al. (2010b) first proposed a task on ECE and constructed a corpus for the task. They summarized seven groups of linguistic cues that could serve as an indicator of cause events. Based on these cues, some studies (Lee et al., 2010a, 2013; Chen et al., 2010; Gui et al., 2014; Li and Xu, 2014; Neviarouskaya and Aono, 2013; Yada et al., 2017) proposed various word-level or clause-level rule-based methods for this task. Gao et al. (2015a,b) presented a rule-based ECE method for microblogs based on cognitive theory. All these rule-based methods suffer from the problem of low coverage. In addition, they can't aware of the boundaries of the cause spans, which causes inferior performance when fine-grained ECE is required.

Gui et al. (2016a) formalized the ECE task as a clause-level binary classification problem and released a benchmark ECE dataset collected from news articles. Based on this corpus, many traditional machine learning methods (Gui et al., 2016a,b; Xu et al., 2017) and deep learning methods (Gui et al., 2017; Li et al., 2018; Yu et al., 2019; Ding et al., 2019; Li et al., 2019; Xia et al., 2019; Xu et al., 2019; Fan et al., 2019; Yan et al., 2021) were proposed. Xia and Ding (2019) introduced a new task called Emotion-Cause Pair Extraction (ECPE) in news articles and many following studies have been proposed on this task (Ding et al., 2020a; Fan et al., 2020; Ding et al., 2020b; Wei et al., 2020; Chen et al., 2020a,b; Wu et al., 2020; Singh et al., 2021). Recently, some works proposed to extract emotion causes at the span level, and pointed out that the span-level cause is more precise than the clause-level cause (Oberländer and Klinger, 2020; Bi and Liu, 2020; Li et al., 2021a,b).

Specific for the social media scenario, Song and Meng (2015) used topic modeling to extract word-level emotion causes in Chinese microblogs. Cheng et al. (2017) constructed a dataset with multiple-user structure for cause detection in Chinese microblogs. They proposed two cause detection tasks for microblogs (current subtweet-based cause detection and original subtweet-based cause detection) and used SVM and LSTM to deal with them. Chen et al. (2018b) presented a joint neural network approach for emotion classification and cause detection to obtain the mutual interaction across these two sub-tasks. Chen et al. (2018a) introduced a hierarchical Convolution Neural Network (Hier-CNN) to incorporate word contextual features and event-based features. Li et al. (2020) proposed a bootstrapping method to extract COVID-19 related triggers of different emotions on Twitter.

Although the learning-based approaches achieve sound emotion cause extraction performance, they suffer from the dependence on significant amounts of domain-specific fine-grained human annotations to reach their full potential.

5 Conclusion

In this paper, we explore how to build an emotion cause analysis system on social media without human annotation. First, we design a dedicated rule-based approach based on explicit causal cue words and constituency parsing tree, and then use it to annotate data on a large-scale corpus. On the basis of this, we introduce a strategy to alleviate the overfitting problem of the rule-annotated dataset and refine the labels during training to improve the generalization and scalability of our model. Experimental results on two datasets demonstrate the effectiveness of our approach.

Limitations

Although we have shown the potential of performing automatic emotion cause extraction (ECE) on social media without human annotation, there are still several limitations in our work.

Firstly, our work only considers the ECE task in Chinese microblogs. It might be interesting to investigate the effectiveness of our framework in social media platforms in other languages.

Secondly, we only focus on extracting the emotion cause expressed in the current post. However, according to Cheng et al. (2017), 37% of the emotion causes exist in the original or historical posts in a conversation thread. Hence, it would be interesting to extend our work to more complex microblog structures in the future.

Ethics Statement

Our methodology encompasses the creation of a Chinese emotion cause dataset using a rule-based approach, accompanied by the development of a human-annotated dataset. Importantly, all the posts we gathered are openly accessible, ensuring transparency and accessibility for fellow researchers. To ensure the privacy and confidentiality of our annotators and experts, we have implemented rigorous anonymization measures, retaining only their ID for research purposes. Moreover, we have ensured fair compensation for the efforts invested by our annotators, recognizing and valuing their contributions to the project.

Acknowledgments

This work was supported by the Natural Science Foundation of China (No. 62076133 and 62006117), and the Natural Science Foundation of Jiangsu Province for Young Scholars (No. BK20200463) and Distinguished Young Scholars (No. BK20200018).

References

Hongliang Bi and Pengyuan Liu. 2020. Ecsp: A new task for emotion-cause span-pair extraction and classification. *arXiv preprint arXiv:2003.03507*.

Xinhong Chen, Qing Li, and Jianping Wang. 2020a. A unified sequence labeling model for emotion cause pair extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 208–218.

Ying Chen, Wenjun Hou, and Xiyao Cheng. 2018a. Hierarchical convolution neural network for emotion cause detection on microblogs. In *International Conference on Artificial Neural Networks*, pages 115–122. Springer.

Ying Chen, Wenjun Hou, Xiyao Cheng, and Shoushan Li. 2018b. 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.

Ying Chen, Wenjun Hou, Shoushan Li, Caicong Wu, and Xiaoqiang Zhang. 2020b. End-to-end emotion-cause pair extraction with graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 198–207.

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 2010)*, pages 179–187, Beijing, China. Coling 2010 Organizing Committee.

Xiyao Cheng, Ying Chen, Bixiao Cheng, Shoushan Li, and Guodong Zhou. 2017. An emotion cause corpus for chinese microblogs with multiple-user structures. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(1):1–19.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6343–6350.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020a. 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, Online. Association for Computational Linguistics.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020b. End-to-end emotion-cause pair extraction based on sliding window multi-label learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3574–3583, Online. Association for Computational Linguistics.

Jingfei Du, Édouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418.

Chuang Fan, Hongyu Yan, Jiachen Du, Lin Gui, Lidong Bing, Min Yang, Ruifeng Xu, and Ruibin Mao. 2019. A knowledge regularized hierarchical approach for emotion cause analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5614–5624, Hong Kong, China. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Kai Gao, Hua Xu, and Jiushuo Wang. 2015a. Emotion cause detection for chinese micro-blogs based on ecocc model. In *Pacific-Asia Conference on*

Knowledge Discovery and Data Mining, pages 3–14. Springer.

Kai Gao, Hua Xu, and Jiushuo Wang. 2015b. A rule-based approach to emotion cause detection for chinese micro-blogs. *Expert Systems with Applications*, 42(9):4517–4528.

Lin Gui, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du. 2017. A question answering approach to emotion cause extraction. *arXiv preprint arXiv:1708.05482*.

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, Austin, Texas. Association for Computational Linguistics.

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

Lin Gui, Li Yuan, Ruifeng Xu, Bin Liu, Qin Lu, and Yu Zhou. 2014. Emotion cause detection with linguistic construction in chinese weibo text. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 457–464. Springer.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

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

Sophia Yat Mei Lee, Ying Chen, Chu-Ren Huang, and Shoushan Li. 2013. Detecting emotion causes with a linguistic rule-based approach 1. *Computational Intelligence*, 29(3):390–416.

Sophia Yat Mei Lee, Ying Chen, Shoushan Li, and Chu-Ren Huang. 2010b. Emotion cause events: Corpus construction and analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Min Li, Hui Zhao, Hao Su, YuRong Qian, and Ping Li. 2021a. Emotion-cause span extraction: a new task to emotion cause identification in texts. *Applied Intelligence*, 51(10):7109–7121.

Weiyuan Li and Hua Xu. 2014. Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications*, 41(4):1742–1749.

Xiangju Li, Shi Feng, Daling Wang, and Yifei Zhang. 2019. Context-aware emotion cause analysis with multi-attention-based neural network. *Knowledge-Based Systems*, 174:205–218.

Xiangju Li, Wei Gao, Shi Feng, Yifei Zhang, and Daling Wang. 2021b. Boundary detection with bert for span-level emotion cause analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 676–682.

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.

Xiaoya Li, Mingxin Zhou, Jiawei Wu, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. Analyzing covid-19 on online social media: Trends, sentiments and emotions. *arXiv preprint arXiv:2005.14464*.

Zhuojin Liu, Zhongxin Jin, Chaodi Wei, Xiangju Li, and Shi Feng. 2021. Coemocause: A chinese fine-grained emotional cause extraction dataset. In *International Conference on Web Information Systems and Applications*, pages 519–530. Springer.

Alena Neviarouskaya and Masaki Aono. 2013. Extracting causes of emotions from text. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 932–936, Nagoya, Japan. Asian Federation of Natural Language Processing.

Laura Ana Maria Oberländer and Roman Klinger. 2020. Token sequence labeling vs. clause classification for english emotion stimulus detection. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 58–70.

Aaditya Singh, Shreeshail Hingane, Saim Wani, and Ashutosh Modi. 2021. An end-to-end network for emotion-cause pair extraction. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 84–91, Online. Association for Computational Linguistics.

Shuangyong Song and Yao Meng. 2015. Detecting concept-level emotion cause in microblogging. In *Proceedings of the 24th International Conference on World Wide Web*, pages 119–120.

Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaxing Zhang. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Sixing Wu, Fang Chen, Fangzhao Wu, Yongfeng Huang, and Xing Li. 2020. A multi-task learning neural network for emotion-cause pair extraction. In *ECAI 2020*, pages 2212–2219. IOS Press.

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, Florence, Italy. Association for Computational Linguistics.

Rui Xia, Mengran Zhang, and Zixiang Ding. 2019. RTHN: A RNN-transformer hierarchical network for emotion cause extraction. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5285–5291.

Bo Xu, Hongfei Lin, Yuan Lin, Yufeng Diao, Liang Yang, and Kan Xu. 2019. Extracting emotion causes using learning to rank methods from an information retrieval perspective. *IEEE Access*, 7:15573–15583.

Ruifeng Xu, Jiannan Hu, Qin Lu, Dongyin Wu, and Lin Gui. 2017. An ensemble approach for emotion cause detection with event extraction and multi-kernel svms. *Tsinghua Science and Technology*, 22(6):646–659.

Shuntaro Yada, Kazushi Ikeda, Keiichiro Hoashi, and Kyo Kageura. 2017. A bootstrap method for automatic rule acquisition on emotion cause extraction. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 414–421. IEEE.

Hanqi Yan, Lin Gui, Gabriele Pergola, and Yulan He. 2021. Position bias mitigation: A knowledge-aware graph model for emotion cause extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3364–3375, Online. Association for Computational Linguistics.

Xinyi Yu, Wenge Rong, Zhuo Zhang, Yuanxin Ouyang, and Zhang Xiong. 2019. Multiple level hierarchical network-based clause selection for emotion cause extraction. *IEEE Access*, 7:9071–9079.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
line 633

□ A2. Did you discuss any potential risks of your work?
Not applicable. not applicable

☑ A3. Do the abstract and introduction summarize the paper's main claims?
line 3-31, line 136-142

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

sec 3.2.1, sec 4.3

☑ B1. Did you cite the creators of artifacts you used?
sec 3.2.1, sec 4.3

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
sec A.2

C ☑ Did you run computational experiments?

sec 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
sec 4.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
sec 4.2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
sec 4.4

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
sec 4.1

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
sec 4.1

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
sec 4.1

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
sec 4.1