A Knowledge Regularized Hierarchical Approach for Emotion Cause Analysis

Chuang Fan[1,5], **Hongyu Yan**[1,5], **Jiachen Du**[1,5], **Lin Gui**[2], **Lidong Bing**[3]
Min Yang[4], **Ruifeng Xu**[1,5,6]*, **Ruibin Mao**[7]
[1]Harbin Institute of Technology (Shenzhen), China [2]University of Warwick, UK
[3]R&D Center Singapore, Machine Intelligence Technology, Alibaba DAMO Academy
[4]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[5]Joint Lab of Harbin Institute of Technology and RICOH
[6]Peng Cheng Laboratory, China [7]Shenzhen Securities Information Co.,Ltd, China

Abstract

Emotion cause analysis, which aims to identify the reasons behind emotions, is a key topic in sentiment analysis. A variety of neural network models have been proposed recently, however, these previous models mostly focus on the learning architecture with local textual information, ignoring the discourse and prior knowledge, which play crucial roles in human text comprehension. In this paper, we propose a new method to extract emotion cause with a hierarchical neural model and knowledge-based regularizations, which aims to incorporate discourse context information and restrain the parameters by sentiment lexicon and common knowledge. The experimental results demonstrate that our proposed method achieves the state-of-the-art performance on two public datasets in different languages (Chinese and English), outperforming a number of competitive baselines by at least 2.08% in F-measure.

1 Introduction

Sentiment analysis has gained increasing popularity in recent years due to many useful applications (Pang and Lee, 2007). The goal of sentiment analysis is to classify the sentiment polarity of a given text as positive, negative, neutral, or more fine-grained classes (Kim, 2014; Li et al., 2015; Yang et al., 2016; Zhou et al., 2016; Chen et al., 2017; Qian et al., 2017; Li et al., 2018b; Yu et al., 2018). Most of these researches have assumed that emotion expressions are already observed and try to identify the emotion categories from text. However, in practice, such as in product comments or political reviews, we may care more about the reason why the customers or critics hold the emotion rather than a simple category label. Because they can improve the quality of the products or services according to the emotion cause provided by users. Emotion cause analysis (ECA) aims to identify the reasons behind a certain emotion expression in an event text, for example:

Ex.1 *When the children saw the gifts I prepared carefully,* (c_{-2})*|they cheered happily and hugged me.* (c_{-1})*| I was full of happiness.* (c_0)

Here, **Ex.1** shows a document with three clauses marked as (c_{-2}), (c_{-1}), and (c_0). The goal of ECA is to determine which clause contains emotion cause (e.g., (c_{-1})) for an emotion word (e.g., *happiness* in (c_0)).

Previous approaches for emotion cause analysis mostly depend on rule-based methods (Lee et al., 2010; Chen et al., 2010) and machine learning algorithms (Ghazi et al., 2015; Gui et al., 2016; Xu et al., 2019). Most of them rely heavily on complicated linguistic rules or feature engineering, which is time-consuming and labor-intensive. Recent studies have focused on solving the task using neural models (Gui et al., 2017; Li et al., 2018a; Chen et al., 2018; Li et al., 2019) with well designed attention mechanism based on local text. Despite the effectiveness of neural models, there are some defects in previous studies. First, they usually consider each clause individually, i.e., ignoring the discourse context information that can impact the semantic expression among different clauses of a document. Second, prior knowledge such as sentiment lexicon and relative position information that can provide crucial emotion cause cues has not been fully exploited in neural models.

To alleviate these limitations, we propose a regularized hierarchical neural network (RHNN) for emotion cause analysis, which combines the discourse context information and knowledge-based regularizations. Our model investigates the following intuitions. Firstly, documents exhibit discourse structure which may carry valuable information about the emotion cause cues. We em-

* Corresponding Author: xuruifeng@hit.edu.cn

ploy a hierarchical learning structure to capture the mutual impacts of semantic expression among the discourse context, to help produce better clause representation. Secondly, in emotion events, emotion causes usually express a certain sentiment polarity by some sentiment words. For example, in the emotion cause of **Ex.1**, the sentiment words *cheered* and *happily* express a positive polarity, also play a crucial role to provoke the emotion *happiness*. Therefore, capturing these sentiment words can enhance the causal connection between learned features and predictions. We approach this issue by designing a regularizer that incorporates linguistic knowledge (e.g., sentiment lexicon) to enlarge the margin of attention weights of sentiment words and non-sentiment words. Besides, it is often the case that humans usually write important points in different sections. For emotion events, emotion causes generally occur on positions very close to the emotion word and occur frequently. **Ex.1** shows an anecdotal example illustrating this behavior that the emotion cause clause c_{-1} adjoins the emotion word *happiness*. To benefit from this phenomenon, we introduce a regularizer biased by relative position information to supervise the representation learning of text and further to revise the **predictive position distribution of emotion causes relative to the emotion word** (in brief, predictive distribution).

To sum up, our contribution includes:

- We propose a novel discourse-aware learning structure with knowledge-based regularizations for emotion cause analysis.
- We empirically evaluate the proposed model on two public datasets in different languages (Chinese and English) and show statistically significant improvements compared to the state-of-the-art methods.
- To make the mechanism of our model clear, we also compare the performance of different combinations by ablation experiments. Extensive analysis on both datasets confirms the feasibility of incorporating discourse information and restraining the parameters by sentiment lexicon and common knowledge.

2 Our Framework

In this section, we first give the task definition. Then, our proposed regularized hierarchical neural network (RHNN), as shown in Fig 1, will be described. The two auxiliary regularizers of RHNN will be introduced in the next section.

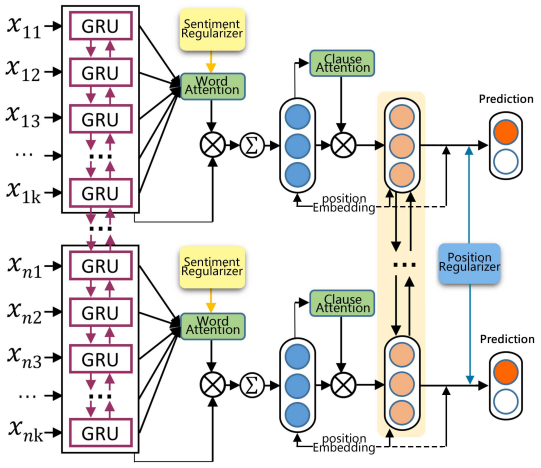

Figure 1: The architecture of our model.

2.1 Task Definition

The formal definition of emotion cause analysis is given in (Gui et al., 2016). Formally, for a document $d = \{c_1, c_2, \ldots, c_n\}$ consisting of n clauses, it contains an emotion word e and at least an emotion cause clause corresponding to this emotion word. Each clause $c_i = \{w_{i1}, w_{i2}, \ldots, w_{ik}\}$ consists of k words and is labeled with emotion cause-oriented labels $\in \{0, 1\}$. We regard ECA as a binary classification task and aim to identify which clause contains emotion cause.

2.2 Hierarchical Attention Network

Documents exhibit discourse structure which can serve as useful information for clause representation generation. One simple but effective approach is to adopt a hierarchical attention network to simulate this structure. Our hierarchical attention network consists of several parts: a word encoder, a word attention layer, a clause attention layer and a clause encoder. The details of each component will be described in the following paragraphs.

Word Encoder Gated Recurrent Unit (GRU) has been widely adopted for text processing (Cho et al., 2014). In this work, we first map each word into a low dimensional embedding space by Word2Vec (Mikolov et al., 2013) and then feed the whole document into a GRU-based word encoder to extract word sequence features. To summarize information from both directions, we use bidirec-

tional GRU to exploit two parallel passes:

$$\overrightarrow{h_{it}} = \overrightarrow{\mathrm{GRU}_w}(x_{it}), t \in [1, k] \quad (1)$$

$$\overleftarrow{h_{it}} = \overleftarrow{\mathrm{GRU}_w}(x_{it}), t \in [k, 1] \quad (2)$$

where x_{it} is embedding vector for the word w_{it} in clause c_i at time step t and k is the length of clause c_i. Then we concatenate hidden states of the two directions $h_{it} = [\overrightarrow{h_{it}}, \overleftarrow{h_{it}}]$ as the representation of each word.

Word Attention We introduce an attention mechanism to extract such words that are important to the meaning of the clause and aggregate the representation of these informative words to construct the clause vector. Specifically,

$$g_{it} = w_m^T(h_{it} \oplus e^E) \quad (3)$$

$$\alpha_{it} = \frac{\exp(g_{it})}{\sum_{t'} \exp(g_{it'})} \quad (4)$$

$$o_{c_i} = \sum_t \alpha_{it} h_{it} \quad (5)$$

where w_m is the parameter for computing attention signals and $\oplus$ is concatenate operation. The embedding of emotion word e is denoted by e^E. α_{it} is the emotion-specific attention signal showing the importance of word w_{it}. o_{c_i} is the weighted sum of word representation based on weights.

Clause Attention Intuitively, different clauses of a document are different informative and should be labeled with different importance. Targeting this problem, we design a clause attention mechanism to indicate the importance of each clause. See it differently, the attention signals can be regarded as some "prior" information to bias the clause encoder toward some content that is more important to extracting the emotion cause.

As for details, we adopt a one-layer MLP to get the attention signals. Apparently, position information plays an important role in capturing the relative distance of the clause to emotion word. Thus, we concatenate the clause vector o_{c_i} and its position embedding as the feature to obtain the attention signal, this yields:

$$\alpha_i = \mathrm{sigmoid}(w_v(o_{c_i} \oplus l_i)) \quad (6)$$

$$o'_{c_i} = o_{c_i} \cdot \alpha_i \quad (7)$$

where w_v is a parameter vector, l_i is the randomly initialized position embedding and keeps unchanged in the training stage, α_i is the weight of clause c_i. Then the clauses with different importance (e.g., o'_{c_i}) are fed into a clause encoder.

Clause Encoder Just as the meaning of a word is determined by its context, the semantic expression of a clause is usually impacted by its discourse context. Based on this observation, we introduce a clause encoder to model the latent semantic relations among different clauses. Analogously, we also append the relative position information to enhance the relations between the clause and its position. Formally,

$$\overrightarrow{h_i} = \overrightarrow{\mathrm{GRU}_c}(o'_{c_i} \oplus l_i), i \in [1, n] \quad (8)$$

$$\overleftarrow{h_i} = \overleftarrow{\mathrm{GRU}_c}(o'_{c_i} \oplus l_i), i \in [n, 1] \quad (9)$$

where n is the number of clauses in a document. Also, the two directional hidden state $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ are concatenated as the final emotion-specific representation $o_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$.

2.3 Model Training

The emotion-specific representation o_i with its position embedding l_i as the final feature for emotion cause prediction and the model is trained by minimizing the cross entropy:

$$y_i = \mathrm{softmax}(W_m(o_i \oplus l_i)) \quad (10)$$

$$\mathcal{L}_{ce} = -\sum_i \widehat{y}_i \log y_i \quad (11)$$

where W_m is a parameter matrix, $\widehat{y}_i$ and y_i are target class distribution and predictive class distribution respectively.

3 Regularizers Based on Sentiment Lexicon and Relative Position

One crucial emotion cause cue is sentiment words since emotion causes usually express a certain sentiment polarity by sentiment words. However, there is no effective mechanism to guarantee that the above module indeed attends the words with sentiment polarity. Beyond this, a straightforward solution to inject position information is to directly incorporate relative word position embedding, which is a weak representation of the relative distance between the emotion word and each clause. To ease these problems, we introduce two auxiliary regularizers:

- **Sentiment Regularizer (SR.):** If a clause contains several words which exist in a sentiment lexicon, the calculated average weights of these words should be properly larger than other words. We approach this issue with an auxiliary hinge loss function.

- **Position Regularizer (PR.):** Relative position is a critical emotion cause indicator: in general the closer a clause is to the emotion word, the higher emotion cause probability it should be assigned. We approach this issue by introducing a proxy distribution and an auxiliary cross entropy function.

Formally, we disassemble the joint loss of emotion cause detection into an original cross entropy loss, a sentiment regularization loss, and a position regularization loss. The new training objective is revised as:

$$\mathcal{J}(\theta) = \mathcal{L}_{ce} + \lambda_1 * \mathcal{L}_{sr} + \lambda_2 * \mathcal{L}_{pr} \quad (12)$$

where $\mathcal{L}_{ce}$ is the original cross entropy loss (§2.3). λ_1 and λ_2 are hyper-parameters. $\mathcal{L}_{sr}$ and $\mathcal{L}_{pr}$ are two auxiliary regularization losses (§3.1 and §3.2). θ is the parameter set.

3.1 Sentiment Regularizer

The weak causal connection between learned features and predictions is a major issue in emotion cause analysis. Even though sentiment words are key to clause representation generation, most existing models do not focus on sentiment words or place less emphasis on them when producing clause representation. In other words, attention is distracted by irrelevant words with less sentiment polarity. To address this issue and sufficiently benefit from linguistic resources, we explicitly encourage the larger margin of attention weights between sentiment words and non-sentiment words using a sentiment lexicon. For sentiment words, the average attention weight is calculated by:

$$Avg_s = \frac{1}{l_s} \sum_{t=1}^{k} S_{w_{it}} \quad (13)$$

$$S_{w_{it}} = \begin{cases} \alpha_{it} & \text{if} \quad w_{it} \in c_i \cap \mathcal{S} \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

where w_{it} is the t-th word of clause c_i, k is the length of c_i and l_s is the number of sentiment words in c_i. α_{it} is the calculated attention weight in §2.2 and $\mathcal{S}$ is a sentiment lexicon. Correspondingly, for non-sentiment words:

$$Avg_{ns} = \frac{1}{l_{ns}} \sum_{t=1}^{k} NS_{w_{it}} \quad (15)$$

$$NS_{w_{it}} = \begin{cases} \alpha_{it} & \text{if} \quad w_{it} \in c_i - \mathcal{S} \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

Our training objective is to lead the model to pay more attention to sentiment words. Thus, the regularization term is formally expressed as:

$$\mathcal{L}_{sr} = \sum_i \max(0, m - (Avg_s - Avg_{ns})) \quad (17)$$

where m is a hyper-parameter for margin.

3.2 Position Regularizer

Empirically, emotion causes usually occur at the positions which are very close to the emotion word. However, another main issue is that the predictive distribution may locates the clauses that are distant from and irrelevant to the emotion word. The goal of this regularizer is to narrow the difference between the predictive distribution and the **true position distribution of emotion causes relative to the emotion word** (in brief, true distribution). Obviously we can not obtain the true distribution. Hence, we assume that it should satisfy the following conditions: (1) It should be a normed function within $[0, 1]$; (2) It should be a symmetric function of a certain value. Based on these conditions, we employ a function defined as follows:

$$q_i = \begin{cases} 1 - \frac{|r_i|}{n} & \text{if} \ -b \leq r_i \leq b \\ 0 & \text{otherwise} \end{cases} \quad (18)$$

where r_i is the relative distance of clause c_i to emotion word, n is the number of clauses in a document, and b is the left and right boundary which limits the scope of emotion cause. Then we apply $q = (q_1, q_2, \ldots, q_n)$ as the proxy distribution to simulate the true distribution.

Simultaneously, from section 2.3, we get the predictive class distribution y_i of clause c_i. Then, the probability for the emotion cause at position i can be calculated as:

$$p_i = y_i(label = 1) \quad (19)$$

Similarity, we can obtain the predictive distribution of emotion causes relative to the emotion word by:

$$p = (p_1, p_2, \ldots, p_n) \quad (20)$$

The goal is to enforce the model to constrain the difference between the p and q, thus, we use cross entropy to measure the difference:

$$\mathcal{L}_{pr} = -\sum_i q_i \log(p_i) \quad (21)$$

Note that the two introduced regularizers work like L_1 and L_2 terms, which **do not introduce**

Item	Number	Item	Number
Chinese Dataset (Chi.)			
Documents	2105	Cause_1	2046
Clauses	11799	Cause_2	56
Causes	2167	Cause_3	3
English Dataset (Eng.)			
Documents	2156	Cause_1	1949
Clauses	16259	Cause_2	164
Causes	2421	Cause_3	32

Table 1: Details of the two datasets. Cause_1, Cause_2 and Cause_3 represent the documents with 1, 2 and 3 cause clauses, respectively.

any new parameters and only influence the training of the standard model parameters. The hyper-parameters λ_1 and λ_2 guide the model to achieve the best trade-off among three types of losses.

4 Datasets and Implementation Details

4.1 Datasets

We select two public datasets from different languages to evaluate the proposed model: Chinese Dataset (Gui et al., 2016) collected from SINA city news[1] and English Dataset (Gao et al., 2017) collected from English novels. Each document of both datasets has only one emotion word and one or more emotion causes. It has been ensured that the emotion and the causes are relevant. The documents are segmented into several clauses manually for emotion cause analysis. The details about the two datasets are summarized in Table 1.

4.2 Implementation Details

For the Chinese dataset, there is no training/test split, we randomly divide the documents into a training/development/test set in a ratio of 8:1:1 and partition the clauses by Jieba[2]. For the English dataset, we randomly select 10% from the original training set as the development set and lowercase, lemmatize all the tokens by NLTK[3]. We evaluate our method 25 times with different splits and then perform one sample t-test on the experimental results by following (Gui et al., 2017). The precision (P), recall (R) and F-measure (F) are employed to measure the performance in this task.

The sentiment lexicon adopted for the Chinese dataset consists of two parts. The first part is selected from HowNet (Dong et al., 2006) sentiment analysis lexicon set[4] and the second part comes from NTUSD (Ku et al., 2006). The combination of the two parts serves as the Chinese sentiment lexicon of this research. The English sentiment lexicon comes from MPQA (Wilson et al., 2005) and we only select the words with high sentiment polarity, because they are less sensitive to contextual information and usually express consistence sentiment polarities from their prior polarity. For both sentiment lexica, we filter out the words that are not in the datasets. Ultimately, 2022 and 1348 sentiment words are selected for the Chinese and English dataset respectively.

Online learning is performed with the Adam optimizer (Kingma and Ba, 2015) and initial learning rate 0.001 is adopted. The number of layers in Bi-GRU is set to 2 and dropout rate 0.5 is used to avoid overfitting. The word vectors are pre-trained by word2vec (Mikolov et al., 2013) and keep unchanged during training stage. We perform grid search over the hyper-parameters m ($\{0.10, 0.15, 0.20\}$), the boundary b ($\{2, 3, 4\}$), the dimensionality of the Bi-GRU ($\{64, 128\}$), λ_1 and λ_2 (both $\{0.25, 0.5, 0.75\}$). For each corpus, the highest F-measure combination of these hyper-parameters is selected using development set.

5 Experiments

In this section, we will compare our RHNN model with the following groups of methods:

- Rule-based and commonsense-based methods: Rule-based method (**RB**) is a traditional rule-based method proposed by Lee et al. (2010). Commonsense-based methods (**CB**) is a knowledge-based method proposed by Russo et al. (2011). It uses Chinese Emotion Cognition Lexicon (Xu et al., 2013) as commonsense knowledge.

- Machine learning method: **SVM** is a SVM classifier trained on unigrams, bigrams and trigrams features (Li and Xu, 2014). **Word2vec** is a SVM classifier trained on word representations pre-trained by Word2vec (Mikolov et al., 2013). **Multi-kernel** represents a document by a syntactic structure and utilizes a modified convolution kernel method to determine which clause contains the emotion cause (Gui et al., 2016).

[1] http://news.sina.com.cn/society/
[2] https://github.com/fxsjy/jieba
[3] http://nltk.org
[4] http://www.keenage.com/html/c_index.html

Method	P	R	F
RB*	0.6747	0.4287	0.5243
CB*	0.2672	0.7130	0.3887
SVM*	0.4200	0.4375	0.4285
Word2vec*	0.4301	0.4233	0.4136
Multi-kernel*	0.6588	0.6972	0.6752
LambdaMART*	0.7720	0.7499	0.7608
CNN*	0.6472	0.5493	0.5915
ConvMS-Memnet*	0.7076	0.6838	0.6955
CANN	0.7721	0.6891	0.7266
HCS	0.7388	0.7154	0.7269
MANN	0.7843	0.7587	0.7706
RHNN	**0.8112**	**0.7725**	**0.7914**

Table 2: Experimental results on the Chinese dataset. Superscript $*$ indicates the results are reported in (Gui et al., 2017) and the rest are reprinted from the corresponding publications (p <0.001).

Method	P	R	F
Word2vec	0.1651	**0.8673**	0.2774
SVM	0.2757	0.6416	0.3856
CNN	0.7218	0.2628	0.3390
ConvMS-Memnet	0.4605	0.4177	0.4381
MANN	**0.7933**	0.4081	0.5328
RHNN	0.6901	0.5267	**0.5975**

Table 3: Experimental results on the English dataset, we follow the results that are implemented in (Li et al., 2019), the only available results on this dataset (p <0.001).

LambdaMART extracts emotion causes using learning to rank methods which based on the emotion-independent and emotion-dependent features (Xu et al., 2019).

- Deep learning method: **CNN** is a convolutional neural network for sentence classification (Kim, 2014). **ConvMS-Memnet** considers emotion cause analysis as a reading comprehension task and designs a multiple-slot deep memory network to model context information (Gui et al., 2017). **CANN** uses a co-attention neural network to identify emotion causes (Li et al., 2018a). **HCS** is proposed by Yu et al. (2019) using a multiple-level hierarchical network to detect the emotion causes. **MANN** is the current state-of-the-art method employing a multi-attention-based model for emotion cause extraction (Li et al., 2019). **RHNN** is our proposed model.

5.1 Main Results

The experimental results on both datasets are shown in Table 2 and Table 3, respectively. RB yields high precision but with low recall. CB has an opposite scenario from RB. A possible reason is that these linguistic-based methods depend on some cue words to identify the emotion cause, different rules or common sense may contain different cue words.

For the machine learning methods, SVM and Word2vec have similar performance on the Chinese dataset, but SVM outperforms Word2vec on the English dataset. The main reason is that the polysemantic phenomenon is more obvious in English expressions. Multi-kernel has better performance by capturing context information through a syntactic tree. LambdaMART, which is based on ranking strategy and global emotion features, performs best among feature-based methods. However, both Multi-kernel and LambdaMART rely on expensive human-based features and lack of expandability on different dataset.

Compared with CNN, ConvMS-Memnet models the context of each word and obtains better performance on both datasets. The co-attention based CANN captures the mutual relations between the emotion clause and each candidate clause, which has a comparable result with hierarchical-based HCS. MANN considers the interaction between the emotion clause and candidate clauses by designing a multi-attention mechanism and achieves the best performance among baselines.

The proposed RHNN model further improves the performance on both datasets as shown in the tables. The improvement is significant with p-value less than 0.001 in one sample t-test. Specifically, RHNN manages to boost the performance by 3.06% in F-measure compared to LambdaMART, which exhibits that by restraining the parameters with knowledge-based regularizations, RHNN is better to identify the emotion cause cues than feature engineering. RHNN also outperforms the current best-performing method MANN by 2.08% on the Chinese dataset and 6.47% on the English dataset in F-measure respectively. Furthermore, for the English dataset, our proposed model has balance performance in precision and recall. The reason for this phenomenon is that RHNN can capture more emotion cue (e.g., sen-

Method	Applying with			Chi.	Eng.
	H.	SR.	PR.	F	F
Base	×	×	×	0.7152	0.5199
H	✓	×	×	0.7553	0.5668
SR	×	✓	×	0.7570	0.5765
PR	×	×	✓	0.7483	0.5683
HSR	✓	✓	×	0.7860	**0.6154**
HPR	✓	×	✓	0.7659	0.5777
SPR	×	✓	✓	0.7600	0.5878
RHNN	✓	✓	✓	**0.7914**	0.5975

Table 4: Effect of different components, i.e., hierarchical structure (H.), sentiment regularizer (SR.) and position regularizer (PR.). The leftmost column is the abbreviation for corresponding sub-models (e.g., SPR denotes the sub-model with SP. and PR. except H.).

Method	Chi.		Eng.	
	Sub.	All.	Sub.	All.
SR	0.7650	0.7570	0.6713	0.5765
HSR	0.7934	0.7860	**0.7084**	**0.6154**
SPR	0.7741	0.7600	0.6904	0.5878
RHNN	**0.7997**	**0.7914**	0.6849	0.5975

Table 5: The F-measure on sub-dataset (Sub.) that only selects the clauses which contain sentiment words, and all-dataset (All.) that experiments on the whole dataset.

timent words) information to optimize the model extracting emotion causes more exactly.

5.2 Detailed Analysis

Ablations of RHNN Model The proposed RHNN model consists of three components, including hierarchical structure (H.), sentiment regularizer (SR.) and position regularizer (PR.). We conduct ablation experiments to reveal the effect of each component. As illustrated in Table 4, all models with the proposed component consistently improve upon the Base model, verifying the effectiveness of the proposed approach. Compare with H. and PR. model, the SR. improve the performance most. The main reason is that there are 55.24% and 66.49% of emotion causes which contain sentiment words on the Chinese dataset and English dataset respectively, enforcing the model to pay more attention to sentiment words can enhance the causal connection between learned features and predictions.

On the Chinese dataset, the RHNN achieves the best performance with a 7.62% improvement on

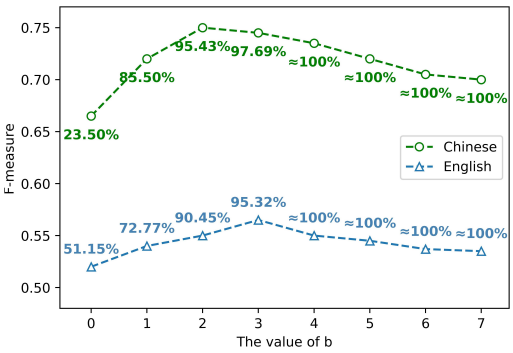

Figure 2: The F-measure of different limited scopes of emotion cause on two datasets.

the F score compared with the baseline. However, on the English dataset, the HSR model performs better than the RHNN model. It may be caused by the overlapping between components. Besides, the performance on the English dataset always lower about 20% in F-measure than that on the Chinese dataset, one possible explanation for this phenomenon is that there are more clause structures in English expressions which is difficult for the model to capture this information without discourse tree.

Effect of Sentiment Regularizer To gain more insights into our proposed model, we conduct further experiments to examine the effectiveness of sentiment regularizer on the subset of two datasets. The experiment results in Table 5 show that: 1) RHNN and HSR model achieve the best performance on the subset of two datasets respectively, similar observations can be found regarding on whole dataset; 2) With sentiment regularizer, the performance is boosted on both subsets compared to that on the whole dataset. This is consistent with our intuition because sentiment regularizer contributes much to pick up the words with sentiment polarity and these words are important causal indicators in clauses. Meanwhile, each clause contains sentiment words in the subsets, resulting in a better performance on the subsets. 3) The performance improvement on the English subset is remarkably higher than that on the Chinese subset, one possibility is that there are more explicit emotion terms in English expression than in Chinese.

Effect of Position Regularizer From Eq.(18), we see that the value of b limits the scope of emotion cause. In this section, we further investigate the effect of different limited scopes. For simplicity and efficiency, here we only apply the PR

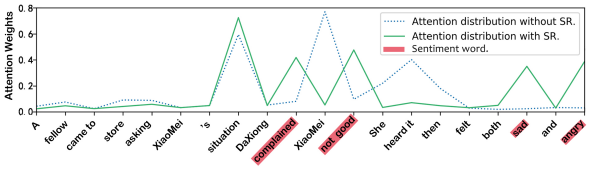

Figure 3: Comparison of attention distribution with or without sentiment regularizer.

model on this experiment. The results are shown in Fig 2, the percentages denote the coverage of emotion cause in datasets. For instance, 72.77% represents that there are 72.77% emotion causes adjoining the emotion word in English dataset. From Fig 2, we can see that the performance trends for the two datasets are similar, and the performance improves with the expanding limited scope of emotion cause. However, when the limited scope of emotion cause is larger than a certain value (2 or 3), the performance decreases. This may be due to the reason that the larger of limited scope, the higher coverage of emotion cause in datasets. Nevertheless, when the limited scope is too large, the model is forced to allocate higher probability to these clauses which are distant from and irrelevant to the emotion word, and then leads to the performance degradation.

5.3 Case Study

Essentially, sentiment regularizer (SR.) aims to enlarge the margin between sentiment words and non-sentiment words. The question is, will the model focus on the words with sentiment polarity? We randomly choose one example (**Ex.2**) from the Chinese dataset to visualize its attention distribution and compare the difference between plusing sentiment regularizer or not.

Ex.2 一个老乡来到店里问起小美的情况, (c_{-2})|大熊抱怨小美不好. (c_{-1})|她听到后感到既伤心又生气. (c_0)

Ex.2 *A fellow came to store asking XiaoMei's situation,* (c_{-2})*|DaXiong complained XiaoMei not good.* (c_{-1})*| She heard it then felt both sad and angry.* (c_0)

In this example, the cause of emotion word *sad* is in (c_0). The visualization results are shown in Fig 3, we can observe that the model without sentiment regularizer most focus on non-sentiment words such as *XiaoMei*, *she* and *heard it* which are inessential to provoke the emotion. However, when we plus the SR. into model, we can see an obvious weights shift on attention distribution.

Emotion cause events	RHNN results
1 ***I was immediately *ashamed* of myself for my vanity, for having assumed that he wanted me to stay with him forever.*** *I'm sorry, that was a little arrogant*	for having assumed that he wanted me to stay with him forever
2 ***He hopes of being *admitted* to a sight of the young ladies,*** *of whose beauty he had heard much. But he saw only the father.*	of whose beauty he had heard much
3 ***I didn't know where in the hell you was,*** *said Ennis, four years, I about *give up* on you.*	**Null**

Table 6: The error instances of RHNN model for emotion cause analysis.

More clearly, the model captures the sentiment words *complained* and *not good* which are crucial to identify the emotion cause. This shows that our model with sentiment regularizer is more effective in extracting the most important keywords relating to the emotion cause. Also, better results are obtained using sentiment regularizer, this is consistent with what we observed in Table 4.

Finally, we perform error analysis to understand what types of errors are introduced with the proposed model, focusing on three cases from the English dataset. The results are listed in Table 6, where the first column depicts the content of the emotion cause events and the second column depicts the emotion causes identified by RHNN. As shown in Table 6, the emotion causes appear in bold and emotion word is labeled between *.

From Table 6, we can find that there are two clauses contain the emotion cause in event 1. However, our model only detects one emotion cause clause. In event 2, our model has an error prediction. One possible reason is that our model is prone to treating clauses which contain sentiment words as emotion cause. RHNN extracts nothing from event 3, it may be due to the reason that the far distance between the emotion word and emotion cause clause, resulting in a difficult understanding of causal relations. Our proposed model is capable of getting rich emotion cause

cues with knowledge-based regularizations. However, it also introduces some noisy into emotion cause analysis.

6 Related Work

Emotion classification is an important fundamental aspect of sentiment analysis. Going one step further, emotion cause analysis (ECA) which aims to discover the reason behind emotions, can be constructive to guide the direction of future work, i.e., improving the quality of products or services according to the emotion causes of comments provided by users. In this section, we describe the related work on emotion cause analysis.

Lee et al. (2010) first gave the formal definition of emotion cause analysis and manually constructed a dataset from the Academia Sinica Balanced Chinese Corpus. Based on this corpus, Chen et al. (2010) designed two sets of features built on six groups of linguistic cues to detect emotion cause. Support vector machines (SVMs) and conditional random fields (CRFs) were investigated to detect cause or non-cause text with extended rule-based features in existing methods (Gui et al., 2014; Ghazi et al., 2015). Other than rule-based methods, Russo et al. (2011) proposed a crowdsourcing method to construct a commonsense knowledge base for emotion cause extraction in Italian newspaper articles. But it is challenging to extend the common-sense knowledge base automatically. Recently, Gui et al. (2016) proposed a multi-kernel based method to identify the emotion cause from a manually annotated emotion cause corpus. Xu et al. (2019) proposed a method based on learning to re-rank candidate emotion cause clauses with extracting a number of emotion-dependent and emotion-independent features. However, these methods are heavily dependent on the expensive human-based features and are too difficult in a real-world application.

Inspired by the success of neural network methods, deep neural models and attention mechanisms have been widely used in emotion cause analysis. Gui et al. (2017) proposed a novel deep neural network which regarded emotion cause analysis as a question-answering task. In this study, a convolution-based memory network was introduced to store the context information. Li et al. (2018a) considered the context around the emotion word as a query instead of only emotion word to model the mutual impacts between each candidate clause and the emotion clause. Cheng et al. (2017) constructed a corpus based on Chinese microblog and proposed to detect emotion cause using multiple-user structures. Besides, Yu et al. (2019) proposed a multiple-level hierarchical network-based clause selection strategy. Li et al. (2019) proposed a multi-attention-based neural model to capture the mutual influences between the emotion clause and each candidate clause, and then generate the representations for the above two clauses separately. This method achieves the current best performance. However, the existing approaches usually focus on the local textural information, ignoring the discourse structure (Zubiaga et al., 2018), and prior knowledge such as sentiment lexicon (Qian et al., 2017) and relative position information, which can provide important emotion cues for emotion cause analysis task.

7 Conclusion and Future Work

In this paper, we provide a regularized hierarchical neural network (RHNN) for emotion cause analysis. The proposed model aggregates discourse context information through a hierarchical learning structure and restrains the parameters with knowledge-based regularizations. We evaluate the proposed model on two public datasets in different languages. The experimental results demonstrate that our proposed method achieves the state-of-the-art performance on both datasets and extensive analysis confirms the feasibility of incorporating the discourse context and knowledge-based regularizations.

To preserve the simplicity of the proposed model, we do not consider document as a tree structure. In the future, we will exploit how to incorporate discourse parse tree or discourse relations into emotion cause analysis task to further improve the performance.

8 Acknowledgements

This work was supported by National Natural Science Foundation of China U1636103, 61632011, 61876053, Shenzhen Foundational Research Funding JCYJ20180507183527919, JCYJ20180507183608379, Key Technologies Research and Development Program of Shenzhen JSGG20170817140856618, Joint Research Program of Shenzhen Securities Information Co., Ltd. No. JRPSSIC2018001, and the EU-H2020 (grant no. 794196).

References

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 452–461.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 646–651.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause detection with linguistic constructions. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 179–187.

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):6.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.

Zhendong Dong, Qiang Dong, and Changling Hao. 2006. Hownet and the computation of meaning.

Qinghong Gao, H Jiannan, X Ruifeng, G Lin, Y He, KF Wong, and Q Lu. 2017. Overview of ntcir-13 eca task. In *Proceedings of the NTCIR-13 Conference*.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II*, pages 152–165.

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, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1593–1602.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1639–1649.

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 *Natural Language Processing and Chinese Computing*, pages 457–464. Springer.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.

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

Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA, March 27-29, 2006*, pages 100–107.

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. Association for Computational Linguistics.

Shoushan Li, Lei Huang, Rong Wang, and Guodong Zhou. 2015. Sentence-level emotion classification with label and context dependence. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1045–1053.

Weiyuan Li and Hua Xu. 2014. Text-based emotion classification using emotion cause extraction. *Expert Syst. Appl.*, 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. *Knowl.-Based Syst.*, 174:205–218.

Xiangju Li, Kaisong Song, Shi Feng, Daling Wang, and Yifei Zhang. 2018a. 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, Brussels, Belgium, October 31 - November 4, 2018, pages 4752–4757.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018b. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 946–956.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing System*, pages 3111–3119.

Bo Pang and Lillian Lee. 2007. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Qiao Qian, Minlie Huang, Jinhao Lei, and Xiaoyan Zhu. 2017. Linguistically regularized LSTM for sentiment classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1679–1689.

Irene Russo, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco. 2011. Emocause: An easy-adaptable approach to extract emotion cause contexts. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 153–160.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, pages 347–354.

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, Chengtian Zou, Yanzhen Zheng, Xu Jun, Lin Gui, Bin Liu, and Xiaolong Wang. 2013. A new emotion dictionary based on the distinguish of emotion expression and emotion cognition. *Journal of Chinese Information Processing*, 27(6):82–90.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1480–1489.

Jianfei Yu, Luís Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1097–1102.

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.

Deyu Zhou, Xuan Zhang, Yin Zhou, Quan Zhao, and Xin Geng. 2016. Emotion distribution learning from texts. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 638–647.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Inf. Process. Manage.*, 54(2):273–290.