Emotion Cause Detection with Linguistic Construction in Chinese Weibo Text

Lin Gui[1], Li Yuan[1], Ruifeng Xu[1,*], Bin Liu[1], Qin Lu[2], and Yu Zhou[1]

[1] Key Laboratory of Network Oriented Intelligent Computation,
Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
{guilin.nlp,yuanlisail}@gmail.com, xuruifeng@hitsz.edu.cn,
bliu@insun.hit.edu.cn, zhouyu.nlp@gmail.com
[2] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
csluqin@comp.polyu.edu.hk

Abstract. To identify the cause of emotion is a new challenge for researchers in nature language processing. Currently, there is no existing works on emotion cause detection from Chinese micro-blogging (Weibo) text. In this study, an emotion cause annotated corpus is firstly designed and developed through annotating the emotion cause expressions in Chinese Weibo Text. Up to now, an emotion cause annotated corpus which consists of the annotations for 1,333 Chinese Weibo is constructed. Based on the observations on this corpus, the characteristics of emotion cause expression are identified. Accordingly, a rule-based emotion cause detection method is developed which uses 25 manually complied rules. Furthermore, two machine learning based cause detection methods are developed including a classification-based method using support vector machines and a sequence labeling based method using conditional random fields model. It is the largest available resources in this research area. The experimental results show that the rule-based method achieves 68.30% accuracy rate. Furthermore, the method based on conditional random fields model achieved 77.57% accuracy which is 37.45% higher than the reference baseline method. These results show the effectiveness of our proposed emotion cause detection method.

Keywords: Emotion cause detection, corpus construction, Chinese Weibo.

1 Introduction

The rise of social media has produced flooded of text such as Blogs and micro-blogging. How to analyze the emotions from these texts is becoming a new challenge for natural language processing researchers. Generally speaking, there are three basic tasks in emotion computation. 1. Emotion analysis, which focuses on how to classify the emotion categories of the texts and extract the holder/target of the emotion; 2. Emotion prediction, which predicts the readers' emotion after they read the given texts; and 3. Emotion cause detection, which extracts the cause of emotion in the text.

* Corresponding author.

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 457–464, 2014.

Currently, most existing works on emotion computing focus on the emotion analysis [1-5] and emotion prediction [6-9]. Few works on emotion cause detection are reported. Meanwhile, the majority of these works follows linguistic based approach [10-12]. As we know, emotion is "physiological arousal, expressive behaviors and conscious experience" [13,14]. This motivated the psychological based emotion cause detection approach which emphasizes the "physiological arousal" rather than "empirical cue" in linguistic based approach..

The lack of emotion cause annotated corpus is a big barrier to emotion cause detection research. Especially, based on our knowledge, there is no open emotion cause annotated corpus on Chinese Weibo text. Therefore, in the first part of this study, an emotion cause annotated corpus is designed and developed. In order to reveal the relationship between emotion expression and emotion cause from the point of view of psychology, the framework which annotates the expression related to emotion cause is designed. Following this framework, the emotion cause annotated corpus corresponding to 1,333 Weibo is constructed.

Based on the observation on this corpus, three emotion cause detection methods are investigated. The first one is a rule based method. 25 rules based on the syntactic and semantic characteristics related to emotion cause expression are manually complied for emotion cause detection. Furthermore, two machine learning based methods are developed. One is based on classification method using support vector machines (SVMs) model and the other one is based on sequence labeling based method using conditional random fields (CRFs) model. They use the same feature space. The experimental results show that the CRFs based method achieves the best accuracy of 77.57% which improves from the reference baseline method for 37.45%. This improvement shows the effectiveness of our proposed emotion cause detection method based on CRFs model.

The rest of this paper is organized as follow: section 2 will introduce the annotation format of our corpus; section 3 is our proposed methods for emotion cause extraction and section 4 is the experiment; section 5 will make a conclusion of this paper.

2 Construction of Emotion Cause Annotated Corpus

In this study, the corpus adopted in NLP&CC 2013 emotion analysis share task (in short NLPCC13 dataset) is selected as the basic annotation resource. NLPCC13 dataset annotates up to two basic emotion categories to each sentence and Weibo. In this dataset, 7 basic emotion categories, namely, *fear, happiness, disgust, anger, surprise, sad* and *like* are adopted. The corpus contains the emotion cauterization annotations for 10,000 Weibos.

Firstly, we select the Weibos with explicit emotion cause as the annotation target. Secondly, according to the relationship between "physiological arousal" and "expressive behaviors" from psychological review, both the emotion expression and emotion cause are annotated. Thirdly, according to the part of speech of the emotion cause, there are two major types of cause namely noun/noun phrase and verb/verb phrase. The examples for the noun/noun emotion causes are given below:

1. 我想,大概没有什么比世博会更能使上海人[自傲]的了。
   I think there is nothing more than ***world expo*** *could make Shanghai citizens* [***feel proud***] *any more.*
2. 非常喜欢这片子的主题,人活着都是因为有梦。
   I [***like***] *the* ***topic*** *of this film very much, every live for their dream.*

In the first example sentence, the cause of the emotion (bolded and underlined) is "世博会*world expo*" which is a noun phrase. For the second example, the cause of the "like" emotion is "主题*topic*", which is a noun. Meanwhile, the emotion expressions are annotated with brackets.

The other kind of emotion cause is verb or verb phrase. Two examples sentences are given below.

3. 刚才打篮球赢了,[太激动了]。
   Just ***win a basketball match****, it is* [***so exciting***]*!*
4. 从前台搜刮了一堆零食,[哈哈]。抱回自己办公室,这有点周扒皮的赶脚
   Plunder a lot of snack*,* [***LOL***]*, get back to my office. It feels like Grandet.*

Here, the cause of emotion in this two sentences are "打篮球赢了*win a basketball match*" and "从前台搜刮了一堆零食*plunder a lot of snack*", which are verb and verb phrase, respectively.

Up to now, we annotated the emotion expression and emotion cause in 1333 Weibos. In which, 722 (54.16% of all) emotion causes are nouns/noun phrases and 611 (45.83% of all) causes are verbs/verb phrases.

The observation on the annotated emotion cause corpus show that 796 causes are in the same sentence with the emotion expression. Meanwhile, 30.10% emotion causes occur before the emotion expression and only 9.49% emotion causes occur behind the emotion expression. The detail distribution information is listed in Table. 1.

Table 1. The distance from the cause to expression of emotion

Distance of cause	Number	Percent
In the same sentence	796	59.71%
Left 1 sentence	282	21.15%
Left 2 sentence	66	4.95%
Right 1 sentence	83	6.23
Right 2 sentence	11	0.82%
Other	95	7.13%

The observation on the number of cause shows interesting results. Most Weibos (93.55% of all) have only one emotion cause while 82 Weibos have two emotion causes and 3 Weibos have four emotion causes. .

We also observe the distribution and characteristics of emotion expression in the corpus. Most emotions (1234 of 1333) are expressed by using the emotion words in Weibo while others use emotion icons.

The length of emotion words is also observed. The detail is shown in Table 2.

Based on the above observation and characteristics analysis, the methods for emotion cause detection are developed

Table 2. Distribution of the length of emotion words

Length	No.	Percent	Example
1	115	11.41%	One character word, such as "好*good*"
2	1033	78.08%	Two character word, such as "漂亮*beautiful*"
3	106	8.01%	Three character word, such as "够爷们*real man*"
4	32	2.41%	Four character word, such as "令人发指*heinous*"

3 Our Emotion Cause Detection Methods

Based on the observation on the annotated emotion cause corpus, we proposed three methods for identifying the sentence which contain the cause of emotion, namely emotion cause detection. One method is rule based and the other two are machine learning based.

The basic idea of our methods is to identify the cause candidate words (CCW) including nouns (noun phrase) and verbs (verb phrase) and determine whether they are cause of emotion. In the rule based method, we utilize the linguistic rules to decide if the CCW is a cause or not. Considering that from the point of view of psychological, the emotion expression should be helpful to emotion cause detection, we incorporate the linguistic clues and the emotion expression characteristics as features for the machine learning based method.

3.1 Rule-Based Emotion Detection

Rule-based method has shown efficient in emotion detection from news texts by Lee [10]. However, Weibo texts have characteristics different from the news texts. Thus, we construct an expanded rule set for emotion cause detection from Weibo text. Here, we firstly define the linguistic clues words for identifying emotion cause from the view of linguistics. Normally, the most linguistic cues words are verbs, conjunctions and prepositions. (Shown in Table 3)

Table 3. Linguistics clue words

Number	categories	Cue words
I	Independent conjunction	因为/because, 因/due to, 由于/because of , etc.
II	Coupled conjunction	,<(之)所以/so,(是)因为/because>,<(之)所以/so,(是)由于/due to> etc.
III	Preposition	为了/for ,以/according to,因/due to,, etc.
IV	Verb of cause	让/make,令/cause,使/let
V	Verb of feeling	想到/think of,谈起/speak of,提到/mention, , etc.
VI	Others	的是/by the fact,是/is,就是/is,的说 /said, etc.

Based on these linguistics clues, 18 rules are complied to determine whether a CCW is a cause of an emotion expression. Here, Rule 1 - Rule 14 are the same as Lee's work [10] . Rule 15- Rule 18 are new that are designed for emotion detection from Weibo texts. The detail of these rules is listed in Table 4.

Table 4. Expanded rules for emotion cause detection

No.	rules
15	verb-object(C,E), E is verb C=the sentence contains CCW and emotion word between which the dependency relation is verb-object
16	E+"的/of"+C(F) C=the focus sentence contains "de" and CCW
17	C(contains "的/of")+E C= the focus sentence contains "de" and CCW
18	C(B/F)+E E=special emotion expression in Weibo C= the sentence contains special emotion word and CCW or the sentence before the focus sentence contains CCW

Here, *C* is the cause of emotion, *CCW* is the cause candidate word, *E* is the emotion expression word, *F* is the sub-sentence contains emotion expression word, *B* is the left sub-sentence of *F*, *A* is the right sub-sentence of *F*, II_1 is the former part of coupled conjunction and II_2 is the later part of coupled conjunction. For each *CCW* in all sub-sentences, we utilize the 18 rules to detect which one is the emotion cause.

3.2 Machine Learning Based Emotion Cause Detection

The above rule based method mainly uses linguistics features. Besides those features, the observation from the psychology view point show that the relation between emotion expression and its cause are also useful. Thus, the linguistics features and psychology based features are incorporated. Firstly, the mentioned 18 rules are converted to linguistic-based Boolean features. If a sub-sentence matches the rule, the value of corresponding feature is 1, otherwise 0. It is observed that the distance between *CCW* and emotion expression is helpful to determine whether *CCW* is the emotion cause. Thus, the distance is selected as a feature. If the *CCW* is in the same sentence with the expression, the value of distance feature is 0. If the sub-sentence of *CCW* is next to the emotion expression, the value of distance is 1, otherwise 2. Furthermore, considering that to the POS of emotion expression is helpful to determine the POS of emotion cause, all of the possible combination of POS patterns are also mapped as Boolean features. The complete feature space is shown in Table 5.

In this study, the SVMs and the CRFs are employed, respectively. For the SVMs based method, the emotion cause detection problem is transferred to a binary classification problem. For each sub-sentence, the SVMs classifier is employed to classify each sub-sentence to emotion cause or not. For the CRFs based method, the cause

detection problem is transferred to a sequence labeling problem. In this method, the Weibo is transferred to a sequence of sub-sentences. The CRFs is applied to label of each sub-sentence to 0-1 label. The 0 means the corresponding sub-sentence has no emotion cause and the 1 stands for the sub-sentence contains emotion cause.

Table 5. Feature space for machine learning based emotion detection

Feature categories	Description	Value
Linguistic feature	18 rules based on linguistic cues	0 or 1
Distance feature	Distance between cause and expression	0, 1, or 2
POS feature	All possible POS pattern combination of cause and expression	0 or 1

4 Experimental Results

The annotated emotion cause corpus on 1333 Weibos is adopted in this study to evaluate the proposed emotion detection method. The evaluation metric is the accuracy.

4.1 Evaluation on the Rule Based Method

In the first experiment, the proposed rule-based method (using 18 rules) are evaluated. Its performance is compared with Lee's rule based method which is regarded as reference baseline method. The achieved performance is listed in Table 6.

Table 6. The performances of rule-based methods

Method	Accuracy
Lee's rules (baseline)	40.12%
18 rules based method	68.30%

It is observed that our method improves 28.18% accuracy from Lee. It shows the effectiveness of the new proposed rules, especially the specific ones for Weibo text.

4.2 Evaluation on the Machine Learning Based Method

For the machine learning method, the 5-fold-validation is employed in the experiment. The achieved performances are shown in Table 7.

Table 7. The performance of machine learning based methods

Method	Accuracy
SVMs based	61.98%
CRFs based	77.57%

It is observed that the machine learning based methods further improve the accuracy of emotion detection. The majority of performance improvement attributes to the new features which considering the relationship between emotion expression and emotion cause. Furthermore, the CRFs based method considers the information between adjacent sentences and thus it achieves a better performance.

4.3 Further Analysis on CRFs for Emotion Cause Detection

According to 4.2, due to the sequence information, it is shown that CRFs achieves better performance. To further analyze the performance of CRFs-based method, the test samples are divided into sub-sentences. For the sub-sentences which contain emotion cause (EC) or no emotion cause (NEC). The achieved precision, recall and F-measure of EC and NEC are listed in Table 8, respectively.

Table 8. The performance of CRFs based method (sub-sentences)

Category	#correct	#total	#proposed	precision	recall	F-measure
EC	165	282	219	58.51%	75.34%	65.87%
NER	715	769	832	92.98%	85.94%	89.32%

It is observed that for the NEC sub-sentences, the CRFs method achieves a high precision and recall performance. The achieved F-measure for NEC sub-sentence classification is 89.32%. However, for the EC sub-sentences, the CRFs based method achieves a precision at 58.51% and 65.87% F-measure. These results mean that many NEC are wrongly classified as EC. It indicates that our feature space could cover the EC samples well, but it is not good enough to classify EC samples for NEC samples.

5 Conclusion

In this paper, an emotion cause corpus on Chinese Weibo text is designed and annotated. It is the first corpus for supporting the research of emotion cause detection from micro-blogging/Weibo text, based on our knowledge. Based on the observation on this corpus, three emotion cause detection methods, namely rule-based, SVMs based and CRFs based method are developed, respectively. The evaluations for these methods show that the CRFs based method achieves the best accuracy of 77.57% which is higher than the baseline method for 37.45%. The major improvement attributes to the new linguistics features which is specific to the Weibo text and the new features which considering the relation between emotion expression and emotion cause.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 61300112, 61370165, 61203378), Natural Science Foundation of Guangdong Province (No. S2012040007390, S2013010014475), MOE Specialized Research Fund for the Doctoral Program of Higher Education 20122302120070, Open Projects Program of National Laboratory of Pattern Recognition, Shenzhen

International Co-operation Research Funding GJHZ20120613110641217 and Baidu Collaborate Research Funding.

References

1. Mohammad, S.: From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 105–114 (2011)
2. Purver, M., Battersby, S.: Experimenting with Distant Supervision for Emotion Classification. In: Proceedings of the 13th Conference of the EACL, pp. 482–491 (2012)
3. Panasenko, N., Trnka, A., Petranova, D., Magal, S.: Bilingual analysis of LOVE and HATRED emotional markers. In: Proceedings of the 3rd SAAIP workshop, IJCNLP 2013, Japan, pp. 15–23 (2013)
4. Vaassen, F., Daelemans, W.: Automatic Emotion Classification for Interpersonal Communication. In: Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (2011)
5. Tang, D., Qin, B., Liu, T., Li, Z.: Learning sentence representation for emotion classification on microblogs. In: Zhou, G., Li, J., Zhao, D., Feng, Y. (eds.) NLPCC 2013. CCIS, vol. 400, pp. 212–223. Springer, Heidelberg (2013)
6. Xu, R., Ye, L., Xu, J.: Reader's Emotion Prediction Based on Weighted Latent Dirichlet Allocation and Multi-Label k-Nearest Neighbor Model. Journal of Computational Information Systems 9(6) (2013)
7. Yao, Y., Xu, R., Lu, Q., Liu, B., Xu, J., Zou, C., Yuan, L., Wang, S., Yao, L., He, Z.: Reader emotion prediction using concept and concept sequence features in news headlines. In: Gelbukh, A. (ed.) CICLing 2014, Part II. LNCS, vol. 8404, pp. 73–84. Springer, Heidelberg (2014)
8. Xu, R., Zou, C., Xu, J.: Reader's Emotion Prediction Based on Partitioned Latent Dirichlet Allocation Model. In: Proceedings of of International Conference on Internet Computing and Big Data (2013)
9. Ye, L., Xu, R., Xu, J.: Emotion Prediction of News Articles from Reader's Perspective based on Multi-label Classification. In: Proceedings of IEEE International Workshop on Web Information Processing, pp. 2019–2024 (2012)
10. Lee, S.Y.M., Chen, Y., Li, S., Huang, C.-R.: Emotion Cause Events: Corpus Construction and Analysis. In: Proceedings of International Conference on Language Resources and Evaluation (2010)
11. Chen, Y., Lee, S., Li, S., et al.: Emotion Cause Detection with Linguistic Constructions. In: Proceeding of International Conference on Computational Linguistics (2010)
12. Lee, S., Chen, Y., Huang, C., et al.: Detecting Emotion Causes with a Linguistic Rule-based Approach. In: Computational Intelligence (2012)
13. Das, D., Bandyopadhyay, S.: Analyzing Emotional Statements – Roles of General and Physiological Variables. In: Proceedings of IJCNLP (2011)
14. Myers, G.D.: Theories of Emotion. Psychology: Seventh Edition, p. 500. Worth Publishers, New York