WUST CRF-Based System at NTCIR-13 ECA Task

Maofu Liu, Linxu Xia, Zhenlian Zhang, Yang Fu
School of Computer Science and Technology, Wuhan University of Science and Technology,
Wuhan 430065, China
liumaofu@wust.edu.cn, 990234477@qq.com

ABSTRACT

This paper describes our work on Emotion Cause Analysis (ECA) task in NTCIR-13. This task aims to detect the emotion cause description when an emotion happened. In this paper, we apply the Conditional Random Field (CRF) classification model to identifying the emotion cause description with a series of features, such as POS features, basic word features, distance features, and contextual features. The system includes three parts, i.e. data preprocessing, feature extraction, and CRF classifier. Experimental results demonstrate that the CRF model is superior to other classification models in the emotion cause description detection.

Categories and Subject Descriptors

I.2.7 [Artificial Intelligence]: Natural Language Processing - text analysis.

I.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing - linguistic processing.

General Terms

Experimentation

Team Name

WUST

Subtasks

Emotion Cause Detection (Chinese)

Keywords

Emotion Cause Analysis; CRF Classifier; Feature Extraction; Natural Language Processing

1. Introduction

In NTCIR-13, ECA is an evaluation task aiming to detect the emotion cause description from the text, which is a common problem shared widely among researchers of Natural Language Processing (NLP) and information access. In this task, there are two subtasks, i.e. the emotion cause detection at clause level and the emotion cause extraction. Our system mainly focuses on the first subtask.

The emotion cause detection at clause level means to identify the clause which contains emotion cause. It can be treated as a clause-level binary text classification problem. The clauses will be classified according to containing emotion cause or not. A Chinese **Example 1** is given below, and our goal is to identify the reason sentence or clause for the emotional keywords.

Example 1:

In Chinese: 在等待救护车到来的过程中，两位民警轮换着为中年男子撑伞。其中一位民警发现一把伞无法遮雨，便脱下身上雨衣盖在中年男子身上，自己被淋得直打哆嗦。两位民警的行为感动了路人。看到这一幕，感觉很温暖。

In English: While waiting for the ambulance, the two policemen held an umbrella for the middle-aged man in turn. When one of the policemen found that an umbrella could not cover the rain, he took off his raincoat and covered the middle-aged man with it, so that he was shivering. The passers-by was moved by the behavior of the two policemen, and it is very warm to see this scene.

The annotation results of **Example 1** are as follows.

Emotional words: “感动(moved)”

Emotion category: “高兴(happiness)”

Direct cause: “其中一位民警发现一把伞无法遮雨，便脱下身上雨衣盖在中年男子身上，自己被淋得直打哆嗦(When one of the policemen found that an umbrella could not cover the rain, he took off his raincoat and covered the middle-aged man with it, so that he was shivering.)”

In this subtask, the emotion expressions in the text and their categories are firstly detected. And then, the direct cause stimulating such an emotion is identified from its context and also annotated correspondingly. In some cases, the direct cause is elaborated in more details in another sentence. For such cases, both the direct cause and the elaborated cause are annotated, which are illustrated in the following **Example 2**.

Example 2:

In Chinese: 孙女士说她丈夫喜欢偷偷翻看自己的手机，当看到手机里有不认识的人时，就会盘问她半天。苏女士对此很无奈，觉得丈夫一点都不信任自己。她逐渐对这段婚姻失望，觉得和这人生活一辈子特别没有意思。

In English: Ms. Sun said that her husband liked to peek at the messages in her cellphone, when there was a stranger, he would question her for a long time. So Ms. Su felt helpless for her husband untrusting her. Finally, she was gradually disappointed their marriage, and felt that it was especially uninteresting to live with this man.

The annotation results of **Example 2** are as follows.

Emotional words: “无奈(helpless)”

Emotion category: “悲伤(sadness)”

Direct cause: “丈夫一点都不信任自己(her husband untrusting her)”

Elaborated cause: “丈夫喜欢偷偷翻看自己的手机(her husband liked to peek at the messages in her cellphone)”

The emotion analysis has been carried out for many years, but most of the researches have focused on emotion recognition and classification[1], and this kind of work is designed to automatically classify the emotional keywords contained in the text[2], for example, Ekman`s [3] six basic emotion categories. On the contrary, the study of emotion cause analysis is still scarce. However, in many cases, we care more about the stimuli or the causes of the emotion, because it is better for us to understanding text information. We can see that it is urgent to mine the deep information of texts with emotion cause analysis.

In this paper, we recognize the emotion cause by CRF classifier. This method regards emotion cause analysis as a sequence tagging problem. And the advantage of this method is that not only combine basic word features, POS features, distance features and other characteristics, but also consider the contextual features. The experimental results show that the proposed sequence annotation model has better recognition performance than other classification models.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 discusses the overview of our system by the system architecture. Section 4 then presents experiments and discussions. Finally, Section 5 concludes this paper.

2. Related work

The text emotion analysis has become a research hotspot in the field of NLP, and it aims to study how to automatically analyze the emotion and its related information expressed in the text. At present, the study of emotion analysis mainly focuses on the emotion recognition and emotion cause analysis.

The emotion recognition focuses on two tasks, i.e. emotion classification and emotion information extraction. The research of emotion classification aims to classify the emotion types expressed in the text, for example, Yang et al [4] proposed a machine learning to determine the emotion category of web blog and they adopted Ekman's [3] emotion classification, namely happiness, sadness, fear, disgust, anger, and surprise. The emotion information extraction is to extract the information related to the emotional expression. Das and Bandyopadhyay [5] used an unsupervised method to extract emotion holders and topics from Bengali blog.

In order to have a deeper understanding of the text which contains the emotion and its cause description, it is necessary to carry out deep excavation of the emotional information, such as the emotion cause analysis in this paper. At present, due to the study of emotion cause analysis being in its infancy, few works on emotion cause analysis are reported. Meanwhile, the majority of these works are based on rules and statistical model, and also have achieved some effect.

Based on the rules of emotion cause analysis, Chen and Lee [2] proposed a rule based method to detect emotion causes. This idea is to make linguistic rules for cause extraction. Gao et al [6,7] extended the rule method to informal text in Weibo text. However, besides rule based methods, Gui et al [8,9,10] presented machine learning based cause detection methods, such as classification based on support vector machines and sequence labeling based conditional random model.

According to the above findings, the method based on statistical model can achieve high accuracy on emotional cause analysis, and it can also be integrated into the multi-dimensional features for emotional cause detection. In this paper, we summarize all of the above methods, and choose the CRF model to solve the problem of the ECA task.

3. System Description

Based on the analysis of the corpus and the problem of emotion cause recognition, the subtasks can be regarded as a classification problem, which mainly judges whether the context clause of a core word contains emotional reasons. Our system includes three main modules, i.e. data preprocessing, feature extraction and CRF classifier. Figure 1 illustrates our system architecture in detail.

3.1 Data preprocessing

In the data preprocessing, the system mainly implements the Chinese word segmentation, which removes the stop words according to the stop word list and part of speech (POS) tagging. We choose Jieba Chinese word segmenter [1] to segment the Chinese word. The following Table 1 is the preprocessing result of the original training data given by NTCIR-13 after POS tagging and setting label of the preceding **Example 1**.

In Table 1, we add label to the sentences around the emotion keywords, the clause position are Left2, Left1, Keywords, and Right1, whether it belongs to the clause of reason or not. If the cause description is in this sentence, we mark the clause "1", otherwise mark "0".

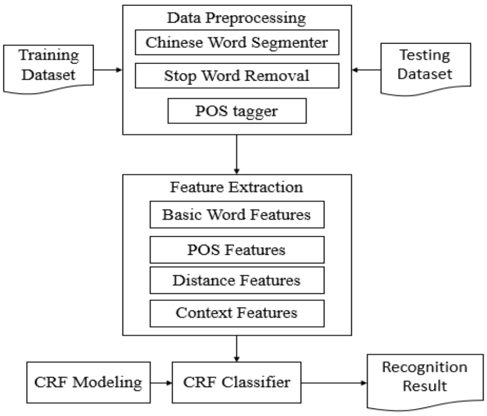

Figure 1. System Architecture

Table 1. POS tagging and setting label of Example 1

Clause Position	Clause	Label
Left2	在等待救护车到来的过程中，两位民警轮换着为中年男子撑伞。	0
Left1	其中一位民警发现一把伞无法遮雨，便脱下身上雨衣盖在中年男子身上，自己被淋得直打哆嗦。	1
Keywords	两位民警的行为感动了路人。	0
Right1	看到这一幕，感觉很温暖。	0

3.2 Feature extraction

In this subsection, we divide feature sets into basic features and contextual features, which are described in Table 2.

Table 2. Basic features

Features	Description
nouns	The nouns in the present sentence, if not, should be filled with "NULL".
verbs	The verbs in the present sentence, if not, should be filled with "NULL".
numbers of nouns	The number of nouns containing in the present sentence.
numbers of verbs	The number of verbs containing in the present sentence.
distance	The distance between present sentence and emotional keyword, and the values represented by Left2, Left1, Keyword, Right1 and Right2 are -2, -1, 0, 1 and 2 respectively.

The basic features of the sentence include nouns, verbs and their numbers. Only nouns and verbs are considered as the emotion cause, which is mainly composed of noun phrases and verbal phrases. Generally speaking, the emotion cause can be found around emotional words, so the distance feature is an important location feature. The contextual features mainly contain the syntactic ones of the previous sentence and the next sentence, and they can be represented by feature templates. In addition, the categories of emotional words can also provide information for the classification model, so we also take the classification features of emotional words into consideration.

[1] http://pypi.python.org/pypi/jieba/

3.3 CRF classifier

The CRF model is a typical discriminant model proposed by Lafferty [11] based on hidden Markov models and maximum entropy model [12], and has been widely used in the field of NLP [14,15,16].

The sequence labeling technology based on CRF theory provides a model with strong discriminating ability in classification, which models feature information and label information of training sets with the purpose of finding a suitable parameter set, so that the maximum conditional probability can be obtained [17]. The conditional distribution of the labels *Y* given the observations *W* has been shown in Formula (1),

$$\mathbf{P}_{\lambda}(\mathbf{Y}|\mathbf{W}) = \frac{1}{Z(W)}\mathbf{exp}\left(\sum_{t\in T}\sum_{k}\lambda_k f_k(y_{t-1},\mathbf{W},t)\right) \quad \textbf{(1)}$$

where $f_k(y_{t-1},\mathbf{W},t)$ is a feature function of vertex and edge, λ_k , λ, $\mathbf{W}$, and $\mathbf{Z}\,(\mathbf{W})$ are the learned weight associated with f_k, the weight set, the sequence of words to be predicted, and the normalization factor separately, and $\mathbf{Z}\,(\mathbf{W})$ is also known as the partition function. $\mathbf{Y} = \{\mathbf{y}_t\}$ is the sequence labeling of output, and $y_t \in \{1,0\}$ represents whether the corresponding sentence is the emotional cause sentence or not.

In fact, the CRF is a supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The following Figure 2 describes the schematic diagram of CRF classification model in our system.

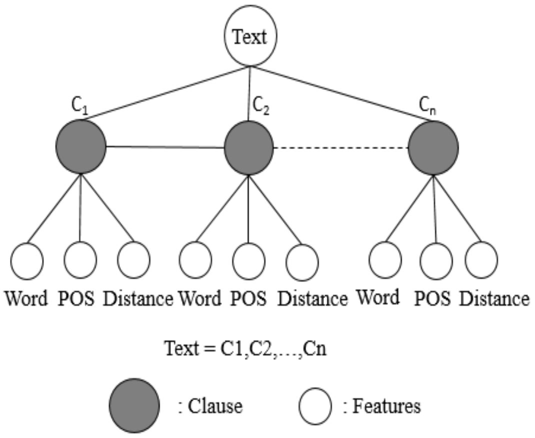

Figure 2. Schematic diagram of CRF classification model

3.4 Template feature selecting

In the experiments, the CRF model is implemented with the CRF++ tools[2]. There are total of 43 features used for the experiments, including both basic features and contextual features, as shown in Table 3.

Table 3. Features used in model training

number	features
1	U00:%x[-2,0]
2	U01:%x[-1,0]
3	U02:%x[0,0]
4	U03:%x[1,0]
5	U04:%x[2,0]
6	U05:%x[-1,0]/%x[0,0]
7	U06:%x[0,0]/%x[1,0]
…	…
41	U60:%x[-2,5]/%x[-1,5]/%x[0,5]
42	U61:%x[-1,5]/%x[0,5]/%x[1,5]
43	U62:%x[0,5]/%x[1,5]/%x[2,5]

4. Experiments

4.1 Experiment settings

In this paper, we use the data sets provided by NTCIR-13. This dataset contains about 2,200 training instances and 10,000 instances for testing. We segment the texts in this data set by the Jieba word segmentation tool in the data preprocessing.

There are three values to evaluate the experimental results, namely precision (*P*), recall (*R*) and *f*-score (*F*) [18].

$$Precision_{clause} = \frac{\#correct\ cause\ relevant\ clauses}{\#detected\ cause\ relevant\ clause}$$

$$Recall_{clause} = \frac{\#correct\ cause\ relevant\ clauses}{\#annotated\ cause\ relevant\ clause}$$

$$F - measure_{clause} = \frac{2 \times Precision_{clause} \times Recall_{clause}}{Precision_{clause} \times Recall_{clause}}$$

4.2 Experimental results

We submitted one system result to NTCIR-13 of emotion cause detection at clause level. The official evaluation results of performance are listed in the Table 4 [19].

Table 4. Performances of the CRF model

Team	P	R	F
WUST	0.6930	0.6399	0.6654

From the Table 4, we can see that the CRF model has achieved good performance. However, for some special instances, the performance is still not satisfactory, and the main reasons are listed as follows.

(1) When the clauses, containing the emotional reason, are too far from the emotional words, just as shown in **Example 3**, the model tends to identify the sentences which are more close to the emotional words, rather than the emotional ones. In **Example 3**, the CRF model wrongly identifies the last sentence as the emotion cause sentence, i.e. "北京一位不愿透露姓名的捐款者向北大医院送上了一份敬意。(An anonymous donor in Beijing sent a tribute to the Peking University Hospital.)".

Example 3:

In Chinese: <cause id = 0>"把方便留给别人，把困难留给自己，把安全留给别人；把危险留给自己，把幸福留给别人，把痛苦留给自己。"</cause>北京一位基层党组织负责人对参加一线战斗的党员干部提出三点要求。"那么多医务人员倒下了，5 万元并不多，哪怕是改善一下他们的生活"。北京一位不愿透露姓名的捐款者向北大医院送上了一份<keyword>敬意</keyword>。

In English: <cause id = 0>"Leave the convenience to others, leave the difficulties to yourself, leave the safety to others, leave the danger to yourself, leave the happiness to others, and leave the pain to yourself."</cause> The head of a grass-roots party organization in Beijing made three demands for party members and cadres involved in the first line of fighting. "So many medical workers have fallen ill, it deserves more than 50 thousands yuan, even if it is to improve their lives." An anonymous donor in Beijing sent <keyword>a tribute</keyword>to the Peking University Hospital.

(2) When the text contains multiple cause clauses, as shown in **Example 4**, the CRF model can only identify one of them. In the following Example 4, the CRF model only identifies the second emotion cause sentence.

[2] https://sourceforge.net/projects/crfpp/

Example 4:

In Chinese: 谢母说，<cause id = 0>小荣不到 4 岁的时候，就到地里弄草皮，然后做成肥料，赚点小钱</cause>。<cause id =1>小学没毕业，就在建筑工地提灰桶打小工，每天挣 1 元 1 角钱</cause>，“想起来就<keyword>寒心</keyword>”。

In English: The mother said that <cause id = 0>when the child was under 4 years old, she went to the fields to collect the sod, and made the fertilizer for some change</cause>. <cause id =1>When he was still in primary school, he went to work on a construction site, earned 1 yuan and 1 jiao a day</cause>, "It <keyword>hurts</keyword> to think about these things".

5. Conclusions

In this paper, we construct the classification model based on conditional random field to recognize emotion cause in Chinese text pair using the hybrid features, including words, POS, distance and context. The experimental results show that these features are of some help for the identification of the emotion causes. However, the effect of the present method is very limited, and the accuracy is only about 70%. We will continue to improve our algorithms for error analysis. Moreover, we mostly consider statistical features in our system, if we add some rule and semantic features, the accuracy may be significantly improved.

ACKNOWLEDGMENTS

The work presented in this paper is partially supported by the Major Projects of National Social Foundation of China under Grant No. 11&ZD189.

REFERENCES

[1] Alm C O, Roth D, Sproat R. Emotions from text: machine learning for text-based emotion prediction. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, Pages:347-354.

[2] Chen Y, Lee S, Li S, et al. Emotion Cause Detection with Linguistic Constructions. Proceedings of International Conference on Computational Linguistics, 2010, Pages:179-187.

[3] Ekman P. Expression and the nature of emotion. Approaches to Emotion, 1984, 3:19-344.

[4] Yang C, Lin H, Chen H. Emotion classification using web blog corpora. Proceedings of International Conference on Web Intelligence, 2007, Pages:275-278.

[5] Das D, Bandyopadhyay S. Identifying emotion holder and Topic from Bengali emotional sentences. Proceedings of the IEE - Part A: Power Engineering, 2010, 108(38):168-172.

[6] Gao K, Xu H, Wang J. A rule-based approach to emotion cause detection for Chinese micro-blogs. Expert Systems with Applications, 2015, 42(9):4517-4528.

[7] Gao K, Xu H, Wang J. Emotion cause detection for Chinese micro-blogs based on ECOCC model. Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2015, Pages:3-14.

[8] Gui L, Yuan L, Xu R, et al. Emotion cause detection with linguistic construction in Chinese weibo text. Proceedings of Natural Language Processing and Chinese Computing, 2014, Pages:457-464.

[9] Gui L, Wu D, Xu R, et al. Event-driven emotion cause extraction with corpus construction. Proceedings of International Conference on Empirical Methods on Natural Language Processing, 2016, Pages:1639-1649.

[10] Gui L, Hu J, He Y, et al. A question answering approach to emotion cause extraction. Proceedings of International Conference on Empirical Methods on Natural Language Processing, 2017, Pages:1594–1603 .

[11] Lafferty J D, Mccallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning, 2001, Pages:282-289.

[12] Eddy S R. Hidden Markov models. Current Opinion in Structural Biology, 1996, 6(3):361.

[13] Hayes B, Wilson C. A Maximum Entropy Model of Phonotactics and Phonotactic Learning. Linguistic Inquiry, 2014, 39(3):379-440.

[14] Tian D. Exploiting PLSA model and conditional random field for refining image annotation. High Technology Letters, 2015, 21(1):78-84.

[15] Dong Y, Chu Q, Ling P. Personalized Facet Recommendation based on Conditional Random Fields. Proceedings of the 3rd International Conference on Machinery, Materials and Information Technology Applications, 2015, doi:10.2991/icmmita-15.2015.41.

[16] Ding Y, Li Q, Dong Y, et al. 2D correlative-chain conditional random fields for semantic annotation of web objects. Journal of Computer Science & Technology, 2010, 25(4):761-770.

[17] Li J, Zhang B. Facial expression recognition based on Gabor and conditional random fields. Proceedings of the IEEE 13th International Conference on Signal Processing, 2016, doi:10.1109/ICSP.2016.7877933.

[18] Han L, Luo S, Chen Q, et al. Fast Chinese syntactic parsing method based on conditional random fields. Journal of Beijing Institute of Technology, 2015, 24(4):519-525.

[19] Gao Q, Hu J, Xu R, et al. Overview of NTCIR-13 ECA task. Proceedings of the 13th NTCIR Conference, 2017.