Decision Tree Method for the NTCIR-13 ECA Task

Xiangju Li
School of Computer Science
and Engineering,
Northeastern University,
Shenyang, China
lixiangju100@163.com

Shi Feng
School of Computer Science
and Engineering,
Northeastern University,
Shenyang, China
fengshi@cse.neu.edu.cn

Daling Wang
School of Computer Science
and Engineering,
Northeastern University,
Shenyang, China
wangdaling@cse.neu.edu.cn

Yifei Zhang
School of Computer Science
and Engineering,
Northeastern University,
Shenyang, China
zhangyifei@cse.neu.edu.cn

ABSTRACT

This paper details our participation in the Emotion Cause Analysis (ECA), which is a subtask of the NTCIR-13. ECA aims to identify the reasons behind a certain emotion expressed in text. It is a much more difficult task compared with traditional emotion analysis. We consider the task as a slight variation of supervised machine learning classification problems. Inspired by rule-based systems for emotion cause detection, the key features are obtained which can serve for training models. Furthermore, this paper adopts the C4.5 method that has been widely used in data mining and machine learning for comprehensible knowledge representation. The effectiveness of our method is evaluated using the official dataset and we have achieved about 0.5445 for F-measure.

Team Name

neuL

Subtasks

Emotion Cause Analysis (Chinese)

Keywords

Emotion Cause Analysis, Rule-based systems, C4.5

1. INTRODUCTION

The team neuL has participated in Emotion Cause Analysis (ECA) task of NTCIR-13. This report describes our approach for solving the ECA problem and discusses official results. Emotion analysis is one of the most important research tasks in natural language processing and public opinion mining [18, 20]. Recently, emotion cause analysis, a new challenging task for emotion analysis, has become a hot research topic for both academic and industrial communities [2, 8, 11]. For detailed introductory information of ECA task, please refer to the overview paper [6].

Example 1. 朱某今年55岁，1979年参加工作时才19岁，已有36年的手艺。**说起自己的荣誉**，朱某很是自豪。
Mr. Zhu is 55 years old. He started working in 1979 as a barber when he was 19, and has 36 years of experience. ***Talking about his honors****, Mr. Zhu is so proud.*

In this paper, we adopt Ekman's emotion classification [4, 19], which identifies six primary emotions, namely happiness, sadness, fear, anger, disgust and surprise, known as the "Big6" scheme in the W3C Emotion Markup Language. As can be seen from Example 1, "proud" is an emotion word, and the type of this emotion word is happiness. That is the emotion category of the clauses in Example 1 is happiness. Meanwhile, the fifth clause which contains the emotion word is called the core clause in Example 1. The purpose of the emotion cause extraction task is to identify the cause behind of an emotion expression. For example, the cause of "proud" is "Talking about his honors" in Example 1.

Emotion cause extraction is a much more difficult compared with traditional emotion classification problem [5, 7]. On the one hand, the size of corpus for emotion cause extraction is usually very small because of the complexity in annotation. On the other hand, emotion cause extraction requires a deeper understanding of document than emotion analysis since it need to identify the relation between the description of an event which causes an emotion and the expression of that emotion [7].

The decision tree representation is a natural way of presenting a decision-making process among numerous approaches since decision trees are simple and transparent for people to understand [16, 17]. They have a wide range of applications such as business, manufacturing, computational biology, etc [3]. These methods aim at training classifiers to maximize the accuracy in many applications. Meanwhile, researchers have been design many new methods based on these decision tree learning strategies in many studies.

The emotion cause extraction task attempts to detect the clause which contains emotion causes [9]. In previous studies, emotion cause extraction can be treated as a binary text classification problem, where the clauses are classified as containing emotion cause or not by a classifier. That is, the instance in training and testing datasets is a clause with label exclusive "Yes" or "No". Following previous studies, in this paper we leverage the decision tree based learning method to solve this task. A dataset which contains 2105 documents is employed to study the effectiveness of our proposed method.

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 introduces the proposed method of this paper. Section 4 presents the results on the official dataset. Finally, Section 5 concludes this paper.

2. RELATED WORK

There are various approaches that focus on emotion recognition or classification given a known emotion context [1, 14]. Mohammad et al. built their system in SemEval-2013 with a number of features like POS tags, hashtags, characters in upper case, punctuations and so on [15]. A hierarchical LSTM model with two levels of LSTM networks is proposed by Huang et al. to model the retweeting process and capture the long-range dependency [10]. McDonald et al. treated the sentiment labels on sentences as the sequence tagging problem, and utilized CRFs model to score each sentence in the product reviews [13]. Lu et al. proposed a method for combining information from different sources to learn context-aware sentiment lexicon [12].

To the best of our knowledge, little research has been done with respect to emotion cause detection. Identifying emotion cause in text is an emerging hot research topic in NLP and its applications. Lee et al. first defined the task of emotion cause extraction and presented a rule based method to detect emotion causes [11]. The basic idea is to make linguistic rules for cause extraction. Chen et al. proposed a multi-label approach to detect emotion causes [2]. The multi-label model not only detects multi-clause causes, but also captures the long-distance information to facilitate emotion cause detection. An emotion cause annotated corpus was firstly designed and developed through annotating the emotion cause expressions in Chinese Weibo Text in [9]. Recently, emotion cause extraction is considered as a question answering (QA) task by Gui et al.. An attention mechanism is further proposed to store relevant context in different memory slots to model context information [7].

3. METHOD

Our method consists of two separate modules: (a) identifying key features which can serve valuable information to classify the clause in our dataset, and (b) obtaining an effective classifier for emotion cause analysis. The framework of our method for NTCIR-13 Emotion Cause Analysis task is shown in Figure 1. In this section, we present the feature extraction procedure for searching emotion cause, and explain the learning procedure based on C4.5 decision tree method for Emotion Cause Analysis.

3.1 Feature Extraction

Rule-based features. Inspired by the rule based method proposed in [11], we manually define a knowledge base that containing seven groups of linguistic cues. In our method, let a_i be the feature which represents whether the clause is in accord with the i-th rule group. That is,

$$a_i = \begin{cases} Y, & \textit{containing any cue word in i-th group}; \\ N, & \textit{otherwise}. \end{cases}$$

The following example will give a detailed description.

Example 2. 劝说过程中，消防官兵了解到，该女子是由于对方拖欠工程款，家中又急需用钱，无奈才选择跳楼轻生。

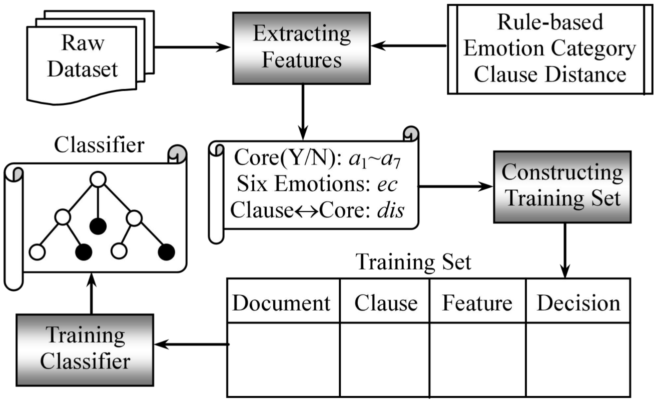

Figure 1: The framework of our method for Emotion Cause Analysis.

During persuasion, firemen realized that the woman attempted suicide ***because of*** *the hold back of wages by the employer, and her family asked for money urgently, she feels helpless and thus.*

There are five clauses in this instance, and the second clause contains the cue word (**"由于"**) which belongs to the first rule group. Therefore, a_2 = "Y" with a_i = "N"($i = 1, 3, \cdots, 7$) for the second clause. And a_i = "N"($i = 1, \cdots, 7$) for the other clauses.

Emotion category feature. The emotion category (ec) is adopted as an important feature for clause classification. In this study, we adopt Ekman's emotion classification, which identifies six primary emotions, namely happiness, sadness, fear, anger, disgust and surprise [4]. As can be seen in Example 2, the emotion word is "helpless" and the type of it is "sadness". That means the emotion category of the clauses in Example 2 is "sadness". Therefore, the values of feature ec are entirely "sadness" in this instance.

Clause distance feature. As we known that the distance between the candidate clause and the clause that expressing emotions (dubbed as core clause) is a very important feature [8]. In this paper, dis denotes this clause distance feature. For example, the core clause is the fifth clause in Example 2. The value of feature dis of the first clause is -4 in this instance.

3.2 Decision System

In data mining and machine learning, the decision system is an important concept and defined as follows.

A decision system is the 5-tuple [22]: $DS = \langle U, C, D, V, I \rangle$, where U is a non-empty finite set of objects called the universe, C is a non-empty finite set of condition features, $D = \{d\}$ is a non-empty finite set of decision features, V: $V = \{V_a\}$ is a set of values for each feature $a \in C \cup D$, I: $I = \{I_a\}$ is an information function for each feature $a \in C \cup D$ ($i.e. I_a : U \rightarrow V_a$).

Table 1 depicts an example of training dataset. doc_1, doc_2 represent the document. x_{ij} denotes the i-th document and the j-th clause. For instance, x_{21} is the first clause in the second document. a_1, $\cdots$, a_7, dis and ec are the features of clauses. d is the decision feature. That is, according to above definition, in this decision system $U = \{x_{11}, x_{12}, x_{13}, \cdots, x_{25}, x_{26}\}$, $C = \{a_1, a_2, \cdots, a_7, dis, ec\}$ and

Table 1: An Example of Training Dataset.

documents	clause	a_1	$\cdots$	a_7	*dis*	*ec*	*d*
doc$_1$	x_{11}	*Y*	$\cdots$	*Y*	-1	*happiness*	*N*
	x_{12}	*N*	$\cdots$	*N*	0	*happiness*	*Y*
	x_{13}	*Y*	$\cdots$	*N*	1	*happiness*	*N*
	x_{14}	*Y*	$\cdots$	*N*	2	*happiness*	*N*
	x_{15}	*N*	$\cdots$	*Y*	3	*happiness*	*N*
doc$_2$	x_{21}	*N*	$\cdots$	*Y*	-2	*disgust*	*N*
	x_{22}	*N*	$\cdots$	*N*	-1	*disgust*	*N*
	x_{23}	*N*	$\cdots$	*Y*	0	*disgust*	*N*
	x_{24}	*Y*	$\cdots$	*N*	1	*disgust*	*Y*
	x_{25}	*N*	$\cdots$	*N*	2	*disgust*	*N*
	x_{26}	*N*	$\cdots$	*N*	3	*disgust*	*N*

$D = \{d\}$. We considered the ECA task as a decision system.

As can be seen in the second row and the third column of Table 1, the value of feature a_1 is "*Y*", which represents that the clause x_{11} contains the cue word in first group. Similarly, the symbol "*N*" in the third row and the third column of Table 1 denotes that there are no first group cue word in the clause x_{12}. The value of feature *dis* is "-2" in Table 1, which means that the clause x_{21} is the second clause in the previous of the core clause. The "*disgust*" in the seventh column means that the emotion category in *doc*$_1$ is disgust. As can be seen in the last column of Table 1, the decision feature *d* has two values: "*Y*" and "*N*", and we can infer that the clauses x_{12} and x_{24} contain the emotion cause.

3.3 C4.5 Decision Tree

C4.5 is a suite of decision tree methods in machine learning and data mining [21]. It learns a mapping from feature values to classes that can be applied to classify new instances. Feature selection is a fundamental process in decision tree induction. The heuristic function in the C4.5 method is

$$GainRatio(a) = \frac{Gain(a)}{Split\_infor(a)}, \tag{1}$$

where

- a is the feature of the given decision system,
- $Gain(a)$ is the information gain of the feature a,
- $Split\_infor(a)$ is the split information entropy of the feature a.

3.4 Method Framework

In this section, we provide a detailed description of our method which is listed in Algorithm 1. It contains two main steps. In the following, we detail each of the steps of the method.

Step 1 contains Lines 1 through 13. We pre-process the raw data and extract features from the data for construction of a decision system.

Step 2 corresponds to Line 14. In this step, we will train the decision system obtained in **Step 1**. We omit the details about the decision tree construction since there are many illustrations in previous works.

4. EXPERIMENT

We will give experiment settings and analyze the results in this section.

Algorithm 1 A C4.5 based Emotion Cause Analysis Method.

Input:
Training data set: S
Seven groups of linguistic cues: Cue
Method: ECA-C4.5
Output: A classifier

```
1: for (document ($doc_i$) in $S$) do
2:   for ($clause_j$ in $doc_i$) do
3:     Get the distance between $clause_j$ and an emotion words: $DS \leftarrow dis$
4:     Get the emotion category of the $clause_j$: $DS \leftarrow ec$
5:     for ($group_k$ cue words in $Cue$) do
6:       if ($clause_j\ contains\ the\ cue\ words\ of\ group_k$) then
7:         feature $DS \leftarrow a_k = Y$;
8:       else
9:         feature $DS \leftarrow a_k = N$;
10:      end if
11:    end for
12:  end for
13: end for
14: Train $DS$ for getting the classification by C4.5 method
15: return classifier
```

- **Dataset**

As of now, there are a few open datasets available for emotion cause extraction. In our work, we employ the dataset provided by NTCIR-13 Emotion Cause Analysis (ECA) subtask. There are 2105 SINA news documents in the dataset for developing effective models. The details of the datasets are described in Table 2.

Table 2: Dataset Used for Developing Model.

Item	Numbers
Documents	2105
Clause	11799
Emotion Causes	2167

- **Evaluation Metrics & Result**

To evaluate the method, the task involves adopting three metrics and they are based on the standard text classification metrics:

$$P = \frac{correct_{num}}{detected_{num}} \tag{2}$$

$$R = \frac{correct_{num}}{annotated_{num}} \tag{3}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{4}$$

where $correct_{num}$ is the number of correct cause relevant clauses, $detected_{num}$ is the number of detected cause relevant clauses, $annotated_{num}$ is the number of relevant clauses whose real class is the cause clause.

A formal run dataset of 2000 samples is provided by the ECA task for examining the method. The experimental results on this dataset are summarized in Table 3. The *aver_value* means the average value of the submitted results. For simply, we list the mean value and our results on the formal run dataset. From this table, we can obtain the following observations. Our method have relatively good performance on detecting the causes. As can be seen from Table 3, the value of R obtained by our method is nearly 0.7. However, the precision of our method still needs to be improved since its value is only 0.4463. The value of F is 0.5445. In the future, we may pay more attention to feature extraction to improve the value of precision.

Table 3: Performances of the Running Results.

Metric	P	R	F
aver_value	0.6026	0.6600	0.6220
Our result	0.4463	0.6984	0.5445

5. CONCLUSIONS

We participated in the NTCIR-13 Emotion Cause Analysis (ECA) task. In this paper, a decision tree method based on C4.5 is proposed for this task. The core parts of our method are the features extraction inspired by the rule based method for emotion cause analysis and the decision tree method which serves for obtaining a classifier. We conducted an experiment with the provided dataset and confirmed that our method have a relatively good recall, but precision of our method should be further improved.

Acknowledgments.

The work was supported by National Natural Science Foundation of China (61772122, 61370074, 61402091).

6. REFERENCES

[1] A. Abbasi, H. Chen, S. Thoms, and T. Fu. Affect analysis of web forums and blogs using correlation ensembles. *IEEE Transactions on Knowledge and Data Engineering*, 20(9):1168–1180, 2008.

[2] Y. Chen, S. Y. M. Lee, S. Li, and C.-R. Huang. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 179–187. Association for Computational Linguistics, 2010.

[3] G. Creamer and Y. Freund. Using boosting for financial analysis and performance prediction: application to s&p 500 companies, latin american adrs and banks. *Computational Economics*, 36(2):133–151, 2010.

[4] P. Ekman. Expression and the nature of emotion. *Approaches to emotion*, 3:19–344, 1984.

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

[6] Q. Gao, J. Hu, R. Xu, and et al. Overview of ntcir-13 eca task. 2017.

[7] L. Gui, J. Hu, Y. He, R. Xu, Q. Lu, and J. Du. A question answering approach to emotion cause extraction. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.*

[8] L. Gui, D. Wu, R. Xu, Q. Lu, and Y. Zhou. Event-driven emotion cause extraction with corpus construction. In *EMNLP*, pages 1639–1649, 2016.

[9] L. Gui, L. Yuan, R. Xu, B. Liu, Q. Lu, and Y. Zhou. Emotion cause detection with linguistic construction in chinese weibo text. In *Natural Language Processing and Chinese Computing*, pages 457–464. Springer, 2014.

[10] M. Huang, Y. Cao, and C. Dong. Modeling rich contexts for sentiment classification with lstm. *arXiv preprint arXiv:1605.01478*, 2016.

[11] S. Y. M. Lee, Y. Chen, and C.-R. Huang. 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, 2010.

[12] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, pages 347–356. ACM, 2011.

[13] R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar. Structured models for fine-to-coarse sentiment analysis. In *Annual meeting-association for computational linguistics*, volume 45, page 432, 2007.

[14] R. Mihalcea and H. Liu. A corpus-based approach to finding happiness. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 139–144, 2006.

[15] S. M. Mohammad, S. Kiritchenko, and X. Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*, 2013.

[16] J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[17] S. R. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[18] K. Song, L. Chen, W. Gao, S. Feng, D. Wang, and C. Zhang. Persentiment: A personalized sentiment classification system for microblog users. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 255–258. International World Wide Web Conferences Steering Committee, 2016.

[19] J. Turner. *On the origins of human emotions: A sociological inquiry into the evolution of human affect.* Stanford University Press, 2000.

[20] Y. Wang, S. Feng, D. Wang, Y. Zhang, and G. Yu. Context-aware chinese microblog sentiment classification with bidirectional lstm. In *Asia-Pacific Web Conference*, pages 594–606. Springer, 2016.

[21] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.

[22] Y. Yao. A partition model of granular computing. *Lecture notes in computer science.*