Expert Systems with Applications 42 (2015) 4517–4528

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A rule-based approach to emotion cause detection for Chinese micro-blogs

Kai Gao [a,b,1], Hua Xu [a,*,1], Jiushuo Wang [a,b,1]

[a] State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[b] School of Information Science and Engineering, Hebei University of Science and Technology, Hebei 050018, China

ARTICLE INFO

Article history:
Available online 7 February 2015

Keywords:
Text mining
Emotion causes
Micro-blog
Cause component proportion

ABSTRACT

Emotion analysis and emotion cause extraction are key research tasks in natural language processing and public opinion mining. This paper presents a rule-based approach to emotion cause component detection for Chinese micro-blogs. Our research has important scientific values on social network knowledge discovery and data mining. It also has a great potential in analyzing the psychological processes of consumers. Firstly, this paper proposes a rule-based system underlying the conditions that trigger emotions based on an emotional model. Secondly, this paper extracts the corresponding cause events in fine-grained emotions from the results of events, actions of agents and aspects of objects. Meanwhile, it is reasonable to get the proportions of different cause components under different emotions by constructing the emotional lexicon and identifying different linguistic features, and the proposed approach is based on Bayesian probability. Finally, this paper presents the experiments on an emotion corpus of Chinese micro-blogs. The experimental results validate the feasibility of the approach. The existing problems and the further works are also present at the end.

1. Introduction

In the era of information explosion, the social network applications present a platform for people to share various news and information sources. As for the micro-blog in China, it has gradually become more popular, and millions of people present and share their opinions every day. The micro-blogs can convey almost all aspects of public opinions, including the description of emergencies, incidents, disasters and some other hot events. Perhaps some of them are full of emotions and sentiments. As a result, many researchers in the field of natural language processing pay more attention to Chinese micro-blog textual emotion processing (Liu, 2012), especially the emotion classification (Desmet & Hoste, 2013; Huang, Peng, Li, & Lee, 2013; Kontopoulos, Berberidis, Dergiades, & Bassiliades, 2013; Yang & Yu, 2013), and the corresponding social relations (Hu, Tang, Tang, & Liu, 2013). Generally, individual emotion generation, expression and perception are influenced by many factors. The emotion cause is considered to be the event or condition that can trigger the corresponding emotion, so emotion cause analysis is essential to mine the public opinion and knowledge discovery. In previous related work, the cause events can be composed of verbs, nominalizations and nouns, and these can evoke the presence of the corresponding emotions by some linguistic cues (Lee, Chen, Huang, & Li, 2013a). For example, as for the sentence (1), the causative verb is *"rang4"* (make), and the emotional keyword is *"Kai1 Xin1"* (happy). With the help of the linguistic rules, we can infer that the emotion cause event is *"Zhe4 Ci4 Chun1 You2"* (this spring outing). Meanwhile, Li and Xu (2014) applied the technology of emotion cause extraction to the micro-blog textual emotion classification.

(1) *Zhe4 Ci4 Chun1 You2 Rang4 Wo3 Hen3 Kai1 Xin1.*[2] (This spring outing makes me very happy.)

In this paper, we combine the relevant knowledge in the field of computer science, emotion psychology and the technology of natural language processing together to explore the emotion causes effectively. Firstly, this paper proposes an emotion model with cause events, and the model describes the conditions that trigger

* Corresponding author.
E-mail addresses: gaokai68@139.com (K. Gao), xuhua@tsinghua.edu.cn (H. Xu), wjs8906@163.com (J. Wang).
[1] Indicates equal contributions from these authors.

[2] It is the original Chinese sentence, and the corresponding literal English translation is shown below.

http://dx.doi.org/10.1016/j.eswa.2015.01.064
0957-4174/

bloggers' emotions in the progress of cognitive evaluation. Meanwhile, all of the sub-events are extracted from micro-blogs. Then this paper detects the corresponding cause events with the help of the proposed rule-based algorithm. Finally, the proportions of different cause components under different emotions are calculated by constructing the emotional lexicon from the corpus and combining different linguistic features.

This paper is organized as follows: Section 2 discusses the related work on various aspects of emotion processing. Section 3 focuses on the method of emotion cause component analysis. Section 4 presents the performance evaluation and the experimental results. In the end, we discuss the remaining challenges and possibilities for the future works.

2. Related work

2.1. Emotion psychological model

As for the cognitive psychology, it focuses on the senior psychological process of human, such as the OCC model which was first proposed by Ortony, Clore and Collins in their book "The Cognitive Structure of Emotions" (Andrew, Clore, & Allan, 1988). The OCC model provided a psychological model of the eliciting conditions of emotions. But it fell short of capturing the logical structure. And Steunebrink, Dastani, and Meyer (2009) proposed a new inheritance-based view of the logical structure of emotions of the OCC model by identifying and clarifying several of the ambiguities. Latter, Steunebrink, Dastani, and Meyer (2012) proposed a formal model of emotion trigger. First, it captured the conditions which triggered emotions in a semiformal way and the main psychological notions used in the emotion model. After that, they proposed a BDI-based framework (belief–desire–intention) to mine the corresponding emotion.

2.2. Emotion classification

In the traditional algorithm on emotion classification, some researchers mainly focus on the following aspects: text processing (e.g., segmentation, part-of-speech tagging, named entity recognition, dependency parsing, etc.), feature extraction and the classification algorithm (e.g., rule-based and machine learning-based methods, etc.). In He (2013), the performances of the three methods (i.e., naive Bayesian, SVM, and SMO) were compared in micro-blog emotion classification. Moraes, Valiati, and Neto (2013) presented an empirical comparison between SVM and ANN for document-level sentiment analysis. Liu, Ren, Sun, and Quan (2013) analyzed the emotion of micro-blog hot events by making use of kernel and SVM method. Wen and Wan (2014) proposed an approach based on class sequential rules to classify the given micro-blog texts into seven emotion types. Li, Li, Li, and Zhang (2014) proposed an approach to generate a multi-class sentiment lexicon by using HowNet, NTUSD and Sina Micro-blog posts. The posts were represented as the lexical vectors based on the lexicon. Then the Semi-GMM and KNN by using symmetric KL-divergence were proposed to classify the lexical vectors for sentiment classification. Liu and Chen (2015) proposed a multi-label classification based approach for emotion analysis, including text segmentation, feature extraction and multi-label classification.

2.3. Construction of the emotional lexicon and multi-language features extraction

As for the emotional lexicon, it can be used to process and identify the emotional words. It usually can be constructed by manual process and acquiring automatically from the corpus (Xu, Liu, Pan, Ren, & Chen, 2008). In the process of emotion analysis, identifying the multi-language features in micro-blog posts is necessary to compute the emotion intensity scores. As for the related work, Zhai, Xu, Kang, and Jia (2011) exploited effective features for Chinese sentiment classification, such as sentiment words, substrings,substring-groups, and key-substring-groups features. Li, Pan, Jin, Yang, and Zhu (2012) expanded a few high-confidence sentiment and topic seeds in target domain by the given RAP algorithm. Ren and Quan (2012) made an analysis on emotion expressions, including the following factors for emotion change: negative words, conjunctions, punctuation marks and contextual emotions. Quan, Wei, and Ren (2013) combined sentiment lexicon and dependency parsing for sentiment classification, and they extracted the evaluation objects based on the dependency and calculated the similarity between the words based on HowNet. Zhang, Xu, and Xu (2015) focused on the semantic features between words by clustering the similar features on the basis of word2vec.

Unlike the related methods, this paper incorporates the method of Chi-squared test, PMI and word2vec into the construction of the emotional lexicon based on Chinese micro-blog corpus.

2.4. Emotion cause analysis

In the aspect of emotion cause analysis, emotions can be invoked by the cause events. Lee, Zhang, and Huang (2013b) presented an event-based emotion corpus to analyze the interaction between event structures and emotions in the text. Lee et al. (2013a) constructed a Chinese emotion cause annotated corpus and presented seven groups of linguistic cues and two sets of generalized linguistic rules for the detection of emotion causes. Li and Xu (2014) proposed and implemented a novel method for identifying emotions of micro-blog posts, and tried to infer and extract the emotion causes by using knowledge and theories from other fields such as sociology. Nguyen, Phung, Adams, and Venkatesh (2013) extracted events using behaviors of sentiment signals and burst structure in social media, and these events often caused the behavioral convergence of the expression of shared emotion. Rao, Li, Mao, and Liu (2014) proposed two sentiment topic models (i.e., Multi-label Supervised Topic Model (MSTM) and Sentiment Latent Topic Model (SLTM)) to extract the latent topics that evoked emotions of readers, then the topics were seen as the causes of emotions as well.

Unlike the above works, this paper presents a novel emotion cause component analysis method for Chinese micro-blog posts. The innovation of this paper lies in mining the micro-blog data by analyzing the corresponding emotion on the basis of an emotion model. And then this paper also presents the subsystem for detecting and extracting the cause events by the designing rules in fine-grained emotions. Thirdly, constructing the emotional lexicon and identifying the multi-language features in micro-blog posts are used to analyze the emotion intensity scores. Finally, this paper presents the proportions of different cause components.

3. Emotion cause component analysis

The main flow on emotion cause component analysis is shown in Fig. 1. On the basis of the corresponding processing, the proportions of different cause components under different emotions will be obtained.

3.1. ECOCC model construction

Based on the cognitive theory, this paper improves the structure about the eliciting conditions of emotions in accordance with the OCC model referred in Andrew et al. (1988) and presents an

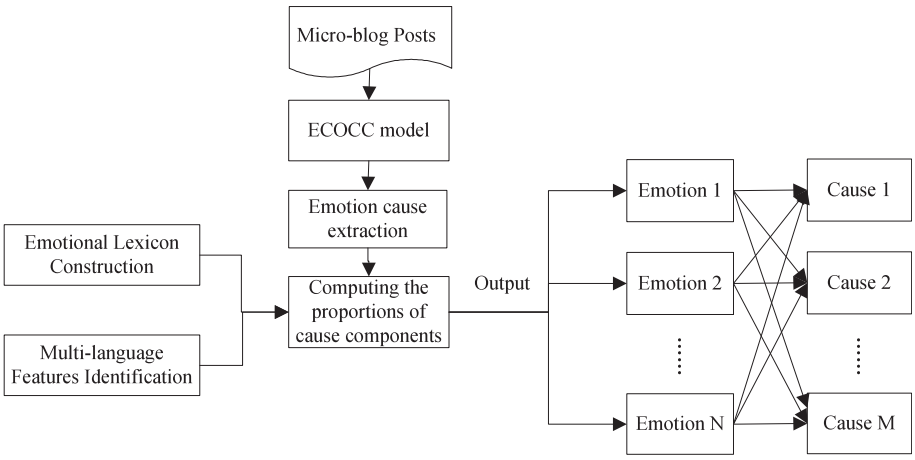

Fig. 1. The framework of the emotion cause component analysis.

emotion model named as ECOCC (Emotion-Cause-OCC) model for micro-blog posts. The ECOCC model combines emotion psychology with computer science to analyze the corresponding emotion cause events. It describes a hierarchy that classifies 22 fine-grained emotions. This hierarchy contains three main branches (i.e., results of events, actions of agents and aspects of objects), and some branches are combined to form a group of the compound and extended emotions. Here, the 22 fine-grained emotions can be divided into the following five groups:

- The emotions in a branch of "results of events": "Hope", "Fear", "Joy" and "Distress".
- The emotions in a branch of "actions of agents": "Pride", "Shame", "Admiration" and "Reproach".
- The emotions in a branch of "aspects of objects": "Liking" and "Disliking".
- The compound emotions in the branches of "results of events" and "actions of agents": "Gratification", "Remorse", "Gratitude"and "Anger".
- The extended emotions: "Satisfaction", "Fears-confirmed", "Relief", "Disappointment", "Happy-for", "Resentment", "Gloating" and "Pity".

On the basis of these main branches, the components of the model matching the emotional rules are divided into six sub-classes (i.e., event_state, event_norm, action_agent, action_norm, object_entity and object_norm). Within the components, the evaluation-schemes can be defined from event_state (i.e., the state that something will happen, the state that something has happened, the state that something did not happen), action_agent (i.e., the emotional agent or others) and object_entity (i.e., the elements of objects). They can be defined as six types (i.e., prospective, confirmation, disconfirmation, main-agent, other-agent and entity). Meanwhile, the corresponding evaluation-standards are defined from the aspects of event_norm (i.e., being satisfactory or unsatisfactory for the event), action_norm (i.e.,being approval or disapproval for the agent) and object_norm (i.e., being attractive or unattractive to the entity). And they can be defined as another six types (i.e., desirable, undesirable, praiseworthy, blameworthy, positive, negative, etc.). In addition, this paper collects the emotional words by using the HowNet,[3] and the National Taiwan University Sentiment Dictionary[4] (where 3505 terms belong to "desirable, praiseworthy, positive" domain, while 9427 terms belong to "undesirable, blameworthy, negative" domain) as the evaluation-standard library for evaluating the relevant events, actions and objects respectively. Finally, we can get the degree of satisfaction of events, approval of agents, and attraction of objects.

The rules (see Table 1) are used to describe the production processes of the 22 types of emotions according to the components in the ECOCC model. Here, "s" means the micro-blog post, "C" represents the cause components that trigger emotions, "→" represents the state that changes from one to the other, "∩" has the same meaning as "and" and "∪" has the same meaning as "or". The particular introduction will be shown in the next subsections.

3.1.1. Production rules of the base emotions on "results of events"

In the "results of events", when the evaluation-scheme of event_state is prospective, and if the evaluation-standard of event_goal is desirable, the corresponding emotion is "Hope". Otherwise, its emotion is "Fear". On the other hand, the corresponding emotions are "Joy" and "Distress", according to the different evaluation-standards of event_state (i.e., confirmation and disconfirmation). For example, in the given Chinese micro-blog post (2), it contains some future characteristics (i.e., *"Ming2 Tian1"* (tomorrow)). As it is a prospective event and the evaluation-standard of event_goal is desirable (*"Wo3 Men Neng2 You3 Yi1 Ge4 Kuai4 Le4 De Ye3 Chui1"* (we can have a happy picnic)), so the emotion is "Hope", and its corresponding emotion cause event is *"Wo3 Men Neng2 You3 Yi1 Ge4 Kuai4 Le4 De Ye3 Chui1"* (we can have a happy picnic).

(2) *Tian1 Qi4 Yu4 Bao4 Shuo1 Ming2 Tian1 Shi4 Ge4 Qing2 Tian1, Wo3 Xi1 Wang4 Wo3 Men Neng2 You3 Yi1 Ge4 Kuai4 Le4 De Ye3 Chui1.* (The weather forecast says tomorrow is sunny, and I hope we can have a happy picnic.)

3.1.2. Production rules of the base emotions on "actions of agents"

Within the "actions of agents", we can analyze the emotion from the perspective of main-agent (e.g., author or blogger) and other-agent. If the evaluation-scheme of action_agent is main-agent and the evaluation-standard of action_norm is praiseworthy (or blameworthy), then its corresponding emotion is "Pride" (or"Shame"). If other-agent has a praiseworthy (or blameworthy) behavior, then its corresponding emotion is "Admiration" (or "Reproach"). For example, in the given Chinese micro-blog post (3), the main-agent is *"Wo3 Men"* (we) and the other-agent is "Alan", and the text shows the "Admiration" emotion and the corresponding emotion cause event is *"Alan Jiu4 Qi3 Le Yi1 Ming2 Luo4 Shui3 Nv3 Hai2"* (Alan rescued a drowning girl).

[3] http://www.keenage.com/.
[4] http://nlg18.csie.ntu.edu.tw:8080/opinion/index.html.

Table 1
The emotional rules.

Classes	The 22 emotional rules
The emotions in the branch of "results of events"	Hope(C)$\stackrel{def}{=}$Prospective(s)∩ Desirable(s)
	Fear(C)$\stackrel{def}{=}$Prospective(s)∩ Undesirable(s)
	Joy(C)$\stackrel{def}{=}$[Confirmation(s)∪Disconfirmation(s)]∩ Desirable(s)
	Distress(C)$\stackrel{def}{=}$[Confirmation(s)∪Disconfirmation(s)]∩Undesirable(s)
The emotions in the branch of "actions of agents"	Pride(C)$\stackrel{def}{=}$main-agent∩Praiseworthy(s)
	Shame(C)$\stackrel{def}{=}$main-agent∩Blameworthy(s)
	Admiration(C)$\stackrel{def}{=}$other-agent∩Praiseworthy(s)
	Reproach(C)$\stackrel{def}{=}$other-agent∩Blameworthy(s)
The emotions in the branch of "aspects of objects"	Liking (C)$\stackrel{def}{=}$ entity ∩Positive(s)
	Disliking(C)$\stackrel{def}{=}$entity∩Negative(s)
Compound emotions	Gratification(C)$\stackrel{def}{=}$Pride(C)∩Joy(C)
	Remorse(C)$\stackrel{def}{=}$Shame(C)∩Distress(C)
	Gratitude(C)$\stackrel{def}{=}$ Admiration (C)∩Joy (C)
	Anger(C)$\stackrel{def}{=}$Reproach(C)∩Distress(C)
Extended emotions	Satisfaction(C)$\stackrel{def}{=}$Joy(C)∩[Prospective(s)→Confirmation(s)]∩ Hope(C)
	Fears-confirmed(C)$\stackrel{def}{=}$Distress(C)∩[Prospective(s)→Confirmation(s)] ∩ Fear(C)
	Relief(C)$\stackrel{def}{=}$Joy(C)∩[Prospective(s)→Disconfirmation(s)] ∩ Fear(C)
	Disappointment(C)$\stackrel{def}{=}$Distress(C)∩[Prospective(s)→Disconfirmation(s)] ∩ Hope(C)
	Happy-for(C)$\stackrel{def}{=}$Joy(C)∩[other-agent∩Praiseworthy(s)∩Desirable(s)]
	Resentment(C)$\stackrel{def}{=}$Distress(C)∩[other-agent ∩ Blameworthy(s) ∩ Desirable(s)]
	Gloating(C)$\stackrel{def}{=}$Joy(C)∩[other-agent∩ Blameworthy(s)∩Undesirable(s)]
	Pity(C)$\stackrel{def}{=}$Distress(C)∩[other-agent∩Praiseworthy(s)∩Undesirable(s)]

(3) *Wo3 Men Wei4 Alan Jiu4 Qi3 Le Yi1 Ming2 Luo4 Shui3 Nv3 Hai2 Gan3 Dao4 Zi4 Hao2.* (We are proud of Alan for rescuing a drowning girl.)

3.1.3. Production rules of the base emotions on "aspects of objects"

As for the "aspects of objects", it is reasonable to classify the given emotions into "Liking" and "Disliking" according to the corresponding evaluation-standard (i.e., positive or negative). For example, in the given Chinese micro-blog post (4), the entity is *"Xiao3 Gou3"* (puppy). According to its characters, the evaluation-standard of the object_norm is positive, so the corresponding emotion is "Liking", and the corresponding emotion cause event is *"Xiao3 Gou3 Tai4 Ke3 Ai4 Le"* (lovely puppy).

(4) *Wo3 Jia1 De Xiao3 Gou3 Tai4 Ke3 Ai4 Le!*. (So lovely of my puppy!).

3.1.4. Production rules of the "compound emotions": Gratification, Gratitude, Remorse and Anger

Some Base emotions can be combined into the compound emotions. In detail, the combination of "Joy" and "Pride" will generate the "Gratification" emotion, while the combination of "Joy" and "Admiration" will generate the "Gratitude" emotion. The combination of "Distress" and "Shame" will generate the "Remorse" emotion, and the combination of "Distress" and "Reproach" will generate the "Anger" emotion. For example, in the given Chinese micro-blog post (5), it presents a blameworthy behavior (*"Ta1 Ba3 Wei2 Yi1 De Shui3 Bei1 Da3 Sui4 Le"* (he broke the only glass)) and an undesirable result (*"Wo3 Men Mei2 Fa3 He1 Shui3 Le"* (we could not drink water)). The event leads main-agent to generate the "Anger" emotion, and its cause event is *"Ta1 Ba3 Wei2 Yi1 De Shui3 Bei1 Da3 Sui4 Le"* (he broke the only glass).

(5) *Ta1 Ba3 Wei2 Yi1 De Shui3 Bei1 Da3 Sui4 Le, Wo3 Men Mei2 Fa3 He1 Shui3 Le.* (He broke the only glass so we could not drink water).

3.1.5. Production rules of the "extended emotions": Satisfaction, Fears-confirmed, Relief, Disappointment, Happy-for, Resentment, Gloating, Pity

"Satisfaction", "Fears-confirmed", "Relief" and "Disappointment" emotions within the ECOCC model are the extended emotions. In detail, while the evaluation-scheme of the event has changed from prospective to confirmation, and the evaluation-standard of the event is desirable (or undesirable), then the corresponding emotion is "Satisfaction" (or "Fears-confirmed"). On the other hand, if the evaluation-scheme of the event has changed from prospective to disconfirmation, and the evaluation-standard of the event is undesirable (or desirable), then the emotion is "Relief" (or "Disappointment"). Similarly, "Happy-for", "Resentment", "Gloating", "Pity" emotions within the ECOCC model are also the extended emotions, which are described by combining results of events and actions of agents. In detail, as for other-agent, when the evaluation-standard of event_goal is desirable, if its base emotion is "Joy" (or "Distress") and the evaluation-standard of action_norm is praiseworthy (or blameworthy), then the emotion is "Happy-for" (or "Resentment"). In addition, when the evaluation-standard of event_goal is undesirable, if its base emotion is "Joy" (or "Distress") and the evaluation-standard of action_norm is blameworthy (or praiseworthy), the emotion is "Gloating" (or "Pity").

3.2. Emotion cause component detection and extraction

The related work has shown that the emotion cause event is consisted of a list of sub-events, which can be formalized as a triple

$U = (nouns, verbs, nouns)$. In this paper, the external events and the internal events are taken into consideration. As for the former, they are considered as the indirect reasons that can trigger emotions and they are also inferred and extracted according to the characters of the micro-blogs. For the latter, this paper proposes the model of extracting sub-events from the results of events, actions of agents and aspects of objects based on the ECOCC model. The Algorithm 1 describes the main process of cause event extraction.

Algorithm 1. The algorithm of the external and internal cause events extraction

```
ECO ← The external events
ECI ← The internal events
M ← The method of emotion cause extraction
for each clause in the micro-blog post do
    topic ← The feature set of #Topic#
    E ← The results of events
    A ← The actions of agents
    O ← The aspects of objects
    if topic is in clause then
        ECO ← Reprocess topic according to Regular Expression
    end if
    if E is in clause then
        ECI ← Extract E according to M
    end if
    if A is in clause then
        ECI ← Extract A according to M
    end if
    if O is in clause then
        ECI ← extract O according to M
    end if
end for
```

3.2.1. Analyzing the external event

Through analyzing the Chinese micro-blog posts, these sentences with the style of "#Topic#" usually contain some social or hot news, which perhaps influence on the corresponding emotion transition tendency. So it is necessary to extract the external event and make it as the emotion cause event. It can be obtained by using the regular expression rule. For example, in the example post (6), by using the regular expression to match the external event, this paper can obtain the external event: *"Li3 Na4 Ao4 Wang3 Du2 Guan4"* (Li Na won the Australian Open).

(6) *#Li3 Na4 Ao4 Wang3 Du2 Guan4# Hen3 Zan4!* (#Li Na won the Australian Open # Great!).

3.2.2. Analyzing the "results of events"

The internal event is usually the direct reason for triggering the change of individual emotion and it can be extracted from the domain of results of events, actions of agents and aspects of objects based on the ECOCC model. As for the domain of results of events, the LTP[5] (Language Technology Platform) is used to set up the model of extracting sub-events based on the named entity recognition, dependency parsing, semantic role labeling, and so on Che, Li, and Liu (2010). The main steps are as follows:

- Labeling the parts of speech of Chinese (e.g., nouns, verbs, adjectives) within the micro-blog posts by using our Chinese segmentation algorithm.
- Identifying the person names, place names and institutions by using Chinese named entity recognition.
- Identifying the core relation of subject–verbs and verb–objects by using dependency parsing.
- Identifying the phrases labeled with the semantic role (i.e., A0) for the actions of the agent, the semantic role (i.e., A1) for actions of the receiver, or four other different core semantic roles (i.e., A2–A5) for different predicates by using the semantic role labeling, respectively.

For example, as for the micro-blog post: *"Wo3 Bang1 Wu2 Bao3 Chun1 Xian1 Sheng1 Pai1 She4 Le Zhe4 Ben3 Shu1 De1 Feng1 Mian4 She4 Ying3, Zhen1 De3 Hao3 Ji1 Dong4 A1!"* (I help Mr. Wu Baochun to take the cover of the book, and it is really very exciting!), the result of semantic parsing is as follows (see Fig. 2):

In detail, this paper first selects the phrases which are labeled as A0, A1, A2–A5 (if it exists) and then combines the above components as the basis of event recognition. Otherwise, the triple $U = (nouns, verbs, nouns)$ will be used as another basis of event recognition. In addition, in the process of the internal event recognition, there are some prospective-events, which stand for the desire for the future events of individual. The processing steps are as follows:

- Determining the feature-words, e.g., *"Jiang1 Lai2"* (in the future), *"Ming2 Tian1"* (tomorrow), *"Ming2 Nian2"* (next year), if they exist.
- Analyzing the lexical features, syntactic features, the characteristics of words, and then generating the feature sets.
- Exploring the relationship between the feature sets and the events by dependency parsing and then confirming the prospective content.

By describing the results of events, it is easy to decide the corresponding emotion and its causes according to the evaluation-schemes and the evaluation-standards.

3.2.3. Analyzing the "actions of agents"

As for the domain of actions of agents, this paper extracts the class of ACT in HowNet which contains a large number of words describing different kinds of actions. If the predicate verb belongs to the class of ACT and there exists an initiative relationship between itself and the agent, then this structure can be used as a kind of agent's action. On the other hand, as for the main-agent, it contains the explicit-agent and the implicit-agent. The former is highlighted in the text and has the subject–predicate relationship with the predicate verb. The latter does not appear in the text, but it is usually expressed by the context and the act of the verb. Finally, on the basis of the description of the actions of agents, it is easy to decide the corresponding emotion and its causes according to the evaluation-schemes and the evaluation-standards.

3.2.4. Analyzing the "aspects of objects"

Meanwhile, the corresponding emotions "Liking" and "Disliking" can describe the reactions of the agent to the corresponding object respectively. For recognizing the characteristic information of aspects of objects, it needs to extract the entities on the basis of the HowNet corpus so as to find the subject–predicate relationship by using dependency parsing. The features of objects are extracted by using the semantic role labeling to confirm the evaluation-standards of object_norm. And then we can get the final emotion and its corresponding cause components.

[5] http://www.ltp-cloud.com/demo/.

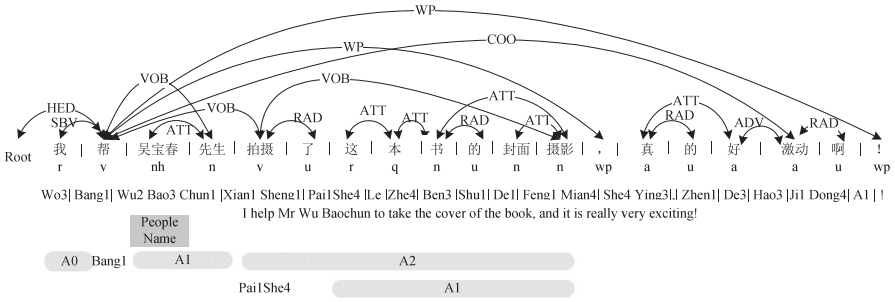

Fig. 2. The result of semantic parsing.

3.2.5. Algorithm on emotion cause extraction

To show the process of the proposed algorithm better, this paper presents the corresponding pseudo-code on cause event extraction. P_i is one of the generation rules of 22 emotions and *E_cause* is the set of emotion causes. *ECI* and *ECO* stand for the internal events and the external events, respectively (see Algorithm 2).

Algorithm 2. The algorithm of emotion cause extraction

```
d ← 1...22
Pi ← Generate rule sets of emotion
Extract cause events by rules in Pi
for each sentence in the micro-blog post do
    ECI ← Extract the internal events according to Pi
    ECO ← Extract the external events
    if ECI == 0  &&  ECO == 0 then
        break
    else
        E_cause ← ECI or ECO
    end if
end for
```

3.3. Emotion cause component analysis

It is clear that several cause events may trigger only one emotion, while one cause event may trigger several emotions. Different cause events have different influences on the production of emotions. The proportion of one cause event in all cause events will also be different. The bigger the proportion is, the greater the probability of triggering emotion is. To achieve the proportion, we commence from the emotion intensities of cause events by constructing the emotional lexicon and identifying the multi-language features in micro-blogs, such as emoticons, negation words, punctuations, conjunctions and degree adverbs, and so on. Finally, the proportion will be calculated based on Bayesian probability.

3.3.1. Emotional lexicon construction

Generally, the emotional lexicon can be constructed manually and automatically from the corpus. Firstly, the standard lexicon can be constructed manually, including the following three steps:

- The choice of emotion types: In our previous work, the ECOCC model describes 22 fine-grained emotions. Each emotion is obtained by analyzing the complex emotions of humanity and the cognitive theory. So it is reasonable to choose the 22 types of emotions as the base emotions.
- The setting of the emotion intensity: As different emotional words have different emotion intensity scores, so the scores can be divided into five level ranges (i.e., 0–1.0, 1.0–2.0, 2.0–3.0, 3.0–4.0 and 4.0–5.0). Among them, 0–1.0 represents the word with the weakest emotion intensity; and 4.0–5.0 represents the corresponding word with the strongest emotion intensity. For example, "Kai1 Xin1 (happy)" belongs to the emotion of "Joy", and its emotion intensity score is 5.0, i.e., the fifth level.
- The choice of standard emotional words: The standard emotional words which belong to 22 different types of emotions are selected manually. And those words are from HowNet, National Taiwan University Sentiment Dictionary, and the Affective Lexicon Ontology (Xu et al., 2008). We give them the corresponding emotion intensities by the setting of the emotion intensity. This work will be completed by three different annotators in order to get a better result.

As for getting the larger capacity and more comprehensive lexicon, it needs to be expanded three times by acquiring automatically from the corpus. The first expansion is completed by chi-square test which can extract the high-frequency and high-class relevance keywords perfectly. This paper chooses the micro-blog posts containing 22 types of emotions as the training data (about 20,000 posts). The emotional keywords will be extracted from the training data by chi-square test, and the emotion intensity of each keyword is marked according to the relevance. The greater the relevance is, the stronger the intensity is. After that, the second expansion is completed by the method based on PMI (i.e., Pointwise Mutual Information). During this processing, the undetermined words are chosen from the corpus, and then the emotion intensities can be determined by the mutual information between the undetermined words and the standard emotional words in the corpus (Xu et al., 2008). The formula (1) of PMI is list as follows. And the parameter k stands for the number of emotion types ($1 \leqslant k \leqslant 22$), WD_k represents the word belonging to the kth emotion in the corpus, ST_{kj} represents the jth word belonging to the kth emotion in the standard lexicon. Finally, the words with the maximum mutual information will be added into the lexicon, and the emotion intensities of the standard words will be as their emotion intensities.

$$PMI(WD_k, ST_{kj}) = \log \frac{P(WD_k ST_{kj})}{P(WD_k)P(ST_{kj})} \tag{1}$$

The third expansion is completed by the word2vec[6] which provides an efficient implementation for computing vector representation of words by using the continuous bag-of-words and skip-gram architectures (Mikolov, Chen, Corrado, & Dean, 2013). Taking full advantage of the tool, we can complete the word clustering and choose the synonyms. Firstly, a large number of posts are collected randomly from Sina Micro-blog website (weibo.com) to constitute a 1.5 G micro-blog dataset. They are transformed to vector

[6] http://word2vec.googlecode.com/svn/trunk/.

representation of words. Then the synonyms of the standard emotional words are chosen as the candidate words. Secondly, it needs to compute the similarity between the candidate word and the standard word. The candidate word with the maximum similarity will be added into the lexicon. Meanwhile, the emotion intensities of the selected words can be defined as the formula (2) below. WD_i represents the ith word in the candidate word list, ST_j represents the jth word in the standard word list, $SIM(WD_i, ST_j)$ represents the maximum similarity between the candidate word and the standard word, and $I(ST_j)$ represents the emotion intensity of the standard word.

$$E_i = SIM(WD_i, ST_j) * I(ST_j) \tag{2}$$

Finally, the emotional lexicon based on 22 types of emotions is shown as follows (see Table 2):

3.3.2. Emoticons analysis

Emoticons can express far more complex emotions. As for the social network domain, according to Twitter statistics, the emoticon "❤" accounts for 10% in the top 100 emoticons. Meanwhile, in previous research works, people tend to classify emoticons into three types: positive, negative and neutral. Some more specific emoticons should be taken into consideration, such as "Happy", "Sadness", "Fear", "Anger", "Disgust" and "Surprise" (Yuan & Purver, 2012). Unlike the related works, this paper will combine the emoticon with the emotion intensity to calculate the proportion of the corresponding cause component.

Table 2
Details of the emotional lexicon.

The 22 types of emotions	The number of words	The number of five level ranges (0–1):(1–2):(2–3):(3–4):(4–5)
Distress	2890	208:704:1262:564:152
Disappointment	1719	628:329:553:177:32
Pity	3670	3268:130:155:70:47
Remorse	3259	2935:139:93:63:29
Fears-confirmed	2200	276:543:808:429:144
Fear	595	155:150:158:96:36
Resentment	276	3:23:113:112:25
Anger	1050	43:283:416:212:96
Disliking	3794	218:1019:1632:724:201
Reproach	13576	1001:3669:5563:2702:641
Shame	555	186:126:154:65:24
Hope	1009	40:284:474:154:57
Admiration	18302	2299:4454:7290:3235:1024
Liking	3134	213:1214:1156:444:107
Gratification	483	39:267:43:96:38
Gratitude	3781	3375:169:159:54:24
Joy	3942	302:1426:1634:482:98
Pride	3426	3070:124:73:116:43
Gloating	952	76:365:472:25:14
Relief	3681	28:1523:2001:82:47
Happy-for	5805	1911:1502:1774:519:99
Satisfaction	2526	54:1078:1270:88:36

Definition 1. The micro-blogs can be formulated as a triple $U = (C, R, T)$, where C means the emoticon list, R is the emotional keyword list, and T represents the micro-blog post list. As for one post, it can be regarded as a triple $u_x = (CV_i, RV_{kj}, t_n)$, where CV_i means the ith emoticon in C, RV_{kj} is the jth emotional keyword in R of the kth ($1 \leqslant k \leqslant 22$) emotion, t_n represents the nth post in T. If CV_i and RV_{kj} appear within t_n at the same time, the corresponding co-occurrence frequency is defined as $|CO(CV_i, RV_{kj})|$.

Definition 2. As for the co-occurrence intensity between the corresponding emoticon and the emotional keyword, it can be represented as $\delta_{ij}(CV_i, RV_{kj})$, see the formula (3), where $|CV_i|$ is the number of CV_i appearing in t_n, and $|RV_{kj}|$ is the number of RV_{kj} appearing in t_n.

$$\delta_{ij}(CV_i, RV_{kj}) = \frac{|CO(CV_i, RV_{kj})|}{(|CV_i| + |RV_{kj}|) - |CO(CV_i, RV_{kj})|} \tag{3}$$

According to the above definitions, it is easy to construct the co-occurrence graph (see Fig. 3). For example, E is the set of center nodes containing emoticons (e.g., CV_1, CV_2, CV_3, etc.), and D is the set of leaf nodes containing emotional keywords (e.g., RV_{11}, RV_{12}, RV_{13}, etc.). P is the set of the edge between E and D, and P represents the degree of closeness between the emotion intensities of CV_i and RV_{kj}. If the edge is longer, then the emotion intensities of both are closer (Cui, Zhang, Liu, & Ma, 2011).

As for the weight of the edge in the co-occurrence graph (which is defined as $W_{ij}(CV_i, RV_{kj})$), it can be set to a value which is equal to the co-occurrence intensity, see the formula (4):

$$W_{ij}(CV_i, RV_{kj}) = \delta_{ij}(CV_i, RV_{kj}) \tag{4}$$

According to the above equations and definitions, the emotion intensity of the ith emoticon (which is called as I_{ICON_i}) can be inferred as follows, see the formula (5), and E_j is the emotion intensity of the jth emotional keyword.

$$I_{ICON_i} = E_j * \max_{1 \leqslant k \leqslant 22} W_{ij}(CV_i, RV_{kj}) \tag{5}$$

3.3.3. Degree adverbs analysis

Besides emoticons, the modification of degree adverbs is also considered in computing the emotion intensities of cause events. In this paper, we use the intensifier lexicon including 219 degree adverbs, which are divided into five levels: "Ji2 Qi2 | extreme"; "Hen3 | very"; "Jiao4 | more"; "Shao1 | -ish" and "Qian4 | insufficiently" (Zhang & He, 2013). Then the influence coefficient is set to x, and the value of x is +0.5, +0.3, −0.1, −0.3 and −0.5, respectively. The "−" has the function of weakening the emotion intensities of the corresponding words while the "+" has the function of strengthening the emotion intensities of the corresponding words. The exponential function e^x is applied to adjust the emotion intensity. Finally, the emotion intensity of the ith emotional keyword (called I_{DA_i}) which is influenced by degree adverbs can be calculated by the following formula (6). Here, E_i is the emotion

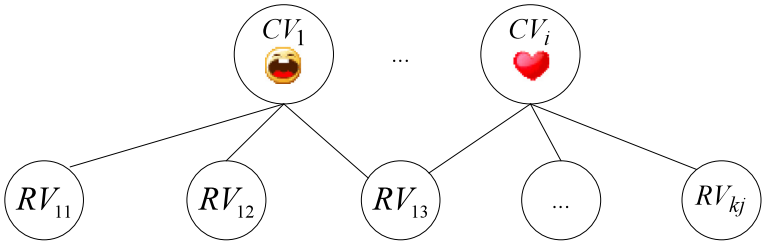

Fig. 3. The co-occurrence graph.

intensity of the ith emotional keyword, and γ is the adjustable parameter.

$$I_{DA_i} = \gamma e^{x} E_i (\gamma \geqslant 1) \tag{6}$$

3.3.4. Negation words recognition and analysis

As negation words can impact the negative transformation of emotions or impact the emotion intensity scores, it is reasonable to analyze this from the following aspects which are common phenomena especially in Chinese micro-blogs.

- Double negation: Double negation has many fixed patterns, such as two negative adverbs appearing in succession or the combination of one negative adverb with one rhetorical question, and so on. Most of them express an affirmative meaning and there exists no transformation of emotion. Meanwhile their emotion intensities become more stronger. For example, "*How could I don't like you?*". This sentence expresses the strong emotion ("Liking"). Here, I_{NEGA_i} is used to express the modified result of the emotion intensity of the ith emotional keyword, see the following formula (7), and the parameter η is the adjustable parameter, E_i is the emotion intensity of the ith emotional keyword.

$$I_{NEGA_i} = \eta E_i (\eta > 1) \tag{7}$$

- The location of the negation word impacts the emotion intensity: If the location of the negation word changes, the emotion and the corresponding emotion intensity will also be changed. For example, the difference between "*not very like*" and "*not like*" is that the former is affirmative and the latter is negative. Table 3 describes the influence of the locations of negation words in three patterns. "NE" means the negation words, "DA" means the degree adverbs, "EW" means the emotional keywords. The parameter α and β stand for the adjustable parameters respectively, and "−" is the sign of the transformation of emotion.

Table 3
Details of the location of negation words.

Patterns	Description	Formalization
NE + DA + EW	Having the same type of emotion with the emotional keyword, but it is weaker than the original emotion intensity	$I_{(NEGA_i)} = \beta E_i (0 < \beta < 1)$
NE + EW	Having the reversed type of emotion with the emotional keyword, but it is same with the original emotion intensity, or it is neutral	$I_{(NEGA_i)} = -E_i$ or $I_{(NEGA_i)} = 0$
DA + NE + EW	Having the reversed type of emotion with the emotional keyword, and the emotion intensity is modified by the degree adverbs	$I_{(NEGA_i)} = -\alpha e^{x} E_i (\alpha \geqslant 1)$

3.3.5. Punctuations processing

As for Chinese micro-blogs, there usually exist a large number of emotional punctuation marks such as the exclamation mark, the interrogation mark and so on. They usually strengthen or promote the emotion intensities of sentences in some manners. On the other hand, the repetitive punctuations can strengthen or promote the emotion intensities more obviously than the single punctuation. The formula (8) is used to compute the emotion intensity which is influenced by punctuations (called I_{PUNC_i}). E_i is the emotion intensity of the ith emotional keyword, and ε is the adjustable parameter that can be confirmed by the degree of repeatability which has a direct relationship with the number of punctuations.

$$I_{PUNC_i} = \varepsilon * E_i (\varepsilon \geqslant 1) \tag{8}$$

Furthermore, with regard to the interrogative sentences, they cannot be analyzed from the above aspects. The reason is that within the Chinese micro-blogs, the frequencies of interrogative sentences under 22 different types of emotions are diverse, and the description of interrogative sentences is shown in Table 4. It is clear that the interrogative sentences mainly appear in the emotions of "Reproach" and "Resentment". This shows that people usually use the interrogation mark to express the two emotions. It also provides an evidence to analyze emotions. Moreover, the interrogation mark has different means in different circumstances, even changes the emotion. For example, as for the sentence: "*This book is very well?*", it is an interrogative sentence which describes a positive attitude to this book literally, but its actual mean is that this book is not good. In another case, "*Who can help me to repair the broken bike?*", which is an interrogative sentence describing a negative attitude, and here, the interrogation mark strengthens this emotion.

3.3.6. Conjunctions processing

There are different kinds of conjunctions such as coordinating conjunctions, adversative conjunctions, causal conjunctions and so on. In some cases, conjunctions may play an emphatic role on the emotional expression. For example, the conjunction *"Dan4 Shi4"* (but) mainly emphasizes the event with a strong emotion behind it. And the influencing parameters are divided into two classes. One is that the conjunction impacts on the emotion intensity of its front event, and the influencing parameter is set to F_{before}. The other is the conjunction impacts on the emotion intensity of its back event, and the influencing parameter is set to F_{after}. If $(F_{before} = F_{after})$, it means that the conjunction has the same effect both on the front event and the back event, so the values of the two parameters are set to 1. If $(F_{before} < F_{after})$, it means that the conjunction has no effect on the front event but strengthens the emotion intensity of the back event, so this paper sets (F_{before}=1) and $(F_{after} > 1)$. If $(F_{before} > F_{after})$, it presents the opposite case with $(F_{before} < F_{after})$, so this paper sets $(F_{before} > 1)$ and $(F_{after} = 1)$. The following Table 5 describes different situations of conjunctions.

Table 4
Distribution of interrogative sentences in 22 types of emotions.

Emotion	Proportion	Emotion	Proportion	Emotion	Proportion
Admiration	0.16	Gratification	0.08	Relief	0.08
Anger	0.18	Gratitude	0.05	Remorse	0.11
Disappointment	0.13	Happy-for	0.09	Reproach	0.34
Disliking	0.09	Hope	0.05	Resentment	0.27
Distress	0.12	Joy	0.10	Satisfaction	0.06
Fear	0.09	Liking	0.08	Shame	0.13
Fear-confirmed	0.13	Pity	0.17	–	–
Gloating	0.09	Pride	0.06	–	–

Table 5
Different situations of conjunctions.

Conjunctions	Examples	The effect of the influencing parameters
Coordinating conjunctions	He2 | and, Yu3 | with, Kuang4 Qie3 | in addition, etc	$F_{before} = F_{after}$
Continuing conjunctions	Ze2 | then, Nai3 | thus, Yu2 Shi4 | hence, etc	$F_{before} < F_{after}$
Adversative conjunctions	Que4 | however, Dan4 Shi4 | but, Ran2 Er3 | nevertheless, etc.	$F_{before} < F_{after}$
Causal conjunctions	Yin1 Wei4 | because, You2 Yu4 | due to, Yin1 Ci3 | hence, Suo3 Yi3 | so, etc	$F_{before} < F_{after}$
Alternative conjunctions	Huo4 | or, Bu4 Shi4...Jiu4 Shi4 | either...or..., etc	$F_{before} = F_{after}$
Comparative conjunctions	Ru2 Tong2 | as, Si4 Hu1 | as if, Bu4 Ru2 | no as, etc	$F_{before} < F_{after}$
Concessive conjunctions	Sui1 Ran2 | although, Jin3 Guan3 | though, Ji2 Shi3 | even though, etc	If the main clause belongs to the front event, $F_{before} > F_{after}$; If the main clause belongs to the back event, $F_{before} < F_{after}$
Progressive conjunctions	Bu4 Dan4...Hai2 | not only...but also, Shen4 Zhi4 | even, etc	$F_{before} < F_{after}$

Therefore, this paper proposes the following formula (9) to express the modified result of the emotion intensity of the ith emotional keyword (called I_{CONJ_i}).

$$I_{CONJ_i} = \begin{cases} F_{before} * E_i & F_{before} > F_{after} \\ E_i & F_{before} = F_{after} \\ F_{after} * E_i & F_{before} < F_{after} \end{cases} \tag{9}$$

3.3.7. The calculation of cause component proportion based on Bayesian probability

The Bayesian algorithm is widely used to many fields such as text classification, word segmentation, information extraction and so on. This paper describes the proportions of cause components under different emotions from the perspective of the prior probability and the conditional probability by combining the characteristics of Bayesian probability model.

Firstly, this paper constructs an emotion cause component matrix $\rho(s)$ for the micro-blog posts, see the formula (10). $E(C_m)$ represents the emotion vector with cause components, m is the serial number of 22 types of emotions, E_{nm} represents the nth emotion cause intensity score of the mth emotion.

$$\rho(s) = (E(C_1), E(C_2), \ldots, E(C_m))^T = \begin{pmatrix} E_{11} & \ldots & E_{1m} \\ \vdots & \ddots & \vdots \\ E_{n1} & \ldots & E_{nm} \end{pmatrix} \tag{10}$$

The proportion of the nth cause component under the mth emotion (which is defined as $P(Emo_m|Cau_n)$) can be computed based on the Bayesian probability, see the formula (11).

$$P(Emo_m|Cau_n) = \frac{P(Cau_n|Emo_m)P(Emo_m)}{\sum_{m=1}^{22} P(Emo_m)P(Cau_n|Emo_m)} \tag{11}$$

Within the above formula (11), the parameter Emo_m is the mth emotion and Cau_n is the nth cause component under Emo_m. The prior probability $P(Emo_m)$ is the probability distribution of Emo_m. It can be calculated by the formula (12) and (13), where the parameter $SCORE(Emo_m)$ is the mth emotion intensity score which can be modified by the multi-language features in micro-blogs. The parameter I_{ICON_i} is the result modified by emoticons in micro-blogs. The parameter I_{DA_i} is the result modified by degree adverbs. The parameter I_{NEGA_i} is the result modified by negation words. The parameter I_{PUNC_i} is the result modified by punctuations. The parameter I_{CONJ_i} is the result modified by conjunctions. All of them are described in the above sections. If there is no any linguistic feature, the modified result will be ignored and set to 0.

$$P(Emo_m) = \frac{SCORE(Emo_m)}{\sum_{m=1}^{22} SCORE(Emo_m)} \tag{12}$$

$$SCORE(Emo_m) = \sum_{i=1} (E_i + I_{DA_i} + I_{NEGA_i} + I_{ICON_i} + I_{PUNC_i} + I_{CONJ_i}) \tag{13}$$

Within the above formula (11), $P(Cau_n|Emo_m)$ is the probability density function of the nth cause component in a known condition of emotion. It can be calculated by the formula (14) and (15), where $SCORE(Cau_n)$ is the emotion intensity score of the nth cause component under the mth emotion. It is also influenced by the multi-language features.

$$P(Cau_n|Emo_m) = \frac{SCORE(Cau_n)}{\sum_{n=1} SCORE(Cau_n)} \tag{14}$$

$$SCORE(Cau_n) = \sum_{i=1} (E_{im} + I_{DA_{im}} + I_{NEGA_{im}} + I_{ICON_{im}} + I_{PUNC_{im}} + I_{CONJ_{im}}) \tag{15}$$

3.3.8. Demonstration

In this section, for better understanding the whole operation of our work, we present a demonstration. The details are shown as follows:

For example, when a sentence is inputted into our system, we suppose that you can get two types of emotions by the method of emotion cause extraction referred in Sections 3.1 and 3.2, namely "*emotion1*" and "*emotion2*". Meanwhile, you can also obtain the two cause events under "*emotion1*" (i.e., "*cause1*" and "*cause2*") and one cause event under "*emotion2*" (i.e., "*cause3*"). Then, by using the calculation method of cause components proportions in Section 3.3, you can get the proportions of different cause components. We suppose that the proportion of "*cause1*" in all cause events under "*emotion1*" is "*p1*", the proportion of "*cause2*" in all cause events under "*emotion1*" is "*p2*", and the proportion of "*cause3*" is "*p3*" under "*emotion2*". The three proportions meet the following condition:

$$p1 + p2 + p3 = 1. \tag{16}$$

4. Experimental results and analysis

In this section, we present some experiments. First, it needs to construct the emotion cause annotated corpus from Chinese Sina Micro-blog (weibo.com). The dataset is crawled by our related work based on simulating browsers behaviors (Gao, Zhou, & Grover, 2014). Then, it needs to do cause extraction experiment based on the above corpus.

4.1. Dataset and metrics

4.1.1. The experimental dataset

In order to obtain the micro-blog dataset, some strategies based on simulating browsers' behaviors are used to obtain more than 18,000 posts from Chinese Sina Micro-blog (weibo.com). Within

the dataset, the micro-blog posts are short, and most of them are less than 140 characters. Users can use the "#" hashtag which marks the hot topics, the emoticons, web links, or pictures in their posts. After the pre-processing (i.e., removing duplicates, filtering irrelevant results and doing some conversions), 16,371 posts are remained in our dataset. Meanwhile, every data needs to be labeled the emotion type and the corresponding cause by some markers manually. As each post may contain different types of emotions, we cannot label it with one kind of emotions, so the proposed 22 kinds of emotions are summed up as five categories (i.e., "*Happiness*", "*Anger*", "*Disgust*", "*Fear*" and "*Sadness*"). In the process of annotating, in order to avoid the ambiguity and simplify the problem, only the micro-blog posts that express explicitly emotion will be labeled with a tag of the corresponding emotion. The labeling task will be completed by three markers in the field of emotion processing. We label the posts according to the following rules:

- If there are more emotions within a post, only the predominant emotion will be labeled then.
- If the post contains the emotion type and the corresponding cause component at the same time, both will be labeled then.
- If the post does not have any cause component, only the type of emotion will be labeled.
- If two of the markers cannot decide which component is the cause, the third marker will label it then.
- If all of the markers do not distinguish which emotion the post belongs to, the post will be marked neutrally.
- After completing the marking tasks, the author must check the results in order to ensure the accuracy of the tag.

The marked results of the experimental dataset are shown in Table 6.

4.1.2. Evaluation metrics

To evaluate the performance, this paper uses the following three metrics: precision (U_P), recall (U_R) and F-score (U_F), see the formula (17)–(19), respectively. S is the set of all posts in the collection. *NEC* is the number of posts with cause components which are identified through our algorithm. *NEM* is the total number of posts with cause components.

$$U_P = \frac{s_i \in S | Proportion(s_i)\ is\ correct}{NEC} \tag{17}$$

$$U_R = \frac{s_i \in S | Proportion(s_i)\ is\ correct}{NEM} \tag{18}$$

$$U_F = \frac{2 * U_P * U_R}{U_P + U_R} \tag{19}$$

As for the correctness of the cause components proportions under different emotions, the following method is used to analyze. Firstly, the proportional range of each cause component is divided into ten levels, and each level is expressed as $\tau_i (1 \leqslant i \leqslant 10) : \tau_1 \in (0\%—10\%)$, $\tau_2 \in (10\%—20\%)$, $\tau_3 \in (20\%—30\%)$, $\tau_4 \in (30\%—40\%)$, $\tau_5 \in (40\%—50\%)$, $\tau_6 \in (50\%—60\%)$, $\tau_7 \in (60\%—70\%)$, $\tau_8 \in (70\%—80\%)$, $\tau_9 \in (80\%—90\%)$ and $\tau_{10} \in (90\%—100\%)$, respectively. The sampled posts with cause components are labeled manually according to the above proportional levels. Secondly, the proportion of the cause component is obtained according to our algorithm, and then it is set to x. If the absolute error between x and τ_i is less than φ (φ is a calibration parameter), the result is correct; otherwise, it is incorrect. The formula is shown as follows.

$$|x - \tau_i| < \varphi. \tag{20}$$

Table 6
Details of the micro-blog data.

Coarse-emotions	Number of posts	Number of posts with causes
Happiness	504	354
Anger	472	452
Disgust	150	137
Fear	140	131
Sadness	304	255
Neutral	14801	N
Total	16371	1329

4.2. Experimental analysis on emotion causes

The goal of our work is to examine the effects of emotion cause extraction. In order to achieve this goal, this paper groups the experiment into two aspects: one is to verify the feasibility of our algorithm in the aspect of emotion cause component detection, the other is to verify whether using the multi-language features effectively can improve the accuracy of calculation of cause component proportion.

As the first point, this paper compares our method with other two methods, and both are discussed in the literature review. In Method I, Lee et al. (2013a) detected emotion causes with a linguistic rule-based approach. And Method II is designed on the rule-based subsystem from the micro-blog posts based on the common social network characteristics and other carefully-generalized linguistic patterns (Li & Xu, 2014). The difference between ours and others is the approach on extracting emotion causes. This paper uses an emotion model to extract emotion causes, and Method I and Method II rely on the linguistic cues to extract emotion causes. This paper uses the same metric as Li's (Li & Xu, 2014) to evaluate the performance.

Although the three experiments use the micro-blog dataset and make the F-score as the evaluation metric, the performance of our method is superior to others compared to Method I and Method II. The results are shown in Table 7. Table 7 shows the performance of our method and the other two methods. The F-score of our proposed approach is improved by 12.95% and 4.21%, respectively. The result shows the feasibility of this approach and lays a foundation for the calculation of cause component proportion.

Latter, the seven groups are organized in the following Table 8 on the same micro-blog dataset. And the corresponding results of the experiment are shown in Table 9.

Table 7
Comparison among the methods.

Our method (%)	Method I (%)	Method II (%)
65.51	52.56	61.30

Table 8
Experiments of the seven groups.

Experiments	Descriptions
Baseline	Calculating the proportions only from the emotion intensities of keywords (which is called "EW").
EW + DA	Calculating the proportions combining EW with degree adverbs.
EW + ICON	Calculating the proportions combining EW with emoticons in Chinese micro-blogs.
EW + NEGA	Calculating the proportions combining EW with negation words.
EW + PUNC	Calculating the proportions combining EW with punctuations.
EW + CONJ	Calculating the proportions combining EW with conjunctions.
EW + ALL	Calculating the proportions combining EW with the five features (DA + ICON + NEGA + PUNC + CONJ).

Table 9
Results of the experiment.

Experiments	Precision (%)	Recall (%)	F-score (%)
Baseline	76.52	64.48	69.99
EW + DA	77.05	64.94	70.48
EW + ICON	79.91	67.34	73.09
EW + NEGA	79.55	67.04	72.76
EW + PUNC	77.50	65.31	70.89
EW + CONJ	77.32	65.16	70.72
EW + ALL	82.50	69.53	75.46

Table 9 shows the results of emotion cause component analysis in different aspects. The precision of the overall experiment is 82.50% which is higher than the other experiments. And the baseline without any linguistic feature has the lowest F-score which is 69.99%. If the experiment is conducted by combining with emoticons, then the F-score increases to 73.09%. Similarly, if we add the other features, the F-score also increases.

Meanwhile, the recall metric is lower in the experiment. There are three reasons leading to this phenomenon. First, the length of micro-blog post is shorter and flexible. If the post only contains an emoticon, it is difficult to extract the emotion cause event. So this will lead to lower recall ratio. Second, from the perspective of micro-blog users, some people tend to express emotions, but perhaps they do not say what triggers their emotions. In this case, we cannot detect the corresponding cause event. Third, there may exist some deviations in the proportional ranges which are labeled and the calculative proportions although the experiment effectively detects the cause components, and as a result, this may also reduce the recall ratio. In the near future, enhancing the recall ratio is necessary. As our current dataset is not very big, increasing the micro-blog dataset is necessary. According to the characteristics of micro-blog, some new features of micro-blog should be taken into consideration.

Fig. 4 presents the comparison among the corresponding baseline experiment and the proposed algorithm with multi-language features in terms of F-score metrics. Obviously, recognizing the linguistic features of emoticons, negation words, punctuations, conjunctions and degree adverbs can be helpful in extracting emotion causes and calculating the proportions of the cause components. We also find that the emoticons and the negation words have a greater influence on calculating the proportion. And the precisions of the two experiments are 79.91% and 79.55%, respectively. That is, people tend to use the two features to express strong emotions in micro-blogs.

On the other hand, even though the performance of the precision and the recall is not extremely high, our method can extract the cause events that trigger different emotions effectively and find the main cause component by the proportions of causes under public emotions. For example, if most people are in "Anger" emotion, we can quickly find out what triggers the generation of the emotion.

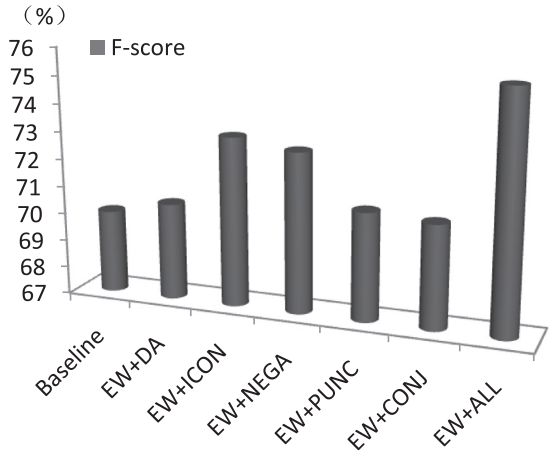

Fig. 4. Comparison of F-score of the seven different experiments.

Our research has significant values. First, as for the application area of crisis management, our work can be useful for dealing with public emergencies. By analyzing the causes of public emotions in social media, it is easy to explore the existing problems of the society. Second, our research is an important branch of emotion mining. It has a profound influence on identifying the implied emotions. Meanwhile, we try to combine the theories of cognitive psychology, emotion psychology and linguistics in our research together. This has important scientific values both on social network knowledge discovery and data mining. Third, mining the relations between the emotions and the corresponding causes is important for product design, and it can also help users to comprehensively analyze the psychological processes of consumers and the psychological mechanisms of implanting advertising.

5. Conclusions and future works

By using the Chinese micro-blog as the experimental dataset, this paper presents a rule-based approach to emotion cause component detection for Chinese micro-blogs. Firstly, this paper presents the ECOCC emotion model and extracts the corresponding cause components in fine-grained emotions. The emotional lexicon can be constructed manually and automatically from the corpus. Meanwhile, the proportions of cause components can be calculated in the influence of the multi-language features based on Bayesian probability. The experiment results show the feasibility of the approach.

Even though some achievements have been made, inadequacies are also existed in the research. There are five significant future directions and they can be meaningful for mining much deeper information for emotion analysis and emotion cause detection in social media.

From the level of research methods, there are two aspects. First, Table 9 obviously presents that different linguistic features play an important role in calculating the proportions of different cause components. Hence, more new features, such as different semantic structures and complicated linguistic patterns, should be explored. Second, we try to use the method based on statistical learning to mine the emotion causes. Meanwhile, exploring the relationship between the time and the cause event will be another meaningful future direction. These research results can help people find the changing rules of the cause event in different periods of time.

From the level of applied research, there are three aspects. First, our research could be applied to precision marketing for product recommendation. For example, we can recommend some cosmetic products to some special customers at the right time. Second, our research result should be used as an approach or method on social management. Through analyzing the causes of public emotions in social media, the government, for example, can find some existing problems. Third, our research can also be used to develop a system of opinion analysis system. It can help people find valuable information and make the right decision. We believe that these research directions are valuable future directions for the research community of expert systems with application.

Acknowledgments

This work is sponsored by National Natural Science Foundation of China (Grant Nos.: 61175110, 61272362) and National Basic Research Program of China (973 Program, Grant No.: 2012CB316301). It is also sponsored by National Science Founda-

tion of Hebei Province (Grant No.: F2013208105), Key Research Project for University of Hebei Province (Grant No.: ZD2014029).

References

Andrew, O., Clore, G., & Allan, C. (1988). *The cognitive structure of emotions*. Cambridge University Press.

Che, W., Li, Z., & Liu, T. (2010). Ltp: A chinese language technology platform. In *Proceedings of the 23rd international conference on computational linguistics: Demonstrations* (pp. 13–16). ACL.

Cui, A., Zhang, M., Liu, Y., & Ma, S. (2011). Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. *Information Retrieval Technology, 7097*, 238–249.

Desmet, B., & Hoste, V. (2013). Emotion detection in suicide notes. *Expert Systems with Applications, 40*, 6351–6358.

Gao, K., Zhou, E.-L., & Grover, S. (2014). Applied methods and techniques for modeling and control on micro-blog data crawler. *Applied Methods and Techniques for Mechatronic Systems, 452*, 171–188.

He, H. (2013). Sentiment analysis of Sina Weibo based on semantic sentiment space model. In *Proceedings of management science and engineering* (pp. 206–211). IEEE.

Huang, S., Peng, W., Li, J., & Lee, D. (2013). Sentiment and topic analysis on social media: a multi-task multi-label classification approach. In *Proceedings of the fifth annual ACM web science conference (WebSci'13)* (pp. 172–181). ACM.

Hu, X., Tang, L., Tang, J., & Liu, H. (2013). Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on web search and data mining* (pp. 537–546). ACM.

Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications, 40*, 4065–4074.

Lee, S. Y. M., Chen, Y., Huang, C.-R., & Li, S. (2013a). Detecting emotion causes with a linguistic rule-based approach. *Computational Intelligence, 39*, 390–416.

Lee, S. Y. M., Zhang, H., & Huang, C.-R. (2013b). An event-based emotion corpus. *Chinese Lexical Semantics, 8229*, 625–644.

Li, Y., Li, X., Li, F., & Zhang, X. (2014). A lexicon-based multi-class semantic orientation analysis for microblogs. *Web Technologies and Applications, 8709*, 81–92.

Li, F., Pan, S. J., Jin, O., Yang, Q., & Zhu, X. (2012). Cross-domain co-extraction of sentiment and topic lexicons. In *Proceedings of the 50th annual meeting of the association for computational linguistics* (pp. 410–419). ACL.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*, 1–167.

Liu, S. M., & Chen, J.-H. (2015). A multi-label classification based approach for sentiment classification. *Expert Systems with Applications, 42*, 1083–1093.

Liu, N., Ren, F., Sun, X., & Quan, C. (2013). Sentiment analysis of Sina Weibo based on semantic sentiment space model. In *Proceedings of 2013 IEEE/SICE international symposium on system integration (SII)* (pp. 233–238). IEEE.

Li, W., & Xu, H. (2014). Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications, 41*, 1742–1749.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the first international conference on learning representations*.

Moraes, R., Valiati, J. F., & Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications, 40*, 621–633.

Nguyen, T., Phung, D., Adams, B., & Venkatesh, S. (2013). Event extraction using behaviors of sentiment signals and burst structure in social media. *Knowledge and Information Systems, 37*, 279–304.

Quan, C., Wei, X., & Ren, F. (2013). Combine sentiment lexicon and dependency parsing for sentiment classification. In *Proceedings of 2013 IEEE/SICE international symposium on system integration (SII)* (pp. 100–104). IEEE.

Rao, Y., Li, Q., Mao, X., & Liu, W. (2014). Sentiment topic models for social emotion mining. *Information Sciences, 266*, 90–100.

Ren, F., & Quan, C. (2012). Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: An application of affective computing. *Information Technology and Management, 13*, 321–332.

Steunebrink, B. R., Dastani, M., & Meyer, J.-J. C. (2012). A formal model of emotion triggers: an approach for BDI agents. *Synthese, 185*, 83–129.

Steunebrink, B. R., Dastani, M., & Meyer, J.- J. C. (2009). The OCC model revisited. In *Proceedings of the 4th workshop on emotion and computing*.

Wen, S., & Wan, X. (2014). Emotion classification in microblog texts using class sequential rules. In *Twenty-eighth AAAI conference on artificial intelligence* (pp. 187–193). Association for the Advancement of Artificial Intelligence.

Xu, L., Liu, H., Pan, Y., Ren, H., & Chen, J. (2008). Constructing the affective lexicon ontology. *Journal of the China Society for Scientific and Technical Information, 27*, 180–185.

Yang, D.-H., & Yu, G. (2013). A method of feature selection and sentiment similarity for chinese micro-blogs. *Journal of Information Science, 39*, 429–441.

Yuan, Z., & Purver, M. (2012). Predicting emotion labels for chinese microblog texts. In *Proceedings of the first international workshop on sentiment discovery from affective data (SDAD 2012)* (pp. 40–47). CEUR.

Zhai, Z., Xu, H., Kang, B., & Jia, P. (2011). Exploiting effective features for chinese sentiment classification. *Expert Systems with Applications, 38*, 9139–9146.

Zhang, P., & He, Z. (2013). A weakly supervised approach to chinese sentiment classification using partitioned self-training. *Journal of Information Science, 39*, 815–831.

Zhang, D., Xu, H., & Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and svm^{perf}. *Expert Systems with Applications, 42*, 1857–1863.