Chinese Emotional Expressions Analysis: Construction of a Blog Emotion Corpus

Changqin Quan, Fuji Ren
Dept. of Info. Science & Intelligent Systems Faculty of Engineering
The University of Tokushima, Tokushima, Japan
E-mail: {quan-c, ren}@is.tokushima-u.ac.jp

The Internet is frequently used as a medium for communication or expression of emotions. In this study a blog emotion corpus is constructed for Chinese emotional expression analysis. This corpus contains manual annotation of emotion category, emotion intensity, emotion holder, emotion target, emotional keyword/phrase, and other linguistic expressions that indicate emotion. There are 198 documents, 5,608 sentences, 135,606 Chinese words contained in this corpus.

1 Introduction

Emotions play important role in human intelligence, rational decision decision making, social interaction, perception, memory, learning, creativity, and more [1]. There is plenty of evidence that emotion analyses have many valuable applications.

Textual affect sensing is becoming increasingly important due to augmented communication via computer mediated communication (CMC) Internet sources such as weblogs, emails, website forums, and chat rooms. Despite the increased focus on analysis of web content, there has been limited emotion analysis of web contents, with the majority of studies focusing on sentiment analysis or opinion mining. In this area, some of the hardest problems involve acquiring basic resources. Corpora are fundamental both for developing sound conceptual analyses and for training these 'emotion-oriented systems' at different levels: to recognise user emotions, to express appropriate emotions, to anticipate how a user in one state might respond to a possible kind of reaction from the machine, and other emotion processing applications.

In this study we propose a relatively fine-grained annotation scheme, annotating emotion in text at 3 levels: document, paragraph, and sentence. In document and paragraph levels, emotion category, emotion intensity, topic keywords and topic sentences are annotated. In sentence level, annotation includes emotion category, emotion intensity, emotional keyword/phrase, degree word, negative word, conjunction, rhetoric, punctuation, objective/subjective, and emotion polarity.

The remainder of this paper is organized as follows. Section 2 presents a review of current emotion corpora for textual emotion analysis. Section 3 describes emotional expression space model in text. Section 4 describes the annotation scheme of this corpus. Section 5 is the conclusions.

2 Related work

Previous approaches to textual emotion analysis have employed some different corpora. Mishne experimented mood classification in blog posts on a corpus of 815,494 blog posts from Livejournal (http://www.livejournal.com), a free weblog service with a large community [2]. Livejournal also was used as data source for finding happiness [3], capturing global mood levels [4], classifing mood ([5], [6]), discovering mood irregularities [7], recognizing affect [8]. A similar emotion corpus in Chinese is Yahoo!'s Chinese news (http://tw.news.yahoo.com), which was used for Chinese emotion classification of news readers [9],

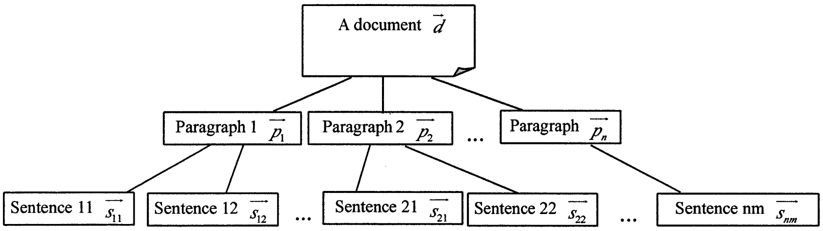

A document $\vec{d}$
Paragraph 1 $\vec{p_1}$
Paragraph 2 $\vec{p_2}$
...
Paragraph $\vec{p_n}$
Sentence 11 $\vec{s_{11}}$
Sentence 12 $\vec{s_{12}}$
...
Sentence 21 $\vec{s_{21}}$
Sentence 22 $\vec{s_{22}}$
...
Sentence nm $\vec{s_{nm}}$

Fig. 1. Hierarchical emotion expression space model in text

emotion lexicon building [10]. These corpora may helpful for analyzing the global moods on a full text, but the inconsistent emotion categories is a problem, and no more labeled information can be exploited from them.

For many applications, identifying emotions only on document level may not be sufficient. A text-based emotion prediction system would benefit from identifying the emotional affinity of sentences. The emotion analysis on sentence level may also be important for more detailed emotion analysis systems. To the best of our knowledge, at present, there's no relatively large corpora annotated with detailed linguistic expressions for emotion in Chinese, and we believe that such corpora would support the development and evaluation of emotion analysis systems.

3 Emotional expression space model in text

This emotional expression space model is hierarchical in consistent with the natural structure of a document. Emotion of a document is represented by a vector

$$\vec{d} = < e_1, e_2, \dots, e_i, \dots, e_n > \qquad (1)$$

Here, e_i $(1 \leq i \leq n)$ is a basic emotional class contained in document d. The values of e_i range from 0.0 to 1.0 (discrete), indicating the intensities of the basic emotional classes.

Similar to a document, emotion of each paragraph and each sentence in a document is represented by an emotional vector. Figure 1 shows the hierarchical model for emotional expression in text.

To decrease confusions on emotion categories' selection and to contain the most common emotional classes in blogs, we select eight emotion classes (expect, joy, love, surprise, anxiety, sorrow, angry and hate) for this manual annotation.

4 Chinese Blog Emotion Corpus Annotation Scheme

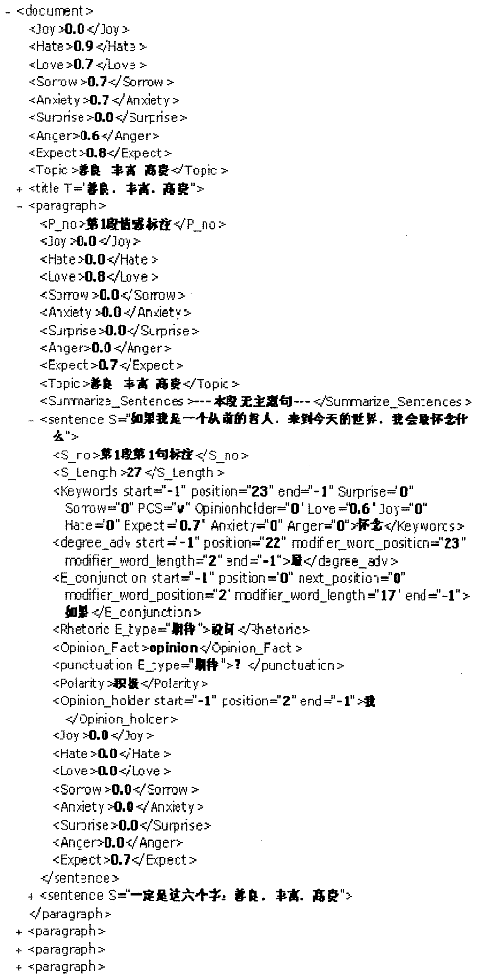
Writing of weblogs suits the recording of facts and the communication of ideas, and their textual basis makes them equally suitable for recording emotions and opinions. So, we select blogs as object and data source for this study.

4.1 The multi-level annotation frame

According to the emotional expression space model in text, the annotation frame includes 3 levels: document, paragraph, and sentence. Sentence level is the basic level for emotion annotation, the annotation include intensities of the eight basic emotion classes, emotion holder/target, emotional keywords/phrases, rhetoric, emotional punctuations, emotional objective/subjective and emotional polarity.

Paragraph level is the upper level of sentence level, the annotation include intensities of the eight basic emotion classes, topic keywords to reflect the topic of a paragraph, and the numbers of topic sentences that can express the main points of this paragraph. Document level is the uppermost level in annotation; its annotation is similar to paragraph level.

The tokenised text files are organized into XML documents, with Chinese segmentation tags and part-of-speech tags included as attributes of the tokens. An example document is listed in Fig. 2.



```
- <document>
    <Joy>0.0</Joy>
    <Hate>0.9</Hate>
    <Love>0.7</Love>
    <Sorrow>0.7</Sorrow>
    <Anxiety>0.7</Anxiety>
    <Surprise>0.0</Surprise>
    <Anger>0.6</Anger>
    <Expect>0.8</Expect>
    <Topic>善良  丰富  高贵</Topic>
  + <title T="善良，丰富，高贵">
  - <paragraph>
      <P_no>第1段情感标注</P_no>
      <Joy>0.0</Joy>
      <Hate>0.0</Hate>
      <Love>0.8</Love>
      <Sorrow>0.0</Sorrow>
      <Anxiety>0.0</Anxiety>
      <Surprise>0.0</Surprise>
      <Anger>0.0</Anger>
      <Expect>0.7</Expect>
      <Topic>善良  丰富  高贵</Topic>
      <Summarize_Sentences>---本段无主题句---</Summarize_Sentences>
    - <sentence S="如果我是一个从前的客人，来到今天的世界，我会最怀念什
        么">
        <S_no>第1段第1句标注</S_no>
        <S_Length>27</S_Length>
        <Keywords start="-1" position="23" end="-1" Surprise="0"
          Sorrow="0" POS="v" Opinionholder="0" Love="0.6" Joy="0"
          Hate="0" Expect="0.7" Anxiety="0" Anger="0">怀念</Keywords>
        <degree_adv start="-1" position="22" modifer_word_position="23"
          modifier_word_length="2" end="-1">最</degree_adv>
        <E_conjunction start="-1" position="0" next_position="0"
          modifier_word_position="2" modifier_word_length="17" end="-1">
          如果</E_conjunction>
        <Rhetoric E_type="期待">设问</Rhetoric>
        <Opinion_Fact>opinion</Opinion_Fact>
        <punctuation E_type="期待">？</punctuation>
        <Polarity>积极</Polarity>
        <Opinion_holder start="-1" position="2" end="-1">我
          </Opinion_holder>
        <Joy>0.0</Joy>
        <Hate>0.0</Hate>
        <Love>0.0</Love>
        <Sorrow>0.0</Sorrow>
        <Anxiety>0.0</Anxiety>
        <Surprise>0.0</Surprise>
        <Anger>0.0</Anger>
        <Expect>0.7</Expect>
      </sentence>
    + <sentence S="一定是这六个字：善良，丰富，高贵">
    </paragraph>
  + <paragraph>
  + <paragraph>
  + <paragraph>
```

Fig. 2 An example annotated document in XML format

4.2 Sentence level annotation

Sentences are basic units for emotional expressions. The central aim of sentence level annotation is to explore as much linguistic expressions for reflecting emotion in Chinese as possible. Besides intensities of the eight basic emotion classes, the annotation include emotion holder/target, emotional keywords and phrases, degree words, negative words, conjunctions, rhetoric, emotional punctuations, emotional objective/subjective and emotional polarity.

4.2.1 Emotion holder and emotion target

We define an emotion holder is the one who holds the emotions, and an emotion target is the object of an emotion holder. As for emotion holder/target identification, little research has been conducted, but we believe it is important for exploring emotional expression and emotion analysis.

Because blogs are personal diaries, in a lot of cases, we can take the writer as emotion holder and all entities in this blog as emotion targets. This condition is not included in this annotation scheme. We annotate distinct emotion holder and emotion target. For instance, In the sentence "我喜欢这个老师。(Wo xi huan zhe ge lao shi; English: I like this teacher.)" , "我 (wo; English: I)" is the emotion holder, and "这个老师。(Zhe ge lao shi; English: this teacher.)" is the emotion target. In this corpus, not every sentence is annotated with emotion holder or emotion target, and emotion holder or emotion target may not appear in pairs in one sentence. As an example, In the sentence "孩子们开心地笑着。(Hai zi men kai xin de xiao zhe; English: Children are laughing happily.)", "孩子们 (Hai zi men; English: Children)" is the emotion holder, no emotion target in this sentence. If one sentence has more than one emotion holder or emotion target, they are all annotated.

4.2.2 Emotional keywords and phrases

For emotion analysis tasks, the function of words is fundamental.

In above sentimental or affective lexicons, the words usually bear direct emotions or opinions, such as happy or sad, good or bad. However, there are a lot of sentences can evoke emotions without direct emotional words, for example:

(1) 我不再需要任何人提醒我、告诉我什么是现实。

(Wo bu zai xu yao ren he ren ti xing wo, gao shu wo shen mo shi xian shi; English: I don't want anybody to remind me and tell me what is realities any more.)

(2) 春天在孩子们的眼里、在孩子们的心里。(Chun tian zai hai zi men de xin li, zai hai zi men de yan li; English: Spring is in children's eyes, and in their hearts.)

We may sense sorrow, angy, or hate in sentence (1). In sentence (2), we may feel joy, love or expect delivered by the writer.

In this annotation scheme, direct affective words and indirect affective words in a sentence are all annotated. In sentence (1), "任何人", "现实" will be lablled with emotional keywords, and in sentence (2), "春天", "孩子们" will be labeled. Some emotional keywords may not bear emotions by themselves, but in a given context, they express emotions, and in similar contexts, they may express similar emotions.

The difference between direct affective words and indirect affective words is reflected by their emotional intensities. An emotional keyword or phrase is represented as a vector to record its intensities of the eight basic emotional classes (expect, joy, love, surprise, anxiety, sorrow, angry and hate). For direct affective words, we annotate the emotions of theirselves, for instance, the vector for the word "喜爱" $\vec{w}=(0.0,0.8,0.8,0.0,0.0,0.0,0.0,0.0)$. For indirect affective words, we annotate their emotional vectors according to their contexts, for example, the vector for the word "任何人" in sentence (1) $\vec{w}=(0.0,0.0,0.0,0.0,0.0,0.2,0.2,0.1)$, the vector for the word "春天" in sentence (2) $\vec{w}=(0.1,0.3,0.3,0.0,0.0,0.0,0.0,0.0)$.

Emotional phrases are combination of words, such as Chinese proverbs, like "世上无难事，只要肯攀登 (shi shang wu nan shi, zhi yao ken pan deng; English: Where there is a will, there is a way) ". For an emotional phrase, the positions of its first character and its last character in a sentence are labelled, and also for emotional keywords if there are Chinese word segmentation mistakes.

4.2.3 Degree words

Degree words are associated with the intensities of emotions. Obviously, the love intensity of the sentence "我非常喜欢这个老师。(wo fei chang xi huan zhe ge lao shi; English: I like this teacher very much.)" is different from the sentence "我喜欢这个老师。(wo xi huan zhe ge lao shi; English: I like this teacher.)" . The former sentence have a degree word "非常 (fei chang; English: very much") enhance love intensity. In Chinese, degree words appear with high frequency.

Degree words and the modifying contents are all labeled. In the above example, "喜欢(xi huan ; English: like)" is annotated as modifying content of degree word "非常 (fei chang; English: very much").

4.2.4 Negative words

In Chinese, negative words can be placed almost everywhere in a sentence to change the meaning of this sentence, also to change its emotions. For example (The negative words are underlined.),

(3) 我喜欢这个老师。 (Wo xi huan zhe ge lao shi; English: I like this teacher.)

(4) 我不喜欢这个老师。 (Wo bu xi huan zhe ge lao shi; English: I don't like this teacher.)

(5) 不是我喜欢这个老师。 (Bu shi wo xi huan zhe ge lao shi; English: It is not I that like this teacher.)

(6) 我喜欢的不是这个老师。 (Wo xi huan de bu shi zhe ge lao shi; English: This is not the teacher that I like.)

(7) 我不是不喜欢这个老师。 (Wo bu shi bu xi huan zhe ge lao shi; English: I like this teacher, but ....)

Sentence (4)-(6) make sentence (3) negative by negative words "不", "不是", "不是" separately, sentence (7) uses double negative to resolve to positive, but the meaning is different from sentence

(3). There are more information about the writer's feeling want to be expressed in sentence (7). Negative words and the content that they modify are all labeled.

Negative words have been frequently used in Chinese. The use of such negative modifying emotional keyword or phrase change completely the original emotion of them. Analysis on negative words will be quite helpful to an accurate emotion analysis.

4.2.5 Conjunctions

There are numerous researches about Chinese conjunctions, in our annotation scheme, we focus on those conjunctions which may influence emotional expressions. For example (The conjunctions are underlined.),

(8) 尽管我们喜欢这个老师，但她已经离开了我们。 (Jin guan wo men xi huan zhe ge lao shi, dan ta yi jing li kai le wo men; English: Despite we like this teacher, she has leaved.)

(9) 我不仅喜欢她的外在美，更喜欢她的内在美。(Wo bu jin xi huan ta de wai zai mei, geng xi huan ta de nei zai mei; English: I like her outer beauty, and I more like her inner beauty.)

Sentence (8) uses the conjuctions "尽管...但...(jin guan...dan...)" express emotions of love and sorrow. Sentence (9) uses the conjuctions "不仅...更...(bu jin...geng...)" enhance love intensity.

Conjunctions and the modifying contents are all labeled. If conjunctions appear in pairs in a sentence, the position of pairing words for each conjunction are also labeled. For the above example (8), conjunctions are annotated as follows (Fig. 3).

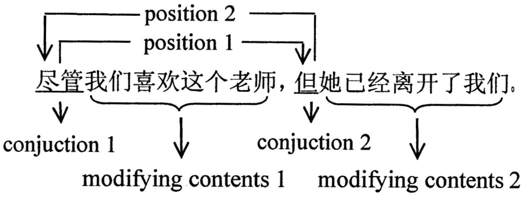


Fig. 3 An example of conjunctions annotation

4.2.6 Rhetoric

Chinese rhetoric has been well studied from the view of linguistics and literature, and the definition of rhetoric category is quite different. We select nine common rhetoric categories to annotate: 比喻 (bi yu. English: metaphor), 借代 (jie dai. English: metonymy), 夸张 (kua zhang. English: exaggeration), 拟人 (ni ren. English: personification), 对偶 (dui ou. English: antithesis or parallel), 排比 (pai bi. English: parallelism sentence), 设问 (se wen. English: rhetorical question), 反问 (fan wen. English: rhetorical question), 重复 (chong fu. English: repeat). We annotate rhetoric category and the corresponding emotion category.

4.2.7 Punctuation

Punctuation is the use of standard marks and signs in writing to separate words into sentences, clauses, and phrases in order to clarify meaning. Some punctuation marks can express our emotion, for example, an exclamation mark (!) is used at the end of a sentence (which may be exclamative, imperative or declarative), or a question mark (?) to show strong emotion. [11] suggest's that people relied on four strategies including punctuation to express happiness versus sadness. Punctuation effect is also shown in [12] to extend to emoticon placement in website text messages.

At sentence level, we annotate punctuation with emotion and the corresponding emotion category.

4.2.8 Objective and subjective

Distinguishing between factual and subjective information could support for many natural language processing applications. Objective and subjective in our annotation scheme is to distinguish between writer's emotion and non-writer's emotion.

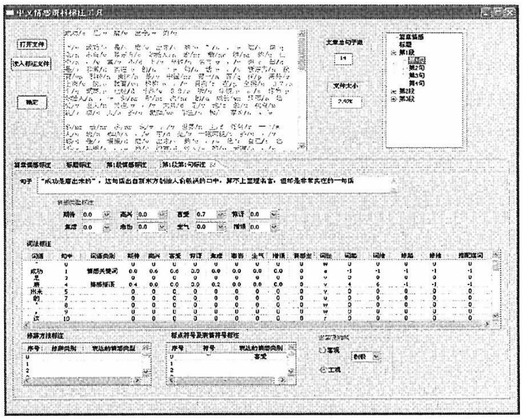
4.2.9 Emotion polarity

There is a positive side or a negative side on emotion. We call this an emotional polarity. We select eight emotion classes (expect, joy, love, surprise, anxiety, sorrow, angry and hate) for this corpus annotation. In most cases, expect, joy, love are positive emotions, while sorrow, angry and hate are negative emotions, the polarities of surprise and anxiety can be positive or negative in different contexts. Emotion polarity of a sentence is determined by integrating its emotions. A sentence without emotion is annotated with neutral.

5 Conclusions

In this study we proposed an emotional expression space model, which is hierarchical in consistent with the natural structure of a document. Based on this model, we described a relatively fine-grained annotation scheme, annotating emotion in text at three levels: document, paragraph, and sentence. In document and paragraph levels, emotion category, emotion intensity, topic keywords and topic sentences are annotated. In sentence level, annotation includes emotion category, emotion intensity, emotional keyword/phrase, degree word, negative word, conjunction, rhetoric, punctuation, objective/subjective, and emotion polarity.

The dataset have consisted of 198 blog articles published at sina blog, sciencenet blog, baidu blog, qzone blog, qq blog, and other blog websites. There are 5608 sentences, 135,606 Chinese words in this corpus. Relative language resources have been extracted from it, such as word lists, sentence lists, but the size seems not enough for large scale textual emotional analysis, a lot of linguistic features are not reflected from it. So more annotation is still ongoing, the number of annotators has increased to ten. The tool for the annotation is written in Java (Fig. 4 is the interface of this annotation tool), and input files are text files with Chinese segmentation and part-of-speech tags [13, 14], the annotated output files are organized in XML documents.



句子 “成功是磨出来的”，这句话出自新东方创始人俞敏洪的口中，算不上至理名言，但却是非常实在的一句话

期待 0.0 高兴 0.0 喜爱 0.7 惊讶 0.0
焦虑 0.0 悲伤 0.0 生气 0.0 憎恨 0.0

Fig. 4 The interface of this annotation tool

Acknowledgements

We are grateful to our annotators: Huana Li, Ye Wu, Lei Chen, Yu Zhang, and Ji Li, and other participants and supporters. This research has been partially supported by Ministry of Education, Science, Sprots and Culture, Grant-in-Aid for Scientific Research (B), 19300029.

REFERENCES

[1] Rosalind Picard: *Affective Computing*, The MIT Press, MA, USA, (1997)

[2] Mishne, G.: *Experiments with Mood Classification in Blog Posts*, Proc. Style2005 in SIGIR'05, (2005)

[3] R Mihalcea, H Liu: *A corpus-based approach to finding happiness*, Proceedings of the AAAI Spring Symposium on Computational, (2006)

[4] De Rijke, M., Mishne, G. A. : *Capturing global mood levels using blog posts* AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs, pp.145-152, (2006)

[5] Yuchul Jung, Hogun Park, Sung Hyon Myaeng: *A Hybrid Mood Classification Approach for Blog Text*, Lecture Notes in Computer Science, pp.1099-1103, (2006)

[6] Yuchul Jung, Yoonjung Choi, Sung-Hyon

Myaeng, *Determining Mood for a Blog by Combining Multiple Sources of Evidence.* Web Intelligence, IEEE/WIC/ACM International Conference on Volume, Issue, 2-5, pp.271-274, (2007)

[7] K Balog, G Mishne, M de Rijke: *Why are they excited? identifying and explaining spikes in blog mood levels*, Proceedings 11th Meeting of the European Chapter of the of the Association for Computational Linguistics, (2006)

[8] Gilly Leshed, Joseph 'Jofish' Kaye: *Understanding how bloggers feel: recognizing affect in blog posts.* Conference on Human Factors in Computing Systems CHI '06 extended abstracts on Human factors in computing systems. pp.1019-1024, (2006)

[9] KHY Lin, C Yang, HH Chen: *What emotions do news articles trigger in their readers?* Annual ACM Conference on Research and Development in Information Retrieval, pp.733- 734, (2007)

[10] C Yang, KHY Lin, HH Chen: *Building Emotion Lexicon from Weblog Corpora,* Proceedings of the ACL 2007 Demo and Poster Sessions, pp. 133–136, (2007)

[11] JT Hancock, C Landrigan, C Silver: *Expressing emotion in text-based communication,* Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 929 - 932, (2007)

[12] RR Provine, RJ Spencer, DL Mandell: *Emotional Expression Online: Emoticons Punctuate Website Text Messages.* Journal of Language and Social Psychology, Vol. 26, No. 3, pp. 299-307, (2007)

[13] Degen Huang, Xiao Sun, Shidou Jiao, Lishuang Li, Zhuoye Ding: *HMM and CRF Based Hybrid Model for Chinese Lexical Analysis*, Processing of Sixth SIGHAN Workshop on Chinese Language, pp. 133-137, (2008)

[14] Ying Qin, Caixia Yuan, Jiashen Sun, Xiaojie Wang: *BUPT Systems in the SIGHAN Bakeoff* , Processing of Sixth SIGHAN Workshop on Chinese Language, pp.94-97, (2008)