Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 24 (2010) 726–749
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

A blog emotion corpus for emotional expression analysis in Chinese

Changqin Quan *, Fuji Ren

Faculty of Engineering, University of Tokushima, 2-1 Minamijosanjima, Tokushima 770-8506, Japan
Department of Computer Science, Huazhong Normal University, Wuhan, China

Received 6 February 2009; received in revised form 15 January 2010; accepted 16 February 2010
Available online 23 February 2010

Abstract

Weblogs are increasingly popular modes of communication and they are frequently used as mediums for emotional expression in the ever changing online world. This work uses blogs as object and data source for Chinese emotional expression analysis. First, a textual emotional expression space model is described, and based on this model, a relatively fine-grained annotation scheme is proposed for manual annotation of an emotion corpus. In document and paragraph levels, emotion category, emotion intensity, topic word and topic sentence are annotated. In sentence level, emotion category, emotion intensity, emotional keyword and phrase, degree word, negative word, conjunction, rhetoric, punctuation, objective or subjective, and emotion polarity are annotated. Then, using this corpus, we explore these linguistic expressions that indicate emotion in Chinese, and present a detailed data analysis on them, involving mixed emotions, independent emotion, emotion transfer, and analysis on words and rhetorics for emotional expression.

Keywords: Emotion analysis; Weblogs; Corpus annotation; Natural language processing

1. Introduction

Emotions play important role in human intelligence, rational decision making, social interaction, perception, memory, learning, creativity, and more (Picard, 1997). There is plenty of evidence that emotion analysis has many valuable applications.

In everyday life people express their emotions through multiple modalities: their linguistic contents, speech, faces and their bodies. All of the cues can be used to convey emotional messages. 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. Especially, blogspace consists of millions of users who maintain online diaries, containing frequently-updated views and personal remarks about a range of issues (Mishne, 2005). Textual emotion analysis also can reinforce the accuracy

* Corresponding author. Tel.: +81 88 656 9684; fax: +81 88 656 6575.
E-mail addresses: quan-c@is.tokushima-u.ac.jp (C. Quan), ren@is.tokushima-u.ac.jp (F. Ren).

0885-2308/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2010.02.002

of sensing in other modalities like speech or facial recognition, and to improve human computer interaction systems. Werry (1996) points out that in Internet relay chat (IRC), linguistic strategies have been adopted to replace the missing intonational and paralinguistic cues of face-to-face discourse (Werry, 1996). This finding is reflected in the use of coordination devices in Hancock and Dunham's (2001) study of computer-mediated task-based interactions (Hancock and Dunham, 2001).

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. Classifying the mood of a single text is a hard task; state-of-the-art methods in text classification achieve only modest performance in this domain (Mishne, 2005). In this area, some of the hardest problems involve acquiring basic resources. Corpora are fundamental both for developing sound conceptual analysis and for training the 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 an emotional expression space model in text, and describe 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 words 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. We explore all of these linguistic expressions that indicate emotion in Chinese, and present a detailed data analysis on them, involving mixed emotions, independent emotion, emotion transfer, POS (part-of-speech) of emotional keywords, multiple emotional keywords and phrases and rhetorics for emotional expression. The annotation scheme has been employed in the manual annotation of a corpus containing 500 documents, with 4004 paragraphs, 12,742 sentences, and 324,571 Chinese words.

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 presents the inter-annotator agreement study. Section 6 describes data analysis on Chinese emotional expression. Section 7 is the discussions. Section 8 concludes this study with closing remarks and future directions.

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 (Mishne, 2005). Livejournal also used as data source for finding happiness (Mihalcea and Liu, 2006), capturing global mood levels (Mishne and Rijke, 2006), classifying mood (Jung et al., 2006; Jung et al., 2007), discovering mood irregularities (Balog et al., 2006), recognizing affect (Leshed and Kaye, 2006). A similar emotion corpus in Chinese is Yahoo!'s Chinese news (http://tw.news.yahoo.com), which is used for Chinese emotion classification of news readers (Lin et al., 2007) and emotion lexicon building (Yang et al., 2007). More and more weblogs have added mood column to record blog users' moods when they read or write blogs. Two merits let them well accepted as emotion corpora: a large number of weblogs contained and moods annotated by blog users. However, there is a great inconsistency on emotion categories given by different websites. Livejournal gives a predefined list of 132 common moods, while Yahoo!'s Chinese news provides readers 8 emotion categories. Too many mood classes may confuse users, and Mishne also pointed out one obvious drawback of the mood "annotation" in this corpora is that they are not provided in a consistent manner; the blog writers differ greatly from each other, and their definitions of moods differ accordingly. What may seem to one person as a frustrated state of mind might appear to another as a different emotional state – anger, depression, and so on (Mishne, 2005). In addition, some words are not fitted to be taken as an emotion class, such as "useful" in Yahoo!'s emotion categories. These corpora may be 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 is not sufficient. A text-based emotion prediction system would benefit from identifying the emotional affinity of sentences. The emotion analysis on