文章编号：1003-0077(2014)05-0083-09

面向微博文本的情绪标注语料库构建

姚源林[1]，王树伟[1]，徐睿峰[1]，刘 滨[1]，桂 林[1]，陆 勤[2]，王晓龙[1]

（1. 哈尔滨工业大学 深圳研究生院，广东 深圳 518055；
2. 香港理工大学 电子计算学系，香港 九龙）

摘 要：文本情绪分析研究近年来发展迅速，但相关的中文情绪语料库，特别是面向微博文本的语料库构建尚不完善。为了对微博文本情绪表达特点进行分析以及对情绪分析算法性能进行评估，该文在对微博文本情绪表达特点进行深入观察和分析的基础上，设计了一套完整的情绪标注规范。遵循这一规范，首先对微博文本进行了微博级情绪标注，对微博是否包含情绪及有情绪微博所包含的情绪类别进行多标签标注。而后，对微博中的句子进行有无情绪及情绪类别进行标注，并标注了各情绪类别对应的强度。目前，已完成 14 000 条微博，45 431 句子的情绪标注语料库构建。应用该语料库组织了 NLP&CC2013 中文微博情绪分析评测，有力地促进了微博情绪分析相关研究。

关键词：情绪语料库；语料库构建；情绪标注；微博文本

中图分类号：TP391　　**文献标识码**：A

The Construction of an Emotion Annotated Corpus on Microblog Text

YAO Yuanlin[1], WANG Shuwei[1], XU Ruifeng[1], LIU Bin[1], GUI Lin[1], LU Qin[2], WANG Xiaolong[1]

(1. Harbin Institute of Technology, Shenzhen Graduate School, Shenzhen, Guangdong 518055;
2. Department of Computing, The Hong Kong Polytechnic University, Kowloon, HongKong)

Abstract: The research on text emotion analysis has made substantial progesses in recent years. However, the emotion annotated corpus is less developed, especially the ones on micro-blog text. To support the analysis on the emotion expression in Chinese micro-blog text and the evaluation of the emotion classification algorithms, an emotion annotated corpus on Chinese micro-blog text is designed and constructed. Based on the observation and analysis on the emotion expression in micro-blog text, a set of emotion annotation specification is developed. Following this specification, the emotion annotation on micro-blog level is firstly performed. The annotated information includes whether the micro-blog text has emotion expression and the emotion categories corresponding to the micro-blog with emotion expressions. Next, the sentence-level annotation is conducted. Meanwhile, the annotation on whether the sentence has emotion expression and the emotion categories, the strength corresponding to each emotion category is annotated. Currently, this emotion annotated corpus consists of 14 000 micro-blogs, totaling 45 431 sentences. This corpus was used as the standard resource in the NLP&CC2013 Chinese micro-blog emotion analysis evaluation, facilitating the research on emotion analysis to a great extent.

Key words: emotion corpus; corpus construction; emotion annotation; micro-blog text

1 引言

文本情绪的识别与分类在文本倾向性分析、舆情分析、事件预测等领域都有着广泛的应用。其识别过程中涉及到了情绪心理学、认知心理学、生活常识、舆论导向等诸多因素，加之机器学习、统计方法等不同的研究手段，正使得文本情绪计算成为自然

收稿日期：2014-06-25　**定稿日期**：2014-08-11

基金项目：国家自然科学基金(61203378, 61300112, 61370165)；高等院校博士学科点专项基金(20122302120 070)；广东省自然科学基金(S2012040007390, S2013010014475)；模式识别国家重点实验室开放课题基金；深圳市基础研究计划(JCYJ20120613152557576, JC201005260118A)；深圳市国际合作计划(GJHZ201206131 106 1217)，百度高校合作项目

语言处理领域的新热点。作为相关研究的基础，遵循统一的标注规范下标注的情绪语料库对具体的语言现象分析以及情绪分类算法的设计和评估都有重要意义。

目前，国内外在情绪标注语料库的构建上取得了一定的进展。Mishne 利用 LiveJournal 博客系统中作者自行标注发布博客时的情绪信息，构建了一个包含 815 494 篇博客的英文情绪标注语料库[1]。该语料库标注了 132 种情绪类别，例如，开心、生气等。Ptaszynski 等人对 50 亿字的日语博客进行了情绪标注[2]。该语料库采用了 10 种情绪类别标注，此外还标注了情绪符号、情感极性等。在中文情绪语料库的构建方面，Quan C. 等人提出了一套细粒度的文本情绪标注方案，该方案采用 8 种基本情绪类别，对 1 487 篇博客进行文档级、段落级以及句子级三个层次的情绪标注[3]。徐琳宏等在小学教材（人教版）、电影剧本、童话故事、文学期刊等语料上进行了句子级别的情绪标注，采用了 7 大类，22 小类的情绪分类体系，完成近 4 万句，100 万字的语料标注[4]。相对于情绪标注语料库，情感倾向性标注语料库的构建则相对较为成熟。Xu R. F 等人针对中文产品评价中倾向性表达特点，设计了一套细粒度倾向性标注方案，分别在词语级、句子级和文档级进行标注。对于每一个倾向性评价，分别标注了观点表达及其对应的产品属性。同时，引入领域本体对评价目标属性进行了概念化规约[5]。Pak A. 等人利用来源于推特（Twitter）的微博建立一个包含了正负面情感的主观文本语料库[6]。

目前情绪标注语料库构建在国内外取得了一定进展，但中文微博文本情绪语料库构建仍处于初级阶段。由于微博文本长度较短，表达较为口语化，网络用语较多，与博客等长文本的情绪表达方式有着较大的差异，导致现有的面向长文本的情绪标注规范不完全适应微博情绪标注的需要。因此，结合微博文本特点设计情绪标注规范，并构建面向微博短文本的情绪标注语料库是十分必要的。

本文选取新浪微博文本作为基础语料进行标注。相较于其他语料库，本语料在选取时充分考虑了中文微博文本的结构、语法和表达特点，诸如表达口语化、情绪多样化、情绪转移多、事件及领域覆盖面广等，从而以符合日常人们表达习惯的特点出发选取数据并建立标注语料库。标注过程中，首先在微博级和句子级上对有无情绪进行判别，然后对有情绪的微博和句子进行 7 种情绪类别的标注，包括快乐、喜好、愤怒、悲伤、恐惧、厌恶、惊讶。此外，在句子级别上增加了包含 3 个情绪强度等级的标注。为保持标注结果的准确性及一致性，建立了相关的评价方法和标注流程管理。目前，该语料库已完成 14 000 条微博、45 431 个句子的情绪标注。其中，有情绪微博 7 407 条，无情绪微博 6 593 条，其中包含有情绪句子 15 688 条，无情绪句子 29 733 条。本语料库为相关科研工作人员分析微博文本的情绪表达特点提供了支持。应用该语料库组织了 NLP&CC2013 中文微博情绪分析评测任务，有效促进了相关领域的研究。

本文组织结构如下：第 2 节介绍微博情绪语料库标注规范；第 3 节介绍语料库构建方法；第 4 节对已构建情绪语料库进行了数据统计以及标注一致性分析。第 5 节简单介绍了应用该语料库组织 NLP&CC2013 的中文微博情绪识别任务评测的情况。第 6 节给出本文结论。

2 微博情绪语料库标注规范

2.1 原始语料选择

本文选取新浪微博文本作为原始标注语料。相对于其他语料库，本语料文本的选择原则是领域无关，事件分布面广。在选取的过程中，从 2011 年至 2012 年共 24 个月的上亿条数据中进行随机选取，同时每个月选取的微博数量大致相同。在对长度较短、含有不规则字符或纯转发的低质量微博过滤后，最终留下格式较为规范的微博作为原始标注语料。

2.2 标注粒度

微博作者要在简短的文字中表达情绪或观点，往往会出现不规范的句子表达以及较为密集的情绪分布，所以相对细致的标注粒度很有必要。为此，本文将情绪标注的粒度划分为微博级和句子级。微博级的标注从微博整体角度出发，标注了微博作者所表达的情绪，而句子级的情绪标注则从微博中每一个句子的角度出发，对作者所表达的情绪进行标注。

2.3 情绪分类体系

目前现有的情绪分类体系存在着不一致的情况，这是由于心理学界对情绪的划分还没有一个公

认标准。较为常用且适合文本情绪分类研究的分类体系是大连理工大学林鸿飞教授提出的中文情感词汇本体[7]。该分类体系是在 Ekman 的 6 大类情绪分类体系，在 6 种情绪类别（“愤怒”、“厌恶”、“恐惧”、“高兴”、“悲伤”、“惊讶”）的基础上，增加了情绪类别“喜好”，对正面情绪进行了更细致的划分。本文采用该方案提出的 7 类情绪体系。

2.4　多标签标注

现有的情绪标注语料库中大多采用单标签情绪标注，也就是认为每一个标注文本对象只包含唯一的情绪类别。但是，在实际表达中，同一条文本作者往往会同时表达多重的情绪，如例 1 所示。

例 1　*“清明节放三天假，但是老师布置了比平时还多的作业，我真是悲喜交加啊。”*

在例 1 中“*悲喜交加*”不仅表达了作者“高兴”的情绪，同时也表达了“悲伤”的情绪。

经过对 500 条抽样微博进行情绪表达统计发现，在有情绪的微博中，仅包含一种情绪的微博占到近 80%，有两种情绪的占到 17%，三种及以上情绪的则只有很小的比例。为此，在标注方案中对微博文本进行了多标签情绪标注。具体的，对每一标注文本标注至多两种情绪，其中一种为主要情绪，一种为次要情绪。主要情绪和次要情绪划分主要遵循如下方法，即首先明确微博或句子所包含的所有种类的情绪，然后对这些情绪在该微博或句子中的强弱程度进行排序，取最强的情绪作为主要情绪，若包含多个情绪，取次强的情绪作为次要情绪。

2.5　情绪强度标注

文本中包含的情绪往往在强度上有很大的差异。如下面两个例子。

例 2　*“这令我伤心欲绝。”*

例 3　*“这令我心情不悦。”*

例 2 和例 3 都表达了“悲伤”的情绪在内，但是“*伤心欲绝*”要比“*心情不悦*”悲伤的强度更大。因此，有必要在情绪类别标注的基础上标注情绪表达强度。

为了更好的体现句子中主次要情绪的纵向对比和微博中句子间情绪的横向对比，标注规范中要求对每个情绪句进行了三个强度等级的标注。分别由 3、2、1 代表强、中、弱。每个情绪最终的强度值通过对多人标注的强度的平均值获得。

3　情绪语料库构建

3.1　微博文本预处理

由于微博的表达方式较为随意，有一些微博不适宜作为最终使用语料，因此在标注前要对微博进行数据筛选。筛选过程分为两个步骤：首先对过短的微博进行筛除，包括单纯转发或@、仅有“转发微博”字样、纯表情符或者标点符号、或字数少于 5 个字的微博，原因在于这些微博对于情绪表达研究意义不大。此外还去除非普通话微博（粤语、英语、日语等）以及各种其他类型怪异微博，如字符画等。

在对微博文本进行情绪标注之前，需要对微博进行分句。分句时原则上按照如下规则：

1）括号及其之内的文本不单独成句。

2）对于较长的句子且仅用空格做分隔符的，用空格作为分句依据。

3）纯标点符号不算做一句话，如全是叹号，尽管表达了一定的情绪，但不作为独立句。

4）因为是中文语料库，所以英文不作为单独的句子，但可以作为一句话中的子句。

在分句阶段，由于微博文本不同于格式规整的新闻文本，很多情况下都是发布者随意发布、格式不一，通过机器提取的规则不足以覆盖所有的微博分句，故需要人工干预分句，以确保准确度。

3.2　标注规则

3.2.1　情绪有无及主客观评价的区分

情绪按照持有者角色属性来说，共分为 4 类，即发出评价者的情绪，所属者或被描述者的情绪，动作、评价、事件、状态受体的情绪，旁观者或者读者的情绪[8]。在本标注体系中，仅考虑的是微博发出者的情绪状态，因而主要从第一类，即发出评价者或微博作者的情绪的角度进行标注。

对事物的评价分为客观评价和主观评价两种。客观评价对客观存在的一种描述，而非表达自己的情绪，所以本研究认为客观评价类的微博是没有情绪的。相反，主观评价类微博是有情绪的，部分主观评价与客观评价比较难判别彼此，通过抽样统计发现，如果形容词前面有程度词或副词修饰的话，则会具有较大的概率被认为这是一种主观评价，如下面三个例子。

例 4　*“宫殿是帝王朝会和居住的地方，规模宏*

大,形象壮丽,格局严谨。”

例5 *“她的咸蛋酥,年糕椰蓉酥,叉烧酥都很不错,超赞哦。”*

例6 *“她看到了这里的风景后,高兴的大呼起来,非常激动。”*

在例4中出现的形如“*规模宏大*、*形象壮丽*”等正面评价词语都是对宫殿的客观、严谨和正式的描述,没有个人情绪蕴含在内,所以不作为情绪句。在例5中则出现了“*不错*”,“*超赞*”这类褒奖词语,含有主观评价的成分在内,则认为是情绪句。而在例6中尽管有着非常明显的情绪表达,但是这个情绪不属于微博作者,而属于对于人物或事件的陈述,在本研究中视为无情绪。

3.2.2 微博整体情绪和句子情绪的关系

通常一条微博由若干句子组成,对应的情绪分布往往有两种情况。即集中分布在一个句子上或散列分布在若干句子上。由此我们也能发现,如果微博整体有情绪的话,微博句子中至少有一个是有情绪。多个有情绪的句子之间可以极性相反,这也符合汉语日常表达,但同样要遵循一个为主要情绪一个为次要情绪。考虑到微博存在转发以及非原创的情况,微博整体无情绪的情况下,允许作为转发或者引用的句子有情绪。

在一条微博中,往往会出现情绪变化的情况,特别是情绪正负极性的变化。例如,

例7 *“不过真好听,一水即兴的 solo,真比原版好听。可惜了,可惜老天不作美”*

例7中第一句有着明显的“喜好”的情绪在内,在第二句则转为了“悲伤”的情绪。对于此类情况,在标注过程中按照其实际情绪进行标注,整体情绪按照微博最大的情绪倾向性标注。确定最大的倾向性首先利用转折关系、篇幅长度来明确微博所要讲述的中心事件,然后确定中心事件的情绪及其强度并作为最终的最大倾向性。所以在标注的过程中,由转折关系可知例7中整体中心事件为“*天公不作美*”,主要情绪为“悲伤”,次要情绪为“喜好”。

3.2.3 反讽,反语情况的标注

反语,反讽的微博文本无论在人工标注还是在机器分类中都有着标准不统一的情况,本文中遵循以下的原则约定反讽的概念。

1) 言非所指。即实际内涵与表面意义相互矛盾。

2) 鲜明性。要避免含糊,具有明确的反语,反讽的倾向性。

3) 按赵毅衡先生在《新批评》一书中的划分,“反讽”分为“克制叙述”、“夸大叙述”、“正话反说”、“疑问式反讽”、“复义反讽”、“悖论反讽”、“浪漫反讽”和人物主题与语言风格上的“宏观反讽”等。

在语料的标注过程中,对于符合上述反讽、反语的微博语句,按照其蕴含的真正的内在情绪并结合上下文语境、常识进行标注。如例8所示。

例8 *“结构主义,我们中国太缺乏这样的思想人士了。还有穷举法,这方法,懂得使用的人太少了。我们国民都太聪明了。所以,各种法规政策总是顾头未顾尾,漏洞百出,而且还死不悔改!”*

在例8中,“*太聪明了*”实际上并不是一个赞扬的语气,作者在此使用了反讽,故而认定为蕴含“厌恶”的情绪在内。

对于不能确定是否为反讽、反语的其他情况则按照文本表面含义进行标注。

3.2.4 表情符的处理

表情符在微博情绪的表达中占有着重要的作用,但在数据的抽样考察中我们发现表情符的使用经常出现字面含义与语境意义不同的情况,例如,在表达特别高兴的情绪的时候,有的微博使用“*[大哭]*”,有的使用“*[大笑]*”等不同情绪极性的表情符。

例9 *“哈哈,我已笑哭…大家走过路过,千万不要错过啊!看看人家多斯文[大哭][大哭]用语多文明[大哭][大哭][大哭]就是靠这样来拉粉的。”*

在例9中,微博自身主要情绪为喜好,次要情绪为高兴,但是在表情的选择时使用了大哭的表情符,借以表达一种强烈的喜好和高兴的情绪,这是微博这一类文本中特有的语言现象,具有一定的情感增强的作用。故而在标注的时候,不能直接使用表情符作为文本情绪的类别,而必须是作为情绪判断的参考,通过上下文的理解确定最终标注的情绪类别。

3.3 标注一致性控制

语料库构建中共有4名标注人员,在标注前进行了统一的培训,但是由于不同人对同一件事物的理解不同,标注结果的差异性很难避免。为了尽量减少标注的不一致,按照如下方式,在三个阶段中进行一致性控制。

1) 将未标注文本分为4份,每人标注一份。

2) 将标注结果随机转至另一名标注人员进行复标,同时记录标注结果不同的文本数量。

3) 将四份文本合并后打乱顺序,再分为两份,每份交予两名标注人员同时讨论复标。

通过如上的方法,保证了每个微博均被标注3

次，同时最少被两名不同人员标注，且该情况下会在步骤 3）双人标注结果比较中再次得到统一，由此可以使误标注数量尽量降到最低。出现 3 次标注均不相同的情况概率很小，如果发现这种情况，则由 4 名标注人员共同讨论并确定最终结果。标注结束后，利用 Kappa 值作为一致性指标的度量。

3.4　完整标注示例

标注完成后，以 XML 格式存储，图 1 显示了一条有情绪微博的标注示例。

```
<weibo id="77" emotion-type1="anger" emotion-type2="disgust">
  <sentence id="1" opinionated="Y" emotion-1-type="anger" emotion-1-weight="3"
emotion-2-type="disgust" emotion-2-weight="2">日本的政府丧心病狂！ </sentence>
  <sentence id="2" opinionated="N">面积小，多地震，核电站的密度世界第一，其数量世界第三。
</sentence>
  <sentence id="3" opinionated="N">美国平均 9 万 km² 有 1 座，日本 Q 67 万 km² 有 1 座，这个密度
相当于北京一地有 2 座，香港 6 座，广东一省 26 座，中国 1 400 座！ </sentence>
  <sentence id="4" opinionated="Y" emotion-1-type="anger" emotion-1-weight="2"
emotion-2-type="none">地震是天灾，核泄漏是人祸，政府别狡辩，不修就不会泄露！ </sentence>
  <sentence id="5" opinionated="Y" emotion-1-type="anger" emotion-1-weight="3"
emotion-2-type="disgust" emotion-2-weight="2">广岛长崎死伤 40 万，老百姓勤奋忍耐，日本政府王八
蛋</sentence>
</weibo>
```

图 1　有情绪微博标注示例及存储格式

4　语料库标注结果分析

目前语料库构建已完成 14 000 条微博、45 431 个句子的情绪标注。在此基础上，对微博情绪表达的语言现象和语言规律进行了一系列的统计和分析。

4.1　情绪占比统计

表 1 和表 2 分别是对微博级和句子级有无情绪的数量统计。

表 1　微博级有无情绪比例

	有情绪微博		无情绪微博
	仅主要情绪	含主次要情绪	
数量	5 414	1 993	6 593
占比/%	38.67	14.24	47.09
合计/%	52.91		47.09

表 2　句子级有无情绪比例

	有情绪句子		无情绪句子
	仅主要情绪	含主次要情绪	
数量	14 032	1 656	29 733
占比/%	30.89	3.65	65.46
合计/%	34.54		65.46

从统计中可以看出，有情绪的微博和无情绪的微博在微博级上比例大致相同。而在句子级别上，在句子级上有情绪和无情绪的比例大致为 1∶2，这与先期对微博原始语料进行抽样统计得到的情况基本符合。

本语料对于微博级和句子级都实现了多情绪标注，表 3 及表 4 是对有情绪的微博及句子进行的数量统计。

表 3　有情绪微博中各情绪分布情况

	微博主要情绪	微博次要情绪
高兴	1 460	359
喜好	2 203	546
愤怒	669	203
悲伤	1 173	269
恐惧	148	61
厌恶	1 392	385
惊讶	362	170
合计	7 407	1 993

从表 3 和表 4 中可以看出，无论是有情绪微博还是有情绪句子中，各类别情绪的分布都有所差异，其中，“喜好”类所占的比例最大，而“恐惧”、“惊讶”类所占的比例则较小。

表4 有情绪句子中各类情绪分布情况

	句子主要情绪	句子次要情绪
高兴	2 804	258
喜好	4 258	373
愤怒	1 899	255
悲伤	2 477	150
恐惧	299	37
厌恶	3 131	497
惊讶	820	86
合计	15 688	1 656

4.2 情绪伴随统计

通过对标注数据的分析我们发现,在同一条微博或句子中,当一种情绪出现后,往往有些其他的情绪伴随出现。例如,出现"喜好"的时候,"高兴"也会有很大的概率随之出现。同一个微博或句子中,每种可能出现的主要、次要情绪的组合,称之为一种情绪的伴随,取值范围为7种基本情绪或无情绪的组合。同时情绪伴随是一个有序的组合,即{高兴,喜好}和{喜好,高兴}是不同的情绪伴随。理论上有情绪微博最多含有49种情绪伴随可能。

我们对所有含有两种情绪的情绪表达进行统计,利用条件概率公式计算伴随情绪的出现概率。

$$P(\text{Emotion2}|\text{Emotion1})=\frac{\#\{\text{Emotion1},\text{Emotion2}\}}{\#\{\text{Emotion1}\}} \quad (1)$$

式(1)中,Emotion1表示主要情绪,Emotion2表示次要情绪。

表5显示了利用式(1)进行的微博级情绪伴随的统计,表6显示了句子级情绪伴随的统计。

表5 微博级别伴随情绪统计

情绪1/% \ 情绪2/%	高兴	喜好	愤怒	悲伤	恐惧	厌恶	惊讶	无
高兴	——	25.07	0.48	3.08	0.41	1.10	2.05	67.81
喜好	13.75	——	0.50	3.90	0.18	2.09	2.00	77.58
愤怒	0.60	1.64	——	7.32	1.49	35.13	3.14	50.67
悲伤	1.45	7.50	2.98	——	1.71	5.12	2.81	78.43
恐惧	2.70	5.41	0.68	10.14	——	5.41	10.14	65.54
厌恶	1.44	3.81	10.49	4.38	0.93	——	1.94	77.01
惊讶	3.04	5.52	0.83	3.59	2.21	5.52	——	79.28

表6 句子级别伴随情绪统计

情绪1/% \ 情绪2/%	高兴	喜好	愤怒	悲伤	恐惧	厌恶	惊讶	无
高兴	——	10.59	0.04	0.53	0.04	0.14	0.71	87.95
喜好	5.43	——	0.00	0.78	0.05	0.19	0.59	92.98
愤怒	0.05	0.05	——	2.26	0.26	22.64	0.68	74.04
悲伤	0.32	1.70	1.53	——	0.44	1.33	0.44	94.23
恐惧	0.67	1.34	0.33	4.35	——	2.68	3.34	87.29
厌恶	0.06	0.38	6.58	1.02	0.22	——	0.22	91.50
惊讶	1.71	2.07	1.10	1.71	1.34	1.71	——	90.37

通过统计可以看出,无论在句子级别还是微博级别,"高兴"和"喜好"经常同时出现,"愤怒"则经常同"厌恶"伴随出现。其他的情绪之间也有一定的关联。

4.3 情绪转移统计

所谓情绪转移指的是同一条微博中,相邻的两个句子之间的情绪变化。分析邻接句间情绪的转移规律往往也能够更好地促进句子的情绪分类。为

此，我们对微博中句子间情绪的转移进行统计。情绪 a 向情绪 b 的转移概率可以利用式(2)计算得出。

$$P(\text{Emotion}_S = b \mid \text{Emotion}_{S_p} = a) = \frac{count[\text{Emotion}_{S_p} = a, \text{Emotion}_S = b]}{count[\text{Emotion}_{S_p} = a]} \quad (2)$$

式(2)中，Emotion_S 表示句子 S 的情绪；Emotion_{S_p} 表示句子 S 的前一句的情绪。

表 7 显示了 7 种情绪以及无情绪之间的情绪转移概率。

表 7　句子间情绪转移统计

情绪$_{Sp}$/% \ 情绪$_S$/%	高兴	喜好	愤怒	悲伤	恐惧	厌恶	惊讶	无
高兴	47.52	13.26	0.37	2.23	0.19	0.99	1.18	34.26
喜好	9.12	50.47	0.53	3.29	0.16	2.30	0.45	33.68
愤怒	0.73	1.03	69.82	4.19	0.22	6.75	0.73	16.52
悲伤	2.51	6.87	2.84	52.84	0.66	3.04	0.92	30.32
恐惧	3.66	5.76	4.71	7.85	36.65	6.81	3.14	31.41
厌恶	0.99	2.87	5.05	2.72	0.74	61.21	0.89	25.53
惊讶	4.89	7.76	4.55	4.72	2.19	5.56	33.22	37.10
无	3.72	6.02	1.45	3.22	0.43	3.58	1.12	80.46

通过上表可以看出，具有相同极性的情绪转移概率往往大于不同极性的情绪之间的转移概率。例如正面情绪“高兴”到正面情绪“喜好”的转移概率要远大于到负面情绪“愤怒”的转移概率。通过分析情绪的转移规律可以更好地指导结合上下文的情绪分类。

4.4　情绪强度统计

针对每个微博句子的情绪，我们设定了 3 个强度等级，分别使用 3，2，1 表示强中弱不同等级的强度。在实际标注中，句子的第一情绪原则上要强于第二情绪的强度，个别情况下可以相等。

表 8　情绪平均强度统计

	作为主要情绪	作为次要情绪
高兴	2.276 047	1.878 431
喜好	2.224 024	1.966 480
愤怒	2.601 165	2.061 475
悲伤	2.201 052	1.802 721
恐惧	2.250 836	1.685 714
厌恶	2.314 168	2.305 547
惊讶	2.253 366	1.869 048

可以看出，对于“愤怒”等情绪，情绪强度较为强烈。而对另一些情绪，例如“悲伤”，情绪强度则相对弱一些，这也是符合人们情绪的客观情况的。

4.5　标注一致性分析

本研究使用 Kappa 值作为标注一致性的检验标准，分别对微博级情绪有无、微博级情绪类别选择、句子级情绪有无、句子级情绪类别选择、句子级情绪强度进行一致性检验。表 9 显示了以上 6 种情况下的 Kappa 值。

通过表 9 可以看出，在情绪有无方面，各标注者的一致性较高，而在细粒度的情绪类别选择方面，各标注者的一致性相对较差一些。在句子级情绪强度标注中得到的一致度约为 0.646，主要原因是对于同一个情绪句，不同标注者的情绪敏感性不同，但整体上一致性仍然达到了较高的水平。

表 9　标注一致性统计

	Kappa 值
微博级情绪有无	0.867 5
微博级情绪类别选择	0.712 5
句子级情绪有无	0.803 5
句子级情绪类别选择	0.694 7
句子级情绪强度	0.646 0

5　微博情绪分析评测

应用本文建立的微博情绪标注语料库，组织了

NLP&CC2013 中文微博情绪分析评测。其中，选择 4 000 条微博作为训练数据，10 000 条微博作为测试数据。该评测任务中，共有 19 支队伍提交了 58 组有效结果，在这些参赛队伍中提出了很多的新的思想和方法。其中，贺飞燕等人结合 TF-IDF 方法与方差统计方法，提出了一种实现多分类特征抽取的计算方法[9]。采用先进行极性判断，后进行细粒度情绪识别的处理方法，构建细粒度情绪分析与判别流程，并将其应用于微博短文本的细粒度情绪识别。张晶等以情绪因子中的常用情绪词和情绪短语为基础构建情绪词典，并针对特殊的情绪表达式，结合标点符号和表情符号在情绪分析中的功能，建立情绪规则库，然后，通过对情绪词典和情绪规则的匹配和计算，实现对中文微博情绪的识别和分类[10]。欧阳纯萍等人针对中文微博的用户情绪分析问题，提出了一种基于多策略融合的细粒度情绪分析方法，首先采用贝叶斯算法对微博的有无情绪分类，然后构建有情绪微博的 21 维特征向量，最后采用 SVM 和 KNN 算法对微博进行细粒度情绪分类[11]。

本次评测分别进行 Close 封闭资源测试和 Open 开放资源测试。其中 Close 封闭资源测试要求各参评单位只使用组织者提供的词典、分词工具等资源；Open 测试则允许参评单位利用现有语言资源开发和训练系统，并用于测试结果生成。评测分别评估了 Close 和 Open 测试所取得的系统性能。

表 10 和表 11 分别列出了 NLP&CC2013 微博级情绪有无断任务 Close 和 Open 评测中性能较优的部分结果。

表 10 NLP&CC2013 微博级情绪有无 Close 评测部分结果

队伍编号	正确率	召回率	F 值
9	0.749 4	0.635 5	0.686 6
15	0.526 3	0.943 6	0.675 7
19	0.667 6	0.798 2	0.727 1

表 11 NLP&CC2013 微博级情绪有无 Open 评测部分结果

队伍编号	正确率	召回率	F 值
4	0.636 3	0.761 6	0.693 3
8	0.642 0	0.842 2	0.728 6
14	0.665 0	0.710 0	0.686 7

通过表 10 和表 11 可以看出，在情绪有无判断任务中，各队伍都取得了较高的分类性能。同时，由于 Open 评测可以充分利用各种外部资源，分类性能相比 Close 评测略高。

表 12 和表 13 分别列出了 NLP&CC2013 微博级情绪类别识别任务 Close 和 Open 评测中性能较优的部分结果。

表 12 NLP&CC2013 微博级情绪类别识别任务 Close 评测部分结果

队伍编号	正确率	召回率	F 值
13	0.257 2	0.230 9	0.243 4
15	0.216 0	0.281 0	0.244 2
19	0.270 4	0.306 4	0.287 3

表 13 NLP&CC2013 微博级情绪类别识别任务 Open 评测部分结果

队伍编号	正确率	召回率	F 值
4	0.284 2	0.348 0	0.312 9
8	0.258 8	0.296 6	0.259 5
14	0.247 4	0.252 8	0.250 1

通过表 12 和表 13 可以看出，相比情绪有无判断任务，对微博级的情绪类别识别任务分类性能相对较弱。同样，Open 评测要比 Close 评测任务性能略高。

表 14 显示了 NLP&CC2013 句子级情绪类别识别任务 Close 和 Open 评测中性能较优的部分结果。

表 14 NLP&CC2013 句子级情绪识别任务部分结果

队伍编号	Close 评测		Open 评测	
	平均精度（宽松指标）	平均精度（严格指标）	平均精度（宽松指标）	平均精度（严格指标）
6	0.299 6	0.289 2	0.332 4	0.320 8
15	0.332 5	0.320 8	0.365 0	0.348 4
17	0.343 9	0.330 5	0.251 6	0.241 0

通过表 14 可以看出，句子级情绪分类性能相比微博级情绪分类性能有了一定的提高。同时，Open 评测比 Close 评测性能也有一定的提升。

应用面向微博文本的情绪标注语料库所组织的 NLP&CC2013 中文微博情绪分析评测有力地促进了中文微博情绪分析相关研究。

5 结论

本文在对微博情绪表达特点进行观察和分析的

基础上，设计了面向微博文本的情绪标注规范。遵循这一规范，建立了微博文本情绪标注规程以及标注一致性控制方案。本文重点介绍了语料库的构建过程和构建规则。在标注过程中，对微博文本首先进行了微博级情绪标注，对微博是否包含情绪及有情绪微博所包含的情绪类别进行多标签标注。而后，对微博中的句子进行情绪标注，在有无情绪及情绪类别进行标注的基础上，增加了情绪强度的标注。经过了对微博情绪标注方案的不断设计和完善，以及对微博语料的多轮标注，该语料库已完成 14 000 条微博，45 431 句子的情绪标注。在此基础上，对语料库进行一系列的数据统计和分析，有助于发现微博情绪表达的语言现象和语言规律。应用该语料库作为 NLP&CC2013 中文微博情绪分析评测任务标准语料，促进了中文微博情绪分析相关研究。

致谢

本文感谢先后参加语料采集、标注和整理的丘桥云、袁丽、汪奕丁、周继云、王赵煜、孔兵、曹宇慧、王帅等同学的辛勤努力。

参考文献

[1] Mishne G. Experiments with mood classification in blog posts [C]//Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access. 2005

[2] Ptaszynski M, Rzepka R, Araki K, et al. Automatically annotating a five-billion-word corpus of Japanese blogs for sentiment and affect analysis [J]. Computer Speech & Language, 2014, 28(1): 38-55.

[3] Quan C, Ren F. Construction of a blog emotion corpus for Chinese emotional expression analysis [C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009: 1446-1454.

[4] 徐琳宏，林鸿飞，赵晶. 情感语料库的构建和分析[J]. 中文信息学报，2008，22(1)：116-122.

[5] Xu R. F, Xia Y. Q.; Wong K. F. and Li W. J. Opinion Annotation in On-line Chinese Product Reviews [C]//Proceedings of Language Resource and Evaluation Conference 2008.

[6] Pak A. and Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining [C]//Proceedings of Language Resource and Evaluation Conference 2010: 1320-1326 .

[7] 徐琳宏，林鸿飞，潘宇，等. 情感词汇本体的构造[J]. 情报学报，2008，27(2)：180-185.

[8] 徐睿峰，邹承天，郑燕珍，等. 一种基于情绪表达与情绪认知分离的新型情绪词典[J]. 中文信息学报，2013，27(6)：82-90.

[9] 贺飞燕，何炎祥，刘楠，等. 面向微博短文本的细粒度情感特征抽取方法 [J]. 北京大学学报，2014，50(1)：48-54.

[10] 张晶，朱波，梁琳琳，等. 基于情绪因子的中文微博情绪识别与分类 [J]. 北京大学学报，2014，50(1)：79-84.

[11] 欧阳纯萍，阳小华，雷龙艳，多策略中文微博细粒度情绪分析研究 [J]. 北京大学学报，2014，50(1)：67-72.

姚源林(1989—)，硕士研究生，主要研究领域为自然语言处理，文本情绪计算。
E-mail：yuanlin. yao@foxmail. com

徐睿峰(1973—)，博士，副教授，主要研究领域为自然语言处理，文本情绪计算。
E-mail：xuruifeng@hitsz. edu. cn

王树伟(1989—)，硕士研究生，主要研究领域为自然语言处理。
E-mail：wangshuwei@live. cn