

# 基于情感要素的中文观点句抽取与极性判别方法

刘楠<sup>1, 2</sup> 何炎祥<sup>1</sup> 方文其<sup>1</sup> 彭敏<sup>1, †</sup> 刘健博<sup>1</sup> 黄佳佳<sup>1</sup>

1. 武汉大学计算机学院, 武汉 430072; 2. 军事经济学院军需系, 武汉 430035;

† 通讯作者, E-mail: pengm@whu.edu.cn

**摘要** 本文提出一种通过建立情感词典和基于情感文本训练语料的来获取句子中的情感要素的观点句抽取与极性判别方法, 通过将单一情感要素和复合情感要素汇总计算, 构建主客观句抽取和情感句极性分析的处理流程。并提供可同时用于两种情感信息的情感信息抽取模型。通过参与公开评测, 证明该方法的有效性。

**关键词** 自然语言处理; 文本情感分析; 观点句抽取; 极性判断; 情感要素

中图分类号 TN914

## Extract the Chinese Opinion Sentence and Identify the Polarity based on Sentiment Elements

Nan Liu<sup>1,2</sup>, Yanxiang He<sup>1</sup>, Wenqi Fang<sup>1</sup>, Min Peng<sup>1,†</sup>, Jianbo Liu<sup>1</sup> and Jiajia Huang<sup>1</sup>

1 School of Computer, Wuhan University, Wuhan 430072, China; 2 Department of Quartermaster, Military Economic Academy, Wuhan 430035, China

† Corresponding Author, E-mail: pengm@whu.edu.cn

**Abstract** In this paper, we propose a new method for the extraction of opinion sentence and the judgment of its polarity in the sentences that contain sentiment element. In this method, the single sentiment element and combination sentiment element will be calculated comprehensively by building sentiment dictionary and utilizing sentiment-based training corpus. Based on this method, we design a model of extracting objective & subjective sentence and analyzing the polarity of the subjective sentence. This model can be used to extracting the sentiment information in two different types of sentiment elements at the same time. We apply this model in the analyzing of micro-blogging sentences and evaluate its effectiveness on the Sentiment Analyzing of Chinese Micro-blogging of CCF. The results are promising..

**Key words** Natural Language Processing; Text Sentiment Analysis; Opinion Sentence Extraction; Polarity identification; Sentiment Element

## 1 引言

随着互联网的日益普及, 尤其是 Web2.0 的蓬勃发展, 用户的参与度大大提高, 互联网上产生了大量

---

本文研究内容受到国家自然科学基金项目“基于反馈式排序框架 F-Rank 的查询导向的更新式多文档自动摘要研究”(批准号: 61070083)和国家科技支撑计划子课题“面向个体的网络行为分析、监管与保护技术研究”(编号: 2011BAK08B03-01)的赞助。

收稿日期: 2012-09-23; 修回日期: ; 网络出版时间:

网络出版地址:

对诸如人物、事件、产品等具有情感倾向的评价性信息。通过对这些情感信息的分析，我们可以了解到大众对某件事物或者某件产品的看法，或是积极，或是消极，或是中立的。微博，作为一种新的社交媒体已大众广泛接收，其发展和增长的速度超出大家的预料，每日交流的数据量呈

爆炸式增长。这为自然语言处理提供了新的研究领域，并提供了大量新形式的评论文本，传统的文本分析大多放在提取新闻文献等格式规范的文本的核心内容及主题上，但对微博这种长度较短，情感强烈，话题单一的短文本，更需要一种技术手段去了解其关注的内容和倾向，尤其是对于公众新闻事件和商品评论，这两个领域能明显的表现出用户情感的极性。面对如此海量的情感信息的微博文本，仅靠人工的方法是难以应对。因而人们迫切地需要找到一种快速对大规模文本进行情感分析的方法。

## 2 相关研究

中文文本情感分析的级别可划分为词语级、短语级、句子级、篇章级。句子是包含情感词、评价对象等信息的最小的完整的评价样本。因此，研究句子情感有较重要的意义<sup>[1]</sup>。而且，在微博等网络媒体中，由于提交内容的字数受到限制，用户会尝试在有限字数内表达完整的信息，从而提供了丰富的短文本样本。当前这些网络评论信息多用于两个领域，一个是商品评论的情感信息，有助于商家和用户了解产品的口碑，具有较大的商业意义和实用性。Kusha等<sup>[2]</sup>通过分析商品相关评论性文章的语义倾向，为其受欢迎程度打分，Liu等<sup>[3]</sup>利用顾客评价资源，对商品的市场反馈进行分析，为生产商和消费者提供直观的商品评价报告。另一个是基于公众话题的新闻评论情感分析，多用于舆情监控和信息预测当中。如Lin等<sup>[4]</sup>主要使用3种分类模型识别有关“巴以冲突”主题的评论文本所表达的观点，是“支持巴方”还是“支持以方”。

使用情感词语和词语间的信息分析文本情感是合理的途径，并且正确率往往较高。但是，该方法容易忽略文本中隐含的内在信息，使得其召回率较低。当前研究人员提出几种不同的解决方式。一种是从词汇之间内在联系入手，Turney等<sup>[5]</sup>使用点对互信息(PMI)方法扩充基准的褒贬词汇表。另一种是寻找具有情感的最小的词素，Yuen等<sup>[6]</sup>尝试采用利用中文情感词中强极性的词素(morphemes)来判断极性。Wang等<sup>[7]</sup>进一步将词素组合为短语，再判断整个句子的极性。还有一种方法是寻找有情感的短语。Wiebe和Wilson<sup>[8]</sup>提出主观表达式(subjective expressions)概念，将n元( $1 \leq n \leq 4$ )词语组合和训练语料中的标准情感短语对比，计算其能成为情感短语的概率。

对于语法内容灵活的网络文本，区别主观句和客观句并非易事，由于情感词典收录的情感词在某些领域表达的并非原意，因此许多客观句也可能会包含评价词。而且在极性判断时发现，有人会选择婉转的表达出情感，使得仅利用情感词有时难以识别该句子的情感极性。为了消除歧义，除了情感词典外，更需要挖掘文本中短语等包含情感信息的组合单元来判断。本文所采用的是利用短语中可能包含的主观情感信息，结合已有的情感词典，将其作为平级的情感要素综合考虑，来抽取观点句和判别极性。刘群等<sup>[9]</sup>提出将HOWNET情感词作为种子，计算词语的情语义相似度，这种方法得到广泛认同，我们将其作为对照方法，编写相应程序并提交公测平台，以验证文中方法的有效性。

## 3 观点句抽取与极性判别方法

### 3.1 构建情感词典

判断一个句子是否包含情感，以及其情感的极性，需要根据句中各个包含情感信息的单元来计算，网络上有一些公开的情感词典，对词汇级的基本单元所包含的情感信息进行了人工标记，但所包含的词汇量仍然较少，而且大多是有较强情感信息的特征词汇。这种采用情感词典来抽取观点句的方法，其效果的好坏会受到词典的规模的影响。一个规模大而准确的情感词典，将很直观的提高识别效果。想提高效果，一方面是增加情感词典的规模，这个需要投入大量的人力物力资源进行准确的标注。另一方面可以通过研究句子各结构里潜在的情感信息关系来寻找各单元之间的联系。

我们在研究的时候，分别在这两方面都做了大量工作。我们收集整理在网络上公开的情感词典，再投入大

量工作对其进行人工筛选，做了人工去重和更正标记的工作。我们根据HOWNET评价词词典<sup>1</sup>、NTUSD简体中文情感词典、武汉大学分级情感词典等多部情感词典来构建自己的情感词典，根据人工对词中情感的理解，标记为积极情感词库ALLWordsPos、消极情感词库ALLWordsNeg、中立情感词库ALLWordsNeu。见表1。

表 1 情感词典对比

词典名称	总词数	积极词数	消极词数	中立词数
HOWNET 评价词词典	6884	3730	3116	38
NTUSD 简体中文情感词典	11130	2810	8320	0
武汉大学分级情感词典	1389	679	660	50
OurWork	13628	4388	9213	27

### 3.2 获取情感信息

情感信息抽取是文本情感分析的基础，是其最底层的任务，通过抽取文本中有意义的信息单元，可将无结构化的情感文本转换为计算机可识别的结构化文本，从而可以做进一步的研究。

在建立词典和实验过程中，我们发觉，情感词典对于公共领域文本的识别有较高的准确率，而对于专业领域的文本，却很难有较好的识别效果。其原因在于，部分词语在通常认识中的词性和专业领域中，有不一致的效果。比如“圆滑”一词，在公共领域中“这个人很圆滑”，应该表达贬义的消极情感。而在电脑等专业领域中“苹果手机外壳很圆滑”只是客观的描述。此外，由于中文词语数量巨大，仅靠情感词典难以覆盖所有内容。因此，需要其他的方法来获取情感信息。

由于我们建立的情感词典是对 HOWNET 情感字典进行的扩充，而大部分词语不在 HOWNET 字典里面，因此，无法利用 HOWNET 字典来计算相似度。我们采用统计方法来获取情感信息，通过统计某一词汇或者短语、固定搭配在标记极性的文本样本中出现的情况，来判断其可能包含的情感信息。比如说，“官二代”本身是一个无情感意义的名词，但是由于常与官宦子弟的违纪行为联系在一起，因此其评论大多带有贬义。当一个句子中提及官二代，那么十有八九这个句子就是个消极情感句。此外，这一词汇是新出现的网络词语，在 HOWNET 等现有词典均并没有收录。

除了基础词汇外，有些句子中的固定搭配也有着其情感信息，这种关系很难在情感词典中列举。另外，一些评论文本中常见的口头语，也有较强烈的情感。比如说“就行了”，“(看、听)不下去了”，如果将其分解“行、下去”这些基本词汇并不包含情感，或情感信息较弱，但组合在一起，就会有较强的情感信息。这种组合出来的最小情感信息单元，我们称之为情感要素 (Sentiment Element)，按照其内部包含的具有词性的文本单元数量，可以分为单一情感要素 (Single Sentiment Element) 和复合情感要素 (Combination Sentiment Element)。单一情感要素如情感词，或者名词、副词等独立词性的词汇，本身具有一定的情感值。复合情感要素则是具有情感信息的最小的组合方式，比如“不下去”、“下去了”、“下去”都无法提供完整的情感信息。我们将单一情感要素和复合情感要素都作为句子中包含情感信息的一个基本单元来计算，通过汇总计算句中所有情感要素，来识别整个句子的极性。当积极情感要素占主要部分时，认为其为积极情感句，当消极情感要素占主要部分时，认为其为消极情感句。当两者数量相当，或其它其它无法明确归为正面或者负面的情况，都认为其为中立情感句。

在由客观句和主观句中的积极情感句、中立情感句、消极情感句共同组成的独立句子样本空间  $S = (s_1, s_2, \dots, s_m)$  中，有  $S = S_{pos} \cup S_{neg} \cup S_{neu} \cup S_{obj}$ 。对于整个样本空间  $D$  中各句包含的情感要素 (Sentiment Elements) 为  $E = (e_1, e_2, \dots, e_n)$ 。每个情感要素的情感级别 (Rank)，或称之权重为  $R = (r_1, r_2, \dots, r_n)$ 。其极性 (Polarity) 为  $P = (p_1, p_2, \dots, p_n)$ 。对于句子  $s_j$ ，其中出现的情感要素  $E_j$ ，有  $E_j \subseteq E, E = E_1 \cup E_2 \cup \dots \cup E_m$ 。在句子  $s_j$  中情感要素是否出现情况 (Find) 有  $F_j = (f_{1,j}, f_{2,j}, \dots, f_{n,j}), f_{i,j} \in \{0,1\}, 0 \leq i \leq n$ 。对句  $s_j$  中情感要素  $E_j$ ：

情感信息量：

<sup>1</sup> <http://www.keenage.com/html/e`index.html>

$$\text{SentiValue}_j = \sum_{k=1}^n (r_k \times f_{k,j}) \quad (1)$$

情感极性:

$$\text{SentiPolarity}_j = \sum_{k=1}^n (r_k \times p_k \times f_{k,j}) \quad (2)$$

为了计算方便,我们暂且将每个情感要素的情感级别都设定为1,即 $r_k = 1, 0 \leq k \leq n, |R| = 1$ ,极性只考虑积极极性(值为1)和消极极性(值为-1),另外有些并不包含情感信息的要素,我们将其设定为0,于是有 $p_i \in \{-1,0,1\}, 0 \leq i \leq n$ 。

$$\text{SentiValue}_j = \sum_{k=1}^n (f_{k,j}) \quad (3)$$

$$\text{SentiPolarity}_j = \sum_{k=1}^n (p_k \times f_{k,j}) \quad (4)$$

从而可得到句子j的情感信息量和情感极性。

### 3.3 观点句识别

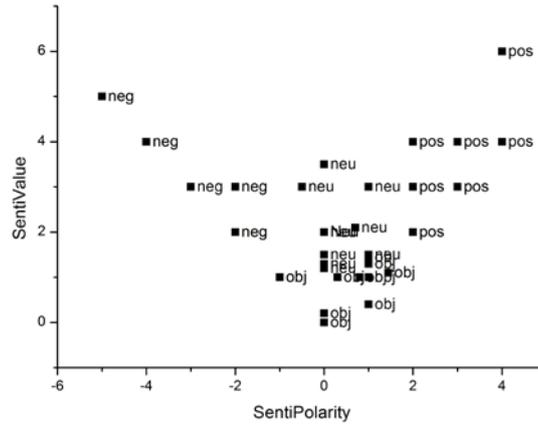


图1 主客观句分布图

我们选用“CCF中文微博情感分析评测测试”所提供的“韩寒方舟子之争”话题微博文本样本进行分析。以 $\text{SentiPolarity}$ 为X轴坐标,  $\text{SentiValue}$ 为Y轴坐标,将评测单位提供的人工标注句子样本采用此方法进行分析,其结果展示在坐标系中,可以明显看出不同种类句子的分布情况(图1)。

积极情感句(pos):具有正值的情感极性,情感词的总数较多。

消极情感句(neg):具有负值的情感极性,情感词的总数较多。

中立情感句(neu):又称为混合情感句,可能具有正值或负值的情感极性,但数值较少,接近于0,情感词的总数较多。

客观句(obj):可能具有正值或负值的情感极性,但数值较少,接近于0,情感词的总数较少。

我们以此作为分区主观句与客观句,以及判断主观句极性的依据。当 $\text{SentiValue}$ 大于主观阈值时句子为主观句,反之为客观句,然后通过计算句中所有情感要素的极性,并考虑其权重。从而得到整个句子的极性,当其极性超过阈值时,则认为其有极性存在,否则为中立句。

$$S_{pos} = \{s_j : s_j \in S, 0 \leq j \leq n, \text{SentiValue}_j \geq \alpha, \text{SentiPolarity}_j \geq \beta\}$$

$$S_{neg} = \{s_j : s_j \in S, 0 \leq j \leq n, \text{SentiValue}_j \geq \alpha, \text{SentiPolarity}_j \leq \gamma\}$$

$$S_{neu} = \{s_j : s_j \in S, 0 \leq j \leq n, \text{SentiValue}_j \geq \alpha, \gamma \leq \text{SentiPolarity}_j \leq \beta\}$$

$$S_{obj} = \{s_j : s_j \in S, 0 \leq j \leq n, \text{SentiValue}_j < \alpha, \gamma \leq \text{SentiPolarity}_j \leq \beta\}$$

其中 $\alpha, \beta, \gamma$ 为阈值,我们根据对已人工标注极性的样本数据进行分析,考虑到网络用户习惯于对积极情感采用较低级别的情感词和较少的情感描述,而对消极情感较易投入较强烈的情感。设定 $\alpha = 2, \beta = 1, \gamma = -2$ 。

### 3.4 训练情感要素的情感值

部分单一情感要素即为情感词典中收录的情感词,这些词汇可以通过直接查询情感词典来获取极性,而对于其他词典中未收录的词语或者是复合情感要素,则需采用机器学习的方法来获取其情感信息。由于网络语言

更新速度快，因此采用人工标记句子情感的方法要比维护情感词典的难度要小很多。此外，采用人工标记的方法便于根据需求快速建立某一专业领域的情感样本库。

这样的情感样本库建立后，便可通过训练来获取情感要素的情感值。当某要素同时在积极情感句集 $S_{pos}$ 和消极情感句集 $S_{neg}$ 中出现时，这一要素很可能是个常见的无情感要素，但如果该要素只在或大多存在于积极情感句集 $S_{pos}$ 中，则认为其很有可能具有积极的情感极性。我们这里只分析了在有情感的样本空间中的情况。

对于已经标注的样本空间 $S_{training}$ ，有 $S_{training} = S_{pos} + S_{neg} + S_{neu} + S_{obj}$ 。

对于情感元素 $e_i$ ，其数量分别为：

$$N_i = \{\sum_{k=1}^m f_{i,k} : s_k \in S_{training}, f_{i,k} > 0\} \quad (5)$$

$$N_{i,pos} = \{\sum_{k=1}^m f_{i,k} : s_k \in S_{pos}, f_{i,k} > 0\} \quad (6)$$

$$N_{i,neg} = \{\sum_{k=1}^m f_{i,k} : s_k \in S_{neg}, f_{i,k} > 0\} \quad (7)$$

$$SentiScore_i = \frac{N_{i,pos} - N_{i,neg}}{N_i} \quad (8)$$

$$SentiCount_i = N_{i,pos} + N_{i,neg} \quad (9)$$

$$PolarityType_i = \begin{cases} +1 & SentiCount_i \geq 0.33 \text{ and } SentiCount_i \geq 3 \\ -1 & SentiCount_i \leq -0.33 \text{ and } SentiCount_i \geq 3 \\ 0 & Others \end{cases} \quad (10)$$

我们采用谭松波博士提供的中文情感挖掘语料ChnSentiCorp作为训练样本<sup>1</sup>，选用去重后的笔记本、酒店、书本三类评论语料，每类中积极情感句和消极情感句各2000篇，未包含中立情感句和客观情感句。通过ICTCLAS2011分词系统对句子进行分词和词性标注，以满足例如d+v+y（副词+动词+语气词）模式的短语的抽取，通过统计该短语在积极文本中出现的次数PosCount和在消极文本中的次数NegCount，来验证该方法的可行性。

该方法中提到两个阈值(33.3%，3)，这是根据如下的实际经验来考虑的：

1. 句子如果只包含一个或多个单一情感时，其值应该为(1, >1)；
2. 句子如果包含两个或极性的多个情感要素时，在不考虑情感强度情况下，哪方向极性的要素数量更多，其情感的值就会偏向该方向。比如包含两个积极情感，一个消极情感，则，其情感值应该为1/3，即33.3%；
3. 情感要素数量设置3是为了避免偶然出现的噪声信息导致的误差。

虽然这个阈值设定的简单。我们在对情感信息计算时，仍然抽取符合常理的结果（表2）。从表中可以看到，这些由多词语搭配组合而成的短语，其中每个词都不包含情感信息，但组合在一起时，就会显现明显的情感信息。比如：“还可以吧”、“还行吧”、“不错了”。其中的“可以”、“行”、“错”都没有明显的情感意义。如果将其在HOWNET等情感词典中查询，无法获得合理的情感值，而通过查询其在积极情感句和消极情感句中出现的频率，可以发觉其经常用以搭配积极情感词，来表达一种积极情感。说明这些是有积极情感的复合情感要素。

“免费升级了”是和我们常识不同的特例，由于训练样本包括笔记本相关评论，这种行为是容易得到消费者的好感，因此也会出现在积极情感句中。由此可见，通过将样本所包含的主题分类，可以为各领域建立独立的情感信息库，以满足对专业领域话题情感信息的识别。

表2 d+v+y（副词+动词+语气词）模式短语的情感值

Phrase	PosCount	NegCount	SentiCount	Polarity	SentiValue	PolarityType
就没什么了	0	3	3	-3	-1	-1
不下去了	1	20	21	-19	-0.905	-1
不知道怎么	3	43	46	-40	-0.870	-1
就走了	1	9	10	-8	-0.800	-1

<sup>2</sup> <http://www.searchforum.org.cn/tansongbo/corpus-senti.htm>

就买了	1	4	5	-3	-0.600	-1
不值了	2	8	10	-6	-0.600	-1
都省了	1	2	3	-1	-0.333	-1
都没有了	2	3	5	-1	-0.200	0
就跑了	2	3	5	-1	-0.200	0
也没有什么	9	11	20	-2	-0.100	0
实在没什么	3	3	6	0	0	0
就无所谓了	4	4	8	0	0	0
就这样了	5	5	10	0	0	0
就够了	7	7	14	0	0	0
就可以了	23	15	38	8	0.211	0
就收到了	7	4	11	3	0.273	0
就关机了	1	0	1	1	1	0
还可以吧	7	3	10	4	0.400	1
还行吧	6	2	8	4	0.500	1
不错了	13	4	17	9	0.529	1
就行了	16	4	20	12	0.600	1
都是有	7	1	8	6	0.750	1
太喜欢了	3	0	3	3	1	1
又有什么	3	0	3	3	1	1
一定会有	5	0	5	5	1	1
不记得了	4	0	4	4	1	1
免费升级了	4	0	4	4	1	1

### 3.5 识别流程

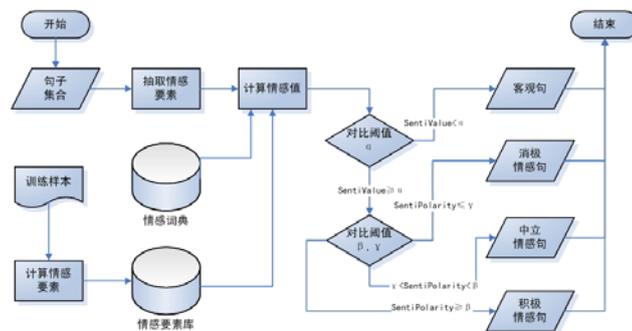


图2 情感句识别流程图

我们结合收集整理的情感词典和训练所获取的情感要素，来区分主观句与客观句，以及主观句中的不同极性的情感句（图2）。除了建立大规模的准确的情感词典以外，通过进一步训练样本的方法建立情感要素库。根据目标文档涉及的内容来选择合适的句子作为训练样本，将会比仅使用公共领域的情感词典有准确的识别效果。对于目标文档，首先对其进行预处理，将划分为最小的单字，再逐步合并，指导其在情感词典情感要素库里找到对应的内容。然后汇总计算单一情感要素和复合情感要素，通过计算整个句子的情感值与阈值SentiValue和SentiPolarity的差值来区分句子类型。若 $SentiValue_j < \alpha$ 则说明其为客观句；若 $(SentiValue \geq \alpha, SentiPolarity \geq \beta)$ 则为积极情感句；若 $(SentiValue \geq \alpha, SentiPolarity \leq \gamma)$ 则为消极情感句；若 $(SentiValue \geq \alpha, \gamma \leq SentiPolarity \leq \beta)$ 则为中立情感句。

现有的方法大多是首先通过判断语句中是否包含情感词来分析文本是否为主观句，然后再进一步分析其情感极性。我们这种方法使其可以把两个步骤合并在一起。通过统一的方法来计算并提取结果。因此简化了步骤，在实验中发现，和对比的两层分析的方法，该方法具有更快的运行速度。

## 4 实验结果与分析

### 4.1 实验设计

我们将根据本文提出的基于情感要素的情感分析方法和传统的基于HOWNET情感词典与词义相似度的情感分析方法同时参加CCF组织的中文微博情感分析评测。该评测数据包括20个微博话题，分别是“菲军舰恶意撞击”、“疯狂的大葱”、“官员财产公示”、“官员调研”、“国旗下讨伐教育制度”、“韩寒方舟子之争”、“假和尚搂女子”、“奖状植入广告”、“90后暴打老人”、“90后当教授”、“六六叫板小三”、“名古屋市长否认南京大屠杀”、“彭宇承认撞了南京老太”、“皮鞋果冻”、“苹果封杀360”、“三亚春节宰客”、“食用油涨价”、“洗碗工剩菜被开除”、“学雷锋被钓鱼执法”、“中国教师收入全球几垫底”，共含有3416条中文微博句子。本文方法为提交结果第39组，传统方法为提交结果第40组

### 4.2 实验结果

此次评测使用准确率（Precision），召回率（Recall）和F值（F-measure）来评价分析结果，微平均以整个数据集为一个评价单元，计算整体评价指标，宏评价以每个话题为评价单元，计算在该话题中的评价指标，最后计算所有话题上各指标的平均值。表3为本文提出的基于情感要素的中文微博情感分析方法和传统的基于HOWNET的情感分析在CCF组织的中文微博情感分析评测中主客观句识别的结果，表4为主观句情感分析的结果。

表 3 CCF 评测反馈的主客观句识别结果

	微平均			宏平均		
	P	R	F	P	R	F
本文方法	0.725	0.632	0.675	0.723	0.618	0.661
传统方法	0.708	0.649	0.677	0.708	0.634	0.663

表 4 CCF 评测反馈的主观句情感分析结果

	微平均			宏平均		
	P	R	F	P	R	F
本文方法	0.809	0.511	0.627	0.799	0.496	0.607
传统方法	0.740	0.480	0.582	0.731	0.465	0.565

从以上实验结果可以看出，在主客观句识别中，采用本文提出的基于情感要素的情感分析方法在准确率上比基于HOWNET的传统方法要高，但在召回率为略微降低。我们分析其原因在于基于HOWNET的传统方法抽取的动词、名词中很大一部分存在噪音，这对情感分析的准确率造成了一定程度上的影响，而本文中的方法由于从情感文本中训练情感要素，因此准确率较高，但由于仅根据常识来选择阈值，因此其召回率略低。如果进一步通过机器学习的方法，选择合适的阈值，和扩大用于训练的情感文本规模，将会有更好的效果。在主观句情感分析中，本文方法在准确率和召回率上都比传统方法要高，这说明该方法对于极性判断，有着较好的效果。此外，与其他评测队伍相对比，我们也并不居劣势，这也证明了该方法的有效性。

## 5 结论

本文研究的问题为如何抽取句子的情感信息和如何识别句子极性，文中通过建立情感词典和基于情感文本训练语料来获取句子中的情感要素，将单一情感要素和复合情感要素汇总计算，从而提出一种可行的观点句抽取与极性判别方法。并建立可同时用于单一情感信息和复合情感信息的情感信息抽取模型。最后，通过参与CCF组织的中文微博情感分析评测，对公测平台上的数据进行分析并提交结果，证明了该方法的有效性。

限于篇幅和参与评测的研究进度，本文的方法还仍有很大改进余地，如：现有方法中情感要素划分的质量受分词和词性标注效果的影响。由于常用分词工具会导致部分未收录的网络新词被分解，因此考虑采用N-Gram

方法来尽可能提取较长短语中包含的情感信息。此外，该方法尚未考虑分句中转折关系对整句情感的影响。在今后的工作中，将对此方法做进一步扩展，使其更加完善。

### 参考文献

- [1] 赵妍妍, 秦兵, 刘挺: 文本情感分析. 软件学报21(8) (2010) 1834 - 1848
- [2] Dave, K., Lawrence, S., Pennock, D.: Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th international conference on World Wide Web, ACM (2003) 519-528
- [3] Liu, B., Hu, M., Cheng, J.: Opinion observer: analyzing and comparing opinions on the web. In: Proceedings of the 14th international conference on World Wide Web, ACM (2005) 342-351
- [4] Lin, W., Wilson, T., Wiebe, J., Hauptmann, A.: Which side are you on?: identifying perspectives at the document and sentence levels. In: Proceedings of the Tenth Conference on Computational Natural Language Learning, Association for Computational Linguistics (2006) 109-116
- [5] Turney, P., Littman, M.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21 (2003) 315-346
- [6] Yuen, R., Chan, T., Lai, T., Kwong, O., T'sou, B.: Morpheme-based derivation of bipolar semantic orientation of chinese words. In: Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics (2004) 1008-1014
- [7] Wang, X., Fu, G.: Chinese sentence-level sentiment classification based on sentiment morphemes. In: Asian Language Processing (IALP), 2010 International Conference on, IEEE (2010) 203-206
- [8] Wiebe, J., Wilson, T.: Learning to disambiguate potentially subjective expressions. In: proceedings of the 6th conference on Natural language learning-Volume 20, Association for Computational Linguistics (2002) 1-7
- [9] 刘群, 李素建: 基于《知网》的词汇语义相似度计算. 中文计算语言学7(2) (2002) 59-76