情报学报 ISSN 1000-0135
第33卷第2期 175-182，2014年2月
JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 33 No. 2, 175-182 February 2014

doi:10.3772/j.issn.1000-0135.2014.02.007

面向中文微博的观点句识别研究[1)]

丁晟春 孟美任 李 霄

（南京理工大学经济管理学院信息管理系 210094）

摘要 中文微博包含了用户对热点话题的观点，对其进行观点挖掘可以实现突发事件预警、舆情监控等。目前，微博研究多数基于英文语料，中文微博观点句的挖掘大多混淆在情感挖掘中少量提及，由于中文微博特殊的语体特征，导致传统中文文本观点挖掘模型无法取得理想效果。区别于已有的情感挖掘工作，本文依据中文微博的语体特征分析结果选取特征，除了选取情感特征外，还加入主张性动词、语气词、程度副词以及固定词性结构等观点句特征，采用CRFs模型进行观点句识别研究。实验结果表明，仅选取情感特征准确率较高，但召回率仅为32.1%，而加入其他观点句特征后，召回率显著提高到61.8%。该方法应用于2012年中国计算机学会（CCF）组织的“观点句识别”测评任务中，取得了很好的效果。

关键词 中文微博 观点挖掘 CRFs模型 观点句识别 语体特征

Study of Subjective Sentence Identification Oriented to Chinese Microblog

Ding Shengchun, Meng Meiren and Li Xiao

(*Department of Information Management, School of Economics and Management, Nanjing University Science and Technology*)

Abstract Chinese Microblog include many opinions about hot topics. Mining opinion can realize early warning and public sentiment monitoring. Most of researches are usually based on English corpus. The existing researches generally confuse opinion mining and sentiment mining. Because of the specific stylistic features of Chinese Microblog, the traditional Chinese text opinion mining models cannot achieve ideal effects. In this paper, the features selections according to the analysis of the specific stylistic features of Chinese Microblog. Selecting declared verb, modal particles, degree adverb and fixed part of speech structures as the experiment features except the sentimental feature, which distinguish from sentiment mining. This paper used a CRFs(Conditional Random Fields) as the classification model. The results showed that recall ratio is only 32.1%, which is only used the sentimental feature. Added the other features, the recall ratio increased to 61.8%. This method was achieved an ideal effect with the opinion mining task of Chinese microblog which is held by China Computer Federation Technical Committee on China Information Technology.

Keywords Chinese microblog, opinion mining, CRFs model, opinion recognition, stylistic features

收稿日期：2013年6月25日

作者简介：**丁晟春**，女，1971年生，南京理工大学信息管理系，副教授，主要研究方向：Web数据挖掘、信息检索、信息系统开发；E-mail：todingding@163.com。**孟美任**，女，1988年生，南京理工大学信息管理系，硕士研究生，主要研究方向：信息检索。**李霄**，女，1989年生，南京理工大学信息管理系，硕士研究生，主要研究方向：数据挖掘。

1） 本文受国家自然科学基金项目“基于文本语义挖掘的商品评论信息可信度分析研究”（71103085）、“突发事件网络舆情演变过程中的人群仿真研究”（71273132）和江苏省高校哲学社会科学重点项目“网络舆情监测与有效引导研究”（2011ZDIXM028）的资助。

1 引 言

微博(MicroBlog)是一种基于用户关系的信息分享、传播及获取的平台,单条博文内容长度通常控制在140字以下。中国互联网数据中心(DCCI)发布的《2012中国微博蓝皮书》指出我国微博用户已达到3.27亿,日发布信息量约为2亿条,如此庞大的微博信息蕴藏着大量重要的用户观点。通过对微博进行观点挖掘,可以及时了解群众对热点话题的看法,帮助政府机构掌握突发事件后的社会群体心理,实现突发事件预警以及舆情监控;还可以作为企业进行市场分析、客户管理、产品升级时的重要依据。由此可见,微博观点挖掘研究具有重要理论与应用价值,但同样也面临着许多问题和挑战。

从近年KDD、WWW等国际会议及重要刊物中可以发现,微博已经逐渐成为学者的研究热点。其研究主要集中于语言层面、文本挖掘以及实际应用三个方面。在语言层面研究中,Ellen对微文本(Microtext)进行了特征分析,发现其具有"短"、"语法不规范"以及"半结构化"等特点[1]。邬智慧专门针对中文微博的语体特征进行了研究,提出中文微博开放性、精炼性、随意性、独特性等特征[2]。这些研究都为研究者进一步进行微博文本挖掘工作提供了重要的特征依据。目前,对于微博的文本挖掘工作主要包括文本分类、聚类;话题抽取;情感分析等几个方面。Davidiv等基于Twitter语料使用机器学习方法,以标签与表情符号作为特征,实现了微博的情感分类[3]。Andreevskaia等提出一种基于词汇的观点情感语义倾向性识别研究[4]。此外,在微博文本挖掘研究成果的基础上,研究者还将研究成果延伸至解决实际应用问题,如Bollen等利用Twitter上用户发布的微博来预测股市走向[5]。部分学者利用有关热点事件的微博进行舆情监控。王林等[6]对微博平台上用户集群行为特征及规律进行了研究,以热点话题"活熊取胆"事件为观测对象,从话题热度及走势、情绪热点及分布变化率、微博影响力路径等方面并对价值性执行意向的规律与感知进行了初步分析,为后期网民集群行为引导及网络营销方案的制定提供了一定的理论和实践指引。史伟等[7]以新浪微博为平台,通过抽取2011年7月23日"动车事故"发生后公众发表的微博并进行情感分析。提取了八类情感(期待,高兴,喜爱,惊讶,焦虑,悲伤,生气和憎恨),构建了用于情感分析的模糊情感本体,建立了微博文本的影响力和情感计算方法,对"动车事故"后的公众情感随事态发展的变化进行了探讨,为政府的舆情控制提供必要的参考。另外,王树义[8]提出了利用Twitter内容的监控来掌控竞争对手行动信息,以及通过Twitter交流可视化来构建竞争对手社交网络图。证明了Twitter能够在竞争情报工作中发挥重要的作用。通过对相关文献进行分析发现,目前在微博观点挖掘研究领域存在以下几个问题:

(1)现有对微博的研究大多是基于Twitter等国外知名微博平台的英文语料。由于语言表达方式、语言结构、语法等多方面的差异,致使国外已有的研究成果不能直接应用于中文微博的挖掘研究。

(2)与英文微博研究相比,基于中文微博的研究尚处于起步阶段。对中文微博观点句的挖掘研究大多都还是混淆在情感挖掘研究中少量提及,并未作为一个独立的研究课题进行研究。观点句指的是表达了对某一事件(事物)的某种评价、意见、态度或者立场的句子,其有别于情感句。单纯地将其视为情感挖掘工作,识别结果就会出现偏差,如忽略掉大量表达人们愿望与期许的观点句,错误识别出单纯自我心情发泄的非观点句等。

(3)中文微博语体特征较为特殊,与传统的中文文本存在较大差异。将以往对传统中文文本观点句挖掘模型直接用于中文微博观点挖掘时,常常会产生数据稀疏等严重问题,对实验结果产生影响。所以在进行中文微博观点挖掘的特征选择时,应依据其自身特殊的语体特征选择实验特征,再进行分类实验。

鉴于此,本文通过借鉴已有研究成果以及对大量中文微博语料进行统计分析,总结出中文微博语体特征,将其作为观点句挖掘实验中特征选取的重要理论依据。除了选取情感特征外,在区别观点句与情感句的基础上,一方面,深入分析观点句自身特点,加入主张性动词、语气词、程度副词作为实验特征;另一方面,采用N-POS文本表示模型选取了固定词性结构特征,采用善于处理短文本分类的CRFs模型进行中文微博观点挖掘实验。该方法应用于2012年中国计算机学会(CCF)组织的"观点句识别"测评任务中,取得了很好的效果。

2 中文微博语体特征分析

鉴于中文微博语体特征的特殊性,本文在借鉴

已有研究成果的基础上，对大量中文微博语料进行统计分析，总结出中文微博语体特征，将其作为观点句挖掘实验中特征选取的重要理论依据。

2.1 内容简短

文献[1]提出了微博具有“短”等特点。一方面是由于微博本身的字数限制，另一方面人们也习惯了采用简短的语句，甚至几个词来表明自己的态度、立场或者情感。所以与传统中文文本内容长度相比，微博内容更加简短。本文对2012 NLP & CC中文微博情感分析评测所提供的来自腾讯微博的参评语料以及本课题组收集的语料集合（语料均为中文微博，下文简称“语料D”）进行统计分析，该语料D共有26 526条微博，51 726个句子，其内容长度统计结果见表1。

从表1中我们可以看出，长度最短的微博只有一个标点，此类微博也表明用户时常喜欢用一个标点来表达自己的心情或立场，如“!”表示震惊，“?”表示疑问等。长度最长的微博为128个汉字，但平均长度仅为36个汉字。此外，内容为单句的微博占总语料集的54.7%，超过了总体的一半，可见微博内容简短的特征。

2.2 情感极性单一

一条微博被大量地转发、评论形成一定的规模，往往是由于其内容备受关注，如一些社会现象、刚刚发生的重大事件、发布者权威性等。针对此类事件发表的微博内容往往包含了群众的愤怒或者赞许。所以与一般的新闻或者商品评论相比，其内容的情感极性往往偏向一边：要么是对该事件非常赞同的正面评价，要么就是对该事件十分气愤的负面评论。对此，本文对语料D进行了情感词的统计分析，结果见表2。

从表2中可以发现，单条微博具有单一极性情感词的占70.8%，说明发布者对某一事件基本持有独树一帜的立场，既褒既贬的态度很少出现。其中单条微博具有负面极性情感词的占绝大多数是由于负面话题更能引起公众的关注。由此可知，含有情感特征的微博往往表明了发布者的观点与立场，但也不乏一些单纯的情感抒发。

2.3 内容口语化

微博作为用户的信息分享、传播以及获取的平台，具有人互动性、实时性。人们往往会使用一些较口语话的方式把自己内心的真实感受直接表达出来，并不会写成传统的书面语言。例如，“日子没法过了!”、“这个手机用起来可得劲儿了”等。这样的句子较为口语化，语法结构并不是非常规范，导致在分词等语料处理环节中容易出现错误。

2.4 隐性评价方式

微博除了内容较为口语化以外，人们往往采用一些歧义词、缩略词等隐性词对事件进行评论。此类隐性词的真正含义往往不能够直接从翻译其字面意思来获取。例如，词语“8错”表示“不错”、“稀饭”表示“喜欢”、“748”表示“去死吧”等。本文在微博平台上收集到此类词语526个，其中246个带有情感极性，连同其情感极性构建了歧义词词表。由此可见，出现词表中词语的句子可能会表明发布者的某种观点或者立场。

3 中文微博观点句的特征选取

本文一方面结合语言学文献以及对大量中文微博语料进行分析，总结出情感词、主张性动词、语气词以及程度副词4个观点句特征；另一方面采用N-POS文本表示模型选取观点句固定词性结构特征3个。

表1 语料D内容长度统计

语料名称	单条最短长度（字）	单条最长长度（字）	单条平均字数（字）	单句所占比例
语料D	0.5	128	36	54.7%

表2 语料D中单条微博情感词比例

语料名称	单条微博只存在褒义词/词组比例	单条微博只存在贬义词/词组比例	单条微博同时存在褒义词、贬义词/词组比例
语料D	14.5%	56.3%	4.5%

F1:情感词

2.2小节中提出了含有情感特征的微博往往表明了发布者的观点与立场,所以本文选取情感词作为观点挖掘实验的特征之一,构建了情感特征词典作为此项特征的标注依据。该词典包括两部分,一部分为HowNet情感词表,另一部分为课题组基于CRFs模型的半监督迭代学习方法识别出的情感词集[9]。本文认为如果一个句子中出现了情感词特征,则认为该句为观点句的可能性就越大。

F2:主张性动词

语法结构较为规范的观点句基本都伴随着一些主张性动词(如"认为"、"声明"等)出现。所以本文认为出现了主张性动词特征的句子为观点句的可能性较大。本文共收集了72个主张性动词,如"表示"、"承认"、"呼吁"、"要求"等。

F3:语气词以及带有情感色彩的标点符号

鲁川在其《知识工程语言学》一书[10]中,在汉语虚词的依附性标记中提出:"语气"标记(如"吗"、"呢"、"吧"等)表达了"发话者"的主观意图和主观认识。在2.1小节中也提到了长度最短的微博只有一个标点,表明"发话者"通常喜欢用语气词或带有情感色彩的标点符号("?"、"!"等)来表达其情感。所以本文认为如果一个句子中出现了语气词或情感标点特征,则认为这个句子是观点句的可能性较高。

F4:程度以及带有语气/态度的副词

在2.2小节中提到负面话题往往更能引起公众的关注,所以发布者对某一事件基本持有独树一帜的立场,情感极性单一。程度以及带有语气/态度的副词可以用来限制、修饰动词、形容词性(如再三、屡次、竟然、未免等),这些词一般会用来修饰情感词、发表看法或意见的动词。因此,此类特征出现时,当前句子为观点句的可能性也会很大。

F5-7:固定词性结构

叶强等[113]用N-POS文本表示模型来描述中文微博观点句固定词性结构特征,该模型的基础是将语料中的词按照词性(Part of Speech,POS)进行分类,再用语句中连续N个词性的顺序组合作为一个项,计算每个项的CHI值,选取排名靠前的项作为特征,对观点句进行判断。最终,主观文本的分类查准率和查全率均达到了76%。张博[12]也参考该方法利用SVM分类器,选取一些基于词语和词性的特征信息,对句子进行二分类。最终观点句准确率、召回率以及F值均达到了85%以上。可见,N-POS文本表示模型在观点句识别问题上表现出色。鉴于此,本文采用2-POS文本表示模型解决中文微博隐性主观句特征的选取,具体实验步骤如下所示:

(1)对语料D采用人工标注的方式筛选出观点句与非观点句

根据4.3小节中的观点句判别条件,采用人工独立全集标注(2人),将两人标注有差异的语料去掉,分别以句子为单位存储为观点句集O以及非观点句集P。

(2)采用中文分词工具进行自动分词、词性标注

本文使用中国科学院计算机所编写的中文分词工具ICTCLAS对语料O和语料P进行分词以及词性标注。

(3)提取分词后的观点句集O中所有的2-POS项

以观点句集O中的一句语料D11的2-POS项提取过程为例,如表3所示:

表3 N-POS模型实例

语料D11	我非常喜欢这款手机!
语料D11词性标注结果	我r非常d喜欢v这r款q手机n!wp
词性标注结果解释	我(代词)非常(副词)喜欢(动词)这(代词)款(量词)手机(名词)!(标点)

那么,D11的2-POS模型即为:"代词-副词、副词-动词、动词-代词、代词-量词、量词-名词",其中"代词-副词"即为一个2-POS项。

(4)从所有2-POS项中选取中文微博观点句挖掘的固定词性结构特征

采用卡方公式(1)计算第(3)步中提取出的全部2-POS项的值,并进行降序排序。本文认为值越高,说明此2-POS项在观点句集合O与非观点句集合P中存在的几率有显著差异,也就是表明该2-POS项越能更好的描述观点句的特征。另外,考虑到特征数量对实验结果的限制,本文只取排名前三的2-POS项作为中文微博观点句挖掘的固定词性结构特征。

$$\chi^2(\text{pattern}_i,0)=\frac{N\times(A\times D-C\times B)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (1)$$

其中,pattern_i表示2-POS文本表示模型;O表示观点句语料集类别;N表示语料D总数;A表示包含pattern_i的观点句语料O的句子频数,B表示包含

$pattern_i$的非观点句语料 P 的句子频数；C 表示不包含 $pattern_i$的观点句语料 O 的句子频数；D 表示不包含 $pattern_i$的非观点句语料 P 的句子频数。

在观点句集 O 中共发现 301 个 2-POS，按照 2-POS 项的值进行降序排序。表 4 列出了前 5 个 2-POS 项，本文实验中选取排名前 3 的 2-POS 项作为中文微博观点句挖掘的固定词性结构特征。

表 4 χ^2值排名前 5 的 2-POS 项

排名	2-POS 项	χ^2值
1	形容词 - 的(助词)(F5)	135.24
2	副词 - 形容词(F6)	128.75
3	的(助词) - 名词(F7)	90.46
4	副词 - 动词	67.20
5	动词 - 形容词	56.65

4 实验及结果分析

本文基于 CRFs 模型对加入观点句特征后的实验与只选取情感特征的实验结果进行对比分析，并将该方法应用于 2012 年中国计算机学会(CCF)组织的“观点句识别”测评任务中。

4.1 研究方法

条件随机场(Conditional Random Fields, CRFs)[13]最早由 Lafferty 等于 2001 年提出的，其模型思想的主要来源是最大熵模型。CRFs 模型不对单个标记归一化，而是通过定义标记序列和观察序列的条件概率 $P(X|Y)$ 来预测最可能的标记序列，避免了标记偏置问题。由于 CRFs 模型良好的学习能力，其已被广泛的应用于序列标记、数据分割、组块分析等自然语言处理任务中。同时，本课题组先前采用 CRF 模型进行了中文商品评论信息评价对象抽取、评论情感极性及其强度计算获得比较好的实验结果[14]，发现了 CRFs 在短文本分类问题上的优势。而中文微博正属于短文本，并且其内容口语化，直白化，甚至会出现只用一个表情或简单几个词来表述观点。鉴于此，本文选取其作为实验分类模型。

4.2 实验方案设计

本文实验数据来自 2012 NLP & CC 微博情感分析评测所提供的来自腾讯微博的参评数据，包括“菲军舰恶意撞击”、“疯狂的大葱”、“官员财产公示”等 20 个话题，共 17 526 条微博，31 726 个句子。设计中文微博观点句挖掘实验流程如图 1 所示。

(1)依据对中文微博语体特征分析结果，选取情感词、主张性动词、语气词、程度副词以及固定词性结构作为本文实验特征，并构建特征词典。在特征词典中，依据中文微博具有口语化、隐形评价方式等特点，特别加入了一部分从语料中收集到的歧义词、缩略词等。

(2)对参评语料进行预处理。句子中与观点句挖掘工作不相关的词以及符号(如量词、乱码等)会影响模型的识别效果。因此，在分词等标准化处理后，还借助相关词表对分词后的参评语料进行废词剔除。

(3)一方面，只通过情感特征库中的特征词进行观点挖掘实验(观点挖掘实验 1)；另一方面，加入主张性动词、语气词、程度副词以及固定词性结构特征再次进行观点句挖掘实验(观点挖掘实验 2)。依据构建的特征词表自动对参评语料进行特征集标注，模型的特征转换详见 4.4 小节。

(4)随机从参评语料集中选取 10% 的语料 S 用来训练模型。其中 20% 作为训练模型语料 S1，80%

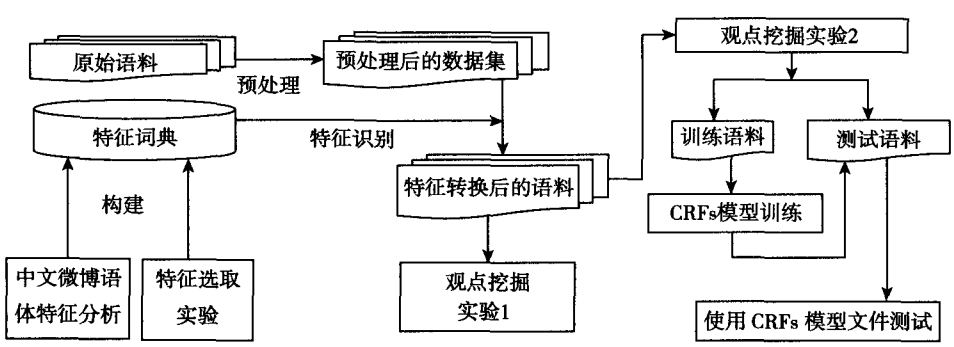

图 1 中文微博观点挖掘研究方案

作为测试模型语料T1。对S1的观点类别(观点句、非观点句)采用人工独立全集标注(2人),结果不同由第3个人判别,以尽量避免由于个人理解不同造成的误差;

(5)使用外部工具包CRFs + + -0.53对标注好的语料S1进行训练,选择识别效果较好的模板,生成模型文件。并使用最终得到训练模型文件对剩余语料进行观点句标注,得到实验结果;

(6)对实验结果分析。本文采用了文本分类分类器中最常用的评测指标——准确率(Precision)、召回率(Recall)和F-measure作为观点挖掘实验结果的评判标准。

4.3 观点句的判别标准

目前学术界对观点句的判别标准还没有统一的定论,本文借鉴国内外学者的研究成果以及近年来CCF会议测评、COAE等测评的答案语料,将观点句的判别标准归纳如下:

(1)表达了对某一事件(事务)的某种评价、意见、态度或者立场的句子为观点句,与发表人无关(第一人称、第三人称均可);

(2)对事件(事务)表达了一些愿望、期许或者预测的句子为观点句;

(3)只是表达自我情感、意愿或心情的句子为非观点句;

(4)包含评价词,但是只是描述某个客观事实的句子为非观点句。

4.4 特征转换

CRFs模型下的特征转换如表5所示。

4.5 评测结果分析

按照2012 NLP & CC中文微博情感评测任务1的观点句识别结果提交格式的标准,本实验将采用CRFs方法进行观点句识别的参评系统识别结果提交给测评委员会,得到如表6所示的测评结果。其中,26号为只选择情感词作为特征的识别结果,27号为使用CRFs模型的评测结果。

表5 CRFs模型下特征集描述

特征类别	特征标记	特征标记描述
情感词(F1)	A/N	当前句含有F1/不含有F1
主张性动词(F2)	B/N	当前句含有F2/不含有F2
语气词以及带有情感色彩的标点符号(F3)	C/N	当前句含有F3/不含有F3
程度以及带有语气/态度的副词(F4)	D/N	当前句含有F4/不含有F4
固定词性结构(F5/F6/F7)	E/F/G/N	当前句含有F5/含有F6/含有F7/均不含有

表6 CRFs方法参评结果

参评标识号	微平均			宏平均		
	准确率	召回率	*F*值	准确率	召回率	*F*值
26(只选用情感词特征)	0.852	0.321	0.481	0.863	0.332	0.494
27(CRFs模型)	0.732	0.618	0.671	0.735	0.606	0.658
准确率最高的参评系统	0.835	0.449	0.584	0.836	0.435	0.557
召回率最高的参评系统	0.645	0.959	0.772	0.649	0.960	0.770
*F*值最高的参评系统	0.671	0.944	0.784	0.674	0.942	0.783
平均值	0.727	0.615	0.647	0.727	0.608	0.634

注:微平均是以整个数据集为一个评价单元,计算整体的评价指标;宏平均是以每个话题为一个评价单元,计算参评系统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。

从表6中能够清晰地发现,本文的实验结果在准确率和召回率上有较出色的表现,根据结果对实验进行总结,主要原因如下:

(1)CRFs模型善于处理短文本分类问题,中文微博正属于短文本。与仅仅使用情感词特征相比,本文通过加入主张性动词、语气词及标点符号、副词、固定型结构等特征,更加全面,所以,观点句识别中准确率较高的现象得以解释。

(2)本文并没有考虑到上下文的关系。如"以后过节不送礼,送礼只送两根葱"此句中"两根葱"从字面意思可以理解成葱很便宜,那么这句话意思就是送礼提倡节约,不要铺张浪费。但是通过上下文发现词微博讨论的是大葱价格上涨,所以"两根葱"表明的是一种抱怨物价上涨的观点。虽然CRFs在识别短文本上有很好的效果,但在这种情况下,CRFs在处理新词的标注上作用很有限,表现一般。并且本次测评要求训练集不能超过整体语料的10%,即只用200条语料作为训练集,有限的训练集也制约了CRFs的表现。

(3)只选用情感词作为特征的实验结果准确率较高,这是由于大多数观点句都具有明显的褒贬情感。但是对于不带有褒贬情感的观点句就起不到较好的识别效果,所以召回率较低。而加入其他观点句特征后,召回率显著提高到60%以上。

(4)微博隐性评价较多,特征库不够全。一方面,我们已经考虑到了一些网络词汇的歧义性和缩略性,在特征库中加入了一定收集来的词。但是由于本次测评的语料为话题微博,主题针对某一事件,所以观点句情感色彩强烈,话语尖锐,出现大量不雅词语(如脏话、粗话),导致词库覆盖不全,通过对语料进行统计发现,共含有6个不雅词语,出现346次。

(5)人们在表达自己的观点和立场时也不直接的使用陈述性语句,而是采用反问的形式、与现实不符甚至毫不相关的言语来进行情感的表达。如:"我们怎么还能再继续使用这样的产品呢?"此句使用反问句的形式表明"我不会再继续使用这样的产品了"的观点,表达了发布者不满的情绪;"三鹿奶粉,后妈明智的选择"其中"后妈"一词带有贬义的色彩,具有"对孩子不好的妈妈"的意思,所以此句表明了"只有对孩子不好的妈妈才会选择三鹿奶粉"的观点,表达了发布者对三鹿奶粉持有的负面立场。表明应研究对反问句等特殊句式的处理方法,以提高模型的效果。

(6)分词错误导致特征并没有被准确识别。如"原配叫板小三",分词时,把"叫"和"板小三"分开,而"叫板"本身带有的贬义含义并没能与词典相匹配。

(7)不同人对观点句的认识程度不同。本文为了正确的标注观点句,已经采用了两人标注,第三人裁决的形式,但是与测评组发布的样例数据答案还是存在一定的差异。

5 结 论

本文的研究对象为中文微博,立足于中文微博的观点句挖掘研究。分析了中文微博的语体特征,并将其作为观点句挖掘实验中特征选取的重要理论依据。在选取情感特征的基础上,还加入主张性动词、语气词、程度副词以及固定词性结构等观点句特征,采用CRFs模型进行中文微博观点挖掘研究,将本文提出的观点挖掘方法应用于2012年中国计算机学会(CCF)组织的"观点句识别"测评任务中,准确率大于73%,召回率保持在60%以上,说明了本文提出的观点挖掘方法的有效性。为中文微博的观点挖掘提供了新的解决思路。

参 考 文 献

[1] Ellen J. All about Microtext-A Working Definition and a Survey of Current Microtext Research within Artificial Intelligence and Natural Language Processing[C]. Rome, Italy:ICAART, 2011:329-336.

[2] 邬智慧. 中文微博的语体特征研究[D]. 华中师范大学,2012.

[3] Davidov D, Tsur O, Rappoport A. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys [C]. Beijing, China: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010: 241-249.

[4] Andreevskaia A, Bergler S. Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses [C]. Trento, IT: Proceedings EACL-06, the 11rd Conference of the European Chapter of the Association for Computational Linguistics, 2006: 209-216.

[5] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2011, 2(1):1-8.

[6] 王林,赵杨,时勘. 集群行为的价值性执行意向微博实验研究[J]. 情报学报,2013,32(1):105-112.

[7] 史伟,王洪伟,何绍义.基于微博平台的公众情感分析[J].情报学报,2012,31(11):1171-1178.

[8] 王树义.基于微博客Twitter的企业竞争情报搜集[J].情报学报,2010(3):545-552.

[9] 丁晟春,文能,蒋婷,等.基于CRF模型的半监督学习迭代观点句识别研究[J].情报学报,2012(10):1071-1076.

[10] 鲁川.知识工程语言学[M].北京:清华大学出版社,2010.

[11] 叶强,张紫琼,罗振雄.面向互联网评论情感分析的中文主观性自动判别方法研究[J].信息系统学报,2007(01):79-91.

[12] 张博.基于SVM的中文观点句抽[D].北京邮电大学,2011.

[13] Lafferty J D, McCallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. CA, USA: Proceeding ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning, 2001:282-289.

[14] Ding Shengchun, Jiang T, Wen N. Research on Sentiment Orientation of Product Reviews Based on Cascaded CRFs Model[C]. Hong Kong, China: Proceedings of 2012 International Conference on Machine Learning and Cybernetics, ICMLC, 2012:1993-1999.

(责任编辑 车 尧)