基于汉语情感词表的句子情感倾向分类研究

王素格[1],杨安娜[1],李德玉[2]
WANG Su-ge[1],YANG An-na[1],LI De-yu[2]

1.山西大学 数学科学学院,太原 030006
2.山西大学 计算机与信息技术学院,太原 030006
1.School of Mathematics Science,Shanxi University,Taiyuan 030006,China
2.School of Computer & Information Technology,Shanxi University,Taiyuan 030006,China
E-mail:wsg@sxu.edu.cn

WANG Su-ge,YANG An-na,LI De-yu.Research on sentence sentiment classification based on Chinese sentiment word table.Computer Engineering and Applications,2009,45(24):153-155.

Abstract: This paper presents the weighted linear combination method for the sentence sentiment classification based on Chinese sentiment word table.In proposed method,firstly,Chinese sentiment word table is constructed by using five existing dictionaries,secondly,automatically identifying method is explored for the sentence sentiment classification using the weighted linear combination method.The experiment results indicate that the *F* value of sentence sentiment classification with word language granularity reaches 78.2%,and adding the negative phrase to language granularity,the *F* value of sentence sentiment classification has increased by 4.14%.

Key words: sentiment word table;weighted linear combination;sentence sentiment classification

摘 要:提出了一种基于汉语情感词词表的加权线性组合的句子情感分类方法。该方法通过已有的五种资源构建了中文情感词词表,并采用加权线性组合的句子情感分类方法对句子进行情感类别判断。实验结果表明,直接利用词汇语言粒度的句子情感分类综合 *F* 值为 78.62%,若加入了否定短语语言粒度后,句子情感分类的综合 *F* 值提高了 4.14%。

关键词:情感词表;加权线性组合;句子情感分类
DOI:10.3778/j.issn.1002-8331.2009.24.045 **文章编号:**1002-8331(2009)24-0153-03 **文献标识码:**A **中图分类号:**TP391

1 引言

从语言学角度,语言粒度从小到大依次为语素、词、短语、句子、段落、篇章。在计算语言学中,利用小粒度语言单元研究较大粒度语言单元是一种基于解析思想的常用方法。作为最小语言粒度的语素,它是最小的音义结合体,其主要功能是构词。词是可以独立运用的最小语言单位,而词义的内容却很丰富,词汇的褒贬义是其词义的重要组成部分,就句子的情感分类而言,词是构成句子的最基本的语言粒度,利用词汇的情感倾向可以确定句子的情感倾向。

所谓句子的情感分类就是识别出一个句子中作者对评价对象所持的态度是肯定还是否定,或者支持还是反对。目前,对句子的情感分类研究,J.Wiebe 等[1-2]将形容词作为判别句子的主客观性的主要依据。Hong Yu 等[3]面向自动问答系统首先采用抽取观点句,然后再对抽取的观点句进行情感分类,判断其极性。Hu 和 Liu[4]通过 WordNet 的同义词-反义词关系,得到情感词汇及其情感倾向,然后由句子中占优势的情感词汇的语义倾向决定该句子的极性。Wang 等[5]选取形容词和副词作为特征,提出了基于启发式规则与贝叶斯分类技术相融合的评论句子语义倾向分类方法。王根、赵军提出了一种基于多重冗余标记的 CRFs[6]汉语句子情感分析方法。该文主要是通过多个已有的与情感倾向相关的词典,建立了一个针对中文文本情感倾向分析用的情感词词表,并利用该情感词表,研究了句子的情感分类问题。通过对 200 篇汽车评论语料中包含情感词汇的句子进行测试,结果表明,该方法是可行的。

2 情感词表的构建

目前,虽然汉语文本或句子的情感倾向性分析开展的如火如荼,但还没有一部像英文的 General Inquirer(GI)(http://www.wjh.harvard.edu/~inquirer/)词典的中文词典,将借助 General Inquirer(GI)词典、《学生褒贬义词典》[7]、知网[8]、《褒义词词

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60573074);山西省自然科学基金(the Natural Science Foundation of Shanxi Province of China under Grant No.2007011042);教育部科学技术研究重点基金(No.2007018);山西省重点实验室开放基金项目;山西高校科技研究开发项目(No.200611002)。

作者简介:王素格(1964-),女,博士研究生,副教授,研究方向:自然语言处理与文本挖掘;杨安娜(1983-),女,硕士研究生,研究方向:自然语言处理;李德玉(1965-),男,教授,博士生导师,研究方向:计算智能与数据挖掘。

收稿日期:2008-05-09 **修回日期:**2008-07-28

典》[9]、《贬义词词典》[10]五种资源构建中文情感词词表,记为SWT。

2.1 情感词表中词条在五种资源中的分布情况

(1)General Inquirer(GI)词典是1966年开发的英文文本情感倾向分析研究中常用的基础资源之一。该词典包含了182个词语类目以及11 788个英语词汇。其中有1 915个词汇标注了“褒义”属性、2 291个词汇标注了“贬义”属性。对于一个词汇的多个义项,词典中将之作为不同条目列出,用于区分某个词语在特定义项或词性上体现的不同褒贬属性。例如,英语词汇kind作为形容词时,具有褒义情感,而作为名词时则不具有这种情感倾向。对英文General Inquirer(GI)词典中的每个词条根据《牛津英汉双解词典》逐条翻译,并与英文情感倾向进行对照标注,得到的词典称为CGI(Chinese General Inquirer)词典。

(2)中文情感词汇已有的相关资源

①《学生褒贬义词典》[7](张伟、刘缙等编纂)。该词典收录了含有褒义或贬义的双音词、成语和惯用语共1 669条,其中褒义730条,贬义939条。该词典的优点在于为每一个词语明确标注了其情感倾向,同一个词语在不同词性时所代表的情感倾向也有明确的标记,并且为每个词语列出了与其褒贬色彩相同的近义词语,称之为学生褒贬义词典扩展。

②知网[8](董振东先生研制)。知网(HowNet)是一个以汉语和英语词语所代表的概念为描述对象,以表示概念与概念之间、以及概念所具有的属性之间的关系为基本内容的常识知识库。在知网中,分为程度级别词语(219个)、负面评价词语(3 116个)、负面情感词语(1 254个)、正面评价词语(3 730个)、正面情感词语(836个)、主张词语(38个)。

③《褒义词词典》[9](史继林、朱英贵编著四川辞书出版社)收录了5 067个词条。

④《贬义词词典》[10](杨玲,朱英贵编著四川辞书出版社)收录了3 495个词条。

利用上述五种资源及《学生褒贬义词典》在同义词意义下的扩展(学生褒贬义词典扩展),得到了情感词表SWT,其词条在五种资源中的分布情况详见表1。

表1 SWT中词条在五种资源中的分布情况

词典类型		褒义词条	贬义词条	总词条
词典无重复(词条仅出现在一部词典)	学生褒贬义词典(PNC)	163	331	494
	学生褒贬义词典扩展(PNCE)	161	264	425
	褒义词词典(PC)	3 061	0	3 061
	贬义词词典(NC)	0	1 949	1 949
	HowNet	2 744	3 009	5 753
词典有重复(词条出现在多部词典)	GI翻译(CGI)	891	836	1 727
	学生褒贬义词典(PNC)	567	608	1 175
	学生褒贬义词典扩展(PNCE)	340	318	658
	褒义词词典(PC)	2 006	0	2 006
	贬义词词典(NC)	0	1 546	1 546
	HowNet	1 688	1 213	2 901

来源于两种以上词典资源词条总数及被标注情感倾向不一致的词条数目统计信息详见表2和表3。

表2和表3中,T表示词条总数,DT表示情感倾向不一致的词条数。DT/T为词条情感倾向不一致率。

由表2的数据,经过比较不一致率,得到五种词典资源对多源词条的情感倾向标注不一致率排序为:CGI>HowNet>NC>PC>PNC。

2.2 词表中词条情感义的确定

设计情感词表(SWT)中词条的数据格式如下:

表2 来源于两种词典资源的词条数目统计

词典	词典与词条数									
	CGI		PNC		PC		NC		HowNet	
	T	*DT*	*T*	*DT*	*T*	*DT*	*T*	*DT*	*T*	*DT*
CGI	1 727	0								
PNC	63	11	2 752	0						
PC	346	37	173	1	5 067	0				
NC	228	5	255	0	5	5	3 495	0		
HowNet	143	9	404	14	774	21	560	18	8 653	0

表3 来源于两种以上词典资源的词条数目统计

词典	词条数		
	T	*DT*	*DT*/*T*×100%
CGI∩PNC∩PC∩NC∩HowNet	1	1	100
CGI∩PNC∩PC∩HowNet	200	6	3.00
CGI∩PNC∩NC∩HowNet	148	11	7.43
CGI∩PC∩NC∩HowNet	1	1	100
CGI∩PNC∩PC	75	8	10.67
CGI∩PNC∩NC	156	8	5.13
CGI∩PNC∩HowNet	54	9	16.67
CGI∩PC∩NC	2	2	100
CGI∩PC∩HowNet	206	22	10.68
CGI∩NC∩HowNet	104	4	3.85
PNC∩PC∩NC	1	1	100
PNC∩PC∩HowNet	220	4	1.82
PNC∩NC∩HowNet	83	4	4.82
PC∩NC∩HowNet	2	2	100

SW	Pos	S	CGI	CGIS	PNC	PNCS	PNCE	PNCES	PC	PCS	NC	NCS	HowNet	HowNetS

数据格式中SW表示词条,Pos表示词性,S表示情感倾向;CGI表示当前词条SW在CGI中,CGIS表示该词条在CGI中的倾向,数据格式中的其余各项类似。

定义1 词条SW的情感倾向一致率(WSSA)定义为:

$$WSSA(SW)=\frac{NUMACC(SW)}{NUM(SW)}\times 100\% \quad (1)$$

这里,$NUM(SW)$表示含词条SW的词典数,$NUMACC(SW)$表示$NUM(SW)$中对词条SW标注某种情感倾向最多的词典数。

SWT中词条SW的情感倾向确定原则:如果词条SW的情感倾向$WSSA(SW)>50\%$,则将$NUMACC(SW)$所指的情感倾向确定为词条SW在SWT中的情感倾向,否则按优先级PNC>PC>NC>HowNet>CGI为其确定情感倾向。

例如:词语“鼓吹”包含在CGI、学生褒贬义词典、贬义词词典,而CGI、贬义词词典的倾向均为“贬义”,故“鼓吹”在情感词汇词典中的情感倾向标注为贬义(反面)。

最终SWT中共收录词条15 886个(正面8 427个,反面7 459个),其中仅来源于一部词典的词条11 682个(正面为6 129个,反面为5 553个)。另有来源于多个词典的词条4 204个(正面为2 298个,反面为1 906个)。

3 基于情感词词表的句子情感分类

在第2章中建立的情感词表SWT的基础上,探索句子的情感倾向判别。

3.1 句子情感分类的评价指标

句子的分类评价指标为准确率(Precision)、召回率(Recall)、F值三种评价指标。对句子进行情感分类的三种评价指

标分别定义如下:

$$Precision=\frac{A}{B}\times 100\% \tag{2}$$

$$Recall=\frac{A}{C}\times 100\% \tag{3}$$

$$F=2\times\frac{Precision\times Recall}{Precision+Recall} \tag{4}$$

这里,A 表示系统判别该类正确的句子数,B 表示系统能够判别该类的句子数,C 表示该类句子的总句子数。

3.2 句子的情感分类

3.2.1 基于词汇语言粒度的句子表示

按照人在表达不同情感时的语言结构,将句子按照词汇语言粒度表示,句子 $sentence$ 的表示如下:

$$v(sentence)=((t_1,o(t_1)),(t_2,o(t_2)),\cdots,(t_r,o(t_r)),\cdots,(t_m,o(t_m)))$$

这里,t_r 表示句子 $sentence$ 中第 r 个词汇语言粒度,$o(t_r)$表示词汇语言粒度 t_r 的情感倾向,用 1 表示词汇语言粒度 t_r 为正面情感倾向,用-1 表示词汇语言粒度 t_r 为反面情感倾向。

3.2.2 句子的分类函数构造

为了刻画各情感词汇语言粒度对句子的情感倾向判别的贡献,采用加权线性组合方法构造句子的分类函数 $f(sentence)$。

定义 2 句子的分类函数:

$$f(sentence)=\sum_{r=1}^{m}\alpha(t_r)\cdot o(t_r) \tag{5}$$

这里,$o(t_r)$表示词汇语言粒度 t_r 的情感倾向,$\alpha(t_r)$表示第 r 个词汇语言粒度 t_r 的权值,$\sum_{r=1}^{m}\alpha(t_r)=1$。

根据句子的组成特点,可以把句子分为简单句和复杂句。简单句为只具有主语和谓语的句子以及短语。复杂句为具有连词连接的句子,或没有连词连接但至少有两个或两个以上的分句构成的句子。复杂句一般分为以下几种:

(1)各分句情感倾向一致的句子。

①一般由连词{也,还,同时,既……又,不但……而且(并且)、更}连接的并列句,各分句平等地排列在一起,分别说明或描写几件事情、几种情况或同一事物的几个方面,情感倾向由各个情感倾向一致的分句所决定,且任一情感倾向鲜明的分句足以代表整个句子的情感倾向。如:

"它不仅仅是提高了撞车事故中车辆的被动安全性,也对车辆的舒适性和操控品质的提高有很大帮助"。

"他们为 NF 御翔安装了加厚侧窗玻璃和车底隔音板,并改进了空调、车内换气系统和发动机风扇"。

②由连词{因为……所以……,由于……因此……,……以致……,之所以……是因为……,于是,从而}连接的因果句,由结果句的情感倾向判定整个句子的情感倾向。如:

"事实上,此前上海通用之所以能够取得 2005 年市场销售的冠军,很大一个原因也在于,其推出的新车都由泛亚汽车技术中心针对中国市场进行了本地化改造"。

"美国通用就因为品牌众多而使研发费用、运营成本居高不下,致使新车研发经费分散,从而使推新车的速度减缓,极大地削弱了通用汽车的市场竞争力"。

③由",",连接的无连词的几个分句构成的复杂句,语义上一般是顺承关系,情感倾向也一致。如:

"新旗云的外形比较新,发动机技术先进"

"旗云的助力适度,比较轻,使用起来较舒服"

"车内很安静,驾驶稳定表现出色,应对颠簸的阻尼处理同样优秀"

(2)各分句情感倾向不一致的句子,由{虽然……但是……,尽管……但是(却)……,不过,然而,只是}等词连接的转折句在语义上,后一分句同前一分句相反或相对,一般更偏重于强调转折词后的分句的倾向。如:

"捷达的发动机虽然比较落后,但最大扭矩来的比较早,所以提速还是比较迅速的"。

"CVT 的高科技没有带来低油耗,起步没有捷达顺,但是在起步后,优势很明显"。

由上述分析可知,句子尾部的情感倾向对整个句子的情感倾向影响较大,因此第 r 个词汇语言粒度 t_r 的权值 $\alpha(t_r)$采用句子末词汇语言粒度占优法,使句子末的情感词汇的权值大一些,其余的权值相等。

定义 3 第 r 个词汇语言粒度 t_r 的权值 $\alpha(t_r)$:

$$\alpha(t_r)=\begin{cases}\dfrac{2}{m} & r=m\\ \dfrac{1}{m} & r=1,2,\cdots,m-1\end{cases} \tag{6}$$

3.2.3 句子的情感分类过程

基于词汇的句子情感分类过程如下:

步骤 1 对每个句子抽取含情感词表 SWT 中的情感词汇;

步骤 2 对每个情感词汇利用情感词表 SWT 标注其情感倾向;

步骤 3 利用公式(5)得到结果,当 $f(sentence)>0$ 时,判定句子 $sentence$ 为正面类别,当 $f(sentence)<0$ 时,判定句子 $sentence$ 为反面类别,否则 $sentence$ 为中性类别。

只对正面和反面句子进行测试。

3.3 实验结果与分析

测试句子采用 200 篇(100 篇正面,100 篇反面)汽车评论语料中 3 190 个句子进行实验,其中正面句子 2 144 个,反面句子 1 046 个。利用第 3.2.3 节中基于词汇的句子情感分类过程,得到实验结果见表 4。

表 4 基于词汇的句子情感分类实验测试结果

情感类别	评价指标		
	精确率/(%)	召回率/(%)	F 值/(%)
正面	77.58	95.90	85.77
反面	83.70	43.21	56.99
综合	78.62	78.62	78.62

由表 4 可以看出词汇对正面句子情感分类的结果优于反面倾向的句子,主要是因为在具有反面倾向的句子中常有"否定词"出现,如:句子"凯越 HRV 驾驶的时候,门窗是不会自动上锁的,这是一个很不安全的设计。"和"存放备胎的地方也发现了不和谐的声音"。如果由句子中所有情感词汇情感类别值的累加表示整个句子的情感类别,导致标注错误,故反面句子召回率比较低。如果句子的情感类别由"否定词+情感词汇"来表示,这样可以提高反面句子的召回率。

将与否定副词连用的情感词的情感倾向取反,得到句子的情感倾向类别的测试结果见表 5。

从上述得到的结果可以看到,各项情感分类评价指标均有

(下转 161 页)