一种基于知网的中文句子情感倾向判别方法 *

党 蕾, 张 蕾
(西北大学 信息科学与技术学院, 西安 710127)

摘 要: 针对基于知网的中文句子情感倾向判别方法中存在的准确率不高的问题,提出采用否定模式匹配与依存句法分析相结合的方法。研究分析了修饰词极性以及否定共享模式,确定修饰词以及扩展极性的定量和否定共享范围,提出依存语法距离的影响因素来计算情感倾向,并且在否定模式匹配后改进句子极性算法。实验结果表明该方法取得了良好的效果。

关键词: 否定词扩展; 否定共享; 依存句法关系; 句子情感倾向; 知网
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2010)04-1370-03
doi:10.3969/j.issn.1001-3695.2010.04.044

Method of discriminant for Chinese sentence sentiment orientation based on HowNet

DANG Lei, ZHANG Lei
(*College of Information Science & Technology, Northwest University, Xi'an* 710127, *China*)

Abstract: This paper proposed a method which combining negation model matching with dependency parsing , to solve the low precision rates in the Chinese sentences sentiment orientation based on HowNet. On one hand, according to analysis the negative polarity of the qualifier and sharing models to determine the polarity of the qualifier, identified the expansion of quantitative and the scope of negation sharing. On the other hand, according to the dependency grammar distance factors to calculate the sentiment orientation, and to improve the polar algorithms of the sentence after the negation model-matching. The experimental results show that the method has achieved good results.

Key words: negative words extended; negation sharing; interdependent syntactic relations; sentence sentiment orientation; HowNet

在网络资源日渐丰富的今天,越来越需要通过一种有效的搜索策略来甄别网络信息,如对新闻时事以及媒介传播的评论,企业产品发布前的市场调研和用户反馈等。对这些信息的褒贬性评估,能够直接反映出人们主观上对该事物的喜好程度或支持与否的态度,这些有价值的信息有利于企业或专家进行更好的应用和管理。而在计算机语言学方面,目前普遍关注的是客观性信息的分析和提取,对主观性信息分析与提取的研究尚处于起步阶段,因此语义倾向具有重要的学术研究价值。

目前对语义倾向的研究分为词汇和句子两方面,关于词汇语义倾向的研究始于最早,而且取得了一定的成果[1,2]。而对句子的语义倾向的研究目前相关工作很少,具体在中文句子的情感倾向研究方面有蔡健平等人[3]提出的基于机器学习的词语和句子极性分析,该方法通过构建极性词典来分析领域极性词,同时采用基于词的方法和 Bayes 方法对网上手机评论文章包含的主观意见进行褒贬挖掘,取得了一定的成果。熊德兰等人[4]提出了基于知网的语义距离和语法距离相结合的句子褒贬倾向性计算方法,利用夹角余弦法对语义倾向进行了改进。

1 中文句子情感倾向的研究

1.1 否定词的扩展

情感倾向的研究中,副词往往对中文句子的褒贬倾向影响较大。否定词的影响也起到了至关重要的作用,一般会有情感语义倾向词汇与语句中否定成分结合在一起形成相反极性的情况。配价移动指示符(VSI)能传递对应情感描述项极性总的量值的相反极性,但是它本身却不具有极性,如"不是""不能"等[5]。因此定义 $O(W)$ 为词汇 W 的原极性,A 为否定词,B 为程度副词,按照分类属性将 B 分为相对副词和绝对副词[4],且表示为 λ_1、λ_2($1>\lambda_1>\lambda_2>0$)。根据相对和绝对的强度定义 $\lambda_1=0.4$,$\lambda_2=0.7$;同时根据极量、高量、中量、低量四个不同的程度级别,将其分别定义为 η_1、η_2、η_3、η_4(其中 $1>\eta_1>\eta_2>\eta_3>\eta_4>-1$),且取值为 $\eta_i=0.8$,$\eta_2=0.6$,$\eta_3=0.1$,$\eta_4=-0.5$。这里对副词的分类参照文献[4]中的副词分类表。定义模式 $A+O(W)$,$B+O(W)$ 的极性值:

$$A+O(W)=\begin{cases} O(W)-1 & \text{当 } O(W)>0 \text{ 时} \\ O(W)+1 & \text{当 } O(W)<0 \text{ 时} \end{cases}$$

$$B+O(W)=\begin{cases} O(W)+(1-O(W))\times\lambda_i\times\eta_i & \text{当 } O(W)>0 \text{ 时} \\ O(W)-(1+O(W))\times\lambda_i\times\eta_i & \text{当 } O(W)<0 \text{ 时} \end{cases}$$

配价移动指示符所代表的极性定义为 $A+O(W)$ 的模式极性,因为这些能愿动词在表达语义倾向的意义层面上并不是起到了反义的效果,只是简单的语义弱化过程。考虑到副词对否定词的影响,将副词进行扩充。具体扩充规则如表 1 所示。

收稿日期: 2009-09-05; **修回日期**: 2009-10-20 **基金项目**: 陕西省教育厅专项科研基金资助项目(HD01302)

作者简介:党蕾(1983-),女,陕西咸阳人,硕士研究生,主要研究方向为人工智能及自然语言理解(purplelianren@163.com);张蕾(1964-),女,陕西西安人,教授,硕导,博士,主要研究方向为人工智能及自然语言理解.

表 1　否定词扩充模板及褒贬取值

扩展方式	例子	模板	褒贬取值
前扩展	否定词:并不,很不,大不	$B+A+W$	if $O(W)>0$ $(O(W)-1)+(1-(O(W)-1))\times\lambda_i\times\eta_i$ if $O(W)<0$ $(O(W)+1)-(1+(O(W)+1))\times\lambda_i\times\eta_i$
	肯定词:莫不,无不	$A+A+W$	$O(W)$
	否定词+能愿动词:不能,不会	VSI+W	$A+O(W)$
后扩展	否定词+副词:不大,不多	$A+B+W$	if $O(W)>0$ $O(W)+(1-O(W))\times\lambda_i\times(-\eta_i)$ if $O(W)<0$ $O(W)-(1+O(W))\times\lambda_i\times(-\eta_i)$
	否定词+语素:不必,不便,不宜	VSI+W	$A+O(W)$

1.2　否定词的共享问题

以往的研究多局限于单句内部的否定辖域,很少涉及到跨标点句的否定辖域,而且主要针对基本否定词,很少论及基本否定词扩充词的辖域。跨标点句的否定辖域是整个跨标点句句法共享问题的一个组成部分。从句子极性的分析来看,否定词是否共享也影响到判断句子情感的倾向。如果 AB 是跨标点句的句法结构,标点句 A 称做原配句,标点句 B 称做续配句[3]。针对否定词的管辖范围以及对否定词扩展的研究,可以将否定词共享问题按照基本否定词与扩展否定词来制定相应模式,否定词的管辖分类情况归纳于表 2。在研究基本否定词“不”的共享与否问题之前,必须去除以固定搭配、词语、习语,“不”做主语,连动结构并列结构中对“不”的运用,如不好意思,经不起,不小心等。只有明确研究的对象,才能对否定词的管辖意义给出比较好的结果。本文中否定词的获取是通过知网实现的,由于处理的基本否定词以及其扩展,在知网中选取具有否定意义的义原[6],如{neg|否},{BeUnable|无能},{impossible|不会},{unsuitable|不宜}等,从中抽取出包含否定义原的概念,经人工过滤得到 18 个否定词,这些否定词不仅包括对基本否定词的义原定义还包含有扩展后否定词的义原,从而方便在语料中对句子进行否定模式匹配。

2　句子的情感倾向判别

在明确否定词及其扩展的修饰词极性后,利用哈尔滨工业大学依存分析器定义的句法关系作为参考来计算句子极性,选择语义倾向词经常出现的动宾关系 VOB、状中结构 ADV、主干中心词(HED),其他修饰词的语义倾向也需考虑在内,包括动补结构 CMP,“的”字结构 DE。

2.1　句子的极性算法

在依存树结构中可以发现,如果主干中心词包含有褒贬倾向的词汇时,离中心词越近的修饰结构,对整个句子的情感倾向影响较大,而离其较远的则影响较小。因此将依存句法分析中的这种距离定义为依存语法距离,即自顶向上搜索依存树结构,获取具有语义倾向的词汇到主干中心词的距离,将其定义为 d。对主干中心词汇的影响因子可以记为 $\frac{\lambda}{\lambda+d}$。其中 λ 为调节因子。句子的极性为式(1)。

$$\text{orientation}(S)=\sum_{i=1}^{k}\frac{\lambda}{\lambda+d_i}O(w_i) \tag{1}$$

表 2　否定词共享模式

否定词	不被共享模式	共享模式
基本否定词“不”	续配句句首出现以下情况: a) conj + s b) pp + s c) np + s d) adv + s (这里除程度副词与范围副词)	neg + phrase_1, phrase_2 phrase_1, phrase_2为四字短语或者成语 neg + v_1+n_1, v_2+n_2 当 $v_1=v_2$时,不被共享 当 $v_1\neq v_2$时,且 $m\geq4$ (m 为 $v+n$ 字数),可以共享 $\text{neg}_1+v_1+n_1$, v_2+n_2
扩展否定词(否定词后扩充标记为 neg_1)	$\text{neg}_1+v_1+n_1$, model verb/adv/morpheme + v_2+n_2(续配句以能愿动词、程度副词以及某些词素开头的,扩展否定词不被共享)	

当修饰词出现时,根据修饰词的极性计算方法来重新计算句子中语义倾向的词汇极性,所以 $O(W_i)$ 不是代表词汇的原有语义倾向值,而是出现修饰词后的极性值。

2.2　句子情感倾向的判别

1)寻找依存树中所有含有 SBV 结构关系对

a)如果谓语是形容词,则直接计算该词汇的原极性,记形容词为 adj_1,计算该词汇的极性记为 $O(\text{adj}_1)$,且该极性不为 0;如果该谓语(predicate)为动词,且为主干中心词,记谓语动词为 verb,计算该词汇的极性记为 $O(\text{verb})$,且该极性不为 0。

b)那么以谓语为主干中心词的节点对依存树进行左遍历,查找含有 predicate 的 ADV(状中结构)关系对,通过计算得到该谓语的修饰后的极性 $O(W_1)$。

c)如果谓语不具有极性,则直接执行下一步。

2)寻找依存树中 VOB 结构关系对

a)如果关系对里宾语为名词(noun),且该名词具有语义倾向,对以该名词为节点的子树进行左遍历搜索。如果搜索为空,则返回该宾语的极性记为 $O(\text{noun})$,转为执行 3),反之,查找含有该名词的 DE(“的”字结构)关系对,记形容词为 adj_2,计算该词汇的极性记为 $O(\text{adj}_2)$;如果关系对里宾语为形容词,记该形容词 adj_3,计算该词汇的极性为 $O(\text{adj}_3)$。

b)同样也要找到含有 adj 的 ADV(状中结构)关系对。根据修饰极性的计算得到 $O(W_2)$、$O(W_3)$。

c)如果宾语不具有极性,则直接执行下一步。

3)继续寻找依存树中 CMP 结构关系对

a)如果关系对里补语为形容词,记该形容词 adj_4,计算该词汇的极性为 $O(\text{adj}_4)$,也要找到含有 adj_4的 ADV(状中结构)关系对。根据修饰极性的计算得到 $O(W_4)$。

b)如果关系对里补语为动词或副词,若其极性不为 0,则计算其原有极性值或修饰后的极性,记为 $O(W_5)$,$O(W_6)$等。

4)查找结束后按照式(2)计算句子的极性值

下面用一个例子来验证该算法的可行性,如例句:“电影的主题概念非常具有创意,唯美的画面场景也让人感到愉悦。”依存分析结果如图 1 所示。

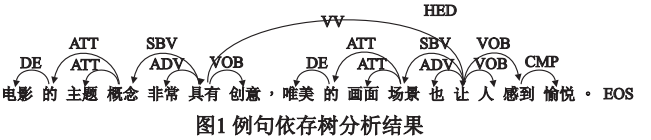

图1 例句依存树分析结果

分析步骤如下:a)主干中心词(HED)为“具有”。SBV 关系对中“具有”为 VSI 转移符,则向左搜索查找 ADV 关系对,得到“非常”的修饰词性;寻找以该动词为中心的 VOB 关系对,其中宾语为名词的“创意”,且以该词为节点向左搜索为空,则返回“创意”的极性值并记为 $O(\text{noun})$,根据修饰词的极性算

法得到 $O(W_1)$。b)这时查找时发现连谓结构“vv”,因此也按照中心词来处理,寻找“的”字结构,其中形容词为“唯美”的极性值记为 $O(adj_1)$;再寻找以该动词为中心的 VOB 关系对,宾语不具有极性值,查找到以该动词为中心的“CMP”关系对,其中补语为形容词“愉悦”,记为 $O(adj_2)$。

通过计算分别得到具体的极性值 $O(W_1)=0.862$, $O(adj_1)=0.743$, $O(adj_2)=0.697$。需要注意的是在续配句中计算句子极性时,权值因子按照与其中心词的依存语法距离来计算。因此两个句子的极性值分别为 $O(S_1)=0.784$, $O(S_2)=0.725$。因为不同的句子情感倾向的描述主题是不尽相同的,从而导致两个句子的情感倾向表达也不尽相同,所以需要明确评价对象是什么,即主题的确定;除此之外,进行否定模式匹配后若出现否定共享时,续配句的极性判别应该怎样进行修正。

2.3　否定模式匹配与主题抽取

从否定管辖范围的研究发现,当原配句为否定句,续配句为肯定句时,否定词不被共享。因此否定模式匹配有两种情况:

a)若否定词或扩展否定词不被共享,按照句子极性算法步骤找出每个句子(即原配句和续配句)中的评价词(即那些具有语义倾向的词汇),然后找出与评价词依存语法距离最近的 SBV、VOB 结构,对两者分别进行语义角色标注,将对应的角色映射为评价对象。例如句子:“影片情节没有扣人心弦,只是配乐很优美。”在查找到“扣人心弦”“优美”两个评价词以及计算出极性后,对其进行语义角色标注,得到“影片情节”为 AGR0,“配乐”为 AGR1 是相对应主题。

b)若否定词或扩展否定词被共享,不仅同样需要同上确定主题词的情况外,还需要对续配句子的极性判别方法进行改进。具体修正方法为:在对句子的评价词进行选择后,在续配句中增加原配句中否定修饰词的极性来计算句子极性值。例如句子:“文艺电影不宜运用突兀的叙事结构,夸张繁杂的拍摄手法。”在对续配句进行极性推断的时候增加修饰词“不宜”的极性,同时依照相同的依存语法距离作为权值因子进行计算。除此之外,语义角色标注后的主题也分别对应为“叙事结构”AGR0 与“拍摄手法”AGR1。

这两种情况都是将否定模式匹配与依存树结合在一起进行分析研究,虽然在实际语料中否定词不被共享的句子占多数,但是从句子语义倾向的研究来看,否定的共享意义在于对句子情感倾向的极性转移,它不仅直接影响到句子极性的判断,而且也是对评价对象褒贬感情的一个度量值。

3　实验结果分析

实验数据来源于专业影评网对电影《贫民窟的百万富翁》的评论文章,随机选取正面文章和负面文章各 84 篇作为影评测试集。剔除文章中客观描述性语句,最终确定 623 个褒奖句和 450 个贬义句为实验语料句。本文按照基本否定词、扩展否定词两类进行分析测评,首先对否定共享的情况给出分析结果。

从表 3 中可以看出,扩展的否定词共享在实际语料中比例明显增多,有利于进行句子极性的判别。

表 3　否定共享分析结果

否定词类型	不共享否定词	共享否定词	不共享否定词的比例
基本否定词	1 014 句	59 句	94.50%
扩展否定词	952 句	121 句	88.72%

在对句子极性进行分析计算后得到其情感判别的结果为褒义(正面)评价和贬义(反面)评价两种。由于不仅要抽取句子主题,还需对评价的倾向性进行评测,采用准确率、召回率以及 F 值作为评测标准。实验采用否定模式匹配与不考虑否定共享两种方法进行,这里根据哈尔滨工业大学 IR-Lab 的语法分析系统进行分词和语法分析,阈值设置为 0,可将结果分类为褒义或者贬义。同时为了说明否定共享在句子极性判别中的作用,将本文方法与文献[4]的方法进行实验比较。

由表 4 的结果可知,文本特征向量距离方法的准确率较低,因为该方法在处理搭配和修饰词方面只简单地依靠语法距离来判定是否加重语气,未考虑词语修饰极性而造成判断出现偏颇。否定共享判别方法由于在依存语法距离与修饰词极性的判断前提下,增加了否定匹配,对句子情感倾向的判断准确性比较理想。在文本内容繁多的情况下,噪声数据会急剧增多,以及对未知的句法结构关系判定不准确从而对分类器的判断有一定的干扰,所以会造成一定的判别误差。尽管如此,本文运用该方法对句子极性的判断有了较准确的结果,符合了大部分主观认知。

表 4　实验分析结果　%

实验方法	准确率	召回率	F 值
文本特征向量距离方法	68.01	78.04	61.31
否定模式匹配判别方法	75.70	74.52	72.29

4　结束语

本文在基于知网的基础之上对中文情感倾向作出一些研究,否定共享以及扩展后的应用在很大程度上对句子情感倾向度的分析起到了积极的作用,在大量的语料评审中发现,有些对情感描述有影响的因素未考虑,如转折词、感叹词等,甚至根据某些特有的句法结构对句子极性的影响来进行分析研究。在句子极性推断时,只是对主要的几种句法结构进行计算,因此造成句子在分析时,遗漏一些对情感判断的句法结构从而使结果产生误差,导致与主观判断不符的实际情况。由于主要实验语料都是通过网站上挑选的影评,繁杂的描述内容会影响对评价对象的主观态度,并且各类评价词对于领域知识方面有较强的依赖性,从而影响到对整个句子情感倾向的判断,可以考虑受限领域的词汇倾向。

参考文献:

[1] TURNEY P, LITTMAN M. Measuring praise and criticism: inference of semantic orientation from association [J]. ACM Trans on Information Systems, 2003, 21(4):315-346.

[2] 朱嫣岚,闵锦,周雅倩,等.基于 HowNet 的词汇语义倾向计算[J].中文信息学报,2006, 20(1):14-20.

[3] 蔡健平,王琳琳,林世平.基于机器学习的词语和句子极性分析[C]//中国人工智能学会第 12 届全国学术年会论文集:上集.北京:北京邮电大学出版社, 2007.

[4] 熊德兰,王爽,张泊平.基于 HowNet 的句子褒贬倾向性计算[C]//中国人工智能学会第 12 届全国学术年会论文集:上集.北京:北京邮电大学出版社,2007.

[5] 姚天　,娄德成.汉语语句主题语义倾向分析方法研究[J].中文信息学报,2007, 21(5): 75-78.

[6] 董振东,董强.知网简介[EB/OL].(2006). http://www.keenage.com/zhiwang/c_zhiwang.html.

[7] 张瑞朋,宋柔.否定词跨标点句管辖的判断[J].中文信息学报,2007, 21(5): 134-135.

[8] 张桂宾.相对程度副词与绝对程度副词[J].华东师范大学学报,1997,2(1):92-96.