

汉语句子语义极性分析和观点抽取方法的研究

娄德成,姚天昉

(上海交通大学 计算机科学与工程系,上海 200240)

lou-de@sjtu.edu.cn; yao-tf@cs.sjtu.edu.cn)

摘要:利用自然语言处理技术,对中文网络评论语句进行语义极性分析和观点抽取。提出了计算词语上下文极性的算法,并且分析主题和修饰成分的匹配关系。最后与手工标注结果进行比较,通过实验证明了该算法的合理性和有效性。

关键词:语义极性;网络评论;句法分析

中图分类号: TP391.1 **文献标识码:** A

Semantic polarity analysis and opinion mining on Chinese review sentences

LOU De-cheng, YAO Tian-fang

(Department of Computer Science & Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: By using natural language management technology, Chinese Web review sentences were semantic studied and opinion mined. A new algorithm for computing the contextual polarity of polar words was proposed, and a framework was introduced for topic identification and feature extraction. The framework provided an innovative solution for the association of extracted opinion to its specific topic. The experiment results show that the algorithm is both reasonable and effective compared with the results of manual annotation.

Key words: sentimental polarity; Web reviews; syntactical parser

0 引言

语义极性倾向分析和观点抽取是指利用计算机技术自动分析带有观点信息的句子或文档,从而提取出用户感兴趣的主体或特征,并分析其语义极性倾向(褒义、贬义或中性)和强度。语义极性倾向的研究工作,从最初对极性词语或极性成分的识别,已经发展到对句子层次和文档层次的分析。

目前已知的研究,要么是针对新闻报道等符合语法规则的语料^[1,2],要么是针对分类明确、格式固定的论坛评论^[3],主要采用的是统计学方法。此外,中文领域的研究还没有考虑到主题或特征词的抽取问题,没有研究主语与极性成分的修饰和匹配关系。

文献[1]利用不同的极性成分组合从句子层次和文档层次进行语义极性分析。所选用的极性成分包括单词、Bigrams、Trigrams 以及极性标注。在此基础上,利用统计学的方法,识别句子的极性,计算文档的极性。

文献[2]对北京、上海、香港、台北 4 个城市的 600 篇新闻报道进行极性分析,从文档层次分析几位著名政界领导人的声誉褒贬。采用的方法是计算极性成分在文章中出现的广度、密度以及强度。并且给出了与手工标注相比较的结果。

然而,新闻报道的评论是由某一个或几个作者发表,没有读者的参与和反馈,无法全面反映广大消费者的观点;统计学方法虽然对结构简单的句子可以取得较好的结果,但是,对于结构较复杂的语句,统计方法在主题抽取、语义极性倾向和极性强度三个方面,都具有很大的不确定性,不能取得理想的效果。例如,句子中含有多个主题,含有否定前缀,以及含有语气强调成分。

为了解决这些问题,不能完全依赖统计学方法,需要研究新的算法,借助自然语言处理技术,对句子的成分和结构进行语法分析,提高语义极性分析的可靠性和准确率。

基于上述考虑,本文改变以往的研究思路,利用自然语言处理技术,通过对句子中词语的依存关系进行深层分析,研究关键成分的依存修饰关系,计算极性词的上下文极性,并且对句子讨论的主题进行观点抽取,对中文语句进行极性分析和观点抽取。并且,我们分析的语料不是新闻报道等规范语料,而是网络上的自由评论语句。

与新闻报道相比,网络评论具有以下显著特点:信息容量大;信息来源丰富;语言表达形式自由,语句成分不完整、语序紊乱、标点符号使用不规范。

1 极性词

极性词是指句子中带有感情色彩的词语。极性词是判断句子极性和文档极性的前提条件和根本依据。早期研究中,Hong Yu 等^[1]挑选出若干极性较强的形容词构建一个种子集合,通过计算新词语与种子集合中某些词同时出现的概率,判断新词的语义极性倾向。

然而,此类算法有两个不足之处:首先,计算复杂度高,对于每个新词都需要花费大量时间计算其极性,无法推广到实际应用中;其次,没有考虑种子词的强度,因此也就无法为计算新词的强度提供合理的依据。我们在此基础上,改进思路和算法:

第一步,引进 HowNet。HowNet 包含 6564 个词条。为每个词条添加两个重要的属性:极性和强度。极性包含 3 个分类:Positive, Negative 和 Neutral。强度分为 5 个等级:±1, ±0.5, 0。

收稿日期:2006-05-30;修订日期:2006-07-18

作者简介:娄德成(1981-),男,江苏徐州人,硕士研究生,主要研究方向:自然语言处理;姚天昉(1957-),男,山西太谷人,副教授,主要研究方向:意见挖掘、信息抽取、机器学习。

由 4 个人独立完成极性和强度的标注,然后取平均值。

这样就保证了绝大多数常用的极性词可以在 HowNet 中直接检索到,降低了时间复杂度。

第二步,对 2454 篇有关汽车的网络评论文章进行分词,从中选出所有的极性词,共 1405 个。由 4 个人独立地进行极性分类,从中取分类一致的结果,而忽略有歧义的词语,从中得到 631 个正极性(positive)的词语,394 个负极性(negative)的词语。对这些极性词,进行强度标注,方式与标注 HowNet 相同。得到正极性种子集合 PositiveSet 和负极性种子集合 NegativeSet。

第三步,对于 HowNet 和两个种子集合中都检索不到的词语,即新词,利用 SO-PMI 算法^[4]计算其极性。

互信息(Pointwise Mutual Information)概率定义:

$$PMI(word_1, word_2) = \log \frac{P(word_1 \& word_2)}{P(word_1) P(word_2)}$$

其中, $P(word_1 \& word_2)$ 表示 $word_1$ 和 $word_2$ 同时出现的概率。

依此为基础,计算一个新的词语与种子词的互信息概率,可以得到该词的极性:

$$SO-PMI(word) =$$

$$\sum_{pword \in Pset} PMI(word, pword) - \sum_{nword \in Nset} PMI(word, nword)$$

由于种子集合中包含种子词的极性,在计算一个新词 $word$ 的 PMI 时,将与之同时出现的种子词的强度赋予 $word$ 。这样,与极性强的词同时出现的词,也将获得比较强的极性。

2 词语的极性分析和句子的观点抽取

2.1 句法分析器

本文利用 DeParser 对句子进行句法分析。DeParser 是由哈尔滨工业大学信息检索实验室开发的汉语句法分析器。其输出结果中,用三行表示一个句子,第一行是输入的原句;第二行对句子进行了分词和词性标注,并在句子中每个词及词性的前面加上序号,句子的末尾增加一个句尾标志“<EOS>”,由其支配全句的核心词;第三行是句子中词与词之间的依存关系。依存关系中,每个关系以一个依存对表示,依存对中的第一个词是核心词,支配第二个词,如:“[2]公司_[1]我(ATT)”这个依存对表示“我”和“公司”存在依存关系 ATT,其中,“公司”是这个关系的核心成分,“我”依存于“公司”。

例 1 后备箱没有内开装置,我觉得很不安全,也很不爽!

分词结果:

[1]后备箱/n [2]没有/vg [3]内/nd [4]开/vg [5]装置/n [6],/wp [7]我/rh [8]觉得/vg [9]很/d [10]不/d [11]安全/a [12],/wp [13]也/d [14]很/d [15]不爽/vg [16]! /wp [17]<EOS>/<EOS>

依存关系对:

[8]觉得_[7]我(SBV) [4]开_[5]装置(VOB)
[15]不爽_[14]很(ADV) [15]不爽_[13]也(ADV)
[11]安全_[10]不(ADV) [11]安全_[9]很(ADV)
[2]没有_[1]后备箱(SBV) [3]内_[2]没有(ATT)
[4]开_[3]内(ADV) [8]觉得_[11]安全(VOB)
[4]开_[8]觉得(VV) [4]开_[15]不爽(VV)
[18]<EOS>_[4]开(HED)

括号中的符号表示两个词之间的修饰关系(详细定义见表 1)。对于关系对“[8]觉得_[7]我(SBV)”,我们称“[8]觉得”为

出现在关系对左侧,“[7]我(SBV)”为出现在关系对的右侧。

表 1 DeParser 词性和依存关系标记注释

词性与 依存关系	标记	词性与 依存关系	标记
定中关系	ATT(attribute)	连动结构	VV(verb-verb)
数量关系	QUN(quantity)	同位关系	APP(appositive)
并列关系	COO(coordinate)	前附加关系	LAD(left adjunct)
后附加关系	RAD(right adjunct)	动宾关系	VOB(verb-object)
介宾关系	POB(preposition-object)	主谓关系	SBV(subject-verb)
比拟关系	SIM(similarity)	核心	HED(head)
关连接构	CNJ(conjunctive)	语态结构	MT(mood-tense)
独立结构	IS(independent structure)	状中结构	ADV(adverbial)
动补结构	CMP(complement)	“的”字结构	DE
“地”字结构	DI	“得”字结构	DEI
“把”字结构	BA	“被”字结构	BEI
独立分句	IC(independent clause)	依存分句	DC(dependent clause)

汉语句法分析器对于网络自由评论的语句的分析效果不是很理想,而对于稍微复杂的句子,分析效果更具有很大的不稳定性。尤其是当句子太长,结构较复杂时,依存关系会出现很多错误。

Deparser 的依存分析结果是严重依赖于其切词效果的,但是它还不能正确识别一些专有名词,比如,蒙迪欧,宝莱等。而海量智能计算技术研究中心开发的中文分词器 HLSeg 可以添加用户词典,因此,本文运用了“替换映射”的方法:

- 1) 利用 HLSeg 切分句子,识别出其中的特征词;
- 2) 用自定义的特殊符号替换句子中的特征词;
- 3) 利用 Deparser 处理替换后的句子,得到句法分析结果;
- 4) 将替换符号还原为原句中的特征词或主题词。

2.2 特征词和主题词的抽取

主题是句子评论的对象,在实际应用中,用户往往关注某类产品的特征,一个特征词需要满足以下三个条件之一^[5]: 1) 是给定主题的一部分; 2) 是给定主题的一个属性; 3) 是给定主题的部分的一个属性。

在研究中,我们选取了三类特征:1) 汽车的品牌和型号,如蒙迪欧、奥迪 A6 等; 2) 汽车的机械部件,如车窗、制动系统等; 3) 汽车的性能指标,如速度、外观等。

但是,如果一个句子中同时出现多个或多类特征词,就需要明确它们之间的从属关系。Ontology 可以解决这一问题。利用 Ontology 针对上述三类特征分别定义一个层次式分类体系,以树的结构表示。通过遍历树,利用父子节点的关系,就可以知道特征词之间的从属关系。

2.3 计算极性词的上下文极性

一个词的极性分为两种情况,一种是原极性,这种极性可以利用 HowNet 以及 PMI 算法得到;另一种是上下文极性。后者可以出现在两种语境中:一是句子含有否定前缀,例如,“不漂亮”;二是句子含有强调前缀,例如,“非常漂亮”。如果采用简单的统计方法计算语义极性,忽略词语的上下文语境,就会导致两种后果:1) 极性倾向判断错误,这是由于没有考虑否定前缀; 2) 虽然极性倾向判断正确,但是强度不够准确,这是由于没有考虑强调前缀。

为了解决这两个问题,本文定义两个字典:否定词字典 Privative Dictionary 和强调词字典 Intensifier Dictionary。后者中,为每个强调词定义一个表示程度强弱的值,如:很 12,非常 13。

对一个句子进行句法分析后,通过研究词与词之间的依

存关系,可以找到否定前缀和强调前缀,从而计算词的上下文极性。出现极性词的关系对有三种:ADV(状中结构)、DE(的字结构)和VOB(动宾结构)。

计算词语上下文极性的算法如下:

1) 利用 HowNet 检索或利用 SO-PMI 算法计算出新词 *word* 的原极性 $PriorPolarity(word)$ 。如果 $PriorPolarity(word) \neq 0$,则进行第2)步。

2) 记 *word* 的上下文极性为 $ContextualPolarity(word)$,且令 $ContextualPolarity(word) = PriorPolarity(word)$ 。

3) 搜索所有依存关系对的左侧,找到含有词 *word* 的关系对 $RelationPair_1$:

① 如果 $RelationPair_1$ 为 ADV(状中结构),其中副词为 adverb,则继续追溯,找到含有 adverb 的关系对 $RelationPair_2$ 。如果 $RelationPair_2$ 中含有修饰 adverb 的强调词或否定词,则:

如果是强调词 $intensifier_1$,其强度为 $Strength([intensifier_1])$,则令 $ContextualPolarity(word) * = Strength(intensifier_1)$ 。

如果是否定词 $privative$,则令 $ContextualPolarity(word) = -ContextualPolarity(word)/2$ 。

这里之所以除以2,是因为否定结构带有语气弱化倾向,如果直接对极性词取反向语义,则不合理。例如,对于“不漂亮”这一短语,如果“漂亮”的原极性为+2,则其反向语义为-2,然而,事实上,“不漂亮”并不表示“丑陋”等强烈情感,不是强烈否定,而是弱化了否定,故除以2。

② 如果为 ATT(定中关系),修饰 *word* 的词语为“的”字,则追溯到含有该“的”字的关系对 $RelationPair_3$,其中的强调词为 $intensifier_2$,则 $ContextualPolarity(word) * = Strength(intensifier_2)$ 。

4) 在所有依存关系对的右侧找到词 *word*:

① 如果依存关系为 DE(“的”字结构),修饰“的”字的词语为 *adj*,则追溯到含有 *adj* 的关系对 $RelationPair_3$,如果 $RelationPair_3$ 中修饰 *adj* 的词语是强调词或否定词,则按3)①的方法处理。

② 如果依存关系为 VOB(动宾结构),宾语为 *object*, $ContextualPolarity(word) = ContextualPolarity(object)$ 。

例2 我觉得马自达6在节油方面真的无法挑剔,我深有体会!

极性词分析结果:

否定关系:挑剔 无法

强调关系:挑剔 真

挑剔: $-0.51 + 1.0$ (-0.5 为原极性, $+1$ 为上下文极性)

2.4 句子的观点抽取

对于一个带有观点的句子,仅仅知道观点的极性褒贬是不够的,用户更希望知道观点的持有者是谁,观点讨论的主题或特征是什么。统计学方法是无法可靠地解决这一问题的,需要对句子进行句法分析。

通过对依存关系的分析,发现 SBV 结构(主谓结构)可以提供主语和谓语的修饰关系等信息。SBV 结构中,主语要么是观点的持有者,要么是讨论的主题或特征。而谓语部分的词性可能有两种情况,一种是形容词,另一种是动词。

观点抽取的 SBV 算法如下:

1) 找到所有含有 SBV 结构的关系对,对每个关系对,记主语为 *subject*,谓语为 *predicate*,执行2)。

2) 如果 $ContextualPolarity(predicate) \neq 0$,一般这种情

况下,谓语是形容词,则 $ContextualPolarity(subject) = ContextualPolarity(predicate)$ 。结束。否则,表示谓语为动词,则执行3)。

3) 如果 $ContextualPolarity(predicate) \neq 0$,则 $ContextualPolarity(subject) = ContextualPolarity(predicate)$;否则,执行4)。

4) 找到含有该动词的 VOB(动宾结构)关系对:

① 如果宾语是形容词 $adjective_1$, $ContextualPolarity(predicate) = ContextualPolarity(adjective_1)$ 。

② 如果宾语是名词 *noun*,则找到含有 *noun* 的 DE 关系对,其中形容词为 $adjective_2$, $ContextualPolarity(predicate) = ContextualPolarity(adjective_2)$ 。

5) 找到含有 *predicate* 的 ADV 关系对,其中的形容词为 $adjective_3$ 的上下文极性赋予谓语动词。

6) $ContextualPolarity(subject) = ContextualPolarity(predicate)$ 。

虽然利用上述算法,根据 SBV 关系对可以找到句子的主语,但是主语并不一定是讨论的主题。例如,“我也不喜欢蒙迪欧单调的内饰。”主语是“我”,而讨论的主题是“内饰”,很显然,我们更关心说话者对内饰的观点。因此需要改进算法。

此外,由于一个长句中会包含子句,而且,网络评论中断句不规范,有时候一个句子太长。而 SBV 分析算法可以覆盖的范围有限。

例3 蒙迪欧就显得比较沉闷,满眼深色的内饰稳重有余却又显得单调。

依存关系对如下:

[5]沉闷[4]比较(ADV)	[9]的[8]深色(DE)
[3]显得[2]就(ADV)	[15]显得[14]又(ADV)
[15]显得[13]却(ADV)	[3]显得[1]F0(SBV)
[15]显得[16]单调(VOB)	[12]有余[15]显得(VOB)
[9]的[7]满眼(DE)	[10]F9[9]的(ATT)(注1)
[3]显得[5]沉闷(VOB)	[12]有余[11]稳重(ADV)
[12]有余[10]F9(SBV)	[3]显得[12]有余(VV)
[18]<EOS>[3]显得(HED)	

注1 利用“替换影射”方法将“蒙迪欧”替换为 F9,利用 SBV 分析算法,“蒙迪欧”的极性强度为 -1.00 ,“内饰”的极性强度为 $+0.42$ 。

可知,算法对于“内饰”的极性分析是不准确的,原因是 SBV 算法分析时,遗漏了一个负极性词“单调”。因此,需要对 SBV 分析算法进行补充。

SBV 算法的补充算法如下:

1) 对每个利用 SBV 算法分析时利用到的主语,如果是主题词(或特征词),则记录下来;对于利用到的极性词,也记录下来。

2) 利用 SBV 分析算法后,继续查找整个句子,找到没有利用过的主题词 *UnusedTopic*,找到包含 *UnusedTopic* 的 ATT(定中结构)关系对,对于 ATT 关系对中出现的词,查找所有关系对,直至找到极性词 $adjective_4$,将 $adjective_4$ 的上下文极性赋予 *UnusedTopic*。

3) 记录以上所有算法步骤中(包括 SBV 分析算法)使用过的极性词,查找没有使用过的极性词 *UnusedPolar*。

∴ 如果 *UnusedPolar* 是一个修饰主题词的前缀,那么在2)中就已经利用到了。

∴ 所以该极性词不可能是前缀词。

因此,向前查找最邻近的主题词 *Topic*,将 *ContextualPolarity*(*UnusedPolar*)作为调整参数,调整 *Topic* 的极性。

运用了调整算法之后,“蒙迪欧”的极性 & 强度为 -1.00 ,“内饰”的极性 & 强度为 -0.46 。

3 实验

3.1 网络自由评论句子

实验从新浪汽车论坛(<http://forum.auto.sina.com.cn>)有关汽车评论的主题帖中,随机选取 100 个带有观点性的句子。所谓观点性的句子是指句子中至少包含一个极性词。

3.2 词语的上下文极性

句法分析的一个重要作用就是可以找到一个词的修饰前缀,包括否定前缀和强调前缀。实验对 100 个句子中的 209 个极性词进行分析,并与手工标注的结果进行比较,结果如表 2。

表 2 极性词的原极性和上下文极性实验结果

极性	算法识别	手工标注	准确率(%)	召回率(%)	
原极性	正	155	165	96.77	93.94
	负	29	44	96.55	65.91
前缀	强调	55	64	92.73	85.94
	否定	19	24	94.74	79.17
上下文	正	140	157	93.57	89.17
极性	负	44	52	79.55	84.62

结果表明,句子含有强调前缀的比例为 30.62% (64/209);含有否定前缀的比例为 11.48% (24/209)。这就表明,如果仅采用统计学方法,简单地统计极性词的个数,依此为根据判断句子的极性,在约 12% 的情况下,会将语义极性方向判断错误;即使极性方向判断正确,也有高达 31% 的情况,对极性的强度判断失真。这就进一步证明了单纯统计学方法的不可靠性,以及本文算法的必要性。

对结果进一步分析,负极性词的召回率偏低(65.91%),原因是算法无法判断词的动态极性。

例 4 车漆方面,我不去说车漆是厚是薄,就侧光看车漆表面像橘子皮一样,不平滑。

例 5 马自达 6 的特点就是发动机声音大、风噪大、路噪大。

其中,例 4 中的“厚”和“薄”,以及例 5 中的“大”无法预定义其原极性。

另外,否定前缀的召回率也不高(79.17%),因此影响了上下文极性的准确率,特别是负极性的准确率(79.55%)。

例 6 内饰用材太节约。

算法分析结果为:

强调关系:节约 太

节约: $+0.51 + 1.5$

算法认为“太”为强调词,强度为 3,而节约的原极性为 $+0.5$,所以,节约的上下文极性为 $+0.5 \times 3 = +1.5$ 。而事实上,这里“太”表示负极性。

3.3 观点抽取

首先,对 100 个带有观点信息的句子进行手工标注,目标是选出讨论的主题,并对每个主题进行极性标注。从以下两个方面,将算法识别结果和手工标注结果作比较:1)识别出的主题;2)主题和修饰成分的对对应关系。实验结果见表 3。

极性词识别的准确率和召回率都会影响观点抽取的效果。然而,通过对每个句子的分析,发现 SBV 分析算法能抽

取到的主题词(特征词)仅占 25.74% (61/237),而通过手工标注得到的结果表明,主语是主题词(特征词)的概率为 38.82% (92/237)。由此可见,仅仅依靠主谓结构是无法抽取全部主题词的。同时,分析得知,切词效果也制约着句法分析的效果。此外,主题和修饰成分组合的识别之所以效果不理想,是由于句法分析时,对句子的结构分析错误,即依存关系误差较严重。主要原因是网络评论语句结构不规范,许多语句成分省略较严重。此外,算法没有处理语气问题。

例 7 领取 2.0MFI 的发动机据说比较差,是吗?

表 3 主题和修饰成分组合实验结果

算法	算法识别	手工标注	准确率(%)	召回率(%)
SBV 算法	61	92	63.38	66.30
补充算法	109	145	74.31	75.17
综合	170	237	75.88	71.73

4 结语

本文提出了汉语自由评论语句语义极性分析的新算法,主要解决了两个问题:一是利用依存关系判断极性词的上下文极性,这对于利用统计方法解决句子和文档的层次的极性是一个较大的改进;二是抽取句子讨论的主题和特征词,解决主题和修饰成分的对对应关系问题。通过与手工标注的结果相比较的实验表明,对于第一个问题的解决效果比较好,这表明尽管网络自由评论(是一种极其口语化的表达形式)的整个句子结构并不规范,但是局部(比如极性短语部分)表达还是比较规范的,因此为语义极性分析提供了突破口。

由于自由评论句子的不规范性,本文的工作仅利用了浅层的句法分析,因而,主题和极性成分的识别算法依然有待于进一步的改进。

致谢 感谢哈尔滨工业大学信息检索实验室提供的汉语句法分析器 Deparser,以及海量智能计算技术研究中心开发的中文分词器 HLSeg,使我们的研究工作得以顺利进行。感谢上海交通大学中德语言技术联合实验室对本研究的资助。

参考文献:

- [1] YU H, HATZIVASSILOPOULOS V. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences[A]. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing[C]. Sapporo, Japan, 2003. 129 - 136.
- [2] TSOU BKY, YUEN RWM. Polarity classification of celebrity coverage in the Chinese press[A]. International Conference on Intelligence Analysis[C]. Virginia, USA, 2005.
- [3] LIU B, HU M. Opinion observer: analyzing and comparing opinions on the web[A]. International World Wide Web Conference Committee (IW3C2)[C]. Chiba, Japan, 2005.
- [4] TURNEY PD, LITTMAN ML. Measuring praise and criticism: inference of semantic orientation from association[J]. ACM Transactions on Information System, 2003, 21(4): 315 - 346.
- [5] YI J, NIBLACK W. Sentiment mining in WebFountain[A]. Proceedings of the 21st International Conference on Data Engineering (ICDE 2005)[C]. Washington, DC, USA: IEEE Computer Society Press, 2005. 1073 - 1083.
- [6] TURNEY PD. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[A]. Proceedings of the Association for Computational Linguistics 40th Anniversary Meeting[C]. Philadelphia, PA, USA, 2002. 417 - 424.