

中文微博评价对象识别研究

张景,牛耘

(南京航空航天大学 计算机科学与技术学院,江苏 南京 210016)

摘要:旨在对中文微博文本的句子中评价对象进行识别。评价对象识别是指识别出评论中情感表达所针对的对象,进行评价对象的识别有助于对事件发展状况进行监控管理。目前,针对中文微博领域评价对象识别的研究较少。由于微博文本的句子简短、语言表达不够规范且表达的观点缺少带情感倾向性的词语(评价词),因而传统的通过评价词来找到评价对象的方法不适用于微博文本。利用词性分析提取和过滤评价对象候选词,并结合语义分析对句子中的候选词进行分类,基于相似的句子有着相似的评价对象的假设,采用候选词的相似性迭代算法识别中文微博文本句子中的评价对象。实验结果表明,通过深入分析微博文本的语言特征提出的方法,提高了对评价对象识别的精度。

关键词:评价对象;候选词提取;语义分析;相似性计算

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2017)01-0006-05

doi:10.3969/j.issn.1673-629X.2017.01.002

Research on Opinion Target Extraction in Chinese Microblogs

ZHANG Jing, NIU Yun

(School of Computer Science and Technology, Nanjing University of Aeronautics and
Astronautics, Nanjing 210016, China)

Abstract: It focuses on extracting opinion targets in Chinese microblogs. Opinion target extraction aims to find the object to which the opinion is expressed, helping to monitor and control the development of events. At present, there are few researches on opinion target extraction in Chinese microblogs. Due to short text span, colloquial writing style and the lack of words with emotional tendency (opinion words) in Microblogs, the traditional approaches rely on opinion words are not suitable for microblogs. In this paper, we use the part-of-speech analysis to extract and filter candidate words, and combine semantic analysis to classify candidate words. Based on the assumption that similar messages may have similar opinion targets, a similarity iterative algorithm of candidate words is proposed to extract opinion targets. Experimental results show that by deeply analyzing language features of Microblogs, the proposed method has improved high accuracy.

Key words: opinion target; candidate extraction; semantic analysis; similarity calculation

0 引言

随着微博、博客、论坛等社交媒体的蓬勃发展,越来越多的用户参与到社交网络平台内容建设的过程中,互联网上产生了大量带有情感色彩的评论信息。为了能够对它们更好地进行加工汇总,找出富有价值的信息,情感分析作为自然语言处理领域中的一个热点问题应运而生。目前,越来越多的研究者将目光转向更细粒度的情感分析任务,如评价对象的识别,具体表现为识别出某段评论中情感表达所针对的对象。评价对象识别的研究能够在某个社会热点事件、电影、品牌等话题方面发现不同用户所关心讨论的各个主题,

更加全面地了解公众对于一个话题意见表达的方方面面,有助于对事件发展状况进行监控管理,甚至对事件发展状况进行预测。因而,评价对象识别的研究具有一定的商业价值及应用前景。

目前,关于评价对象识别的研究大部分集中于新闻、产品或电影评论领域。传统的研究方法中,大多先将评价对象限定在名词或名词性短语的范畴内,进而通过带有情感倾向性的词语(评价词)来帮助对评价对象的识别。然而,在中文微博领域关于评价对象识别的研究很少。一方面,微博文本具有口语化程度强、表达情感强烈而理性评价淡化、观点表达隐晦、评价对

收稿日期:2016-03-11

修回日期:2016-06-15

网络出版时间:2017-01-04

基金项目:国家自然科学基金资助项目(61202132)

作者简介:张景(1991-),女,硕士研究生,研究方向为自然语言处理;牛耘,副教授,CCF会员,研究方向为自然语言处理。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20170104.1028.054.html>

象在句子中不直接出现、语言不够规范等特点^[1]。另一方面,微博文本句子简短、口语色彩浓重,表达观点的句子并不总是包含评价词。故传统的利用评价词对评价对象识别的方法并不适用于中文微博文本。所以,针对中文微博文本评价对象识别的研究具有很大的挑战和意义。

通过深入分析微博文本的语言特征,结合词性及语义分析提取和过滤评价对象候选词,基于相似的句子有着相似的评价对象的假设,提出了一种候选词的相似性迭代算法来识别评价对象。实验结果表明,该方法提高了对评价对象识别的精度。

1 相关工作

评价对象的抽取是识别出评论中情感表达所面向的对象。现有的研究大部分集中于产品评价领域中评价对象的抽取。主要有两种基本方法:非监督和有监督的方法。

1.1 基于非监督的评价对象识别

在非监督的学习方法中,倪茂树等^[2]使用关联规则挖掘的方法找出频繁出现的候选评价对象,继而使用两种剪枝方法去除错误样例。随之发现在评论文本中,人们评论相同的评价对象时会使用相同的评价词。故通常根据评价对象和评价词之间的依赖关系迭代地进行引导提取。Qiu 等^[3]利用依存句法分析制定了八条启发式的语法规则并采用双传播方法迭代地提取出评价对象和评价词。Zhang 等^[4]拓展了 Qiu 的方法,增加了两条规则(如部分-整体和否定规则),来增加召回率,采用 HITS 算法对评价对象候选词打分并排序,提高了准确度。为了降低句法分析工具带来的误差,Liu 等^[5-6]通过翻译模型中的词对齐方法来捕捉评价对象和评价对象间的关系。此后,Liu 等^[7]进一步进行研究,考虑了候选词间的语义关系并结合句法关系提取评价对象和评价词。高磊^[8]、文坤梅等^[9]通过对微博文本内容进行句法依赖关系分析结合情感词典得到成对的<情感词,情感对象>关系,进而抽取情感对象。

1.2 基于有监督的评价对象识别

在有监督的学习方法中,评价对象的抽取可以看作是信息抽取问题中的一个特例。信息抽取的研究中提出了很多监督学习算法。其中主流的方法有隐马尔可夫模型、支持向量机和条件随机场。由于这些方法是监督学习技术,所以事先需要有标记数据进行训练。Jakob 等^[10]基于 CRF 模型将评价对象识别的问题建模成信息抽取的任务。其使用的特征有词、词性、最短依赖路径、词距离,验证了此模型在不同领域的评价对象识别任务中取得了较好的结果。王荣洋等^[11]通过大量实验,系统地比较研究了各类特征在基于 CRFs

的评价对象识别系统中的选择对性能的影响,将特征归纳为词法、语法、相对位置、语义四大类别。实验结果表明,重点引入的语义角色标注特征对评价对象识别起到了很好的指示作用。郝志峰等^[12]在传统的 CRF 序列标记模型上增加情感对象的全局节点,有效地结合上下文信息、句法依赖以及情感词典,从而可以识别出微博中的情感对象。

2 中文微博评价对象的识别

与产品评价领域不同,微博文本通常围绕某一话题标签表达情感、阐发意见并进行讨论,因而形成了带有一个鲜明主题的话题型微博群。话题型微博由话题标签(hashtag)和正文(content)两部分组成,如:“#90 后暴打老人#公德丧失!这是教育的失败”。其中,“#”之间即为话题标签,除话题标签外即是正文,由一条或多条句子构成。文中目标是识别出话题型微博文本观点句中的评价对象。首先,在一个话题下,结合词性及语义分析提取和过滤评价对象候选词。然后,基于相似的句子有着相似的评价对象的假设,利用候选词的相似性迭代算法对候选词打分。最后,对于每一个句子,提取出得分最高的候选词作为当前句子的评价对象。

2.1 评价对象候选词的提取及过滤

观察语料,发现评价对象大多以名词或名词性短语的形式存在,而且不同词间的语义差异有助于过滤句子中的评价对象。所以,利用词性信息和语义分析对评价对象候选词进行提取和过滤。

(1) 提取候选词。

利用中文分词工具 ICTCLAS 进行分词及词性标记后,针对微博正文中的每一个句子,提取出名词或名词性短语作为当前句子中的显性评价对象候选词。而对于没有显性评价对象候选词的句子,若其所在微博的前一个句子存在显性评价对象候选词,则以前一个句子中的显性评价对象候选词作为当前句子的隐性评价对象候选词。同时,对于微博中的所有句子,提取话题标签内的名词或名词性短语作为隐性评价对象候选词。其中,名词性短语只包含名词和汉字“的”,由定语和中心词构成。中心词是名词,其定语是名词或“的”字短语(如“中国的”)。在“中国/ns 人/n 的/ude1 尊严/n 何在/vi ? /ww”一句中,“中国人的尊严”作为名词性短语被提取出作为评价对象候选词。

(2) 过滤候选词。

观察语料发现,仅仅依靠词性提取候选词,则忽略了词类内部不同词之间的语义差异。有些词抛开上下文语境显然不能单独担任评价对象,这些词被称为评价对象绝缘词,如“详情、时候、大家”等。分析语料人

工筛选出 13 个词并结合周红照等^[13]筛选出的 81 个词共 94 个作为评价对象绝缘词对候选词进行过滤,去掉评价对象候选词中的评价对象绝缘词。

2.2 评价对象的识别

发现在一个话题下,相似的句子有着相似的评价对象,如表 1 所示。

表 1 同一话题下相似的句子

话题标签	相似的句子	评价对象
90 后当教授	1. 太厉害了吧……	90 后
	2. 90 后太厉害了!	
非军舰恶意撞击	1. 政府还是不够强硬。	政府
	2. 政府为何不能强硬一些?	

表中每个话题下的两个句子都是相似的,因为它们拥有相同的一个或数个词。如相似的句子:“太厉害了吧……”和“90 后太厉害了!”两句中有三个相同的词“太”、“厉害”、“了”。文中方法是通过计算句子间的相似性,对每一个句子中的评价对象候选词打分,得分最高的候选词作为当前句子的评价对象。

已有的研究中,Zhou 等^[14]也是根据相似性计算提出了非监督标签传播算法(Unsupervised Label Propagation, ULP)。通过句子间的相似性计算更新微博文本的句子中评价对象候选词分值,来确定句子中的评价对象。他们的方法存在以下不足之处:第一,Zhou 等没有从语义上对句子进行分析,忽略了词位置和词语搭配对识别评价对象的重要指示作用;第二,对于相似句子数很少的句子,候选词的分值在更新过程中变化不大,Zhou 等并没有对候选词初始分值的设定做深入的分析。所以,文中基于相似的句子有着相似的评价对象的假设,针对以上问题,提出了以下方法。

2.2.1 句子的向量表示

为了表示句子中的候选词成为评价对象的可能性大小,文中将微博正文中的句子用向量表示。其中,向量的每一维对应了一个候选词,每一维的权重表示该候选词成为评价对象的可能性大小。

令句子 v 中的候选词集合为 C_v , 则一个话题下的全部候选词集 $CT = \cup C_v$ 。统计候选词的个数 $M = |CT|$ 。句子的向量的每一维代表的候选词与集合 CT 中的候选词一一对应,句子 v 的向量记为 $Y_v \in R_+^{1 \times M}$, 表示如下:

$$(Y_v)_k = \begin{cases} w, & CT_k \in C_v \\ 0, & CT_k \notin C_v \end{cases} \quad (1)$$

其中, w 表示句子 v 中的候选词 CT_k 成为评价对象的可能性大小,不属于当前句子中的候选词对应的向量权值为 0。

发现评价对象候选词在句子中与不同词语的搭配

及不同的出现位置,影响了候选词成为评价对象的可能性不同。故先对一个话题下的候选词进行分类,帮助设置句子的向量的初始权重。

观察语料发现,与某些词语搭配及出现在句首与标点符号后的候选词,更有可能是当前句子的评价对象。首先,对以这两种方式出现的候选词做深入分析,如下:

(1)评价触发词之后的词。根据语用习惯,评价对象经常和一些特定的词语搭配且紧跟在这些词语之后,这些词往往是一个评价的触媒,称之为“评价触发词”,主要有以下四种类型。

· 连词。如:由于(连词)金基范版段誉(评价对象)太磕碜,所以显得张檬的王语嫣就不是太对不起观众了。

· 动词。如:我觉得(动词)这种行为(评价对象)太过分了!

· 副词。如:其实(副词)韩寒(评价对象)真的没什么文学天赋,只是长得好看而已。

· 话语标记词。话语标记词是由数个不同词性的一元词构成,有助于语篇的连贯性与条理性,并起到一定的指示作用。如:客观说(话语标记词)《魔境仙踪》(评价对象)很一般。

分析语料人工筛选出 12 个词并结合周红照等^[13]筛选出的 46 个词共 58 个作为评价触发词,记紧跟触发词之后的评价对象候选词为“搭配词”。

(2)句首及标点符号后的词。评价对象经常会出现以下两种位置:句首和标点符号之后。如:“360 真是有种攀龙附凤的感觉。”“刚还看了直播,不错,这小子有前途。”分析语料记出现在句首和标点符号之后的评价对象候选词为“位置词”。

其次,综合以上两种类型的候选词和句中候选词的提取方式,句子中的候选词分类如图 1 所示。

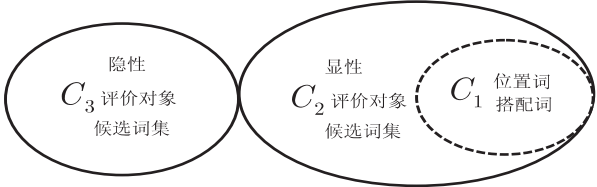


图 1 句子中的候选词分类

记 C_1 为句子中的位置词和搭配词集, C_2 为句子中显性评价对象候选词中除位置词和搭配词之外的词的集合, C_3 为句子中隐性评价对象候选词集。

对于句子 v , 文中根据候选词的分类,对句子的向量初始权重设置为:当 $CT_k \in C_1$ 时, $w = 1.5$;当 $CT_k \in C_2$ 时, $w = 1$;当 $CT_k \in C_3$ 时, $w = 0.5$ 。

通过对句子的向量的深入分析,发现当句子 v 的相似句子数少于等于 10 时,其向量权值在计算更新过

程中变化不大,故对向量初始权值进行重新设定以帮助找到正确的评价对象。认为除搭配词和位置词之外的集合中显性和隐性候选词的重要性程度一致,对于相似句子数少于等于 10 的句子 v 的向量初始权重设置为:当 $CT_k \in C_1$ 时, $w = 1.5$; 当 $CT_k \in C_2$ 或 $CT_k \in C_3$ 时, $w = 1$ 。

2.2.2 构造无向图

为了能够直观表示一个话题下微博正文中句子间的关系,为每一个话题构造了一个无向图 $G = \langle V, E, W \rangle$ 。其中,节点 $v \in V$ 表示微博正文中的一个句子。相似的两个句子之间相互连通构成一条边 $e \in E$,边上的权值表示两个句子间的相似度,故无向图中所有边上的权值构成一个相似性矩阵 W 。句子 u 和 v 间的相似度利用向量空间模型计算如下:

$$W_{uv} = \cos(\mathbf{T}_u, \mathbf{T}_v) = \frac{\mathbf{T}_u \cdot \mathbf{T}_v}{\|\mathbf{T}_u\| \cdot \|\mathbf{T}_v\|} \quad (2)$$

其中, \mathbf{T}_u 和 \mathbf{T}_v 分别表示句子 u 和 v 的词频向量。

2.2.3 评价对象的确定

由于一个句子中有一个或多个评价对象候选词,因此基于无向图 G 及相似的句子有着相似的评价对象的假设,可以得出相似的两个句子中实际为评价对象的候选词之间的相似度很高,所以可以通过计算句子中所有候选词间的相似性来帮助识别评价对象。

首先,定义候选词间的相似性。若两个评价对象候选词有相同的汉字,则认为这两个候选词是相似的。根据一个话题下所有候选词间的相似度构造候选词间的相似性矩阵 S 。候选词 CT_i 和 CT_j 间的相似性计算如下:

$$S_{ij} = \frac{|A(CT_i) \cap A(CT_j)|}{|A(CT_i) \cup A(CT_j)|} \quad (3)$$

其中, $A(CT_i)$ 表示构成第 i 个候选词的字的集合。

其次,为了保证句子 v 的向量 \mathbf{Y}_v ,不属于当前句子的候选词对应维的权值始终为 0,则构造对角矩阵 $\mathbf{F}_v \in R_+^{M \times M}$,计算如下:

$$(\mathbf{F}_v)_{kk} = \begin{cases} 1 & (\mathbf{Y}_v)_k > 0 \\ 0 & (\mathbf{Y}_v)_k = 0 \end{cases}, 1 \leq k \leq M \quad (4)$$

最后,对句子 v 的向量迭代计算更新过程如下:

$$\mathbf{D}_v = \sum_{u \in V, u \neq v} W_{uv} (\hat{\mathbf{Y}}_u \times \mathbf{S}) \times \mathbf{F}_v \quad (5)$$

$$\hat{\mathbf{Y}}_v = p^{\text{inj}} \times \mathbf{Y}_v + p^{\text{cont}} \times \mathbf{D}_v \quad (6)$$

其中, $\hat{\mathbf{Y}}_u$ 表示前一次迭代中计算出的句子 u 的向量; p^{inj} 和 p^{cont} 分别表示在更新句子的向量过程中,向量初始值和式(5)计算得到的向量值的权重。

直至向量 $\hat{\mathbf{Y}}_v$ 权值中的最大值出现,停止迭代,记最

大值对应那一维的候选词作为当前句子的评价对象。

3 实验

3.1 实验数据

文中利用第一届自然语言处理与中文计算会议(NLP&CC 2013)面向中文微博的情感分析评测提供的 20 个话题的微博测试语料。据统计,每个话题中已被人工标注有观点句和情感对象标识的大约有 100 条微博。且在所有话题下,2 152 条观点句中标记了 2 357 个评价对象,平均每个观点句有 1.09 个评价对象,表明对观点句仅抽取一个候选词作为评价对象的方法是合理的。

3.2 实验设置

文中采用严格评价和宽松评价两种方式,均使用精确率(P)、召回率(R)以及 F 值(F)作为评价标准。需要指出的是,比较提交的评价对象正确与否是根据评价对象在整条微博中的起始位置和终止位置与标注的结果是否一致来判断的。

一个句子中,严格评价要求提交的评价对象的起止位置和标注完全一致。而宽松评价,首先定义提交的评价对象的起止区间 s 和标注的评价对象的起止区间 s' 之间的覆盖率 c :

$$c(s, s') = \frac{|s \cap s'|}{|s'|} \quad (7)$$

其中, $|*|$ 表示计算区间的长度。

假设提交的评价对象结果的集合为 S ,标注的结果集合为 S' ,两个结果集合之间的覆盖率 C 定义为:

$$C(S, S') = \sum_{s_i \in S} \sum_{s'_j \in S'} c(s_i, s'_j) \quad (8)$$

则精确率、召回率和 F 值为:

$$P = \frac{C(S, S')}{|S'|} \quad (9)$$

$$R = \frac{C(S, S')}{|S|} \quad (10)$$

$$F = \frac{2 \times P \times R}{P + R} \quad (11)$$

其中, $|*|$ 表示计算集合中元素的个数。

3.3 实验结果及分析

由于只有观点句中才会有评价对象,故只对标注出的观点句中识别出的评价对象进行对比实验。在已标注的 20 个话题型微博测试语料上实现了 ULP 算法^[2],并与 CSI1 和 CSI2 进行比较。文中方法 CSI1 代表对于句子的向量初始权值设定上,仅考虑了语义分析提出位置词和搭配词对识别评价对象的指示作用,相似句子数很少的句子并未另作考虑;CSI2 代表既进行语义分析提出位置词和搭配词的指示作用,又对相似句子数很少的句子加以考虑。实验结果如表 2、表 3

所示。关于式(6)中的参数,设置为 $p^{inj} = p^{cont} = 0.5$ 。

表 2 评价对象识别结果(严格评价方式)

方法	严格评价		
	<i>P</i>	<i>R</i>	<i>F</i>
ULP	45.8	41.8	43.7
CSII	47.2	43.0	45.0
CSI2	48.1	43.9	45.9

表 3 评价对象识别结果(宽松评价方式)

方法	宽松评价		
	<i>P</i>	<i>R</i>	<i>F</i>
ULP	57.9	50.5	53.9
CSII	58.5	51.3	54.7
CSI2	58.5	52.3	55.2

由表 2、表 3 可以看出,从严格评价和宽松评价两个方面来看,文中方法对观点句中评价对象的识别效果均优于 ULP。严格评价上,CSII 和 CSI2 相对于 ULP 在精确率上分别提高了 1.4%、2.3%,在召回率上分别提高了 1.2%、2.1%;宽松评价上,CSII 和 CSI2 相对于 ULP 在精确率上均提高了 0.6%,在召回率上分别提高了 0.8%、1.8%。实验结果表明,文中结合语义分析并对句子中的候选词初始分值的设定作深入分析,从而对评价对象的识别有了明显提升。

同时,比较文中方法 CSII 和 CSI2,严格评价上,CSI2 相对于 CSII 在精确率和召回率上均提高了 0.9%;宽松评价上,召回率提高了 1%。实验结果表明,对句子的向量初始权值设定时,考虑相似句子数很少的句子情况,有助于提高评价对象识别的准确度。

4 结束语

文中基于相似的句子有着相似的评价对象的假设,提出了改进的非监督标签传播算法来识别话题型微博中的评价对象。文中方法进行了语义分析,并考虑了相似句子数很少的句子情况。将文中方法与 ULP 对比发现,文中方法提高了对评价对象识别的精度,但并未考虑句子间相似性计算与候选词相似性计算上的误差。下一步将结合语料分析针对这两个问题进行改进,以更好地识别评价对象。

参考文献:

[1] 侯敏,滕永林,李雪燕,等. 话题型微博语言特点及其情感分析策略研究[J]. 语言文字应用,2013(2):135-143.

[2] 倪茂树,林鸿飞. 基于关联规则和极性分析的商品评论挖掘[C]//第三届全国信息检索与内容安全学术会议. 出版地不详;出版者不详,2007:628-634.

[3] Qiu Guang, Liu Bing, Bu Jiajun, et al. Expanding domain sen-

timent lexicon through double propagation[C]//Proceedings of twenty-first international joint conference on artificial intelligence. Pasadena, California, USA: [s. n.], 2009: 1199-1204.

[4] Zhang Lei, Liu Bing, Lim S H, et al. Extracting and ranking product features in opinion documents[C]//Proceedings of the 23rd international conference on computational linguistics. Posters; Association for Computational Linguistics, 2010: 1462-1470.

[5] Liu Kang, Xu Liheng, Zhao Jun. Opinion target extraction using word-based translation model[C]//Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Jeju Island, Korea; Association for Computational Linguistics, 2012: 1346-1356.

[6] Liu Kang, Xu Liheng, Liu Yang, et al. Opinion target extraction using partially-supervised word alignment model[C]//Proceedings of the twenty-third international joint conference on artificial intelligence. [s. l.]: AAAI Press, 2013: 2134-2140.

[7] Liu Kang, Xu Liheng, Zhao Jun. Extracting opinion targets and opinion words from online reviews with graph co-ranking[C]//Proceedings of the 52nd annual meeting of the association for computational linguistics. Baltimore, Maryland, USA: [s. n.], 2014: 314-324.

[8] 高磊,李斌,戴新宇,等. 基于依存分析和褒义指向的微博情感短语抽取方法[C]//自然语言处理与中文计算会议. 北京:出版者不详,2012.

[9] 文坤梅,徐帅. 基于句法依存关系的微博情感分析方法[C]//自然语言处理与中文计算会议. 北京:出版者不详,2012.

[10] Jakob N, Gurevych I. Extracting opinion targets in a single- and cross-domain setting with conditional random fields[C]//Proceedings of the 2010 conference on empirical methods in natural language processing. [s. l.]: Association for Computational Linguistics, 2010: 1035-1045.

[11] 王荣洋,鞠久朋,李寿山,等. 基于 CRFs 的评价对象抽取特征研究[J]. 中文信息学报,2012,26(2): 56-61.

[12] 郝志峰,杜慎芝,蔡瑞初,等. 基于全局变量 CRFs 模型的微博情感对象识别方法[J]. 中文信息学报,2015,29(4): 50-58.

[13] 周红照,侯明午,颜彭莉,等. 语义特征在评价对象抽取与极性判定中的作用[J]. 北京大学学报:自然科学版,2014,50(1): 93-99.

[14] Zhou Xinjie, Wan Xiaojun, Xiao Jianguo. Collective opinion target extraction in Chinese microblogs[C]//Proceedings of the 2013 conference on empirical methods in natural language processing. Seattle, Washington, USA: [s. n.], 2013: 1840-1850.