第 42 卷 第 7 期 计 算 机 工 程 2016 年 7 月
Vol. 42 No. 7 Computer Engineering July 2016

·人工智能及识别技术· 文章编号：1000-3428(2016)07-0173-08 文献标志码：A 中图分类号：TP311

基于词性特征与句法分析的商品评价对象提取

邱云飞[a]，陈艺方[a]，王 伟[a]，邵良杉[b]

（辽宁工程技术大学 a. 软件学院；b. 系统工程研究所，辽宁 葫芦岛 125105）

摘 要： 针对中文在线评论中语言不规范以及多样性导致评价对象识别错误的问题，提出基于词性特征与句法分析的商品评价对象提取方法。根据中文语言特点，利用形容词、副词、动词的词性特征构建规则提取评价词。通过子句序列的句法树结构提取候选评价对象并进行过滤。基于核心句法路径筛选评价搭配，以减少提取过程中引入的评价对象以及评价词噪声，从而提取出真正的评价对象。实验结果表明，引入句法树结构与核心句法路径使得商品评价对象识别的 *F* 值达到 80% 以上。

关键词： 中文评价词；评价对象；句法树结构；词性特征；句法路径

中文引用格式： 邱云飞，陈艺方，王 伟，等. 基于词性特征与句法分析的商品评价对象提取[J]. 计算机工程，2016，42(7)：173-180.

英文引用格式： Qiu Yunfei，Chen Yifang，Wang Wei，et al. Commodity Opinion Target Extraction Based on Part of Speech Feature and Syntactic Analysis[J]. Computer Engineering，2016，42(7)：173-180.

Commodity Opinion Target Extraction Based on Part of Speech Feature and Syntactic Analysis

QIU Yunfei[a]，CHEN Yifang[a]，WANG Wei[a]，SHAO Liangshan[b]

（a. School of Software；b. System Engineering Institute，Liaoning Technical University，Huludao，Liaoning 125105，China）

【Abstract】 Aiming at the wrong recognition of opinion target in online Chinese comments caused by diverse and non-standard language，a method of commodity opinion target extraction based on the part of speech features and syntactic analysis is proposed. According to the characteristics of Chinese language，rules are constructed to extract evaluation words by part of speech features of adjectives，adverbs，and verbs. Through the syntactic tree structure of clause sequence，opinion target candidates are extracted and filtered. Evaluation collocation is screened based on the core syntactic paths to reduce the noise of opinion target and polarity word introduced in the process of extraction，thus，extracting the real opinion target. Experimental results show that the introduction of syntactic tree structure and core syntactic path makes the *F* value of commodity opinion target recognition over 80% .

【Key words】 Chinese evaluation word；opinion target；syntactic tree structure；part of speech feature；syntactic path

DOI： 10.3969/j.issn.1000-3428.2016.07.029

1 概述

随着 Internet 的飞速发展，电子商务在开放的环境下逐渐渗透到人们的消费、工作以及学习生活中，因其具有普遍性、方便性、协调性、集成性等特点，使得 B2C 电子商务网站在入驻商家的同时吸引更多消费者进行商品体验，继而发布自己对于商品的真实观点。调查显示：消费者中有 71% 阅读、书写产品评论，用户评论成为继内部搜索功能后最重要的网站功能[1]。但是面对数量如此庞大、非结构化的评论信息，通过人工方法抽取评论中的有用信息将耗费大量的人力、物力，因此，自动地评价对象的提取显得异常重要。

本文结合词性特征、句法树结构、句法路径，不依赖于外部词典，提出并实现一种基于词性特征与句法分析的商品评价对象提取方法。该方法借助词

基金项目： 国家自然科学青年基金资助项目“二向性反射分布函数的先验知识耦合式融合方法研究”(61401185)；辽宁省高等学校杰出青年学者成长计划基金资助项目(LJQ2012027)；辽宁省教育厅一般基金资助项目(L2013131，L2013133)。

作者简介： 邱云飞(1976-)，男，教授、博士、CCF 会员，主研方向为数据挖掘、情感分析；陈艺方，硕士研究生；王 伟，讲师、硕士；邵良杉，教授、博士。

收稿日期： 2015-07-17 **修回日期：** 2015-09-04 **E-mail：** keyman_cyf@163.com

性特征提取评价词,利用子句序列的句法结构树结构提取候选评价对象,根据句法成分相同、句法功能相同的特点提取核心句法路径并扩充筛选评价搭配,进而抽取真正的评价对象。

2 研究背景

评价对象是指评论文本中描述的核心内容,多为产品的整体、部件、应有的属性以及相关概念的特征。在英文评论领域中,文献[2]通过限定评价对象与评价词之间的距离 k,提取评价词、评价对象,该方法会过滤掉距离较远的评价对象与评价词,同时将距离相等但不是评价对象的词引入,因此,评价对象提取的准确率受限于 k 值的选取;文献[3]应用评价词典以及主题词典抽取(评价对象、评价词)对,其方法在英文语料中能有效提高召回率,捕获评价对象与评价词之间的关系;文献[4]利用词性标注抽取评价对象、评价词,通过 LSA 特征降维过滤并生成评价词->评价对象规则;文献[5]采用词性标注,根据频繁词性标签集提取产品特征,并借助情感词典识别情感倾向,使用 LSA 进行产品特征降维提取评价对象;文献[6]基于 CRFS 进行评价对象的提取,在词法特征、依存关系特征、相对位置特征、语义特征的基础上引入 SRL 特征,使得准确率有所提高并有效提高了系统性能;文献[7]不借助外部情感词典,利用词性特征,限定评价语句评价词为形容词,进而识别评价对象,但其忽略了具有指示倾向性的动词,在中文评价对象提取中效果不是很理想;在词性特征的基础上,文献[8]提出利用句法分析中丰富的句法信息代替仅利用字典的词序列方法,经过浅层语义分析提取评价对象;文献[9]使用词典对大量语料进行标注,通过统计构建常见句法路径模式库,基于编辑距离对句法路径进行模糊匹配来抽取情感评价单元,但基于编辑距离的句法路径方法会引入出现频率小但与句法路径库中路径相似的情感单元。在中文评论领域中,文献[10]提出一种综合使用词性、词型模板不依赖外部资源,采用模糊匹配、剪枝的方法提取评价对象;文献[11]结合句法分析以及依存关系进行评价对象的提取,句法分析的引入避免了仅停留在词表面特征抽取评价对象的错误,并取得一定研究成果。

3 评价词提取

在英文领域中,多采用 Hownet 评价词汇表中的程度级别词语、负面评价词语、正面评价词语作为评价词,但由于网络评论中语言的不规范性、新兴网络词语的使用,使得部分(评价对象、评价词)对被过滤出去。文献[12]指出,可以使用形容词来判别句子是否具有情感倾向。因此,本文根据词性特征进行评价词及评价短语的提取。

鉴于词性角度分析,评价词的词性包括单独的形容词、形容词性名词、副词修饰的形容词以及副词修饰的动词。其中,副词作为形容词以及动词的修饰词,起到了增强情感强度的作用,而形容词、动词能更好地指示其情感倾向。本文分别对于形容词、形容词短语、动词短语进行评价词的提取,评价词提取规则如表 1 所示。

表 1 评价词提取规则

评价词	词性规则	优先级	评价词/评价词短语
二元评价词	形容词 + 形容词	1	两者皆为评价词
	副词 + 形容词	3	组合作为评价词短语
	副词 + 动词	4	组合作为评价词短语
一元评价词	名词 + 形容词	2	形容词作为评价词
	单独的形容词	2	形容词作为评价词
	形容词 + 名词	2	形容词作为评价词

按照表 1 中评价词提取规则的优先级进行评价词的提取。若词同时符合多种优先级规则且包含规则 1 时,只按照规则 1 进行提取,不考虑其他几种规则;若词同时包含具有相同优先级的多种规则时(包含规则 6、规则 4、规则 5 中一种以上提取规则时),仅考虑其中一种即可,根据以上提取规则获得的评价词建立评价词集。因为候选评价对象只有被评价词修饰才能成为真正的评价对象,所以将含有评价词的句子提取出,方便进行评价对象的抽取。

4 候选评价对象提取

评价对象作为情感分析的一个子任务,在电影领域[13]、汽车评论领域[14]已经取得了较好成果。从词性角度来说评价对象多为名词或者名词短语,文献[15]的研究工作充分说明了其可行性。因此,在研究过程中把名词、名词短语作为评价对象进行提取。文献[10]仅从词性序列角度将评价对象定义为形如 n,n n,n n n 的名词短语,未对其能否构成短语进行句法分析。为了准确识别多个词语构成的单个评价对象(例如名词组合、形容词性名词等),本文在此基础上充分考虑名词短语的句法树结构来判断其是否可以构成名词短语来提取候选评价对象。候选评价对象抽取框架如图 1 所示。

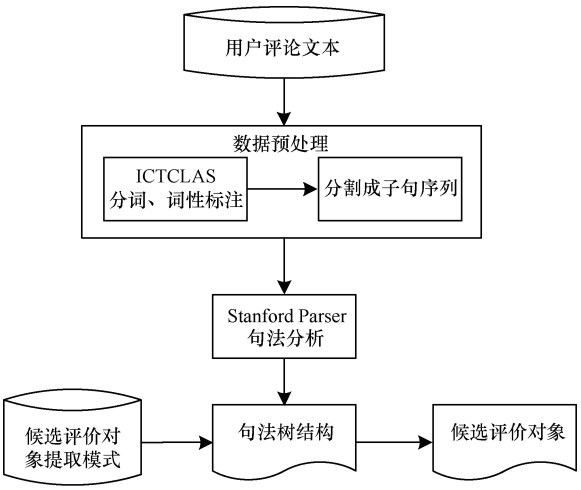

图 1　评价对象提取框架

4.1　数据预处理

结合 ICTCLAS, Stanford Parser 将中文评论从非结构化的文本演变成半结构化、具有词性标注的文本;鉴于传统的句法分析方法直接对文本进行子句、短语结构、词之间关系的识别,严重影响其分析性能,因此,将句子划分为子句序列后进行句法分析。

4.2　基于句法子树结构的候选评价对象提取

中文评论在经过数据预处理后,进行句法分析,得到根据概率上下文无关文法表示的相应句法树,从句法树中可以发现词性标注为 NN 节点的叶子节点即作为候选评价对象的名词,而这些被标记为 NN 的名词在上下文环境中进一步组合,可形成名词短语加入候选评价对象库中。针对此类名词短语,将组合判别引入到句法树结构中,根据训练集得到的模式提取名词、名词短语,解决名词短语组合的不确定性问题,进而优化分词和词性标注处理的结果。

定义 1　完整 NP 句法子树是指句法结构树中一棵子树,该子树中所有叶子节点的公共祖先为 NP,其中可包括最小完整 NP 句法子树。

定义 2　最小完整 NP 句法子树是指以 NP 节点为祖先的子树只有一个孩子节点且其为 NN。

根据已定义的完整 NP 句法子树、最小完整 NP 句法子树建立以下提取模式,进行候选评价对象的提取。

候选评价对象提取模式:

(1)在句法结构树中,完整 NP 句法子树叶子节点的父节点中包含 1 个 ~3 个 NN。

1)形式如 NN + NN + NN 或 NN + NN。NN + NN如图 2 句法子树所示,即该 NP 节点所有叶子节点的父节点皆为 NN,则合并所有叶子节点,将该 NP 构成的名词短语作为候选评价对象。

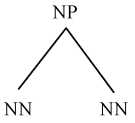

图 2　词性标注为 NN + NN 的名词短语句法子树

2)形式如 NN + NN + DEG + NN 或 NN + DEG + NN 或 NN + DEG + NN + NN。NN + DEG + NN + NN如图 3 句法子树所示,则将该 NP 构成的名词短语作为候选评价对象。

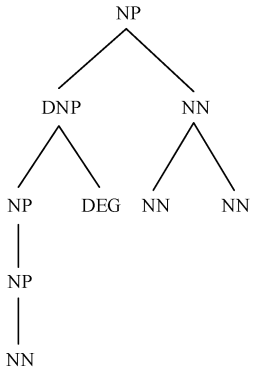

图 3　词性标注为 NN + DEG + NN + NN 的名词短语句法子树

3)形式如 JJ + NN + NN 或 CD + NN + NN 或 JJ + JJ + NN + NN。JJ + NN + NN 如图 4 句法子树所示,则将该 NP 构成的名词短语作为候选评价对象。

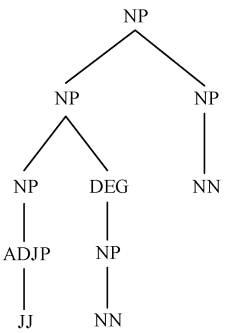

图 4　词性标注为 JJ + NN + NN 的名词短语句法子树

4)形式如 NN + NN + CC + NN。NN + NN + CC + NN 如图 5 句法子树所示,则将名词短语 NN + NN 以及名词 NN 作为候选评价对象。

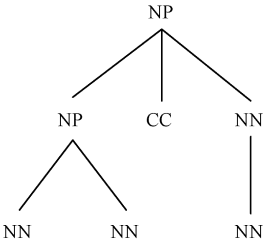

图 5　词性标注为 NN + NN + CC + NN 的名词短语句法子树

5)形式如 NN + CC + NN。NN + CC + NN 如图 6 句法子树所示,则将每个 NN 作为候选评价对象。

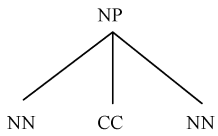

图 6 词性标注为 NN + CC + NN 的名词短语句法子树

6)形式如 JJ + NN 或 NT + JJ + NN。NT + JJ + NN如图 7 句法子树所示,则将 JJ + NN 构成的名词短语作为候选评价对象。

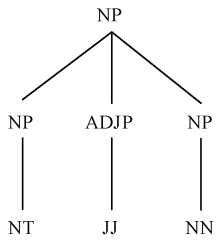

图 7 词性标注为 NT + JJ + NN 的名词短语句法子树

(2)在句法结构树中,最小完整 NP 句法子树中仅包含一个 NN,即分词后的词作为最小单位,如图 8句法子树所示,则将 NN 标注的名词直接提取作为候选评价对象。

图 8 词性标注为 NN 的最小完整 NP 句法子树

5 评价对象识别

鉴于出现在情感句中的候选评价对象并不一定是评价对象,只有当该评价对象被评价词修饰时才是真正的评价对象。因此,本文对提取出的候选评价对象、评价词互相搭配提取出评价搭配,即(候选评价对象,评价词)对,通过扩充后的核心句法路径识别评价对象。

5.1 评价搭配抽取

中文商品评价文本中长句较多,每个商品评价文本对商品体验的多个方面进行评价,导致长句中候选评价对象和候选评价词的数目较多,它们互相搭配产生的评价搭配数目大大增加,从而导致正确评价搭配所占比例减少。若一个长句中包含 M 个候选评价对象、N 个候选评价词,则会产生 $M \times N$ 个评价搭配,与此同时,对整个句子进行句法分析时,需要同时识别子句、词与词、词与短语、短语与短语之间的关系,降低其性能;而在子句序列中,每个子句仅对商品的一个方面进行评价,有效降低了评价搭配的数目且提高了正确评价搭配所占比例,更说明了子句序列划分的必要性。因此,对于子句序列抽取出的候选评价对象、评价词互相搭配,产生(评价对象,评价词)对。

5.2 句法路径库的建立

鉴于评价搭配中规律的句法关系,根据评价搭配生成的句法路径建立句法路径库。其中,句法路径是指句法结构树中,2 个节点之间的路径信息。由于本文提取的评价对象、评价词既有短语又有单独的词,因此句法路径中评价对象、评价词所对应的节点既可以是句法结构树中叶子节点的父节点,也可以是句法子树结构的最小公共祖先。对于子句“介绍的早餐和正餐也物美价廉”,其句法树结构如图 9 所示。其中,根据表 1 中规则 2 提取出的评价词“也物美价廉”所对应的节点为其句法子树结构的最小公共祖先“VP”,根据 4.2 节中模式 1 的第 5)种形式提取出的评价对象“早餐”、“正餐”,根据 4.2 节中模式(2)提取出的候选评价对象“介绍”,其所对应的节点为叶子节点的父节点“NN”。因此,评价搭配(介绍,也物美价廉)的句法路径为 NN,NP,DNP,NP,IP,VP;评价搭配(早餐,也物美价廉)、(正餐,也物美价廉)的句法路径为 NN,NP,NP,IP,VP。

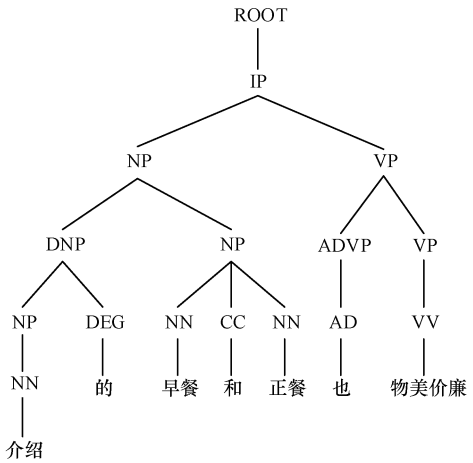

图 9 句法树结构

5.3 基于核心句法路径的评价对象提取

不同用户的表达方式不同,没有统一的规范,导致句法路径信息多样,直接对句法路径进行精确匹配会引入大量噪声,使得抽取结果中准确率降低。因此,对句法路径库中句法路径进行分类获取核心句法路径,并对核心句法路径最佳组合进行扩充,以识别评价对象。

对句法路径库中句法路径进行分类。在句法路径中若本层节点类型与下层节点类型相同或者句法功能相同,则将该节点纳入待合并项中,提取句法路径类,将提取句法路径类作为核心句法路径,使得距离较远的评价搭配也可以通过核心句法路径被识别。由于本文中评价对象为名词或名词短语,所以

词性标记为 NN,NP 的句法功能相同,除此之外,评价词中 VV,VP,ADVP,AD 也具有相同的句法功能。然而,有限类别的核心句法路径并不能精确地抽取出商品评论中的评价对象,由于本文中评价词是根据词性特征进行提取,Stanford Parser 生成句法树的同时导致部分评价词词性标记错误,因此根据评价词词性对核心句法路径进行扩充,进而提取评价对象。对于"介绍的早餐和正餐也物美价廉"中的评价搭配对(介绍,也物美价廉),其核心句法路径为 NN,DNP,NP,IP,VV,对于评价搭配(正餐,也物美价廉)、(早餐,也物美价廉),其核心句法路径为 NN,IP,VV。

5.4　评价对象识别步骤

评价对象识别步骤如下:

步骤 1　将评论文本 S 进行数据预处理后得到子句序列 $S=\{S_1,S_2,\cdots,S_p\}$,其中,p 为子句序列个数;i 代表评论语句第 i 个子句;n 为 S_i 子句中单词个数。

步骤 2　对子句 $S_i(i=1,2,\cdots,m)$ 按照表 1 中的提取规则,提取评价词 $Sense_i=\{s_1,s_2,\cdots,s_l\}$,其中 l 为评价词个数。

步骤 3　对子句 $S_i(\mathrm{i}=1,2,\cdots,m)$ 进行句法分析,得到子句的句法结构树 $Tree_i$,从句法结构树 $Tree_i$ 中提取完整 NP 句法子树,使用名词短语提取模式 1)~模式 4)、模式 6)进行名词短语候选评价对象的提取;从句法结构树 $Tree_i$ 中提取完整 NP 句法子树,使用提取模式 4)、模式 5)进行名词候选评价对象的提取,从句法结构树 $Tree_i$ 中提取最小完整 NP 句法子树,使用(2)中的模式进行名词候选评价对象的提取。得到初步候选评价对象集合 $Target\_Options_i=\{n_1,n_2,\cdots,n_t\}$,其中 t 为子句 S_i 中候选评价对象个数。由于子句中出现的名词及名词短语即候选评价对象,只有当候选评价对象被评价词修饰时才是评价对象,因此候选评价对象的个数大于评价对象的个数,即 $t\geqslant k$。

步骤 4　候选评价对象集合 $Target\_Options_i=\{n_1,n_2,\cdots,n_t\}$,若该句中名词候选评价对象包含在名词短语候选评价对象中,则只保留名词短语候选评价对象。例如模式 1)中,该句中只保留名词短语作为候选评价对象,而不保留构成名词短语的每个名词。

步骤 5　对构成名词短语的名词进行共现权重的分析,初步过滤名词短语。

对于名词短语中的每个名词 $w_1,w_2,\cdots,w_l$,其共现权重为 $weight(w_1,w_2,\cdots,w_l)$,其中,$tf(w_1\cap w_2\cap\cdots\cap w_l)$ 代表名词短语的出现次数,$tf(w_i)(i=1,2,\cdots,l)$ 为构成名词短语中的词项出现次数。

$$weight(w_1,w_2,\cdots,w_l)=\frac{tf(w_1\cap w_2\cap\cdots\cap w_l)}{tf(w_1)+tf(w_2)+\cdots+tf(w_l)}\quad(1)$$

对构成名词短语的名词,统计该单词在文本中词频的方法进行过滤,且直接过滤掉单个字成为评价对象的情况,得到评价对象集合 $Target_i=\{n_1,n_2,\cdots,n_k\}$,其中,$i$ 为该评论语句中第 i 个子句,k 为该句中评价对象个数。

步骤 6　将子句中筛选后的候选评价对象与评价词互相搭配,获得评价搭配(n_i,s_i)并提取句法路径。

步骤 7　根据 5.3 节生成核心句法路径,对训练集中生成的核心句法路径组合得到的最佳组合进行扩充,进而提取评价对象。

步骤 8　处理下一个子句序列,转到步骤 1。

6　实验结果与分析

本文利用聚焦式网络爬虫在携程网、去哪网、大众点评网等网站爬取了酒店领域的评论文本,取 500 条评论划分成 2 099 个子句序列作为训练集,220 条作为测试集,并将其划分成 1 002 个子句序列进行实验。利用 ICTCLAS 对实验数据进行分词处理,Stanford Parser 对每个子句序列进行句法分析。

6.1　基于词性特征的评价词提取实验

根据第 3 节中的评价词规则,提取出评价词 765 个,将其构成评价词集,构造的评价词集的部分如:好,温馨,很喜欢,干净,愉快,热情,好客,整洁,满意,不错,合理,给力,高大上,优越,方便,舒心,便宜,适中,糟糕,丰富,贵,近,陈旧,到位,脏,周到,不耐烦,不足,繁华,宽敞,亮堂,差,潮湿,简陋,便捷,安静,压抑,沉闷,感激,新鲜,推荐,顺利,齐全,遗憾……

6.2　基于句法子树结构的候选评价对象提取实验

根据 4.2 节中候选评价对象提取模式,提取出候选评价对象 972 个,每个模式下提取的部分候选评价对象如表 2 所示。

表 2　候选评价对象提取

句法子树	模式	候选评价对象
完整 NP 句法子树	1	地理位置,电脑配置,酒店价格,酒店位置,服务态度,酒店设施,早饭品种,配套设施……
	2	酒店的早饭,前台的工作人员,酒店餐厅的东西,酒店的地理位置,前台的服务员……
	3	具体酒店环境,一个购物广场,三诺音响……
	4	酒店装修,房间,笔记本电脑,音响……
	5	面积,氛围,早餐,正餐,时间,费用……
	6	装潢,繁华地区,老街,古香,对面车站,小城市,快捷酒店……
最小完整 NP 句法子树	1	房间,设施,酒店,服务,床,高铁,环境 ,客栈,卫生,饭菜……

6.3 评价搭配实验

根据生成的句法子树模式提取的候选评价对象与基于词性特征提取的评价词互相搭配,获得评价搭配。实验结果如表 3 所示。

表 3 评价搭配统计情况

统计项	酒店领域评论文本
候选评价对象	972
评价词	765
单位句子含有评价对象数	4.42
单位子句含有评价对象数	0.97
单位句子含有评价词数	3.48
单位子句含有评价词数	0.76
长句中评价搭配数	4 523
子句中评价搭配数	1 024

表 3 中的实验结果表明,中文评论长句中包含了一个至多个候选评价对象、一个至多个候选评价词,因此针对整句进行评价搭配的获取,得到很多无效的候选评价搭配对;然而划分成子句序列后,一个子句序列中约包含一个候选评价对象、一个候选评价词,因此针对子句序列进行评价搭配的获取会显著减少候选评价搭配的数目。

6.4 句法路径库的建立

在训练集中获得的评价搭配基础上抽取出句法路径,建立句法路径库,若测试集中句法路径不在句法路径库中,则添加该句法路径。其中句法路径库中部分句法路径如表 4 所示。

表 4 句法路径库

编号	句法路径
1	NN,NN,NN,IP,VV
2	NN,NN,IP,VV
3	NN,NN,NN,VV,VV
4	NN,NN,VV,VV
5	NN,IP,VV,VA
6	NN,NN,IP,VV,VA
7	NN,NN,NN,ADJP,JJ
8	NN,NN,DNP,NN,IP,VV
9	NN,VV,IP,VV,PP,LCP,NN,ADJP
10	NN,PP,VV,IP,CP,NN,IP,VV,VA

6.5 基于核心句法路径的评价对象识别实验

在句法路径库的基础上进行核心句法路径抽取,获得句法路径类,前 10 类按出现几率排序如表 5 所示。

表 5 核心句法路径

类别编号	核心句法路径	出现几率/%
1	NN,IP,VV	31.1
2	NN,IP,VV,VA	11.5
3	NN,VV	8.1
4	NN,VV,IP,VV	3.5
5	NN,PP,VV	2.8
6	NN,ADJP,JJ	2.7
7	NN,IP,VV,IP,VV	1.6
8	NN,PP,VV,IP,VV	1.4
9	NN,CP,IP,VV	1.3
10	NN,DNP,ADJP	1.1

提取高频核心句法路径(前 10 条句法路径)进行组合提取评价对象,如表 6 所示,其中准确率、召回率、F 值的计算如式(2)~式(4)所示。实验结果表明,随着 N(N 为前 N 条核心句法路径,$3 < N < 10$)的增大,准确率 P 呈现下降趋势、召回率 R 及 F 值呈现上升趋势;其中较好的一组组合路径是编号 5,即引入前 7 条核心句法路径。从表 6 可以看出,在增加核心句法路径 4~路径 6 时,由于将人工标注不是评价对象的候选词引入,增加了噪声,使得性能降低;核心句法路径 5、路径 8 的引入,使得性能降低,说明若候选评价对象、评价词所在的句法子树是介词短语结构,则该子句中的名词、名词短语不是评价对象;核心句法路径 7、路径 9、路径 10 的引入,可提高性能。

表 6 不同核心句法路径组合的评价对象提取对比 %

编号	方法	准确率	召回率	F 值
1	1+2+3 组合	77.4	80.3	78.8
2	1+2+3+4 组合	75.2	81.0	78.0
3	1+2+3+4+5 组合	73.9	82.0	77.7
4	1+2+3+4+5+6 组合	73.0	84.9	78.5
5	1+2+3+4+5+6+7 组合	73.1	86.8	79.4
6	1+2+3+4+5+6+7+8 组合	72.7	87.0	79.2
7	1+2+3+4+5+6+7+8+9 组合	72.5	88.0	79.5
8	1+2+3+4+5+6+7+8+9+10 组合	72.1	88.9	79.6

因此,结合核心句法路径 1+2+3+7+9+10(编号 3)提高系统性能,并与引入全部前 10 条核心句法路径(编号 2)以及整体性能较好的前 7 条路径(编号 1)进行对比,实验结果如图 10 所示。

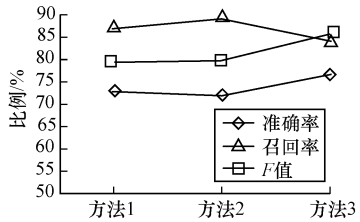

图 10　3 种方法的评价结果对比

实验结果表明,前 7 条核心句法路径的引入取得了较高的性能;前 10 条核心句法路径的引入,使得性能稍微提高,但由于其部分核心句法路径中引入噪声使得准确率降低,同时也说明核心句法路径并非引入越多越好;相比之下,核心句法路径 1 +2 +3 +7 +9 +10 的引入使得系统性能显著提高,并取得较高的准确率 P。因此,本文将句法路径 1 ~ 路径 3、路径 7、路径 9、路径 10 引入作为核心句法路径组合,并依据评价词词性进行核心句法路径组合的扩充,抽取评价对象。

6.6　实验评价

为了充分验证本文基于词性特征与句法分析的方法对于商品评价对象提取的有效性,实验过程中采用最近距离方法与精确匹配 2 种方法进行对比。2 种基线方法如下:

(1)最近距离方法

在情感分析系统中,多数人认为距离评价词最近的名词就是评价对象[16],因此,本文拟定最近距离方法。首先,利用词性特征,选取形容词、动词作为评价词;其次,选取距形容词、动词 4 个距离之内的 n,$n+n$,$n+n+n$ 作为评价对象。

(2)精确匹配方法

对依据本文方法提取出的评价搭配采用核心句法路径精确匹配的方法进行评价对象的提取。

为了验证本文句法分析过程中在核心句法路径精确匹配方法的基础上进一步采用核心句法路径组合的方法对于评价对象提取任务的积极作用,将核心句法路径精确匹配的方法作为基线方法,而为了验证本文基于词性特征与句法分析的方法整体的有效性,将以往基于词性特征采用最近距离匹配的方法作为基线方法进行进一步对比。

利用准确率 P、召回率 R 以及 F 值对评价对象 A 的识别进行评价。通过表 7 对准确率、召回率、F 值进行定义[17],得到测试集的数据关系如表 8 所示。

表 7　实验评价数据关系

人工标注	系统识别类型		
	识别为 A	未识别	识别为空
A	a	b	c

表 8　测试集的数据关系

实验方法	人工标注	系统识别数量		
		识别为 A	未识别	识别为空
最近距离方法	A	282	143	107
精确匹配方法	A	332	140	60
本文方法	A	357	108	67

$$P=\frac{a}{a+b} \tag{2}$$

$$R=\frac{a}{a+c} \tag{3}$$

$$F=\frac{2\times P\times R}{P+R} \tag{4}$$

通过表 8 的测试集数据关系,根据准确率、召回率和 F 值的定义,不同实验方法的对比情况如图 11 所示。实验结果表明,由于在最近距离方法中将 $n+n$,$n+n+n$ 作为名词短语进行评价对象的提取,并未从句法方面进行分析,引入名词短语噪声,且限制评价对象与评价词之间的距离,过滤掉距离较远的评价对象,导致基线方法中准确率较低;句法路径精确匹配的方法虽然可以取得较好的效果,但其准确率较低;相比之下,本文将句子划分为子句序列,利用句法树结构,充分考虑了构成名词短语的模式,采用扩充后的核心句法路径组合信息提取评价对象,使得系统性能提高了 4 个百分点,充分说明了该方法的有效性。

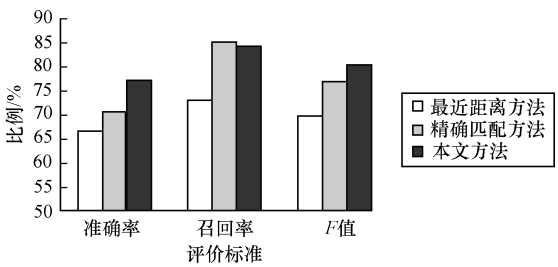

图 11　本文方法与基线方法对比

7　结束语

本文对爬取到的用户评论文本进行子句序列的划分,不借助外部词典,利用词性特征提取评价词;通过句法树结构模式提取候选评价对象,弥补 ICTCLAS 分词与 Stanford Parser 词性标注的不足;在核心句法路径的基础上进行组合及扩充,对(评价对象,评价词)搭配进行过滤,提取出评价对象。实验结果显示,商品评价对象识别的 F 值达到 81.4%,表明本文研究方法是合理有效的。下一步将结合上下文特征作进一步研究,从而更准确地提取商品评价对象。

参考文献

[1] Deloitte. Industry Outlook[EB/OL]. (2008-05-28). http://www.deloitte.com.

[2] Kim S M, Hovy E. Extracting Opinions, Opinion Holders, and Topics Expreeed in Online News Media Text[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Berlin, Germany: Springer, 2006: 1-8.

[3] Li Binyang, Zhou Lanjun, Feng Shi, et al. A Unified Graph Model for Sentence-based Opinion Retrieval[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. New York, USA: ACM Press, 2010: 1367-1375.

[4] Lal M, Asnani K. Aspect Extraction & Segmentation in Opinion Mining[J]. International Journal of Engineeing and Computer Science, 2014, 3(5): 5873-5878.

[5] Samha A K, Li Yuefeng, Zhang Jinglan. Aspect-based Opinion Extraction from Customer Reviews[EB/OL]. (2014-04-10). http://arxiv.org/pdf/1404.1982.

[6] 王荣洋,鞠久鹏,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.

[7] Liu Bing, Hu Minqing, Cheng Junsheng. Opinion Observer: Analyzing and Comparing Options on the Web[C]//Proceedings of the 14th International Conference on World Wide Web. New York, USA: ACM Press, 2005: 342-351.

[8] Li Shoushan, Wang Rongyang, Zhou Guodong: Opinion Target Extraction Using a Shallow Semantic Parsing Framework[C]//Proceedings of the 26th AAAI Conference on Artificial Intelligence. Berlin, Germany: Springer, 2012: 1671-1677.

[9] 赵妍妍,秦 兵,车万翔,等. 基于句法路径的情感评价单元识别[J]. 软件学报,2011,22(5):887-898.

[10] 宋晓雷,王素格,李红霞. 面向特定领域的产品评价对象自动识别研究[J]. 中文信息学报,2010,24(1):89-93.

[11] 王卫平,孟翠翠. 基于句法分析与依存分析的评价对象抽取[J]. 计算机系统应用,2011,20(8):52-57.

[12] Wiebe J, Wilson T, Bruce R, et al. Learning Subjective Language[J]. Computational Linguistics, 2004, 30(3): 277-308.

[13] Niklas J, Iryna G. Using Anaphora Resolution to Improve Opinion Target Identification in Movie Reviews[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2010: 263-268.

[14] 姚天昉,聂青阳,李建超,等. 一个用于汉语汽车评论的意见挖掘系统[C]//中国中文信息学会二十五周年学术会议论文集. 北京:清华大学出版社,2006:260-281.

[15] 何婷婷,闻 彬,宋 乐,等. 词语情感倾向性识别及观点抽取研究[C]//第一届中文倾向性分析评测委员会会议论文集. 上海:复旦大学出版社,2008:89-93.

[16] Hu Minqing, Liu Bing. Mining and Summarizing Customer Reviews[C]//Proceedings of ACM International Conference on Knowledge Discovery & Data Mining. New York, USA: ACM Press, 2004: 168-177.

[17] Ye Qiang, Shi Wen, Li Yijun. Sentiment Classification for Movie Reviews in Chinese by Improved Semantic Oriented Approach[C]//Proceedings of the 39th Annual Hawaii International Conference on System Sciences. Washington D. C., USA: IEEE Press, 2006: 53.

编辑 顾逸斐

(上接第 172 页)

[11] Graham R L. An Efficient Algorithm for Determining the Convex Hull of a Finite Planar Set[J]. Information Processing Letters, 1972, 1(4): 132-133.

[12] 邓 瑞,周玲玲,应忍冬. 基于 Kinect 深度信息的手势提取与识别研究[J]. 计算机应用研究,2013,30(4):1263-1265.

[13] Hu M K. Visual Pattern Recognition by Moment Invariants[J]. IEEE Transactions on Information Theory, 1962, 2(8): 179-187.

[14] 王秀琴,夏洪洋. 不变矩算法的改进与人耳识别技术[J]. 黑龙江科技学院学报,2008,18(1):51-57.

[15] 张素莉,潘 欣. 一种新颖的基于马氏距离的文本分类方法的研究[J]. 长春工程学院学报:自然科学版,2011,12(2):102-105.

[16] 曹雏清,李瑞峰,赵立军. 基于深度图像技术的手势识别方法[J]. 计算机工程,2012,38(8):16-18,21.

[17] 董立峰,阮 军,马秋实,等. 基于不变矩和支持向量机的手势识别[J]. 微型机与应用,2012,31(6):32-35.

编辑 金胡考