# 融合核心句与依存关系的评价搭配抽取

陶新竹[1,2],赵 鹏[1,2],刘 涛[1,2]
(1. 安徽大学 计算机科学与技术学院,安徽 合肥 230601;
2. 安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230601)

**摘 要**:评价搭配抽取是情感分析的基础任务之一。目前大部分抽取方法都是以依存句法分析为基础,但依存分析对中文评论文本的分析结果不稳定。针对此问题,提出了融合核心句抽取与依存关系的评价搭配抽取方法。该方法利用核心句抽取规则简化评论句结构,在此基础上进行依存句法分析,根据人工构建的依存关系模板进行评价搭配的抽取,并引入潜在评价搭配抽取规则抽取文本中省略评价对象的评价搭配。在中文酒店评论语料中进行试验,与基于依存分析的方法相比,该方法的 $F$ 值提高约 7%,证明了该方法的有效性。

**关键词**:核心句抽取;依存关系;评价搭配;潜在评价搭配
**中图分类号**:TP391 **文献标识码**:A **文章编号**:1673-629X(2014)01-0118-04
doi:10.3969/j.issn.1673-629X.2014.01.030

# Extraction of Evaluation Collection of Merging Kernel Sentence and Dependency Relation

TAO Xin-zhu[1,2],ZHAO Peng[1,2],LIU Tao[1,2]
(1. School of Computer Science and Technology,Anhui University,Hefei 230601,China;
2. Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education,Anhui University,Hefei 230601,China)

**Abstract**:Extraction of evaluation collection is one of basic tasks in the area of sentiment analysis. Currently,most extraction methods are based on dependency parsing,but the result of dependency analysis is unstable. For this problem,a novel method merging kernel sentence and dependency relation is proposed. This method used kernel sentence extraction rule to simplify the sentence structure,on this basis,did the dependency parsing and extracted evaluation collection according to artificially constructed dependencies template,in addition,introduced the extraction rules of implied evaluation collection to extract the evaluation collection which omitted the opinion object. Experimental results on Chinese hotel reviews domains show that compared with the method based on the dependency parsing,F-value increased by about 7% to verify the validity of the method.

**Key words**:kernel sentence extraction;dependency relation;evaluation collection;implied evaluation collection

## 1 概 述

随着电子商务的快速发展,网络上的评论文本呈爆炸式增长,情感分析任务成为研究者们关注的热点,评价搭配抽取是情感分析的子任务之一。评价搭配是指评价对象与其对应的修饰词的组合[1],也称为情感评价单元或特征-观点对,表现为二元组<评价对象,评价词>。第三届中文倾向性分析评测(COAE2011)将评价搭配抽取作为新增的要素级评测任务[2]。

近几年,面向英文文本的研究主要采用基于句法分析建立模板或规则的方法[3-8]抽取评价搭配。文献[3]在依存句法分析的基础上,首先使用限定词性的方法建立评价搭配候选集合,再使用最大熵模型的方法对候选评价搭配进行筛选,得到最终的评价搭配集合。这种方法挖掘了评价对象和评价词的结构关系,相比文献[4]使用的最近距离进行匹配的方法有了很

**收稿日期**:2013-03-05 **修回日期**:2013-06-10 **网络出版时间**:2013-09-29
**基金项目**:安徽省教育重点资助项目(KJ2009A001Z);安徽省重大科技专项资助项目(08010201002);安徽大学青年科学研究基金资助项目(2009QN004A)
**作者简介**:陶新竹(1989-),女,硕士研究生,研究方向为文本倾向性分析;赵 鹏,副教授,硕士生导师,研究方向为智能信息处理;刘 涛,硕士研究生,研究方向为文本倾向性分析。
**网络出版地址**:http://www.cnki.net/kcms/detail/61.1450.TP.20130929.1541.035.html

大改进,但是该方法限定了评价对象和评价词的词性,并且假设评价对象和评价词在一个单句中,对于评价对象和评价词分别在两个句子中的情况,无法做出正确处理,如句子“Don't go to that hotel,it's too bad!”文献[5]使用的是模式匹配的方法,首先通过短语句法分析获取评价对象与评价词的句法路径,再通过泛化、筛选等过程得到句法路径库,以此作为模式库采用匹配算法进行评价搭配的抽取,该方法在英文语料上取得不错的结果,但是由于句法路径的扁平化特征,失去了评价对象与评价词的结构特征,在处理长句较多的中文语料时,正确率不理想[9]。

国内对于中文评论文本的评价搭配抽取方法主要有基于最大熵模型以及基于模板或规则两种方法。

章剑锋[10]等人提出了基于最大熵模型的方法自动抽取评价搭配,但评价对象和评价词需要事先确定。方明等人[11]做了类似的工作,证明了将评价词语的类别作为语义信息特征,能一定程度上提高识别性能,另外还将评价对象与评价词的依存关系加入到特征模板中。

王素格等人[12]首先利用依存句法分析结果分别建立了名词、动词及形容词组块规则,从而分别抽取出评价对象和评价词,在此基础上进一步利用词与词之间的搭配关系,设计评价对象与评价词的搭配算法。顾正甲[13]等人利用 SBV 极性传递法识别需抽取的评价对象和评价词(极性词),并引入 ATT 链算法以及互信息法确定评价对象的边界,进一步挖掘了评价对象与评价词的语义关系。文献[12]和文献[13]类似,都是在依存句法分析的基础上,建立一系列规则进行评价搭配的抽取。

但是由于依存句法分析器在处理文本时受文本规范性影响较大,而中文网络评论往往较为复杂,所以这两种方法的处理结果并不稳定,对依存分析的结果依赖程度较大。

由以上的分析可知,依存句法分析在评价搭配抽取的任务中占有重要地位。

然而,由于中文网络评论文本的不规范性,以及中文本身的复杂性,依存句法分析的处理结果存在一定的误差。

为解决这一问题,设想如果能够删除句子中一些与评价搭配无关的成分,简化评论句的结构,再交由依存句法分析器处理,这样既可以提高处理的准确率,也能够一定程度上提高其召回率,基于这种假设,文中提出了融合核心句与依存关系的方法进行评价搭配抽取。

此外,还研究了省略评价对象时,评价搭配的抽取规则,即潜在评价搭配的抽取。

## 2　融合核心句与依存关系的评价搭配抽取

### 2.1　核心句抽取

#### 2.1.1　冗余成分删除

核心句抽取包括冗余成分删除与评价搭配句抽取两部分。冗余成分删除指的是按照一定的规则去除句子中某些冗余成分,得到新的句子。张莉等人[14]提出了一些核心句抽取的规则,除此之外,文中还发现带有假设性倾向的句子应予以删除,另外句首的一些词语如“就是”、“居然是”也会影响依存关系分析结果。通过对大量语料的观察,最终确定给出了以下 6 条规则,如表 1 所示。

表 1　冗余成分删除表

| 序列 | 内容 |
|---|---|
| 规则 1 | 删除句子中的句首状语成分,如:“(根)据……所说”、“(根)据……报道”、“以(从)……来说”、“从……来看”、“(当)(在)……时候”序列 |
| 规则 2 | 删除带有假设性倾向的句子,如:“如果……,(那么)……”、“希望……”、“但愿……”、“建议……”、“除非……,……”、“相信……”等 |
| 规则 3 | 若句子中含有冒号,则删除冒号以及冒号前面的部分,如果冒号前面部分为“举……为例”、“例如”、“举例来说”等等,则同时删除冒号前面部分和冒号后面的一句话 |
| 规则 4 | 删除句子中的括号及括号内的文字 |
| 规则 5 | 如果句子以名词、名词性短语、名词词组开头,后面的动词为“说”或者“感觉”、“认为”等主张词,则将这部分从句子中删去。其中主张词使用知网的主张词词典 |
| 规则 6 | 删除句首的“就是”、“居然是”、“记得”等词 |

#### 2.1.2　评价搭配句抽取

评价搭配句抽取是在冗余成分删除的基础上识别含有评价搭配的句子。文中将评价搭配句分为两种,即显性评价搭配与潜在评价搭配。

显性评价搭配句是指含有完整<评价对象,评价词>对的句子,如:“房间设计简约而明亮”,潜在评价搭配句是对评价搭配抽取的一种补充,指的是表达了主观情绪或倾向但却省略了评价对象的评论句,如:“竟然没有热水”、“还算干净”、“不能够上网!”依据两种不同的评价搭配句的特征,文中制定了 4 条识别规则,如表 2 所示。

表 2　分句规则评价搭配句识别

| 序列 | 内容 |
|---|---|
| 规则 1 | 单句中存在与模板匹配的结构,则将这个句子保留,识别为评价搭配句(随后按照模板进行评价搭配抽取) |
| 规则 2 | 如果没有匹配的模板,但含有 ADV(d,a)结构,则将此句与前一句合并,考察其是否含有 SBV 结构。若含有,则识别为评价搭配句(随后按照模板进行抽取) |
| 规则 3 | 如果不符合以上规则,但句中含有“没有”、“不能够”等词,则将这个句子识别为潜在评价搭配句 |
| 规则 4 | 如果不符合以上规则,且句子中不含有 SBV 结构、总词数小于 5 个且含有形容词,则将此句识别为潜在评价搭配句 |

### 2.2 依存关系模板与潜在评价搭配抽取规则

文中使用两种方法抽取评价搭配,对于潜在的评价搭配句,建立了潜在评价搭配抽取规则,由表 3 所示;对于显性评价搭配依存关系模板进行匹配抽取,依存关系模板是通过哈工大语言技术平台(LTP)[15]对句子进行句法分析的基础上由人工制定的,如表 4 所示。

表 3 潜在评价搭配抽取规则

| 序列 | 内容 |
|---|---|
| 规则 1 | 若句子以“没有”开头后接名词或者以“不能(够)”开头后接动词,则将此名词或动词作为评价对象,将“没有”或“不能够”作为评价词提取 |
| 规则 2 | 若句子中含有形容词,则在已抽取的评价搭配集合中进行搜索,将匹配频率最高的评价对象作为该词对应的评价对象,若没有匹配的评价对象,则将领域词(如在对酒店的评论中,将“酒店”作为领域词)作为该词对应的评价对象 |

表 4 依存关系模板

| 序号 | 依存关系模板 | 评价搭配/<评价对象,评价词> |
|---|---|---|
| 1 | (n $\xleftarrow{ATT}$)n $\xleftarrow{SBV}$a($\xrightarrow{ADY}$d) | <(n+)n,(d+)a> |
| 2 | v $\xleftarrow{ATT}$)n $\xleftarrow{SBV}$a($\xrightarrow{ADY}$d) | <(v+)n,(d+)a> |
| 3 | (n $\xleftarrow{ATT}$nd $\xleftarrow{ATT}$)n $\xleftarrow{SBV}$a($\xrightarrow{ADY}$d) | <(n+nd+)n,(d+)a> |
| 4 | (n $\xleftarrow{ATT}$)n $\xleftarrow{SBV}$v $\xleftarrow{SBV}$a($\xrightarrow{ADY}$d) | <(n+)n+v,(d+)a> |
| 5 | n $\xleftarrow{ATT}$v $\xleftarrow{SBV}$a($\xrightarrow{ADY}$d) | <n+v,(d+)a> |
| 6 | i $\xleftarrow{VOB}$v $\xleftarrow{SBV}$a($\xrightarrow{ADY}$d) | <i+v,(d+)a> |
| 7 | n $\xleftarrow{SBV}$v $\xrightarrow{ADV}$d<br>ADV↓<br>a | <n,d+a> |
| 8 | (n $\xleftarrow{ATT}$)n $\xleftarrow{SBV}$v $\xleftarrow{VOB}$a | <(n+)n,a> |
| 9 | (n $\xleftarrow{ATT}$)n $\xleftarrow{SBV}$v $\xleftarrow{VOB}$n | <(n+)n,n> |
| 10 | (n $\xleftarrow{ATT}$)n $\xleftarrow{SBV}$v $\xleftarrow{VOB}$v | <(n+)n,v> |
| 11 | v $\xleftarrow{ADV}$a $\xleftarrow{VOB}$d | <v,a+d> |
| 12 | (d $\xleftarrow{ATT}$)a $\xleftarrow{DE}$u $\xleftarrow{ATT}$v | <d+a,v> |
| 13 | (d $\xleftarrow{ATT}$)a $\xleftarrow{DE}$u $\xleftarrow{ATT}$n | <d+a,n> |

以上模板中箭头代表依存关系,箭头两端的小写英文字母代表具有依存关系的两个词的词性,箭头上方的标识(如“SBV”)表示两个词的依存关系类型,括号中的部分代表有可能存在的成分。

### 2.3 基于 PMI 的评价搭配筛选

通过以上方法抽取的评价搭配存在一定的噪音,候选搭配集合中包含了一些领域无关的部分,因此需要进行评价搭配筛选。在商品评论中,评价对象通常是领域相关的,而评价词的领域性并不强,所以文中采用筛选评价对象的方式间接筛选评价搭配。

点互信息算法(PMI)可以用来计算两次词之间的相关度,在一定的文本中,两个词 $w_a$ 与 $w_b$ 的 PMI 值计算公式如下:

$$PMI(w_a,w_b) = N(w_a,w_b)/N(w_a) * N(w_b) \quad (1)$$

其中,$N(w_a)$ 表示仅包含 $w_a$ 的文本数;$N(w_b)$ 表示仅包含 $w_b$ 的文本;$N(w_a,w_b)$ 表示既包含 $w_a$ 又包含 $w_b$ 的文本数。

文中利用 PMI 算法计算评价对象的领域相关度,首先选择一个具有代表性的领域词,然后通过百度搜索引擎分别得到包含评价对象($w_o$)与评价词($w_t$)的文本数,从而计算 $w_o$ 与 $w_t$ 的 PMI 值,最后通过设定阈值的方法进行评价对象的过滤。

### 2.4 算法描述

此节内容在使用 LTP 对评论语料进行分句的基础上,融合核心句抽取、依存关系模板等的完整评价搭配抽取算法。

算法描述如下:

输入:评论句集合 StcSet0 = { $s_1,s_2,\cdots,s_m$ }

输出:评价搭配集合 CombinSet = { < $obj_1$,$evalu_1$ >, < $obj_2$,$evalu_2$ >,…, < $obj_n$,$evalu_n$ > }

Step1:利用 LTP 对 StcSet0 中的每一个句子进行分词及词性标注,并按照表 1 的规则进行冗余成分删除,得到新的句子集合 StcSet = { $s_1,s_2,\cdots,s_n$ }。

Step2:对 StcSet 中的整句 $s_k(k = 1,2,\cdots,n)$,按照“,”、“:”、“;”分成单句集合{$sk_1,sk_2,\cdots,sk_p$},对于每个 $sk_t(t = 1,2,\cdots,p)$,按照以下步骤进行处理:

Step2.1:若句子符合表 2 的规则 1 或规则 2,则将句子按照表 4 的依存模板进行评价搭配抽取,将抽取结果加入候选评价搭配集 CandiCombinSet 中。

Step2.2:若句子符合表 2 的规则 3 与规则 4,则将句子按照表 3 的规则进行评价搭配抽取,抽取结果加入候选评价搭配集 CandiCombinSet 中。

Step3:重复 Step2,直至处理完 StcSet 中的最后一个句子。

Step4:计算 CandiCombin 中每个评价对象与领域词的 PMI 值,将高于阈值的评价对象所对应的评价搭配保留,得到评价搭配集合 CombinSet。

Step5:算法结束。

## 3 实验结果与分析

### 3.1 数据集与评价标准

文中的实验语料采用谭松波提供的酒店评论语料 2 000 篇,正反面各 1 000 篇,将其中的 1 200 篇作为依存关系模板的训练语料,其余 800 篇作为测试集 $S$。手工标注所有语料中的评价搭配作为实验的对比标准,包括省略评价对象的评价搭配(潜在评价搭配),并记为 < 评价对象,评价词 > 的格式,其中评价词部分包含了修饰评价词的程度副词及否定词。

实验的性能评估基于以下三个指标:召回率($R$),表示正确识别的搭配占测试语料中实际存在搭配的百分比;准确率($P$),表示正确识别的搭配占实际识别搭

配的百分比;综合指标 $F$ 测试值($F$),$F = 2PR/(P + R)$。

### 3.2 实验结果

为了验证核心句抽取与潜在评价搭配抽取规则的有效性,文中设置了两组实验:第一组实验仅采用表 4 中的依存关系模板进行评价搭配抽取作为 baseline 实验;第二组实验结合核心句抽取与潜在评价搭配抽取规则。文中随机选取测试语料的 30%、60% 作为新的测试集 $S_1$、$S_2$,在 $S_1$、$S_2$ 与 $S$ 上进行重复实验。实验结果如表 5 所示。

表 5　baseline 方法实验结果　　%

| 测试集 | $R$ | $P$ | $F$ |
|---|---|---|---|
| $S_1$ | 73.17 | 72.62 | 75.91 |
| $S_2$ | 73.57 | 72.52 | 73.04 |
| $S_3$ | 73.43 | 72.29 | 72.86 |
| 平均值 | 73.39 | 72.48 | 72.93 |

为了分别验证核心句抽取规则与潜在评价搭配抽取规则对抽取效果的影响,分次加入核心句抽取规则与潜在评价搭配抽取规则,实验结果如表 6 所示。

表 6　文中实验方法　　%

| 方法 | 召回率 | 准确率 | $F$ 值 |
|---|---|---|---|
| 依存模板 | 73.39 | 72.48 | 72.93 |
| 依存模板 + 冗余成分删除 | 74.51 | 79.80 | 77.06 |
| 依存模板 + 冗余成分删除 + 潜在评价搭配(文中方法) | 78.98 | 80.75 | 79.86 |

表 6 的实验结果数据表明,当增加冗余成分删除步骤时,评价搭配抽取的准确率得到显著提高,当在此基础上再增加潜在的评价搭配抽取规则时,召回率增加了约 6 个百分点,整体的 $F$ 值增加了约 7 个百分点。由此可见,文中提出的方法能够有效地提高以依存句法分析为基础的评价搭配抽取性能。

## 4 结束语

文中针对情感分析的子任务—评价搭配抽取任务,提出了融合核心句抽取与依存关系模板的方法,并且加入了潜在评价搭配抽取规则,考虑了省略评价对象的评价搭配抽取。并通过实验证明了:

(1)提高了核心句抽取能够提高句法分析器识别的正确性,从而有效地提高了算法的性能;

(2)潜在评价搭配的抽取能够提高评价搭配抽取的召回率。

文中仍有一些评价搭配无法识别,分析原因如下:

(1)对于一些特殊句式如比较句还没有好的解决办法;

(2)对于潜在的评价搭配挖掘的不够全面,这主要是因为潜在评价搭配的表现形式多种多样,制定的规则过于简单。

因此,在未来的研究中,应该进一步对隐含的评价搭配进行挖掘,以及对中文各种形式的句型进行研究。

**参考文献:**


[1] 赵妍妍,秦　兵,刘　挺. 文本情感分析[J]. 软件学报,2010,21(8):1834-1848.

[2] 许洪波,孙　乐,姚天昉. 第三届中文倾向性分析总结报告[R]. 北京:中国中文信息学会信息检索专业委员会,2011.

[3] Somprasertsri G,Lalitrojwong P. Mining feature-opinion in online customer reviews for opinion summarization[J]. Journal of universal computer science,2010,16(6):938-955.

[4] Liu B,Hu M,Cheng J. Opinion observer:Analyzing and comparing opinions on the Web[C]//Proc of the 14th international conference on World Wide Web. New York: ACM Press,2005:342-351.

[5] 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.

[6] Ojokoh B,Kayode O. A feature-opinion extraction approach to opinion mining[J]. Journal of Web engineering,2012,11(1):51-63.

[7] Popescu A M,Etzioni O. Extracting product features and opinions from reviews[C]//Proc of natural language processing and text mining. London:Springer,2007.

[8] Balahur A,Montoyo A. Semantic approaches to fine and coarse-grained feature-based opinion mining[C]//Proc of natural language processing and information systems. [s. l.]:[s. n.],2010:142-153.

[9] 黄亿华,濮小佳,袁春风,等. 基于句法树结构的情感评价单元抽取算法[J]. 计算机应用研究,2011,28(9):3229-3234.

[10] 章剑锋,张　奇,吴立德,等. 中文观点挖掘中的主观性关系抽取[J]. 中文信息学报,2008,22(2):55-59.

[11] 方　明,刘培玉. 基于最大熵模型的评价搭配识别[J]. 计算机应用研究,2011,28(10):3714-3716.

[12] 王素格,吴苏红. 基于依存关系的旅游景点评论的特征-观点对抽取[J]. 中文信息学报,2012,26(3):116-121.

[13] 顾正甲,姚天昉. 评价对象及其倾向性的抽取和判别[J]. 中文信息学报,2012,26(4):91-97.

[14] 张　莉,钱玲飞,许　鑫. 基于核心句及句法关系的评价对象抽取[J]. 中文信息学报,2011,25(3):23-29.

[15] 语言技术平台. LTP 哈尔滨工业大学社会计算与信息检索研究中心[DB/OL]. 2011. http://ir.hit.edu.cn/ltp/.

# 融合核心句与依存关系的评价搭配抽取

作者： 陶新竹， 赵鹏， 刘涛， TAO Xin-zhu， ZHAO Peng， LIU Tao

作者单位： 安徽大学 计算机科学与技术学院，安徽 合肥 230601；安徽大学 计算智能与信号处理教育部重点实验室，安徽 合肥 230601

刊名： 计算机技术与发展

ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2014(1)