基于依存关系的旅游景点评论的特征-观点对抽取*

吴苏红[1]，王素格[2,3]

[1]山西大学 数学科学学院，山西 太原 030006

[2]山西大学 计算机与信息技术学院，山西 太原 030006

[3]山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006

E-mail: wsg@sxu.edu.cn

摘 要：特征-观点对的抽取是观点挖掘中重要的研究课题之一，本文利用依存语法对句子的分析，研究了评论文本中特征-观点对的抽取。利用词对间的依存关系，构建了用于获取含情感倾向组块的规则以及候选评价对象的识别算法，在此基础上，设计了具有情感倾向的特征-观点对的抽取算法。本文对山西旅游景点评论语料进行了特征-观点对的抽取，实验结果表明，整体的F1值达到了87.10%，验证了算法的有效性。

关键词：特征-观点对；旅游景点评论；依存关系；组块；情感倾向

Feature-opinion Extraction in Scenic Spots Reviews Based on Dependency Relation

Wu Suhong[1], Wang Suge[2,3]

[1] School of Mathematics Science, Shanxi University, Taiyuan 030006

[2] School of Computer & Information Technology, Shanxi University, Taiyuan 030006

[3] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006

E-mail: wsg@sxu.edu.cn

Abstract: Feature-Opinion Extraction is one of the key researches in the area of opinion mining. This paper studies the method to extract the feature-opinion in review texts based on dependency grammar. By using the dependency relation between word and word, we construct the rules to obtain chunks which contain sentiment orientation and the algorithm to identify candidate evaluation objects. We design an algorithm to extract feature-opinion with sentiment orientation. In this paper, we extract the feature-opinion from the relevant scenic spots reviews of Shanxi. Experimental results show that the whole F1-measure can achieve 87.10%, and prove the designed algorithms are effective.

Keywords: feature-opinion; scenic spots reviews; dependency relation; chunk; sentiment orientation

1 引言

随着人民生活水平的提高，旅游已成为人们生活的重要组成部分。许多游客利用论坛、博客和旅游点评网等空间发表有关旅游景点的评论。与此同时，对于一般游客，在出游之前，可以通过网上评论了解其他游客对景点的看法，规划自己的旅游行程。而景点管理商可以通过景点评论了解游客对景点的意见和态度，以便提高服务质量。但是，人工地逐篇阅读海量的评论，需要花费大量的时间和精力，阅读者可能会“迷失”其中，无法识别和利用其中有价值的观点信息。为了准确、高效地挖掘出游客感兴趣的观点信息，特征-观点对的抽取是需要解决的关键问题之一。

特征-观点对是指评价对象及其观点词语之间的搭配，表现为二元对（评价对象，观点词语）。Kobayashi 等[1]实现了一种半自动方法来抽取评价表达式，该方法通过构造八种同现模板来描述评价对象和观点词语之间的修饰关系。Popescu[2]构建了一个无监督的信息抽取系统 OPINE，首先通

* 基金项目：国家自然科学基金资助项目(60875040, 60970014)；教育部高等学校博士点基金(200801080006)；山西省自然科学基金资助项目(2010011021-1)；太原市科技局明星专项(09121001)。

过计算名词或名词短语与一定的区分符之间的点互信息值来获取产品特征，然后利用 MINIPAR Parser 手工构建 10 条规则来识别与特征相关的观点词。Somprasertsri 等[3]基于句法信息和语义信息提出一个可以挖掘产品特征和观点的方法，通过采用依存关系提取特征-观点对，对文本进行观点综述。庄丽等[4]提出一种基于 multi-knowledge 的方法来进行电影评论挖掘，利用关键词和依存关系相结合挖掘文本中的特征-观点对。宋光鹏[5]首先将情感词和一些能改变情感倾向强度的连词、副词、否定词结合，然后再与所评价的对象组合在一起构成短语搭配模式，以此用于判断评价对象的情感。

本文利用依存关系，研究了评论文本中特征-观点对的抽取方法。首先利用依存关系制定一系列用于获取含情感倾向的组块的规则，在此基础上，进一步利用句子中词与词之间的依存关系，设计了评价对象与特征-观点对的识别算法，实现了具有情感倾向的特征-观点对的抽取。

2 情感倾向组块的获取

2.1 情感倾向组块获取规则

为了获取具有情感倾向的组块，本文在李素建等人提出的组块定义[6]基础上，定义了三种类型的情感倾向组块：名词组块、动词组块和形容词组块。其中单独的一个名词、动词或形容词都不在组块构成范围内，并列结构中的词语连同连接词均包含在相应组块中。本文采用哈工大信息检索研究中心[7]提供的中文依存句法分析工具，该工具共涉及到 24 种依存结构标记。

（1）名词组块：是由中心词为名词的 ATT、COO 或 QUN 结构构成。ATT 结构的中心名词的修饰词个数可以是一个或者多个。若“的”字结构作修饰成分时，连同所修饰的中心名词一起构成一个名词组块。对于数量结构，当数量词为数字时，不包含在名词组块中。

（2）动词组块：是由中心词为动词的 ADV、VOB、CMP、VV、MT 或 COO 结构构成。中心动词的对象宾语和后置修饰成分补语也包含在动词组块中。趋向动词、助动词与其前面的中心动词构成动词组块。当“地”字结构作修饰成分时，连同中心动词被划分为一个动词组块。

（3）形容词组块：是由中心词为形容词的 SBV、ADV、ATT、QUN、MT 或 COO 结构构成。需要说明的是，名词组块或动词组块内部的形容词组块不用标记。“的”字结构与其所修饰的中心形容词构成一个形容词组块。形容词加助词也可以组成形容词组块。

为了获得三类组块，利用词与词之间的依存关系，构建获取具有情感倾向组块的相关规则。建立的规则集记为 *RuleSet1*，如表 1 所示。以下规则除特殊说明外，大部分只限于相邻词之间的依存关系。其中表中的 parent.pos 表示关系中支配词的词性，child.pos 表示关系中从属词的词性。

利用 *RuleSet1* 中的规则得到含情感倾向组块，发现所获取的部分组块中含有评价对象和观点词语。例如，利用规则 N1 获取的组块“不错的历史博物馆”、“独特的建筑格局”等，该类组块的共同点都含有名词与其修饰成分，利用这类组块很容易获得特征-观点对。为此，本文在 *RuleSet1* 的基础上，对部分规则的条件做进一步限定，得到 *RuleSet2*，如表 2 所示。

2.2 基于规则的组块获取算法

由于 *RuleSet2* 比 *RuleSet1* 的规则限定条件更多，因此，在获取情感倾向组块时，*RuleSet2* 优于 *RuleSet1* 中的规则。利用 *RuleSet1* 和 *RuleSet2* 中的规则获取情感倾向组块的算法如下。

算法 1：基于规则的组块获取

输入：经过依存句法分析后格式为 XML 的评论句集合 $RSSet=\{s_1,\ldots,s_n\}$，组块集 $ChunkSet1=\Phi$，$ChunkSet2=\Phi$；

输出：ChunkSet1 和 ChunkSet2；

表 1　RuleSet1

类型	规则	条件
N1	DE+ATT+ATT	第二个 ATT：parent. pos = "n/ns"；DE：child. pos ≠"nd"
N2	ADV+DE+ATT	ATT：parent.pos = "n/ns"
N3	ATT+DE+ATT/ ATT+ATT	第一个 ATT：child.pos≠"u/q"；第二个 ATT：parent.pos="n/ns"
N4	DE+ATT/QUN+ATT	ATT：parent.pos="n/ns"； DE：child.pos≠"nd"
N5	ATT	parent.pos="n/ns" and child.pos≠"u/q/r"
N6	LAD+COO /COO	COO：parent. pos= "n/ns" and child.pos="n/ns"
N7	QUN	parent.pos= "n/ns"
V1	VOB+VOB/ ADV+VOB	第一个 VOB：parent. pos="v"
V2	DI+ADV（不一定相邻）	ADV：parent. pos="v"
V3	ADV+CMP/ CMP+DEI（不一定相邻）	CMP：parent.pos="v"
V4	VOB/ ADV/ CMP/ MT	parent.pos="v"
V5	ADV+ADV	第二个 ADV：parent. pos="v"
V6	LAD+COO/ COO	COO：parent. pos= "v" and child.pos="v"
V7	LAD	child ="所" and parent.pos="v"
V8	VV	parent.pos= "v" and child.pos="v"
ADJ1	ADV+SBV/ SBV	SBV：child. pos="n/ns" and parent. pos="a"
ADJ2	ADV+ADV	第二个 ADV：parent. pos="a/i"
ADJ3	LAD+COO/ COO	COO：parent.pos= "a" and child.pos="a"
ADJ4	ADV	parent.pos="a/i" and child.pos≠"q"
ADJ5	DE+ATT/ATT	ATT：parent. pos="a/i"
ADJ6	DE	child.pos="a/i" and parent.pos="u"
ADJ7	MT/ QUN	parent.pos="a"

表 2　RuleSet2

序号	规则	限定
Rule 1	DE+ATT+ATT/ DE+ATT+ATT+ATT	N1 上限定 DE：child.pos="a/v/i"；若第二个 ATT：parent.pos="a/v/i"，则提取第二个 ATT 后的部分
Rule 2	ADV+DE+ATT	同 N2
Rule 3	DE+ATT	N4 上限定 DE：child. pos="a/v/i"
Rule 4	ATT	N5 上限定 parent.pos="a/v/i"
Rule 5	ADV+SBV/ SBV	同 ADJ1
Rule 6	DE+ATT	同 ADJ5
Rule 7	ADV+DE+ATT+ATT	在 N1 上加以扩展，条件同 N1
Rule 8	ADV+ADV+SBV	在 ADJ1 上加以扩展，条件同 ADJ1
Rule 9	ADV+ADV+DE+ATT	在 N2 上加以扩展，条件同 N2
Rule 10	ADV+SBV+VOB	VOB：parent.pos="v"and child.pos="a"

Step1 利用 RuleSet2 中的规则 $Rule_i$（i=1,...,10），对 RSSet 中的句子进行组块获取，得到候选组块集 CanChunkSet1；

Step2 对于 *CanChunkSet1* 中采用 *Rule5* 和 *Rule8* 获取的组块，若组块 $\{Chunk_i^{'}\}_{i=k_1}^{k_p}$ 所在的句子中，组块与前面相邻的词为 ATT 或 DE（*child. pos*≠"r/u"）关系时，则将这些词与该组块合并生成 $\{Chunk_i\}_{i=k_1}^{k_p}$，$ChunkSet1 = ChunkSet1 \cup \{Chunk_i\}_{i=k_1}^{k_p}$，$CanChunkSet1 = CanChunkSet1 \setminus \{Chunk_i^{'}\}_{i=k_1}^{k_p}$；

//$\{Chunk_i^{'}\}_{i=k_1}^{k_p}$ 为原组块，$\{Chunk_i\}_{i=k_1}^{k_p}$ 为最终的组块；

Step3 对于 *CanChunkSet1* 中的每个组块，若组块 $\{Chunk_i^{'}\}_{i=j_1}^{j_r}$ 所在句子中存在与其有 COO 关系的词时，则将这些词与该组块合并生成 $\{Chunk_i\}_{i=j_1}^{j_r}$，$ChunkSet1 = ChunkSet1 \cup \{Chunk_i\}_{i=j_1}^{j_r}$，$CanChunkSet1 = CanChunkSet1 \setminus \{Chunk_i^{'}\}_{i=j_1}^{j_r}$；//$\{Chunk_i\}_{i=j_1}^{j_r}$ 为最终组块，$\{Chunk_i^{'}\}_{i=j_1}^{j_r}$ 为原组块；

Step4 $ChunkSet1 = ChunkSet1 \cup CanChunkSet1$；

Step5 利用 *RuleSet1* 中的规则 $Rule_j(j=1,...,22)$进行组块获取，得到组块集 *ChunkSet2*；

Step6 算法结束。

3 特征-观点对抽取

为了抽取特征-观点对，在 *ChunkSet2* 的基础上，利用词与词之间的依存关系给出了候选评价对象和候选观点词语的识别方法，然后在 *ChunkSet1* 的基础上，给出了候选特征-观点对的获取方法，最后从候选特征-观点对中筛选含有情感倾向的特征-观点对。

3.1 候选评价对象与观点词语的识别

（1）候选评价对象的识别

由于 *ChunkSet1* 中的组块含有评价对象和观点词语，则利用这些组块可构成句子中的部分候选特征-观点对。在算法 1 获得 *ChunkSet1* 基础上，再利用词与词之间的依存关系，对抽取组块后的句子设计候选评价对象的识别算法。

算法 2：识别句子中的候选评价对象

输入：除去含有 *ChunkSet1* 中组块的句子 $RRSSet=\{r_1,...,r_n\}$，候选评价对象集 $CanEOSet=\Phi$，*ChunkSet2*，$k=1$；

输出：候选评价对象集 *CanEOSet*；

Step1 对于 *RRSet* 中的 r_k，如果存在 SBV 关系且关系从属词 *W* 的词性为名词（“话”字除外）/代词（仅包括指示代词和第三人称代词）/动名词，则，如果从属词 *W* 在 *ChunkSet2* 的组块中，则 $CanEOSet=CanEOSet \cup \{ChunkW\}$，否则 $CanEOSet=CanEOSet \cup \{W\}$，转 Step3；//从属词 *W* 所在组块为 *ChunkW*；

Step2 对于 *RRSet* 中的 r_k，如果存在 VOB 关系且关系从属词 *W* 的词性为名词（“话”字除外）/代词（仅包括指示代词和第三人称代词）/动名词，则，如果从属词 *W* 在 *ChunkSet2* 的组块中，则 $CanEOSet=CanEOSet \cup \{ChunkW\}$，否则 $CanEOSet=CanEOSet \cup \{W\}$；

Step3 如果 $k<n$,则 $k=k+1$,转 Step1；

Step4 对于 *CanEOSet* 中的每个候选评价对象 $EO_i(i=1,\ldots,n)$，若部分候选评价对象 $\{EO_i^{'}\}_{i=m_1}^{m_k}$ 所在句中还存在与其有 COO 关系的成分时，也将其 COO 关系成分与该候选评价对象合并生成 $\{EO_i\}_{i=m_1}^{m_k}$，$CanEOSet=CanEOSet \cup \{EO_i\}_{i=m_1}^{m_k}$，$CanEOSet=CanEOSet \setminus \{EO_i^{'}\}_{i=m_1}^{m_k}$；//$\{EO_i\}_{i=m_1}^{m_k}$ 为最终的候选评价对象，$\{EO_i^{'}\}_{i=m_1}^{m_k}$ 为原候选评价对象

Step5 算法结束。

（2）候选观点词语的识别

对于(1)中所识别出的候选评价对象，它们所属的评论句中都有相应的观点词语。J.Wiebe[8]将观点词语的词性局限于形容词词性，而忽略了其他词性的观点词语。通过对评论语料的观察，动词、成语也常作为观点词语，而有些观点词语是以短语形式出现，因此，本文将 2.2 节中所获取的 *ChunkSet2* 中的形容词组块和动词组块也作为候选观点词语。最终选用的候选观点词语为形容词、动词、形容词组块、动词组块、成语。例如“漂亮”、“太不方便”、“值得去”等。

3.2 特征-观点对的抽取

根据候选特征-观点对中特征与观点之间的关系，将候选特征-观点对的抽取过程分为两部分：第一部分直接利用 2.2 节所获取的 *ChunkSet1*。第二部分利用 3.1 节中所识别出来的候选评价对象和候选观点词语构造特征-观点对，当句子中出现一个以上的评价对象和观点词语时，采用邻近法[9]确定候选观点词语与候选评价对象之间的相关性。最后从候选特征-观点对集中选出含有情感倾向的特征-观点对，获取特征-观点对集合。这里的情感词来源于 SWT[10]情感词表、《知网》情感词语集以及与旅游评论相关的情感词。

算法 3：特征-观点对的抽取

输入：除去含有*ChunkSet1*中组块的句子*RRSSet*=$\{r_1,...,r_n\}$，候选评价对象集*CanEOSet* ,候选特征-观点对集*CanFOSet*=Φ,特征-观点对集合*FOSet*=Φ ；

输出：特征-观点对集合 *FOSet* ；

Step1 利用 *ChunkSet1* 中的组块所构成候选特征-观点对，将其加入到 *CanFOSet*；

Step2 对于 r_k 中的候选评价对象 EO_{k_i} ,若 EO_{k_i} 与动词 W_v 之间是 SBV 关系（不相邻），则动词 W_v 不作为观点词语；若 EO_{k_i} 与形容词 W_a 之间是 SBV 关系（不相邻），则形容词 W_a 与候选评价对象构成候选特征-观点对，加入到 *CanFOSet*；

Step3 若候选观点词语和候选评价对象存在于同一个片段，则转 Step4，否则转 Step6；

Step4 若一个句子片段中只出现一个候选评价对象，则此候选评价对象与该片段中的所有候选观点词语构成不同的特征-观点对，加入到 *CanFOSet* 中，转 Step7；

Step5 若一个句子片段中出现多个候选评价对象，则候选观点词语选择邻近的候选评价对象构成候选特征-观点对，加入到 *CanFOSet* 中，转 Step7；

Step6 若候选观点词语和候选评价对象分别存在于不同的片段，则候选观点词语与邻近的候选评价对象构成候选特征-观点对，加入到 *CanFOSet* 中；

Step7 对 *CanFOSet* 中的候选特征-观点对，若包含情感词，则将其加入到 *FOSet*，否则，对于不能直接判定其情感倾向的特征-观点对，使用词语搭配倾向判别方法[11]来判别这类潜在语义特征-观点对的情感倾向，若有情感倾向，则加入到 *FOSet*。

Step8 算法结束。

上述算法中句子片段为以逗号隔开的子句。

4 实验结果与分析

实验数据采用互联网上的论坛、博客、旅游点评网等有关山西省 11 个地级市的 180 个景点的相关评论作为语料库，共 618 篇评论，平均每篇评论大致包含 2~3 个句子。为了衡量特征-观点对的抽取结果，本文采用三个评价指标：精确率（查全率）、召回率（查准率）和 F1 值。

4.1 组块获取的结果

对于旅游景点评论，利用算法 1 得到含评价对象和观点词语的组块集 *ChunkSet1*，共 915 个组块；含情感倾向的三类组块集 *ChunkSet2*，共 3985 个组块，其中名词组块 1742 个，动词组块 1871 个，形容词组块 372 个。例如，评论句“山西历史很悠久。”，依存句法分析结果如图 1 所示。

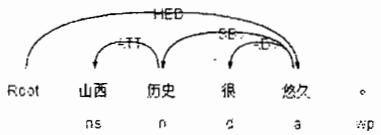

图 1 依存句法分析示例

这条评论句中，利用 *RuleSet2* 中的 ADV+SBV 规则获取组块“历史很悠久”，由于该组块前面词出现 ATT 关系，则应把词“山西”也识别在组块中，得到新的组块“山西历史很悠久”。

4.2 特征-观点对抽取的实验结果与分析

利用算法 1 得到的组块，以及算法 2-3，分别对正面、反面、全部的旅游评论进行特征-观点对抽取，共抽取出 1758 对。例如，对“悬空寺绝对是个一定要去的地方，精致奇特。”这句话进行特征-观点对抽取，依存句法分析结果如图 2 所示。

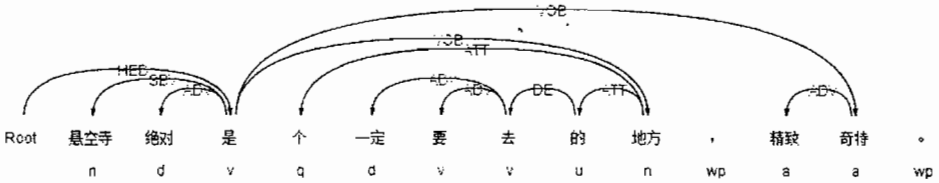

图 2 依存句法分析示例

由 Rule9 抽取组块“一定要去的地方”，获得候选特征-观点对（地方，一定要去），利用算法 2 识别候选评价对象为“悬空寺”，最后利用算法 3 获取候选特征-观点对（悬空寺，精致奇特）、（悬空寺，绝对是），在此基础上，得到特征-观点对（地方，一定要去）、（悬空寺，精致奇特）。

采用以上三个评价指标对特征-观点对抽取实验进行评价，其结果如表3所示。

表 3 特征-观点对抽取实验结果

评价指标 / 类 别	精确率	召回率	F1 值
正 面	88.28%	89.73%	89.00%
反 面	82.68%	86.63%	84.61%
全 部	85.84%	88.40%	87.10%

从表 3 中可以看出，本文的方法在精确率上达到预期的效果。其中，对正面评论进行特征-观点对判别时，精确率、召回率、F1 值都优于反面评论。主要原因是反面评论含有的否定词、程度副词较多，致使反面评论的判别结果错误率高于正面评论，从而影响了实验结果。

另外，对识别错误的结果分析发现，（1）有 80.07%的错误来自评价对象的识别错误，当利用规则抽取含评价对象和观点词语的组块时，句中的评价对象可能被抽掉，致使识别评价对象时出现错误。（2）有 14.76%的错误来自于观点词语的识别错误，该错误主要是由组块获取错误引起的。

5 结束语

本文利用词对间的依存关系，构建了用于获取含情感倾向组块的规则以及候选评价对象识别算法，在此基础上，设计了具有情感倾向的特征-观点对的抽取算法。本文对山西旅游景点评论语料进行了特征-观点对的抽取，整体的 F1 值达到了 87.10%，验证了算法的有效性。但仍存在一些特征-观点对无法正确识别，尤其是评价对象的识别，约有 80.07%的错误由评价对象的判别错误所引起。因此，在未来的工作中，应进一步开展评价对象识别方法的研究。

致谢：感谢哈尔滨工业大学信息检索研究中心提供的“语言技术平台 LTP”。

参 考 文 献

[1] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto. Collecting Evaluative Expressions for Opinion Extraction. In Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)[C]. 2004: 584-589.

[2] Ana-Maria Popescu, Oren Etzioni. Extracting Product Features and Opinions from Reviews. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)[C]. 2005: 32-33.

[3] G. Somprasertsri, P. Lalitrojwong. Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization. Journal of Universal Computer Science[J]. 2010, 16(6): 938-955.

[4] Li Zhuang, Feng Jing, Xiaoyan Zhu. Movie Review Mining and Summarization. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management[C]. 2006: 43-50.

[5] 宋光鹏. 文本的情感倾向分析研究[D]. 北京邮电大学. 2008.

[6] 李素建, 刘群. 汉语组块的定义和获取[C]. 全国计算语言学联合学术会议(SWCL2003)论文集. 2003: 110-115.

[7] 语言技术平台 LTP. 哈尔滨工业大学信息检索研究中心. http://ir.hit.edu.cn/

[8] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, Melanie Martin. Learning Subjective Language [J]. Computational Linguistics. 2004, 30(03): 277-308.

[9] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of the Conference on Knowledge Discovery and Data Mining(KDD) [C]. 2004: 168-177.

[10] 王素格, 杨安娜, 李德玉. 基于汉语情感词表的句子情感倾向分类研究[J]. 计算机工程与应用. 2009, 45(24): 153-155.

[11] 王素格, 杨安娜. 基于混合语言信息的词语搭配倾向判别方法[J]. 中文信息学报. 2010, 24(03): 69-74.