第 36 卷 第 2 期 计 算 机 工 程 2010 年 1 月
Vol.36 No.2 Computer Engineering January 2010
·博士论文· 文章编号：1000—3428(2010)02—0004—03 文献标识码：A 中图分类号：TP391

基于最大熵模型的观点句主观关系提取

樊 娜，蔡皖东，赵 煜
(西北工业大学计算机学院，西安 710072)

摘 要：提出一种提取中文观点句中评价对象和评价词主观匹配关系的方法。分析观点句中评价词和评价对象的词性、词语位置，通过句法分析获取语义特征，将 2 类特征应用于最大熵模型，提取观点句的主观关系。实验结果证明，与取距离评价词语最近的词作为评价对象的 Baseline 方法相比，该方法大幅度提高了准确率和 F 测试值。

关键词：评价对象；主观关系；最大熵；句法分析

Extraction of Subjective Relation in Opinion Sentences Based on Maximum Entropy Model

FAN Na, CAI Wan-dong, ZHAO Yu
(College of Computer, Northwestern Polytechnical University, Xi'an 710072)

【Abstract】 This paper presents a novel method of extracting subjective relation between opinion targets and opinion-bearing words in Chinese opinion sentences. This method analyzes lexical and part of speech information in the sentence. Syntactic analyzing is adopted to achieve syntactic path information which is regerded as semantic feature. The two kinds of feature are both applied in maximum entropy model. According to this model, all subjective relations in the sentence are extracted. Experimental results show that this method is better than Baseline method in precision rate and F-measure.

【Key words】 opinion target; subjective relation; maximum entropy; syntactic analysis

近年来，文本情感分析作为一个新的研究领域正日益受到研究人员的关注，在网络信息安全、产品评价、客户关系管理等方面得到广泛的应用。

1 相关研究

情感倾向主要通过文本中的观点句来表达。观点句是基于断言或评论并且带有个人情感和意向的抒发。以往的研究主要集中在识别文本中的观点句、识别观点持有者、识别观点主题、判断观点句子的情感词等方面。

在一个观点句中存在多个评价对象以及多个情感词语是很普遍的语言现象。带有情感倾向修饰评价对象的词语，称为评价词。一个句子中的不同的评价词语与各个评价对象之间的匹配对应关系称为主观关系。这种主观关系是正确分析句子的情感倾向、识别句子主题的前提和基础。

目前在面向英文的研究中，研究者已经提出了一些判断句子中主观关系的方法。文献[1]通过建立语言模板提取主观关系，虽然该方法具有较高的准确性，但是需要手工根据不同领域建立大量模板。

文献[2]基于模板建立规则库提取主观关系，通过规则定义模板各部分与评价对象和评价词语之间的对应关系，但是该方法只能处理简单句型。

文献[3]提出先找产品特征词，然后找距离该特征词最近的形容词作为评价词的方法。但在实际的观点句子中，动词、名词、形容词都可以作为评价词语。

文献[4]提出基于 FrameNet 的提取方法，先找出评价词语，然后进行语义角色标注，分析观点句子的语义结构，将语义角色映射为评价对象。该方法可以分析复杂结构的句子，但是由于 FrameNet 中词语资源的有限性，有些评价词语没有包括在内，因此这类词语与评价对象的关系无法提取。

语言特点的差异使这些方法无法直接应用到中文文本中。本文针对这一情况，提出了提取中文句子主观性关系的方法，应用最大熵模型，以词性、位置和语义等特征进行主观关系提取，为解决中文情感分析中的这一基础问题提供了可用的方法。

2 评价词与评价对象主观关系的提取

本文采用最大熵模型，通过在模型中应用基本特征和语义特征，获取观点句子中评价词与评价对象之间的主观关系。

2.1 最大熵模型

最大熵模型广泛应用于分词、词性标注、词义排歧、机器翻译等自然语言处理的各个领域中。最大熵的主要思想是找到一个概率分布，它满足所有已知的事实，且不受任何未知因素的影响。

最大熵模型的目标是对于给定上下文 c，计算出 m 的条件概率，即对 $p(m|c)$进行评估，期望能够求出符合 c 条件的 m 的概率分布[5]。最大熵模型要求 $p(m|c)$在满足一定的约束条件下，必须使下面定义的熵取得最大值：

基金项目：国家自然科学基金资助项目(60803151)
作者简介：樊 娜(1978－)，女，博士研究生，主研方向：网络信息安全，自然语言处理；蔡皖东，教授、博士生导师；赵 煜，博士研究生
收稿日期：2009-08-07 **E-mail：**fnsea@mail.nwpu.edu.cn

$$H(p) = -\sum_{c,m} p(m|c)\mathrm{lb}p(m|c) \quad (1)$$

最大熵的条件概率可以用式(2)计算：

$$p(m|c) = \frac{1}{Z(c)}\exp(\sum_{i=1}^{n}\lambda_i f_i(c,m)) \quad (2)$$

$$Z(c) = \sum_{m}\exp(\sum_{i=1}^{n}\lambda_i f_i(c,m)) \quad (3)$$

式(3)是归一化因子，其中，f_i 是模型的特征；λ_i 是 f_i 的参数，即每个特征函数的权值；特征 f_i 是一个二值函数，每个特征包含了上下文的各种信息。参数 λ_i 的值并不能直接得到，需要通过迭代的方式计算其近似值。目前，使用最广泛的是 GIS 迭代算法和 IIS 迭代算法。本文采用 GIS 算法实现，迭代次数为 100。

2.2 模型特征的选取

最大熵模型的关键在于如何针对特定的任务为模型选取特征集合。模型特征的选取需要通过特征选择算法加以解决。假定所有特征的集合是 F，特征选择算法要从中选择一个活动特征集合 S，活动特征集合要尽可能准确反映样本信息，只包括那些期望可以准确估计的特征。为得到集合 S，通常采用逐步增加特征的方法，每一次增加哪个特征取决于样本数据。例如，当前的特征集合是 S，满足这些特征的模型是 $C(S)$，增加一个特征，新的模型集合可以定义为 $C=(S\cup f)$。在特征选择过程中，活动集合越来越小，模型集合越来越大。

由于评价词与评价对象的词特征反映了评价词与评价对象的匹配关系，因此不同的评价对象与评价词之间的修饰关系是有一定规律可循的，例如，“身材”与“高”、“矮”、“胖”、“瘦”之类的词语搭配，而不会与“便宜”、“昂贵”这些修饰“价格”的评价词语搭配。根据修饰关系具有的匹配性，对于一个观点句子，首先提取该句中所有的评价词语，形成集合$\{E_1, E_2, \cdots, E_n\}$，然后提取该句中所有的评价对象，形成集合$\{O_1, O_2, \cdots, O_N\}$。对于评价词语集合中的每一个 E_i，根据特征函数 $F_k(O,\{O_1, O_2, \cdots, O_N\}, E_i)$计算其条件概率 $p(O \mid \{O_1, O_2, \cdots, O_N\}, E_i)$。

通过式(4)可以确定与评价词语 E_i 之间具有对应匹配关系的评价对象 O：

$$O = \arg\max[\sum_{i=1}^{n}\lambda_i F_i(O,\{O_1,O_2,\cdots,O_i\},E)] \quad (4)$$

提取观点句子的主观关系实际上可以看作是对句子中的评价词语进行主观关系标注的过程。这个标注过程被看作是一个事件，因此，由当前评价词及其上下文环境来确定一个事件的特征集合。

根据影响当前评价词主观关系标注的各种因素，定义特征空间为：

(1)词性。当前评价词及其前后各 2 个词的词性。

(2)词。当前评价词的前后各 2 个词。

(3)当前评价词及其前后各 2 个词的语法语义信息。

根据这个特征空间，定义了模型训练中应用的 2 大类特征：(1)基本特征，主要描述词语本身的特性。这类特征包括词语特征、词性特征以及评价词语与评价对象之间的距离特征等。(2)语义特征。

表 1~表 3 分别描述了基本特征包括的词语特征、词性特征和距离特征。基本特征不仅考虑了评价词和评价对象本身及其词性，同时还将它们前后的 2 个词语都纳入特征考虑范围内，因为其左右邻词在一定程度上体现了该词是否具有主观含义。这样的特征选取可以有效解决否定词以及程度副词对评价词语的影响，因为通常起修饰作用的否定词和程度副词都位于评价词语前后 2 个词语范围的位置上。

表 1　词语特征

特征名称	特征具体描述
WE	评价词
WE1	评价词左前第 1 个词
WE2	评价词左前第 2 个词
WE-1	评价词右后第 1 个词
WE-2	评价词右后第 2 个词
WO	评价对象
WO1	评价对象左前第 1 个词
WO2	评价对象左前第 2 个词
WO-1	评价对象右后第 1 个词
WO-2	评价对象右后第 2 个词

表 2　词性特征

特征名称	特征具体描述
P(WE)	评价词性
P(WE1)	评价词左前第 1 个词性
P(WE2)	评价词左前第 2 个词性
P(WE-1)	评价词右后第 1 个词性
P(WE-2)	评价词右后第 2 个词性
P(WO)	评价对象性
P(WO1)	评价对象左前第 1 个词性
P(WO2)	评价对象左前第 2 个词性
P(WO-1)	评价对象右后第 1 个词性
P(WO-2)	评价对象右后第 2 个词性

表 3　距离特征

特征名称	特征具体描述
P(E-O)	评价对象和评价词语的前后顺序关系
N1	评价词语和评价对象之间间隔的评价对象的个数
N2	评价词语和评价对象之间间隔的评价词的个数
D(E-O)	评价对象和评价词语间隔的词的个数

将程度副词作为特征考虑既可以在一定程度上反映评价词语主观性的强弱程度，又可以有效缩短评价词和评价对象之间的距离，因为距离越短，越容易进行正确的判断，而否定词语可以改变评价词语的情感极性，对主观关系产生影响，所以，必须考虑在内。距离特征主要描述评价词语和评价对象在一个观点句子中所处的位置关系，距离越近，两者之间越有可能存在主观匹配关系。

第 2 类语义特征主要描述词语在句子中的句法语义信息。通过对观点句子进行句法分析，获得评价词语和评价对象在观点句子中的语义信息，将其作为语义特征应用到模型中。

首先，采用哈工大信息检索研究室的中文句法分析器对观点句子进行分析，获得该句子完整的句法结构树，从中提取评价对象与评价词之间的句法路径信息作为特征。这种路径信息描述了评价词语与评价对象在句子语法结构中的位置以及修饰关系，有助于正确判断句子中存在的主观关系。

在句法结构树中，任意 2 个节点之间的路径并不是唯一的。同一句子中的评价词语和评价对象在句法树中可能存在多条不同的路径，如果将这些路径信息全部引入模型训练中，会大大影响模型的训练效率。为了解决这个问题，将评价对象和评价词之间完整的路径信息划分为 3 个部分：路径 Path(E~O)，路径 Path(E)和路径 Path(O)。为了明确表示这些路径信息，需要定义句法分析树中的几个重要节点：E 表示

评价词节点，O 表示评价对象节点，Head(E~O)代表同时覆盖 E 和 O 的节点。

表 4 为语义特征的具体描述。

表 4　语义特征

特征名称	特征具体描述
Path(E-O)	节点 Head(E-O)到其左右孩子节点的路径(选取的孩子节点同时是 E 和 O 的父节点)
Path(E)	节点 E 到其任何一个祖先节点的路径(选取的祖先节点同时是 Head(E-O)的孩子节点)
Path(O)	节点 O 到其任何一个祖先节点的路径(选取的祖先节点同时是 Head(E-O)的孩子节点)

在模型训练中，将这 3 条路径信息作为语义特征应用于模型中，既保留了有效的路径信息，又将评价词与评价对象之间的路径唯一化，提高了模型的训练效率。

语义特征在模型中的应用充分考虑了评价词语与评价对象在句子结构中的位置信息、语法语义信息，能更准确地描述句子中存在的主观关系。

3　实验结果与分析

3.1　数据集与评价标准

本文实验中最大熵模型采用的语料为中文手机产品评论文本。首先从手机产品评论网(http://product.it168.com/newpinglun/cSpace_pl.asp? cType_code =0302)搜集整理手机评论文本，并对所有评论认真审查，去除语言不规范的文本，最终选出 1 600 篇文本。将所有语料分为 2 个部分，其中，1 200 篇作为训练语料，其余 400 篇作为测试语料集合 *T*。同时手工标注所有语料文本中的主观关系，表示为<评价对象，评价词语>的形式，作为实验对比标准。

性能评估基于 3 个重要指标：查全率(*R*)，即正确识别出的主观关系数与应被识别出的主观关系数之比；查准率(*P*)，即正确识别出的主观关系数与识别出的所有主观关系数之比；F 测试值(*F*)，即综合衡量指标，$F=\frac{2PR}{P+R}$。

3.2　实验结果

本文实验采用 Baseline 方法先找出观点句子中的评价词语，然后选择距离评价词语最近的名词实体作为评价对象。实验目的是对比 Baseline 方法与本文方法在提取观点句子主观关系时性能上的差异。

(1)Baseline 方法实验

分别随机选取测试语料的 30%, 60%形成 2 个新的测试集 *T*1 和 *T*2，分别在这 2 个测试集和全部测试语料集 *T* 上进行 Baseline 方法的实验。

表 5 为 Baseline 方法的实验结果。

表 5　Baseline 方法实验结果　(%)

语料集	查准率	查全率	*F*
*T*1	38.56	77.75	51.55
*T*2	38.98	77.69	51.91
T	39.50	77.67	52.37

表 5 中实验数据显示，Baseline 方法的查全率较高，但是查准率偏低。3 次实验的平均查准率仅为 39.01%，平均查全率达到 77.70%，平均 *F* 值为 51.94%。而在实际应用中，相对于查全率，查准率更重要。

(2)本文方法实验

为了探测各类特征对性能的影响，首先在 Baseline 方法的基础上逐步增加本文提出的基本特征：词特征，词性特征，距离特征，分别进行实验，实验结果见表 6。

表 6　逐步增加基础特征最大熵结果　(%)

特征	查准率	查全率	*F*
Baseline+词	64.73	63.50	64.11
Baseline+词+词性	64.92	63.64	64.23
Baseline+词+词性+距离	65.32	63.69	64.49

表 6 的实验数据表明，随着模型中使用特征的增加，虽然查全率有所降低，但是查准率及 *F* 值都有明显的提高，表明主观关系提取的性能在逐渐增强。当模型应用全部基本特征时，查准率及 *F* 值的提高最明显，准确率提高了约 15 个百分点，*F* 值提高了约 12 个百分点。

在增加基本特征的基础上，在模型训练中再引入语义特征，即本文提出的结合基本特征和语义特征的方法。实验结果如表 7 所示。

表 7　本文方法与 Baseline 方法性能比较　(%)

方法	查准率	*F*
Baseline 方法	39.01	51.94
Baseline＋基本特征	65.32	64.49
本文方法	71.23	68.32

实验结果表明，与 Baseline 方法相比，增加基本特征后，模型提取的查准率和 *F* 值均有大幅度的提高，而在此基础上增加语义特征后，模型的查准率达到 71.23%，*F* 达到了 68.32%，性能提高显著，综合评价指标 *F* 与 Baseline 方法相比提高了约 17 个百分点。

上述实验数据及分析表明，与 Baseline 方法相比，本文方法不仅考虑了评价词和评价对象本身的词汇信息，还根据句法分析将语义信息作为特征应用到了最大熵模型中，能更准确有效地提取观点句子的主观关系。

4　结束语

本文提出了基于最大熵模型的中文观点句主观关系的提取方法。通过句法分析，得到观点句子中评价对象和评价词的语义信息，将其作为语义特征，同时将词语自身的信息以及评价词语和评价对象之间距离等信息作为基本特征，将这 2 类特征结合共同应用到最大熵模型中，从而获取观点句子中评价对象和评价词语之间的匹配关系。主观关系的提取和分析对于中文情感分析和意见挖掘等研究具有重要意义。在后继的研究中将对句子进行进一步的语法、语义分析，寻找更有效的特征以提高性能。

参考文献

[1] Nasukawa T, Jeonghee Y. Sentiment Analysis: Capturing Favorability Using Natural Language Processing[C]//Proc. of K-CAP'03. Sanibel Island, Florida, USA: [s. n.], 2005: 23-35.

[2] Popescu A M, Etzioni O. Extracting Product Features and Opinions from Reviews[C]//Proceedings of EMNLP'05. Vancouver, Canada: [s. n.], 2005: 145-153.

[3] Liu Bing, Hu Minqing, Cheng Junsheng. Opinion Observer: Analyzing and Comparing Opinions on the Web[C]//Proceedings of the 14th Internationa1 Conference on World Wide Web. [S. l.]: IEEE Press, 2006: 221-229.

[4] Kim Soo-Min, Hovy E. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text[C]//Proc. of Conf. of Association for Computational Lingustics. [S. l.]: IEEE Press, 2007: 318-327.

[5] Chen S F, Rosenfeld R. A Gaussian Prior for Smoothing Maximum Entropy Models[D]. School of Computer Science, Carnegie Mellon University, 1999.

编辑　张正兴