计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(11):254−259 [doi: 10.15888/j.cnki.csa.006050]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于条件随机场的评价对象抽取①

夏　圆, 张　征

(华中科技大学 自动化学院, 武汉 430074)

摘　要: 评价对象抽取是情感分析的重要组成部分, 针对在线商品中文评论非正规化、网络化的特点, 本文提出一种基于句法分析和条件随机场的评价对象的抽取方法, 通过实验分析不同模板与不同特征组合对评价对象提取的F值的影响. 在系统实现上, 主要利用哈工大语言技术平台 (LTP) 的开放接口和 CRFs 开源工具对评论数据集进行训练和测试. 最终使两类数据集的评价对象抽取的 F 值达分别达到到 82.98% 和 83.50%.

关键词: 评价对象; 情感分析; 句法分析; 条件随机场; 特征组合

引用格式: 夏圆,张征.基于条件随机场的评价对象抽取.计算机系统应用,2017,26(11):254–259. http://www.c-s-a.org.cn/1003-3254/6050.html

Objects Extraction of Comment Based on Conditional Random Field

XIA Yuan, ZHANG Zheng

(School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: Extracting object of comment is an important part of emotional analysis. In view of the irregularity of language and the characteristics of network in Chinese online comment, this paper presents a method of extracting objects based on the syntactic analysis and conditional random field. It analyzes experimentally the effect of different templates and different combinations of features on the F value. In the implementation of the system, this paper uses Harbin Institute of Technology language platform open interface and CRFs open source tools to train and test on comment data sets. Finally, the F values of the two types of data sets have reached 82.98% and 83.50% respectively.

Key words: comment object; sentiment analysis; syntactic analysis; conditional random field; combinations of features

1　引言

随着 Web 2.0 技术的快速发展, 在线评论在互联网上正以指数级的速度增长, 成为继内部搜索功能后最重要的网站功能. 对于电商平台, 评论信息影响到消费者的购买决策[1]. 从海量的在线评论中挖掘用户关心的信息可以把用户从中解脱出来, 因为将可视化的结果提供给商家, 可以帮助其改善服务质量, 为顾客带去舒适的购物体验; 提供给用户, 以帮助其做出最有效益的购物决策. 如何从以指数级增加的文本内容中抽取有用的信息、分析规律是一个急需解决的问题. 借助自然语言处理技术分析某一条评论所包含的评价对象以及顾客对产品某个属性的感情倾向性在很大程度上可以解决以上的问题. 评价对象指某段评论中所讨论的主题, 具体表现为评论中评价词所修饰的对象[2]. 评价对象的抽取是文本情感分析的关键, 现有的评价对象的抽取方法主要有基于关联规则和基于统计两种.

Liu Bing 最先提出评价对象抽取的问题, 将有着较高频率的名词以及短语视为评价对象, 把距离评价对象最近的形容词视为其评价短语[3]. 邱云飞、陈艺方等人提出根据中文语言的特点, 利用词性特征与句法分析提取商品评价对象的方法[4]. 张建华、肖中正也进行了基于词性特征和依存句法分析的方法抽取评价对象的研究[5]. 基于关联规则的方法主要是根据文本本身的特点, 结合评价对象所具有的特点, 制定相应的规则

① 收稿时间: 2017-02-21; 修改时间: 2017-03-09; 采用时间: 2017-03-16

或者模板, 用于识别某些领域的评价对象.

Niklas Jako 等人提出使用条件随机场模型来抽取评价对象, 将评价对象抽取任务建模成序列标记任务, 使用有限的特征和单一的模板来抽取评价对象[6]. 金丽君等人进行了基于 SVM 的搜索型商品评论有用性自动识别方法的研究[7]. 刘玮楠使用 HNC 理论研究网络评论情感倾向性, 将评论对象和情感特征统称为文本特征[8]. 基于统计的方法主要是通过训练生成统计模型来识别评价对象.

基于关联规则的方法在处理非规范性、半结构化或者非结构化网络评论文本的时候, 往往会引入非评价对象, 模板的限定范围比较固定, 对于特殊的评价对象或者流行的网络用语并不能较好适应, 需要经常更新规则或者模板, 并且泛化能力较弱; 而基于统计的方法往往会忽视句子间的内部结构[9], 对特定用法的评价对象的识别度不高. 因此, 本文将基于规则/模板的方法与基于统计模型的方法相结合, 综合考虑句法结构信息、词与词之间的依存关系, 并借助条件随机场统计模型来识别评价对象. 另外, 本文对 CRFs 的特征组合和模板的定义进行了一系列的实验, 最终使评论语料的各项实验指标获得最优.

2 基于条件随机场的评价对象抽取

2.1 评价对象抽取方法的总体框架

本文提出的基于条件随机场的评价对象抽取的方法的整体框架如图 1 所示, 并做如下介绍:

① 预处理, 主要是评论语料的获取, 以及分词、词性标注和句法分析.

② 将预处理的语料分为训练集和测试集, 选择不同特征组合和模板来进行训练.

③ 将测试集输入到训练模型中进行识别, 并分析实验结果.

2.2 语料的预处理

本文的语料来自淘宝网, 通过网络爬虫得到两种不同类别 (电脑与服饰) 的商品的中文评论. 预处理主要包括以下几个步骤:

例句 (1): 电脑很薄, 系统流畅, 发货快.

① 分词. 将评论文本通过 ltp 的分词接口, 进行分词, 例句 (1) 的分词结果:

电脑很薄, 系统流畅, 发货快.

② 序列标记. 将评论文本中的所有元素为 3 类: 评价对象 (Comment Object, CO)、评价内容 (Comment Content, CC), 其他 (Other, OT), 标记结果:

电脑/CO 很/OT 薄/CC, /OT 系统/CO 流畅/CC, /OT 发货/CO 快/CC. /OT 将所有的评论文本都处理成这种的格式.

③ 词性与句法分析. 将评论文本通过 LTP 的词性标注与句法分析接口, 返回数据为 json 数据格式, 从中提取相关特征.

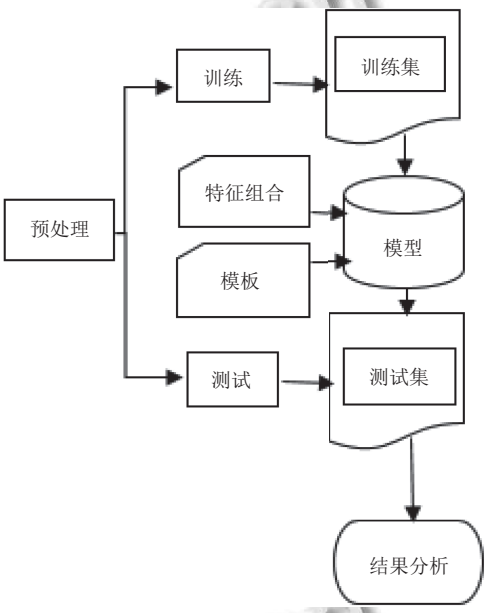

图 1 评价对象抽取总体框架

2.3 条件随机场

条件随机场是一种用于序列标注的概率统计模型, 由 Lafferty 等[10]于 2001 年首次提出, 它结合了最大熵模型和隐马尔可夫模型的特点, 是一种无向概率图模型, 具有表达长距离依赖性和交叠性特征的能力, 能够较好地解决标注偏置等问题的优点, 而且所有特征可以进行全局归一化, 能够求得全局最优解. 条件随机场序列标注模型能较好的捕捉上下文信息[2], 近年来在分词、词性标注和命名实体识别等序列标记任务中取得了很好的效果.

条件随机场模型在用于评论文本评价对象识别时, 输入观察序列, 即经过分词的评论文本 $\mathbf{X}=\{x_1,x_2,x_3,\ldots,x_n\}$, 就可以计算所有可能状态序列的条件概率 $\mathrm{Y}=y_1,y_2,y_3,\ldots,y_n$, 并将最大概率作为序列的输出状态. 计算公式如下:

$$P(y|x)=\frac{1}{Z(x)}\prod_{t=1}^{T}\exp\left\{\sum_{k=1}^{K}\theta_k f_k(y_t,y_{t-1},x_t)\right\} \quad (1)$$

其中, $Z(x)$ 是归一化因子, 它是所有 Y 状态的概率和, 使用它作为分母, 可以确保所求的的概率小于 1, 计算公式为:

$$Z(x)=\sum_{y}\prod_{t=1}^{T}\exp\left\{\sum_{k=1}^{K}\theta_k f_k(y_t,y_{t-1},x_t)\right\} \quad (2)$$

特征组合的定义和模板的选择对评价对象抽取的性能有着重的影响, 本文选择的是 CRFs++_0.58 开源工具包, 第 2.4-2.5 节对特征和模板进行详细的说明.

2.4 特征概述

特征一. 词特征-cont, 是指当前词的字符串特征. 例句 (1) 中的“电脑”、“薄”、“系统”等都是词特征.

特征二. 词性特征-pos, 指当前词特征对应的词性. 例句 (1) 中的“电脑”为名词短语, “发货”为动词, 都是需要提取的评价对象.

特征三. 父词词性特征-ppos, 指父词对应的词性, 例如“电脑“的父词”薄“为形容词.

特征四. 依存关系-relate, 指当前词和父词之间的句法关系, 单个词语组成短语, 再结合成评论句子, 短语中的各单词之间存在诸如动主谓结构 (SBV)、动补结构 (CMP) 等关系, 在大部分 SBV 结构中, 主语要么是意见的持有者, 要么是主题[11]. 例如“电脑薄“、”系统流畅“为主谓结构, ”发货快“为动补结构.

特征五. 领域相关特征-field, 将训练集中的评价对象生成字典, 对于测试集中的词如果在字典中存在, 则对应该特征为 1, 否则为 0, 为布尔类型.

将例句 (1) 进行特征提取的结果如表 1(具体含义可以参照哈工大语言技术平台官网).

表 1 特征提取举例

cont	pos	ppos	relate	tag
电脑	n	a	SBV	CO
很	d	a	ADV	OT
薄	a	-1	HED	CC
,	wp	a	WP	OT
系统	n	a	SBV	CO
流畅	a	a	COO	CC
,	wp	a	WP	OT
发货	v	a	COO	CO
快	a	v	CMP	CC
.	wp	a	WP	OT

2.5 模板定义

条件随机场中的模板定义反映了评论文本中分词之后的相互关系, 模板用于控制窗口的大小和特征组合. 窗口过小, 所包含的信息过少, 不能很好体现随机场中前后单词对当前识别的作用, 理论上, 窗口越大, 可利用的上下文信息就越多[9], 但是窗口过大, 引入信息过多, 可能会出现过拟合, 降低了运行效率和系统性能. 本文定义七种模板, 如下所示:

template_1=(0), 表示只取当前词作为特征, 窗口大小为 1.

template_b1=(-1, 0), 表示以当前词为中心, 同时考虑当前词的前一个词, 窗口大小为 2.

template_a1 =(0, 1), 表示以当前词为中心, 同时考虑当前词的后一个词, 窗口大小为 2.

template_b1_a1=(-1, 0, 1) , 表示以当前词为中心, 同时考虑当前词的前后各一个词, 窗口大小为 3.

template_b2=(-2, -1, 0), 表示以当前词为中心, 同时考虑当前词的前两个词, 窗口大小为 3.

template_a2=(0, 1, 2), 表示以当前词为中心, 同时考虑当前词的后两个词, 窗口大小为 3.

template_b2_a2=(-2, -1, 0, 1, 2), 表示以当前词为中心, 同时考虑当前词的前后各两个词, 窗口大小为 5.

template_b3_a3=(-3, -2, -1, 0, 1, 2, 3), 表示以当前词为中心, 同时考虑当前词的前后各三个词, 窗口大小为 7.

以模板 template_b1_a1 为例, 具体定义如表 2 所示.

表 2 template_b1_a1 模板定义

模板定义	模板含义
U00:%x[-1, 0]	指当前词的前一个词
U01:%x[0, 0]	指当前词
U02:%x[1, 0]	指当前词的后一个词
U03:%x[-1, 0]/%x[0, 0]	指当前词与其前一个词的组合
U04:%x[0, 0]/%x[1, 0]	指当前词与其后一个词的组合
U04:%x[-1, 0]/%x[1, 0]	指当前词的前后两个词的组合

3 实验及结果分析

在实验部分, 针对特征相同而模板不同、模板相同而特征不同两种情况, 对电脑和服饰两类评论数据集进行了评价对象抽取的 F 值的对比与分析, 以选择性能最优的模板以及最有效的特征组合. 本文选择交叉验证, 可以有效提高选择性集成方法的性能[12], 本文取 K=10.

3.1 实验评价指标

本文选用自然语言处理中常用的评价标准, 准确率 (P)、召回率 (R)、F 值, 其中准确率和召回率是广泛用于信息检索领域和统计学分类领域的两个度量值, 用来评价结果的质量, F 值则为准确率和召回率的调和平均值, 它们的计算公式为:

$$P = \frac{\text{抽取正确的评价对象的数量}}{\text{抽取出评价对象的数量}} \quad (3)$$

$$R = \frac{\text{抽取正确的评价对象的数量}}{\text{数据集中的评价对象总数}} \quad (4)$$

$$F = \frac{2 * P * R}{(P + R)} \quad (5)$$

3.2 模板对系统性能的影响

在“电脑”、“服饰”两个领域的数据集上, 分别验证 template_1–template_b3_a3 共七个模板上的评价对象抽取性能, 实验结果如表 3 所示, 并绘制两类数据集的 F 值的变化趋势如图 2 所示.

表 3 不同模板的抽取结果 (单位: %)

指标模板	电脑类			服饰类		
	P	R	F	P	R	F
template_1	78.30	64.40	70.67	89.66	57.38	69.25
template_b1	84.44	66.69	73.94	87.06	58.35	69.87
template_a1	83.93	64.86	72.43	86.04	53.11	65.67
template_b1_a1	84.40	67.87	74.73	83.90	63.55	71.75
template_b2	84.26	64.74	72.30	85.72	59.62	69.93
template_a2	81.84	65.96	72.42	86.68	65.21	73.82
template_b2_a2	85.84	68.13	75.96	82.08	69.05	74.51
template_b3_a3	81.34	65.06	71.74	82.84	67.43	73.74

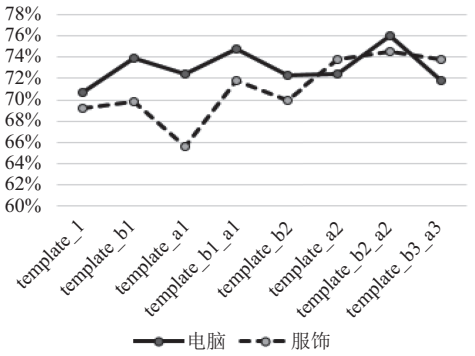

图 2 不同模板的 F 值变化趋势

从折线图中可以看出, 窗口大于 1 的模板的系统性能普遍优于窗口为 1 的性能, 电脑类最高增加 5.29%, 服饰类增加 5.26%. 同时也说明对称模板的识别率要高于非对诚模版的识别率, 即模板中存在相同数量的前后单词更有利于评价对象的识别, 当考虑前后两个单词系统取得最佳的性能, 当窗口大小为 7, 即模板中有前后三个单词时, F 值开始下降, 因为在线商品评论中, 评价单元跨度为 7 个单词的情况较少, 评价单元是指评价对象及其被修饰的评价词的组合[13].

3.3 特征对系统性能的影响

通过不同模板对系统性能影响的实验结果分析可以知道, 窗口过小, 特征利用就不充分, 窗口过大, 会造成系统性能的下降, 并且对称模板的结果优于非对称模板, 因此在验证不同特征对系统性能影响的实验中, 选择 template_b1_a1 和 template_b2_a2 两个模板, 对两类评论数据集进行实验, 以验证单个特征和特征组合对系统性能的影响, 实验结果如表 4, 同时绘制对应的评价对象抽取的 F 值变化趋势 template_b1_a1 如图 3, template_b2_a2 如图 4.

表 4 特征组合的抽取结果 (F 值)(单位: %)

模板特征	template_b1_a1		template_b2_a2	
	电脑	服饰	电脑	服饰
cont	74.73	71.75	75.96	74.51
pos	61.81	67.19	66.39	68.67
ppos	30.06	36.48	34.55	47.67
relate	57.02	65.99	57.98	63.42
cont+pos	77.78	77.56	78.76	78.92
cont+ppos	75.98	78.39	76.07	76.86
cont+relate	76.88	80.49	78.08	77.29
cont+pos+relate	78.08	81.30	78.60	80.67
cont+pos+ppos+relate	76.67	79.36	76.30	78.31
cont+pos+relate+field	82.31	82.93	82.98	83.50

从表 4 中可以看出当单个特征为词自身内容时, 评价对象抽取的 F 值最大, 这也是第一个实验中验证模板不同时的影响只取词特征为模板的基本特征的原因, 在特征组合中, 可以看出任意其他三个特征与词特

征相结合而产生的系统性能都要优于两个特征单独使用时的性能, 其中父词词性产生的相对影响较大, 在商品评论文本中, 评论单元多位主谓结构, 像“电脑薄”、“系统流畅”、“服务态度好”等都是 SBV 结构, 评价对象“屏幕”、“系统”、“服务态度”的父亲词“薄”、“流畅”、“好”都为形容词, 在主谓结构较多的评论文本中, cont+ppos 的组合产生的性能提升较 cont 和 ppos 单独作用时明显. 在特征组合中, 当 cont+pos+ppos+relate 的性能会比 cont+pos+relate 低, 表明特征组合过多会引入噪声, 使系统性能降低.

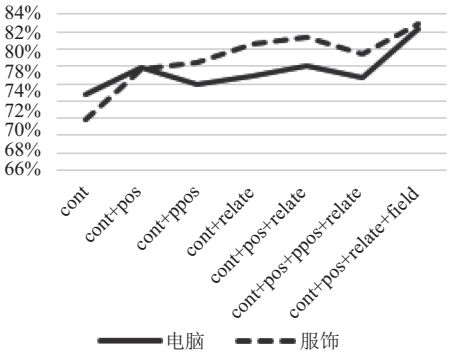

图 3 template_b1_a1 模板 F 值变化

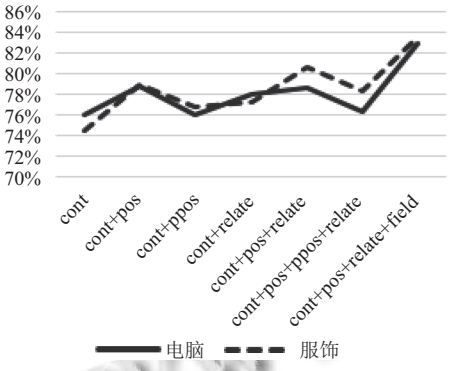

图 4 template_b2_a2 模板 F 值变化

受基于规则的抽取方法的启发, 本文增加了领域相关特征, 使两类数据集的评价对象提取的 F 值最优时分别达到到 82.98% 和 83.50%. 综合分析 3.2 与 3.3 两个实验结果, 基于句法分析与条件随机场的评价对象抽取方法, 针对不同类别的评论语料, 在文本内部语法依存关系的基础上, 充分考虑到评论文本中各单词之间的上下文影响, 较好解决了基于规则方法的泛化问题, 和基于统计方法对文本内部结构的忽视问题.

4 结语

评价对象的正确识别是抽取评价单元的关键, 也是对评论文本进一步分析的基础, 尤其是对于电商平台上相关商品的评论的评价对象的提取, 能够让用户和商家同时受益, 商家可以以此为基础产生可视化报告了解用户关心的产品特征以及第三方提供的物流售后服务情况; 普通用户也可以针对评价对象的提取来辅助自己网上购物, 进而做出更好的购买决策.

本文结合传统的文本分析的两种方法基于关联规则和基于统计分析, 利用哈工大语言技术平台和 CRFs 开源工具, 通过对不同领域的评论文本数据集的实验分析, 得到较优的条件随机场的窗口大小和特征组合. 此外, 在实验中存在因为未能正确分词而导致标记失败的情况, 从而影响了系统的性能, 这也是今后需要改进的地方.

参考文献

1 李丕绩, 马军, 张冬梅, 等. 用户评论中的标签抽取以及排序. 中文信息学报, 2012, 26(5): 14–19, 45.
2 王荣洋, 鞠久朋, 李寿山, 等. 基于 CRFs 的评价对象抽取特征研究. 中文信息学报, 2012, 26(2): 56–61.
3 Hu MQ, Liu B. Mining and summarizing customer reviews. Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA, USA. 2004. 168–177.
4 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. 计算机工程, 2016, 42(7): 173–180.
5 张建华, 肖中正. 结合词性规则和依存句法分析的评价对象抽取方法. 计算机与现代化, 2016, (4): 16–20.
6 Jakob N, Gurevych I. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, Massachusetts. 2010.
7 金丽君. 基于 SVM 的搜索型商品评论有用性自动识别方法研究[硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2013.
8 刘玮楠. 基于 HNC 理论的网购评论情感倾向性分析研究[硕士学位论文]. 大连: 大连理工大学, 2013.
9 杨云. 基于句法结构的评价对象抽取方法研究[硕士学位论文]. 长春: 东北师范大学, 2015.
10 Lafferty J, McCallum A, Pereira FCN, *et al*. Conditional

random fields: Probabilistic models for segmenting and labeling sequence data. Proc. of the 18th International Conference on Machine Learning. Williams College, Williamstown, MA, USA. 2002. 282–289.

11 姚天昉, 娄德成. 汉语语句主题语义倾向分析方法的研究. 中文信息学报, 2007, 21(5): 73–79.

12 胡局新, 张功杰. 基于 K 折交叉验证的选择性集成分类算法. 科技通报, 2013, 29(12): 115–117. [doi: 10.3969/j.issn.1001-7119.2013.12.039]

13 刘丽, 王永恒, 韦航. 面向产品评论的细粒度情感分析. 计算机应用, 2015, 35(12): 3481–3486, 3505. [doi: 10.11772/j.issn.1001-9081.2015.12.3481]