# 面向中文微博的评价对象与评价词语联合抽取

刘全超,黄河燕,冯 冲
(北京理工大学计算机学院,北京 100081)

**摘 要:** 深入挖掘微博内容中评价对象与评价词语的词法特征、句法特征、语义特征以及相对位置特征,提出评价对象与评价词语的序列化联合抽取模型.进一步结合微博间转发关系特性提出基于转发关系的联合抽取优化算法.并与相关算法进行实验对比,对实验结果进行了综合分析,证明了方法的可行性和优越性.

**关键词:** 观点挖掘; 信息抽取; 社交网络; 评价对象; 评价词语; 微博
**中图分类号:** TP391 **文献标识码:** A **文章编号:** 0372-2112 (2016)07-1662-09
**电子学报 URL:** http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.07.021

# Co-Extracting Opinion Targets and Opinion-Bearing Words in Chinese Micro-Blog Texts

LIU Quan-chao, HUANG He-yan, FENG Chong
(*Department of Computer Science and Technology, Beijing Institute of Technology, Beijing* 100081, *China*)

**Abstract:** Using lexical, syntactic, semantic and relative position features to extract opinion pairs <opinion target, opinion-bearing word> in micro-blog, we put forward the co-extracting model, and then give co-extracting opinion pairs optimization algorithm based on forwarding between micro-blogs. According to the experimental results, our two-stage approach greatly improves the performances of co-extracting opinion pairs.

**Key words:** opinion mining; information extraction; social network; opinion target; opinion-bearing word; micro-blog

## 1 引言

观点信息抽取是情感分析的最底层任务,其目的是获取情感评论文本中有意义的信息单元,如观点持有者(opinion holder)、评价对象(opinion target)、评价词语(opinion-bearing word)等情感要素.目前的已有工作主要是利用监督和无监督机器学习算法实现.无监督方法主要是利用词频统计或规则实现评价对象和评价词语抽取,如Hu等人[1]、Popescu等人[2]、Zhang等人[3]以及刘等人[4~8],他们认为评价对象往往是名词或名词短语,而评价词语往往是形容词,并且评价对象与评价词语之间具有评价关系.如果一个词是一个评价对象,则与之具有评价关系的形容词很可能就是一个评价词语;反之,如果一个词是一个评价词语,则与之具有评价关系的名词、名词短语很可能就是一个评价对象.他们将研究重点放在了评价关系发现,进而实现情感要素的抽取.目前相比较无监督方法而言,有监督学习方法取得较好性能.Wilson等人[9,10]提出基于分类器方法来抽取评价词语并进一步判定其情感倾向性,但此类基于分类器的抽取方法,是独立的抽取评价词语或评价对象,缺乏评价对象和评价词语之间的类别对应关系.Jin等人[11]提出基于隐马尔科夫模型的序列标注算法来抽取评价对象和评价词语,并给出情感倾向性.此方法考虑了句子间的序列关系,但马尔科夫模型是产生式模型,并不适合于充分利用数据的高维特征.

后来不少研究者发现,情感要素的组合搭配对情感分析有着更直接的帮助.赵等人[12]利用二元评价搭配进行了情感极性消歧的任务.吕等人[13]利用评价对象和评价词之间的修饰关系进行了在线产品评论用户满意度综合评价研究.另外正确抽取评价对象、评价词和它们之间的对应关系,可以生成便于用户阅读的基于评价对象和评价词语的文章摘要[14].

当前评价搭配抽取既是热点也是难点,尤其是非受限领域微博内容.在传统媒体中,如新闻类型的报道中进行观点持有者抽取是有意义的,因为不同的媒体拥有不同的观点信息.然而面对微博这种新媒体是没有必要进行观点持有者抽取的,因为发帖者往往是观点持有者,所以本文我们重点进行了面向微博的评价

收稿日期:2015-01-29;修回日期:2015-07-02;责任编辑:李勇锋
基金项目:国家973重点基础研究发展计划(No.2013CB329605)

对象和评价词语的联合抽取研究,即抽取微博内容中的二元评价搭配结构 < 评价对象,评价词语 > . 通过分析和研究,我们将该任务看作是序列化标注任务,提出多特征融合的评价搭配抽取算法,并结合微博的传播特性对该算法进行了优化. 将实验结果与施[15]的方法进行了实验对比,并对实验结果进行了综合分析.

## 2　条件随机场模型

### 2.1　条件随机场原理

条件随机场(Conditional Random Fields,CRFs)模型是由 Lafferty 等人[16]于 2001 年,在最大熵模型和隐马尔科夫模型的基础上,提出的一种判别式概率无向图学习模型,是一种用于标注和切分有序数据的条件概率模型. 较简单且最常用的 CRFs 模型是一阶链式结构模型,如图 1 所示,(*a*)与(*b*)均是 CRFs 模型的图表示模型,着色节点表示观察节点,未着色节点表示状态节点.

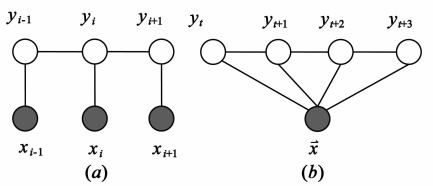



**图1　一阶链式结构CRFs模型**

若用 $X=(x_1,x_2,\cdots,x_n)$ 表示一个观察序列,$Y=(y_1,y_2,\cdots,y_n)$ 表示为状态(标注)序列,则在给定一个观察序列的情况下,一阶链式结构的 CRFs 模型定义为:

$$P(Y|X)=\frac{1}{Z(x)}\exp\Big(\sum_{t=1}^{n}\sum_{k}\lambda_k f_k(y_{t-1},y_t,X,t)\Big) \quad (1)$$

其中 $Y$ 是字符串的标记序列,$X$ 是待标记的字符,$f_k(y_{t-1},y_t,X,t)$ 是一个任意的特征函数,$\lambda_k$ 是对应的特征函数的权重,而 $Z(X)$ 是归一化因子,使得上式成为概率分布,其中

$$Z(X)=\sum_{y}\exp\Big(\sum_{t=1}^{n}\sum_{k}\lambda_k f_k(y_{t-1},y_t,X,t)\Big) \quad (2)$$

CRFs 最早是针对序列数据分析提出的,本文使用一阶链式结构 CRFs 模型来描述句子中单词出现的序列化结构关系,将评价对象和评价词语的联合抽取任务看作是序列标注任务. 目前基于 CRFs 模型的主要系统实现有 CRF、FlexCRF① 以及 CRF ++,我们采用著名的条件随机场开源工具包 CRF ++0.58②,它是目前综合性能最佳的 CRFs 工具之一. 其使用过程大概分为四个步骤:(1)数据预处理,包括数据清洗和生成 CRF ++所规定的数据格式;(2)生成特征模板(template_file);(3)训练;(4)测试. 工具包 CRF ++0.58 使用之前,必须将训练和测试数据转换成其所规定的数据格式,并事先指定特征模板(template_file),该文件描述了训练和测试时用到的特征情况. 实验过程中可以设计多种类型的特征模板进行实验操作,数据格式以及特征模板请参考 CRF ++ 官方网站 http://crfpp.sourceforge.net/.

### 2.2　特征模板

CRF ++是一个泛用的工具,使用时必须事先指定一个特征模板. 我们在实验过程中设计了三种类型的特征模板:T0、T1 和 T2. 我们将 T0 设为实验所用的 CRF ++默认模板,具体如图 2 所示. 其中,"#"后面的部分属于注释内容.

```
#Unigram
#Token                          #词汇特征
U01:%x[-1,0]                    #上个词
U02:%x[0,0]                     #当前词
U03:%x[1,0]                     #下个词
U04:%x[-1,0]/%x[0,0]            #上个词和当前词
U05:%x[0,0]/%x[1,0]             #当前词和下个词
#POS                            #词性特征
U11:%x[-1,1]                    #上个词的词性
……                              ……
#Bigram
B
```

**图 2　默认模板 T0**

为了对比不同特征模板对实验性能的影响,我们又设计了特征模板 T1 和 T2. T1 中将 T0 中上下文信息的条件项去掉,只保留 T0 中当前词的各个特征,即 U02:%x[0,0],U12:%x[0,1],U22:%x[0,2],U32:%x[0,3]和 U42:%x[0,4],生成新模板 T1,如图 3 所示. 特征模板 T2 充分考虑了特征信息的组合信息,在 T0 的基础上,针对特征列表(Token/POS/DDR/SRL/WD)设计了一个新模板 T2,增加了如图 4 所示的特征组合信息.

```
#Unigram
#Token                          #词汇特征
U02:%x[0,0]                     #当前词
#POS                            #词性特征
U12:%x[0,1]                     #当前词的词性
#DDR                            #句法特征
U22:%x[0,2]                     #当前词的句法特征
#SRL                            #语义特征
U32:%x[0,3]                     #当前词的语义特征
……                              ……
# Bigram
B
```

**图 3　特征模板 T1**

① http://sourceforge.net/projects/flexcrfs/

② http://code.google.com/p/crfpp/downloads/list

```
#Unigram
#Token + POS + DDR                                          #词汇 + 词性 + 句法特征组合
U51:%x[0,0]/%x[0,1]/%x[0,2]                                 #当前词的特征组合信息
#Token + POS + SRL                                          #词汇 + 词性 + 语义特征组合
U61:%x[0,0]/%x[0,1]/%x[0,3]                                 #当前词的特征组合信息
#Token + POS + WD                                           #词汇 + 词性 + 位置特征组合
U71:%x[0,0]/%x[0,1]/%x[0,4]                                 #当前词的特征组合信息
# Token + POS + DDR + WD                                    #词汇 + 词性 + 句法 + 语义特征组合
U81:%x[0,0]/%x[0,1]/%x[0,2]/%x[0,4]                         #当前词的特征组合信息
# Token + POS + SRL + WD                                    #词汇 + 词性 + 语义 + 位置特征组合
U91:%x[0,0]/%x[0,1]/%x[0,3]/%x[0,4]                         #当前词的特征组合信息
# Token + POS + DDR + SRL                                   #词汇 + 词性 + 句法 + 语义特征组合
UA1:%x[0,0]/%x[0,1]/%x[0,2]/%x[0,3]                         #当前词的特征组合信息
# Token + POS + DDR + SRL + WD                              #词汇 + 词性 + 句法 + 语义 + 位置特征组合
UB1:%x[0,0]/%x[0,1]/%x[0,2]/%x[0,3]/%x[0,4]                 #当前词的特征组合信息
# Bigram
B
```

**图 4 模板 T2 中增加的特征组合信息**

## 3 基于统计方法观点信息联合抽取

句子的序列化结构关系,对评价对象及其评价词语的标签类别判断有直接的帮助. 在获得评价对象与评价词语的词法特征、句法特征、语义特征以及相对位置特征后,利用一阶链式结构的条件随机场模型,来描述句子中评价对象与评价词语出现的序列化结构关系,实现面向微博句子级的评价对象和评价词语的联合抽取.

### 3.1 词法特征抽取

(1)词汇特征. 词汇(token)是自然语言中最小的有意义的构成单位,在信息抽取和情感分析方面具有十分重要的作用. 然而在中文中,"词"是没有清晰地界限的[17],分词便成为了观点信息抽取的首要工作. 微博自媒体的出现伴随着大量的未登录词和网络用语,因此在观点信息抽取过程中,对未登录词和网络用语的识别十分重要. 本文重点在文本分词后的评价对象抽取过程,对分词系统不做深入研究,所以直接采用哈工大社会计算与信息检索研究中心提供的语言技术平台(Language Technology Platform,LTP)开源工具包①进行微博内容分词. LTP 提供了一整套自底向上的丰富而且高效的中文语言处理模块,主要包括分词、词性标注、命名实体识别、依存句法分析以及浅层语义标注等中文自然语言处理技术. LTP 的分词模块是基于机器学习框架、且模型中融入了用户词典策略,使得 LTP 分词模块可以很便捷地加入新词信息,利于微博分词,且对网络用语识别具有较好的效果,如"屌丝"、"斑竹"等网络用语.

在评价对象与评价词语的联合抽取过程中,使用词汇特征训练测试 CRFs 模型得出预测结果,从而避免了因识别词而导致的错误,词汇特征的选用能使联合抽取达到不错的效果.

(2)词性特征

词性也叫词类,是根据一个词的本意及在短语或句子中的作用划分的,主要用来描述一个词在上下文中的作用. 词性标注(Part-of-Speech tagging,POS tagging)是指对句子中的每个词指派一个合适的词性,即确定每个词是动词、副词、形容词、名词或其他词性的过程,又称为词类标注. LTP 词性标注模块中使用支持向量机[18]进一步提升了词性标注的准确率,并针对数据稀疏问题,特别是分词阶段的新词,引入了汉字特有的偏旁部首特征进一步提高了词性标注的泛化能力. 本文采用 LTP 进行微博内容词性标注. 如"比亚迪是非常节能的"LTP 词性标注结果为"比亚迪/nh 是/v 非常/d 节能/v 的/u./wp".

词性特征表明了一个词在句子中的作用. 评价对象往往是名词或名词短语(如比亚迪/nh),评价词语则通常是形容词或动词(如节能/v),利用词性信息训练测试 CRFs 模型测试评价对象与评价词语的联合抽取,能够取得较好的效果.

### 3.2 句法特征抽取

依存句法分析是指将一个线性序列的句子转化为一棵结构化的依存分析树,通过依存弧上的关系标记反映句子中词汇之间在句法上的语义相关联的搭配关系. 如"酒店位置很不错,交通很方便,是一个不错的酒店.",其依存句法分析结果如图 5 所示. 其中"位置"与"不错"、"交通"与"方便"均有主谓关系(SBV),"不错"与"酒店"有定中关系(ATT). 由此可见,评价对象

① https://github.com/HIT-SCIR/ltp

和评价词语之间往往存在着直接依存关系(Direct Dependency Relation,DDR),这种句法特征有利于评价对象和评价词语的联合抽取.

利用依存句法分析进行评价对象抽取,其核心思想是:在句子依存句法分析结果中,首先依据情感词典定位评价词语,其次以评价词为中心寻找恰当的依存关系,最后在依存关系中获得评价对象. 如图 5 所示例句,获得评价词语“不错”、“方便”以及 SBV、ATT 依存关系后即可得到评价对象“位置”、“交通”以及“酒店”.

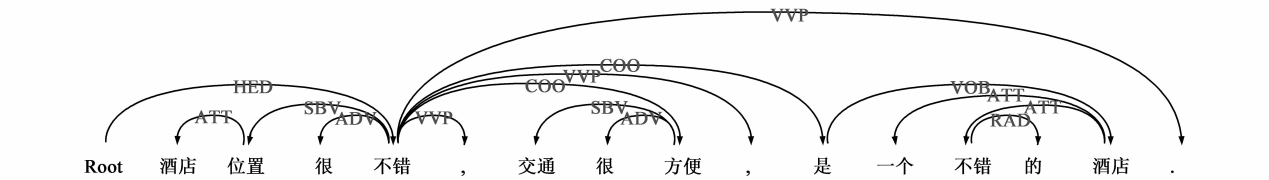



图5　依存句法分析示例

句法特征抽取的任务是从句法分析结果中抽取出合适的依存关系特征信息,主要抽取的依存关系有主谓关系(SBV)、动宾关系(VOB)以及定中关系(ATT). 每条微博的 LTP 依存句法分析返回结果中,每个词的句法信息占一行,每一行独占三列:第一列为依存句法分析的孩子节点信息,由“节点名 + 下划线 + 词 ID”组成;第二列为依存句法分析的父亲节点信息,由“节点名 + 下划线 + 词 ID”组成,如果没有父亲节点,则由“ - 1”表示;第三列为具体的依存句法分析关系. 句与句的信息之间用两个换行分割.

### 3.3　语义特征抽取

语义分析是根据句子的句法结构和句中词汇的词义,推导出能够反映句子意义的某种形式化表示,对句子进行语义分析有利于评价对象和评价词语的联合抽取. 语义角色标注(Semantic Role Labeling,SRL)是目前语义分析的一种主要实现方式,采用“谓语动词-角色”的结构形式,即针对给定谓语动词,标注句子中某些短语承担的语义角色,每一个语义角色被赋予一定的语义含义.

通过对微博数据集的人工统计分析发现,评价对象往往会担任某个谓语动词的语义角色——施事者或受事者,因此利用 SRL 语义特征能够较好地捕捉评价对象. 仍然采用 LTP 进行语义角色标注. LTP 中核心的语义角色为 A0-5 六种,A0 通常表示动作的施事者,A1 通常表示动作的受事者,A2-5 根据谓语动词不同会有不同的语义含义. 评价对象抽取过程中只考虑施事者(A0)和受事者(A1). 如“比亚迪是非常节能的”,其 SRL 标注图示如图 6 所示. 句中“比亚迪”是谓词“是”和“节能”的施事者(A0),从而获得带有评价词的二元评价搭配 < 比亚迪,节能 >.

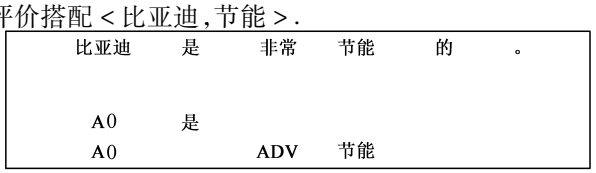



图6　SRL标注示例

评价对象抽取核心思想:在句子 SRL 标注结果微博集中,首先定位谓词(动词、名词等)是评价词的 SRL 标注,以谓词为中心寻找当前句中的施事者(A0)作为评价对象,施事者不存在时选取受事者(A1)作为评价对象.

在语义角色标注过程中,施事者或受事者有时包含多个词汇,我们选取其中的名词作为 SRL 标注结果. 在例句“酒店位置很不错,交通很方便,是一个不错的酒店.”中,名词“酒店”和“位置”均标为 A0,如图 7 所示.

谓语动词的语义角色信息对评价对象抽取有着重要的作用,我们的任务是从 SRL 标注句子中抽取出合适的语义角色特征信息,主要抽取的角色类型有施事者(A0)、受事者(A1)以及谓词信息(PRD). 每条微博的 LTP 语义角色标注返回结果中,每个谓词的语义角色信息独占一行,如果一个句子中有多个谓词,那么占用多行,且行与行之间用换行分割,句与句之间用两个换行分割.

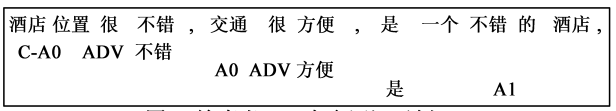



图7　施事者(A0)中多词汇示例

### 3.4　相对位置特征

在用自然语言文本表达观点时,句子中评价对象与评价词语的距离(Word Distance,WD)往往是比较近的,尤其是对于内容短小、撰写不规范的微博来说,当句法特征和语义特征不存在时,位置特征尤为重要. Yi 等人[19]以及 Hu 等人[20]曾利用依存关系特征进行产品评论中评价对象的识别,然而 Jakob 等人[21]为了防止评价对象和评价词语不存在依存关系而遗漏信息,进一步使用了相对位置信息作为 CRFs 模型的另一特征.

评价对象经常出现在评价词的前边或者后边,为了防止词法特征、句法特征以及语义特征不存在时而漏掉评价对象信息,进而利用评价对象与评价词的相

对位置特征作为 CRFs 模型的另一特征,实现微博中评价对象与评价词语的联合抽取.

## 4 基于转发关系的联合抽取优化

微博有两个主要功能:一方面是认识更多的人,另一方面是维系当前的朋友关系.《2014 年中国社交类应用用户行为研究报告》①调查结果显示,微博的分享/转发信息功能使用比例高达 66.6%.其中,新浪微博用户对微博主要功能的使用率较高,60.3% 的新浪微博用户主动分享/转发信息.另据实验室两位人员分别用自己的新浪账户统计,均发现超过 70% 的微博拥有转发关系.

微博内容短小、简洁,常含有隐性评价对象,不利于评价搭配 < 评价对象,评价词语 > 抽取.通过对带有转发关系的微博统计分析,发现利用微博间转发关系可以克服这一难点.来自新浪网页的微博示例如图 8 所示.用户"Gary713"发布原始博文"杰出科学家余凯将来我校做作报告",而用户"Toby_BIT"对其进行了转发并发布"威武"博文表示对原始博文内容的认同.根据第 3 部分的描述可知原始博文存在评价搭配 < 科学家余凯,杰出 >,转发微博同样存在评价搭配 < 科学家余凯,威武 >.


图8 新浪微博示例

带有转发关系的微博提供了更加丰富的信息,一般来说,"转发"往往意味着对原始微博内容以及用户观点的赞同,所以我们做出如下假设,其中评价对象记作 $OT$,评价词记作 $OW$.

**假设** 当转发微博只含有评价词 $OW$ 时,计算 $OW$ 与原始微博评价搭配 $< OT_i, OW_i >$ 中 $OW_i$ 的语义相似度,取得 $\max Sim(OW, OW_i)$ 时的 $OT_i$ 作为转发微博中 $OW$ 的评价对象,即 $< OT_i, OW >$.

评价词的语义相似度计算采用 Hownet API② 实现.Hownet 通过用一系列的义原,利用某种知识描述语言来描述一个概念,而这些义原通过上下位关系组织成一个树状义原层次体系.对于两个评价词语 $OW_1$ 和 $OW_2$,如果 $OW_1$ 有 $n$ 个义元(概念):$S_{11}, S_{12}, \cdots, S_{1n}$,$OW_2$ 有 $m$ 个义元(概念):$S_{21}, S_{22}, \cdots, S_{2m}$,我们规定,$OW_1$ 和 $OW_2$ 的语义相似度为各个概念相似度中的最大值,即:

$$Sim(W_1, W_2) = \max_{i=1\cdots n, j=1\cdots m} Sim(S_{1i}, S_{2j}) \quad (3)$$

通过基于转发关系的隐性评价对象抽取,进而提高评价搭配联合识别的性能.

## 5 实验及其结果分析

### 5.1 数据集及人工标注

是实验数据由两部分组成.一部分来自 CCF TCCI 主办的"自然语言处理与中文计算会议(NLP&CC)"提供的标注数据集,标记为 DataSet-1.包含四个话题:"毁容案"、"Ipad"、"抗日神剧"以及"科比",共 405 条微博消息.这些数据集中拥有转发关系的微博较少,所以我们依据第 4 部分描述的 70% 原则,人工对 405 条中的 280 条内容较接近的微博添加了一层转发关系"forward",即微博间只存在一次转发.保留标注为"opinionated = Y"的 602 句主观句用于 CRF 模型训练与测试,并将标注为"target_word"的词语和评价词典(来自文献[22])作为分词系统的用户词典,分词后生成评价搭配 < 评价对象,评价词语 > 用于实验性能评估,最终获得 806 个标注结果.另一部分数据集来自新浪微博.为了扩充实验数据规模以及转发关系的真实性,爬取话题为"iphone5"和"袁隆平"的 2200 条微博,具有多层转发关系.通过人工去除只有表情符号、图片或超链接的微博,保留其中较完整的 2000 条微博,标记为 DataSet-2.人工标注其中的评价搭配 < 评价对象,评价词语 >,且包含了转发微博的评价搭配,如标注图 8 中示例,会有 < 科学家余凯,杰出 >、< 科学家余凯,威武 >.对于标注有争议的评价搭配暂不保留,最终获得 3632 个标注结果.

### 5.2 数据预处理

实验前对实验数据进行预处理,定义如下过滤规则使微博内容更加规整:

规则 1:对微博内容按照转发关系"//"进行划分,并且使微博内容顺序翻转,这样保证转发微博是基于原始微博进行分析的;

规则 2:对微博内容中用户名进行删除,即删除"@+用户名"结构,且删除如"http://t.cn/h87oy"等超链接;

规则 3:对连续出现多个标点符号情况,如"。。。。。。","!!!!"等,采用第一个标点符号进行替换,并去除微博内容中的表情符号;

规则 4:对于微博内容中含有"#话题#"的情况,则把"话题"直接作为候选评价对象;

经过上面的规则预处理,在结构上进行了调整,利于微博内容进行更加深入的分析.

### 5.3 实验设置

实验采用有监督学习方法,为避免过学习或欠学

① http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/

② http://www.keenage.com/html/e_index.html

习状态发生,我们采用 5 折交叉验证,即随机将数据集分成 5 份,取其中 4 份训练用,另一份测试用,重复 5 次,最后取平均值. 实验中的正确率、召回率和 F 值均是 5 折交叉验证的结果.

另外,把 602 句标注训练集分成大小不同的数据集合,分别进行 5 折交叉验证以观察数据规模不同条件下的实验结果差异性,并进一步验证不同的特征模板对实验性能的影响.

### 5.4　标注集

利用 CRF 模型进行训练和测试过程中,需要将数据转换为 CRF ++所需要的数据格式,共六列,分别代表了词汇特征、词性特征、句法特征、语义特征、相对位置特征以及标注结果. 其中如果最后一列的标记过于复杂,容易导致特征稀疏现象,所以设计了相对简单的标注集,如表 1 所示. 在标注结果序列中如果有出现连续相同标注,我们则判定其为同一对象.

**表 1　标注集及相关说明**

| 标注集 | 相关说明 |
|---|---|
| OT | 评价对象 |
| OW | 评价词语 |
| BG | 其他无相关词 |

以“刚/BG 在/BG 网上/BG 看到/BG iphone5/OT 概念机/OT,/BG 忒/BG 漂亮/OW ！/BG”为例,通过标注我们可以清楚地分析出这句评论中的评价搭配 < iphone5 概念机,漂亮 >.

### 5.5　实验结果

发掘微博中评价对象和评价词间的多种特征,改进 CRF 模型的特征模板,结合微博转发关系特性进行句子级的评价对象与评价词语联合抽取,我们做了三组实验.

(1)基准系统

Hu 等人[20]认为词性是判断情感信息的重要依据,采用词法特征作为我们的基准系统. 实践中经过多次实验,发现当特征模板中特征窗口大小为 2 时整体性能表现较好,所以实验中特征窗口阈值均为 2. 由于基准系统只考虑词汇特征和词性特征,故将默认模板 T0 中其他特征信息删掉,如 DDR、SRL、WD 等特征信息.

我们对 DataSet-1 中的 602 句主观句和 405 条微博以及 DataSet-2 分别进行了实验,在同一默认模板 T0、不同规模数据集条件下进行了 5 折交叉验证,实验结果如表 2 所示.

从表 2 可知,当随着数据规模不断增大时,联合抽取的整体性能也在不断递增. 另外,实验结果中正确率均要比召回率好一些,这说明在联合抽取过程中,漏掉了一些评价搭配,需要我们引入更多特征信息进行进一步挖掘.

实验暂不考虑评价搭配与句子观点倾向性的关系,即评价搭配既可以出现在观点句中,也可以出现在非观点句中. 如非观点句“因此苹果向物流公司提供最可观的费用”中即存在评价搭配 < 费用,可观 >. 为了验证转发关系在联合抽取过程中的作用,需要对 DataSet-1 中 602 句观点句以外的其它 335 句非观点句进行评价搭配标注,最终获得 895 个标注结果. 然后对 DataSet-1 中的 405 条微博和 DataSet-2 中的 2000 条微博进行实验,有无转发关系的实验结果如表 3 所示.

**表 2　602 句主观句实验结果**

| 词法特征 | DataSet-1(句) | 结果 | | |
|---|---|---|---|---|
| | | P(%) | R(%) | F(%) |
| Token + POS | 100 | 68.7 | 61.0 | 64.6 |
| | 200 | 72.7 | 63.4 | 67.7 |
| | 400 | 72.6 | 69.9 | 71.2 |
| | 602 | 75.8 | 72.5 | 74.1 |

**表 3　有无转发关系的实验对比结果**

| 词法特征 | 数据集 | 无转发关系 | | | 有转发关系 | | |
|---|---|---|---|---|---|---|---|
| | | P(%) | R(%) | F(%) | P(%) | R(%) | F(%) |
| Token + POS | DataSet-1 | 70.8 | 68.5 | 69.6 | 71.9 | 69.1 | 70.5 |
| | DataSet-2 | 59.3 | 42.0 | 49.2 | 74.0 | 64.7 | 69.0 |

从表 3 可知,微博间的转发特征对评价对象和评价词语的联合抽取有着重要的作用. 这种转发关系对 DataSet-1 的效果不明显而对 DataSet-2 的效果却显著,并且 DataSet-2 的抽取结果整体性能(F 值为 69%)不如 DataSet-1 的效果好. 我们人工分析对比了 DataSet-1 和 DataSet-2 两个数据集,导致这种现象的主要原因是因为 DataSet-1 数据集经过人工处理,且其间的转发关系是人工随机添加,缺乏内容上的衔接,而 DataSet-2 是来自新浪微博的原始数据,内容连贯. 然而,这却是造成 DataSet-2 无转发关系时联合抽取性能较低的主要原因.

(2)引入句法特征、语义特征、位置特征后的系统实验

在基准系统基础之上,我们逐步引入句法特征、语义特征、位置特征,进行多特征融合的联合抽取实验. 句法特征是布尔型特征,指示当前词与评价词语是否有直接的 SBV、VOB 或 ATT 依存关系,有直接的依存关系记为 1,否则为 0. 语义特征是识别出事件的施事者和受事者,保留最小语义角色单元的 Arg0 和 Arg1 两种信息. 相对位置特征是布尔型特征,指示句子中与评价词语距离最近的名词或名词短语. 按照 CRF 模型处理文件的格式要求,将句法特征、语义特征、位置特征处理后的结果添加到 CRF 模型训练与测试文件中进行实验.

此次实验中,将默认特征模板 T0 复原,按照下述 Rule-X 规定,依次恢复其它特征信息,对去除转发关系的 DataSet-1 和 DataSet-2 进行多特征融合的联合抽取实验对比,结果如表 4 所示. 为了书写方便,内容简洁,我们做了如下规定以表相应特征组合:

Rule-1:Token + POS;

Rule-2:Token + POS + DDR;

Rule-3:Token + POS + DDR + SRL;

Rule-4:Token + POS + DDR + SRL + WD.

从表 4 中可以看出词法特征、句法特征、语义特征以及位置特征对联合抽取性能的影响. 总体来说,基于多特征融合的联合抽取性能要好于我们的基准系统,也说明 CRFs 模型过度依赖于特征,特征选取的好与坏直接影响到识别效果. 不过实验过程中,在加入 DDR 特征后,性能反而略低于基准系统,经过分析联合抽取结果发现,主要错误出现在评价对象的精确识别方面,即评价对象的边界识别影响了算法性能,这主要是由以下两个方面原因造成的:一方面是评价对象本身的构成较为复杂,组成不规范,例如训练语料中标注的评价对象是"黄渤的肢体",而我们往往得到的却是"肢体". 另一方面是由分词带来的噪音. 因此,如何解决复杂结构短语以及分词的纠错是我们今后研究的主要改进方向.

**表 4 多特征融合的联合抽取实验结果**

| 特征 | 602 句主观句 | | | 405 条微博 | | | 2000 条微博 | | |
|---|---|---|---|---|---|---|---|---|---|
| | P(%) | R(%) | F(%) | P(%) | R(%) | F(%) | P(%) | R(%) | F(%) |
| Rule-1 | 75.8 | 72.5 | 74.1 | 70.8 | 68.5 | 69.6 | 59.3 | 42.0 | 49.2 |
| Rule-2 | 73.4 | 71.8 | 72.6 | 69.1 | 66.7 | 67.9 | 64.0 | 68.7 | 66.3 |
| Rule-3 | 77.8 | 80.1 | 78.9 | 73.9 | 74.7 | 74.3 | 73.9 | 84.1 | 78.7 |
| Rule-4 | 80.2 | 84.6 | 82.3 | 74.8 | 78.9 | 76.8 | 77.2 | 87.1 | 81.9 |

同时我们进行了基于转发关系特性的联合抽取性能优化实验,将带转发关系的 DataSet-1 和 DataSet-2 数据集在 Rule-4 特征选择下,进行实验对比,结果如表 5 所示.

**表 5 基于转发特性的性能优化实验结果**

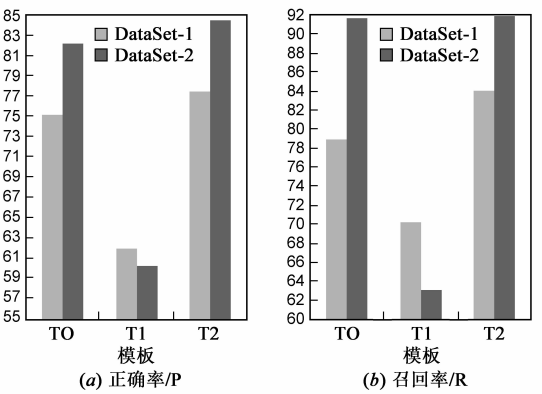
| 特征 | 数据集 | 无转发关系 | | | 有转发关系 | | |
|---|---|---|---|---|---|---|---|
| | | P(%) | R(%) | F(%) | P(%) | R(%) | F(%) |
| Rule-4 | DataSet-1 | 74.8 | 78.9 | 76.8 | 75.1 | 79.0 | 77.0 |
| | DataSet-2 | 77.2 | 87.1 | 81.9 | 82.1 | 91.7 | 86.6 |

从表 5 可知,在 Rule-4 特征情况下,转发关系仍然起到了积极作用,而且对 DataSet-2 的实验效果有了比较大程度的提高. 再次证明转发关系特性在微博观点信息联合抽取中的重要作用. 从 DataSet-1 和 DataSet-2 的数据构成来说,基于转发关系的性能优化不仅仅依赖于选取的微博特征,还依赖于拥有共同关注点的微博用户群体.

(3)不同特征模板条件下的系统实验

通过上述(1)和(2)的实验我们发现,总体来说 CRFs 模型中引入某一个独立特征信息均使整体性能有一定程度的提高. 那么是否特征的组合信息又会同独立的特征信息一样,对微博观点信息联合抽取起到积极作用呢? 我们对带有转发关系的 DataSet-1 和 DataSet-2 进行了最优特征模板选择实验,即在相同的数据条件、不同特征模板下进行,具体实验结果如图 9 所示.

图 9 显示,模板 T0 和模板 T2 在召回率指标上十分接近,但在准确率指标上模板 T2 要优于模板 T0. 而模板 T1 的性能最差,因为我们在设计特征模板 T1 时没有考虑特征的上下文信息,在设计特征模板 T0 和 T2 时考虑了上下文信息,由于词语所处的特征上下文的类别标签、以及特征信息组合的类别标签对目标词类别标签的判断具有十分重要作用,所以模板 T0 和 T2 的实验结果也相对更好. 说明特征的组合信息同样会对微博观点信息的联合抽取起到积极作用.




(*a*) 正确率/P (*b*) 召回率/R

**图9 CRFs模型不同特征模板实验结果**

另外,我们在 F 值综合指标上,与施[15]的算法进行了实验对比,即在相同的数据条件下,采用不同特征进行的实验对比. F 值计算如下:

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

施用到 Token + POS + SRL 特征及其相应的特征模板,记作 Algorithm-1. 我们采用 Token + POS + DDR + SRL + WD 特征以及模板 T2,记作 Algorithm-2. 对 DataSet-1 中 602 句观点句进行实验,实验结果如图 10 所示.

综合分析图 10 中的实验结果,本文算法的综合指标性能要好于施的算法,说明在 CRFs 模型中往往多个特征的综合应用效果会更好,也再次证明 CRFs 模型比较依赖于特征选取.

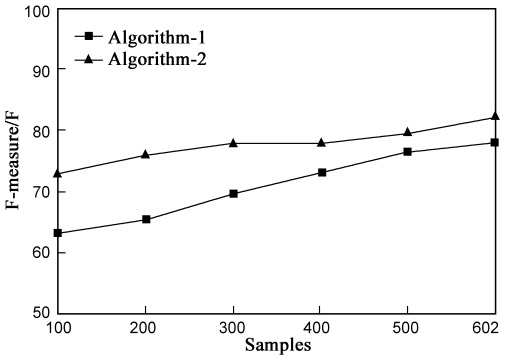



**图10　F值综合指标对照图**

(4)与其他评价搭配抽取算法对比实验

我们对基于转发关系的评价搭配抽取性能进行了验证,并与 Popescu 等人[23]和姚等人[24]的评价搭配抽取方法进行了对比. Popescu 等人利用 MINIPAR parser 手工构建了 10 条依存句法抽取规则来获取隐性评价搭配,姚等人同样利用依存句法分析总结出六组"上行路径"和"下行路径"匹配规则来识别评价搭配.

对比实验过程中我们构建了 SBV、VOB、ATT、ADV、APP、COO、DE、DI、DEI、IC 十组规则,分别表示语法关系主谓、动宾、定中、状中、同位、并列、"的"字结构、"地"字结构、"得"字结构和独立分句等,并对数据集 DataSet-2 使用了统一的情感词典和网络用语库资源,在文献[24]方法验证过程中并未使用本体知识,实验结果如下表 6 所示.

**表 6　评价搭配抽取算法对比实验结果**

| 算法 | DataSet-2 | | |
|---|---|---|---|
| | P(%) | R(%) | F(%) |
| 文献[23] | 76.9 | 81.5 | 79.1 |
| 文献[24] | 72.4 | 83.8 | 77.7 |
| Algorithm-2 | 84.3 | 91.8 | 87.9 |

从上表 6 实验结果可以看出,基于转发关系的评价搭配联合抽取性能达到了最优,尽管 Popescu 和姚的工作融入了较多对评价对象和评价词语之间深层关系的挖掘,但是由于匹配规则的制定存在过多的人工参与,且对微博的覆盖率较低,Popescu 和姚的方法并不能有效识别出微博中隐性评价对象的评价搭配关系.

## 6　结论

本文创新点在于,提出一种多特征融合的中文微博观点信息联合抽取方法,并进一步利用微博转发关系特性对抽取结果进行了优化,在同类算法对比实验中取得了较好的效果. 将面向非受限领域微博内容的观点信息联合抽取看作是序列标注任务,这样做有以下三个益处:第一,相较于隐马尔科夫模型来说,CRF 模型不需要严格的独立性假设条件,可以容纳任意的上下文信息,特征设计比较灵活. 由于识别用户生成内容中的评价对象和评价词语是一项比较复杂的任务,与其关联的因素也有很多,例如词汇原型、词性、所扮演的语义角色以及词的情感先验等等. 标记序列的分布条件属性,可以让 CRF 模型很好的拟和现实数据,在这些数据中,标记序列的条件概率信赖于观察序列中非独立的、相互作用的特征. 第二,可以充分利用评价对象和评价词语之间的句法关系、语义关系和位置关系. 在序列标注模型中评价对象和评价词语的抽取不是独立的,因此可以有效利用它们之间的关系来提升系统的性能. 第三,序列标注模型能有效利用句子的语言学结构,CRF 模型是概率图模型的一种,它可以利用概率图的结构有效表达句子的语言学结构,充分利用句子的语言学结构提升系统的性能.

## 参考文献

[1] Hu M, Liu B. Mining opinion features in customer reviews [A]. Proceedings of the Nineteenth National Conference on Artificial Intelligence [C]. AAAI, 2004. 4 (4): 755 – 760.

[2] Popescu A M, Nguyen B, Etzioni O. OPINE: Extracting product features and opinions from reviews [A]. Proceedings of HLT/EMNLP on Interactive Demonstrations [C]. ACL, 2005. 32 – 33.

[3] Zhang L, Liu B, Lim S H, et al. Extracting and ranking product features in opinion documents [A]. Proceedings of the 23rd International Conference on Computational Linguistics: Posters [C]. ACL, 2010. 1462 – 1470.

[4] Liu K, Xu L, Zhao J. Opinion target extraction using word-based translation model [A]. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning [C]. ACL, 2012. 1346 – 1356.

[5] Xu L, Liu K, Lai S, et al. Mining opinion words and opinion targets in a two-stage framework [A]. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics [C]. ACL, 2013. 1764 – 1773.

[6] Xu L, Liu K, Lai S, et al. Walk and learn: a two-stage approach for opinion words and opinion targets co-extraction [A]. Proceedings of the 22nd International Conference on World Wide Web Companion [C]. International World Wide Web Conferences Steering Committee, 2013. 95 – 96.

[7] Liu K, Xu L, Zhao J. Extracting opinion targets and opinion words from online reviews with graph co-ranking [A]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics [C]. ACL, 2014. 314 – 324.

[8] Liu K, Xu L, Zhao J. Co-extracting opinion targets and opinion words from online reviews based on the word alignment model[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(3): 636 – 650.

[9] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis[A]. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing[C]. Vancouver, British Columbia, Canada: ACL, 2005. 347 – 354.

[10] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis[J]. Computational Linguistics, 2009. 35(3): 399 – 433.

[11] Jin W, Ho H H, Srihari R K. Opinion miner: a novel machine learning system for web opinion mining and extraction[A]. Proceedings of KDD' 2009[C]. KDD, 2009. 1195 – 1204.

[12] Zhao Y, Qin B, Liu T. Collocation polarity disambiguation using web-based pseudo contexts[A]. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning[C]. ACL, 2012. 160 – 170.

[13] 吕品,钟珞,唐琨皓. 在线产品评论用户满意度综合评价研究[J]. 电子学报,2014,42(4):740 – 746.
Lv P, Zhong L, Tang K H. Customer satisfaction degree evaluation of online product review[J]. Acta Electronica Sinica, 2014, 42(4): 740 – 746. (in Chinese)

[14] 庄丽. 评论性信息挖掘研究[D]. 北京:清华大学,2007.
Zhuang L. Studies on review mining[D]. Beijing: Tsinghua University, 2007.

[15] 施寒潇. 细粒度情感分析研究[D]. 苏州:苏州大学,2013.
Shi H X. Research on fine-graind sentiment analyis[D]. Suzhou: Soochow University, 2013.

[16] Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: probabilistic models for segmenting and labeling sequence data[A]. Proceedings of ICML' 2001[C]. ICML, 2001. 282 – 289.

[17] 黄昌宁,赵海. 中文分词十年回顾[J]. 中文信息学报,2007,21(3):8 – 19.

[18] 刘挺,车万翔,李正华. 语言技术平台[J]. 中文信息学报,2012,25(6):53 – 62.

[19] Yi J, Nasukawa T, Bunescu R, Niblack W. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques[A]. Third IEEE International Conference on Data Mining (ICDM 2003)[C]. IEEE, 2003. 427 – 434.

[20] Hu M, Liu B. Mining and summarizing customer reviews[A]. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. ACM, 2004. 168 – 177.

[21] Jakob N, Gurevych I. Extracting opinion targets in a single-and cross-domain setting with conditional random fields[A]. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics[C]. ACL, 2010. 1035 – 1045.

[22] 刘全超,黄河燕,冯冲. 基于多特征微博话题情感倾向性判定算法研究[J]. 中文信息学报,2014,28(4):123 – 131.

[23] Popescu A M, Etzioni O. Extracting product features and opinions from reviews[A]. Natural Language Processing and Text Mining[M]. London: Springer, 2007. 9 – 28.

[24] 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[A]. 中文信息处理前沿进展-中国中文信息学会二十五周年学术会议论文集[C]. 北京:清华大学出版社,2006. 260 – 281.

## 作者简介

**刘全超** 男,1982 年生于河北,博士研究生,主要研究方向为情感计算、信息抽取.

**黄河燕** 女,1963 年生于湖南,教授,博士生导师,主要研究领域为自然语言处理、机器翻译.