ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 2,Mar. 2016,pp. 296－306
DOI:10.16337/j.1004-9037.2016.02.009

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

最长评价短语及其情感评价搭配抽取方法

刘全超　黄河燕　王亚珅　冯　冲

(北京理工大学计算机学院,北京,100081)

摘　要:提出一种统计和规则相结合的最长评价短语自动识别算法。将评价短语的识别问题转化为序列标注问题,结合条件随机场模型进行简单结构的评价短语识别,在此基础上进一步建立和应用规则库,自动识别结构复杂的最长评价短语,其测试的 F 值达到 72.38%。在最长评价短语自动识别的基础上,构建用于评价对象抽取和情感评价单元抽取的规则库,提出基于规则的评价搭配自动抽取算法,实现评价对象和最长评价短语搭配的自动抽取,在网易汽车门户网站进行了系统测试,得到了较高的准确率。

关键词:情感分析;观点挖掘;评价短语;条件随机场

中图分类号:TP391.1　　**文献标志码**:A

Method of Extracting Maximal-length Evaluation Phrase and Appraisal Expression

Liu Quanchao, Huang Heyan, Wang Yashen, Feng Chong

(Department of Computer Science and Technology, Beijing Institute of Technology, Beijing, 100081, China)

Abstract: An algorithm based on statistics and rules is proposed to automatically identify maximal-length evaluation phrase. The identification of evaluation phrase is taken as sequence tagging problem. Then conditional random field model is used to recognize evaluation phrase with simple structure. Therefore, rule database is established and maximal-length evaluation phrase with complex structure is identified automatically. F-measure value reaches 72.38%. Based on the above work, rule base is constructed for extracting opinon target and appraisal expression. Rule-based extracting appraisal expression is proposed to automatically extract opinion target and maximal-length evaluation phrase. Experiments were conducted at netease car portal and got a higher precision.

Key words: sentiment analysis; opinion mining; evaluation phrase; conditional random fields

引　　言

当今互联网每天都在产生大量数据和信息,这些海量信息表达了人们的各种情感色彩和情感倾向性,如喜、怒、哀、乐和批评、赞扬等。如何从海量数据和信息中进行情感分析[1]是目前所面临的难题。最长评价短语(Maximal-length evaluation phrase, MEP)是一定范围内所有连续的、具有评价意义的词语的组合,表达一种观点、情感、估计、立场和猜测等。简单结构最长评价短语(Simple structure maxi-

基金项目:国家重点基础研究发展计划("九七三"计划)(2013CB329605)资助项目。
收稿日期:2014-09-12;**修订日期**:2014-11-17

mal-length evaluation phrase, sMEP)是指程度副词及其修饰的评价词语的搭配,主要出现在定语、状语和补语位置上,结构比较固定且简单,可以单独作为最长评价短语,也可以经过扩展、被包含在其他最长评价短语内部。

最长评价短语能够全面反映出观点持有者对某一评价对象的情感信息,准确、全面地识别最长评价短语对于情感分析等自然语言处理任务有着积极的促进作用。另外,从繁琐的文本中,抽取出基于最长评价短语的评价搭配,并把这种评价搭配简洁地呈现给用户,可以使用户更直观和便捷地了解相关评论信息和掌握情感倾向。

在情感分析的基础任务研究中,目前大部分的研究工作集中在文档级别,然而在类似于问答系统、自动文摘、舆情导向信息抽取及产品反馈信息挖掘等任务中需要句子级别甚至短语级别的情感分析。一些研究者致力于评价短语的识别和分类。评价短语具体指连续出现的一组评价词语[2],如"非常平稳"。文献[3]提出在评价短语的基础上进行文本观点分类,认为情感分析的原子单元应该是评价短语而不是独立的词语;文献[4]提出一个短语层次的观点分析方法,该方法首先要求判断一个表达是否带有情感色彩,然后再对带有情感色彩的表达进行褒贬分类。

评价对象和评价词之间的搭配信息对于情感分析起着至关重要的作用。文献[5,6]提出情感评价单元<评价对象,评价词语>二元搭配,早期的部分研究者将这项任务分为两步:首先获取情感句中的评价对象,然后选择距离评价对象窗口为 k 的评价词语。例如文献[7]选取距离评价对象最近的形容词作为其对应的评价词语;文献[8]在距离评价对象窗口为 k 的范围内选择评价词语。但是这些方法经验性过强,导致系统性能有限,因为很多时候评价词语和评价对象距离较远。随后,部分研究者将对评价对象和评价词语的识别合并为一个独立的任务,提出基于规则[2]或基于模板[9,10]的方法来识别情感评价单元。文献[10]基于依存句法分析的结果,人工制定了 31 条句法规则来描述评价词语和评价对象之间的关系,通过这些规则对测试集进行验证,抽取情感评价单元;文献[11]利用 MINIPAR Parser 手工构建了 10 条依存句法抽取模板来获取情感评价单元。文献[9]同样基于依存句法,用上行序列和下行序列来表示句法路径,以建立句法路径规则库来完成情感评价单元的抽取。在前人工作基础之上,本文设计了一种基于统计和规则相结合的最长评价短语及其情感评价搭配的识别方法。

1 统计和规则相结合的最长评价短语自动识别算法

最长评价短语是一个完整的语义单元,其识别任务是统计和规则相结合,分为两个递进的子任务。首先是基于统计的 sMEP 识别任务;其次是在 sMEP 识别任务基础上,基于规则的 MEP 识别任务。前一阶段中的 sMEP 有可能独立作为最终结果的 MEP,也有可能被包含在其他结构相对复杂的 MEP 中。为了保证最终结果的 MEP 都是具有一定评价意义的短语,只有包含情感词和评价词词典中词语的 MEP 才会被保留,作为最终的 MEP。

1.1 基于统计的 sMEP 自动识别

将 sMEP 识别问题转换成为序列标注问题,采用条件随机场(Conditional random fields, CRF)模型,根据模型标注结果识别出 sMEP。令输入的句子的形式为 $S=W_1/P_1,W_2/P_2,\cdots,W_n/P_n$。其中 W_i 为词语,P_i 为 W_i 的词性标注,$i=1,2,\cdots,n$。令 M_i 为 CRF 模型中词语 W_i 所对应的序列标注,$i=1,2,\cdots,n$,$M_i\in\{B,I,O\}$。则标注序列可以表示为 $Q=M_1,M_2,\cdots,M_n$,满足 $Q^*=\arg\max(Q|S)$。本文采用 BIO 标注法,其中标注为 B 的词语表示以 sMEP 开始的词语,标注为 I 的词语表示 sMEP 中间位置的词语,标注为 O 的词语表示该词不属于 sMEP。

在基于 CRF 的自动识别器中,特征模板的选择对于识别结果起着至关重要的作用。sMEP 的边界

相对比较明确，而且有一定规律可循，这正是将 sMEP 识别问题转换成为序列标注问题的原因之一。边界分布信息和内部结构组合只是为 sMEP 的正确识别提供强有力的支持，CRF 特征模版如表 1 所示。

表 1 CRF 特征模板

Tab. 1 CRF feature templates

模板名称	模板描述	符号表示
模板 1	当前词以及前后各两个词的词和词性。	$W_i, P_i, i=-2,-1,0,1,2$
模板 2	当前词以及前后各两个词的词和词性，及二元复合特征信息。	$W_i, P_i, i=-2,-1,0,1,2$ $W_{i-1}W_i, P_{i-1}P_i, i=-1,0,1,2$
模板 3	当前词以及前后各两个词的词和词性，以及二元复合特征信息，三元复合特征信息。	$W_i, P_i, i=-2,-1,0,1,2$ $W_{i-1}W_i, P_{i-1}P_i, i=-1,0,1,2$ $W_{i-1}W_iW_{i+1}, P_{i-1}P_iP_{i+1}, i=-1,0,1$

本文从某门户网站选取 15 326 个句子进行手工标注作为训练语料，选取其他 30 篇完整文章作为测试语料进行开放测试，其中包含 2 815 个 sMEP。表 2 为不同特征模板下对 sMEP 的自动识别结果，其中 N_1 表示语料中实际存在的 sMEP 数量，N_2 表示系统自动识别出的 sMEP 数量，N_3 表示系统正确识别出的 sMEP 数量，P 表示正确率，R 表示召回率，F 表示 F-measure 值。

表 2 特征模板对比试验及结果

Tab. 2 Contrast experiment and results of feature templates

模板名称	N_1	N_2	N_3	P/%	R/%	F/%
模板 1	2 815	2 918	2 489	85.30	88.42	86.83
模板 2	2 815	3 043	2 584	84.92	91.79	88.22
模板 3	2 815	3 009	2 694	89.53	95.70	92.52

从表 2 可以发现，在目前所进行的实验中，模板 3 所达到的效果最为理想，所以本文后续试验均采用模板 3 进行实验。

1.2 统计和规则相结合的 MEP 自动识别

图 1 是统计和规则相结合的 MEP 自动识别体系结构，分为两部分：基于统计的 sMEP 自动识别模块和基于规则的后处理模块，即在得到 sMEP 的基础上，利用规则识别结构相对复杂的 MEP。本文借鉴文献[12]的组块级别理论和文献[13]对组块的定义和分类研究，在基于规则的最长评价短语自动识别过程中，通过构建相关规则识别表达情感倾向的组块作为最长评价短语。

1.2.1 MEP 自动识别规则库

sMEP 能够表达情感色彩，句式比较简单。然而在很多情况下，一些复杂的句式也能表现出情感色彩，表达出某种观念或者某种情感倾向，对于后续情感分析有很好的帮助作用。本文总结了 3 种复杂句式规则用于 MEP 识别：括号规则、介词组块规则和副词组块规则。

(1)括号规则。括号是一种标号，包括对前文的注释、列举、说明和评价等。有的是针对前文中的一个词，有的是针对前文的一个句子。括号在应用中有内用和外用的区别：内用

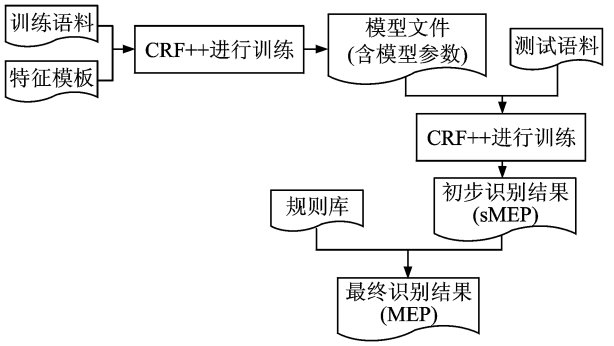


图 1 MEP 识别框架结构图

Fig. 1 Architecture of MEP identification

是用在句子内部,外用是用在句子外部。其中,内用比较常见,本文中括号模板是针对括号内用,所以对于注释、说明和评价作用的括号内容可以被抽取出作为一个评价短语,而且该评价短语负责修饰其前面(即括号左侧)的评价对象,可以认为是一定范围内对评价对象起修饰作用的最长评价短语。

(2)介词组块规则。介词组块是一种重要的组块类型,在汉语句子中占有很大比重。有研究者从2000年某日报语料中随机抽取26 551个句子进行统计,发现句子中平均每100个词有15.1个出现在介词组块中[14]。介词组块的一般规律是介词位于组块左边界;介词组块出现的位置也有一定特点,如经常出现在动词前作为状语,出现在动词后面做作为补语,出现在"的"字前面作为定语。本文利用这些特征通过匹配规则的方法来识别介词组块。步骤描述如下:

(a)对每个分句中的介词组块识别均是从分句的右端开始的。(b)从分句右端往左端搜索,如果发现词性标注为介词词性的词语,则对该词语及其右边内容进行分析:匹配规则,若这一系列词语符合规则,则识别,并将这些词语合并以及进行重新词性标注(标注为MEP);若不符合规则,则不识别。(c)继续向左搜素,重复上述过程,直到搜索到分句的最左端。

从分句右端往左分析的目的在于,由于存在介词组块嵌套或者分句中存在连续介词组块(例如句子"在/p 我/rr 心中/s 国家/n 比/p 个人/n 重要/a"),应该从小粒度的介词组块开始识别。介词组块模式规则库中共有12条基本规则,其中"p"表示介词,"rr"表示人称代词,"s"表示处所词,"n"表示名词,"a"表示形容词,"M"表示最长评价短语,"d"表示"副词","v"表示动词,"f"表示方位词。部分规则描述为:

规则1 p+n+a 。对于介词右侧同一分句的每个词语,若顺序出现(不一定相邻)名词性词语(或词语组合)和形容词词性词语(或词语组合),则抽取并合并介词及其右侧同一分句的所有词语,重新标记词性为MEP(如:"动力/n 上/f 将/d 会/v 与/p 传祺/nz 轿车/n 相似/a 。/wj")。

规则2 p+n+M 。对于介词右侧同一分句的每个词语,若顺序出现(不一定相邻)名词性词语(或词语组合)和词性标记为MEP的词语(或词语组合),则抽取并合并介词及其右侧同一分句的所有词语,重新标记词性为MEP(如:"新款/nz 设计/vn 比/p 老款/nz 更加简洁/MEP 。/wj")。

关于规则库中的12条基本规则有以下几点说明:

(a)规则中的名词性词语可以替换成为代词性词语,规则依然成立。(b)可以通过如下途径扩充相关规则,得到介词组块短语模式规则库的扩展规则:(I)名词性词语前面添加词性标记为MEP的词语或者形容词词性词语;(II)动词性词语前面添加副词性词语。(c)规则中的名词性词语包括以顿号(词性标记为wn)和并列连词(词性标记为cc)相连的并列名词组合的情况。(d)规则中的介词(词性标记为p)位置可换成"相比""对比"等表示比较意义的词语,规则依然成立。

(3)副词组块规则。在实际应用当中,副词是在时间、方式、程度和语气等诸多方面对动词、形容词或整个句子进行修饰和限定的一类词,对于表达一定的情感色彩及观点倾向有着重要作用。

本文主要研究以副词开头的副词动词组块和副词名词组块在句子中做谓语和补语的情况,副词组块短语模式规则库中共有11条基本规则,部分规则描述如下所示:

规则1 d+n 。如果副词右侧紧邻名词性词语,则抽取并合并副词及其右侧的名词性词语,重新标记词性为MEP(如:"新款/b 车/n 的/ude1 外观/n 设计/v 得/ude3 很/d 中国/ns 。/wj")。

规则2 d+v 。对于副词右侧同一分句的每个词语,若顺序出现(不一定相邻)动词性词语,则抽取并合并副词及其右侧同一分句的所有词语,重新标记词性为MEP(如:"在/p 做工/n 方面/n 还/d 有待/v 提高/v 。/wj")。

使用介词组块规则库中的扩充规则方法实现副词组块规则的扩展。

1.2.2 基于规则的后处理算法

在基于统计的sMEP识别结果基础上,对已经经过标注的文本进行规则匹配处理,将符合规则的一系列连续的词语识别成为MEP并重新进行标注。基于规则的MEP识别算法流程如图2所示。

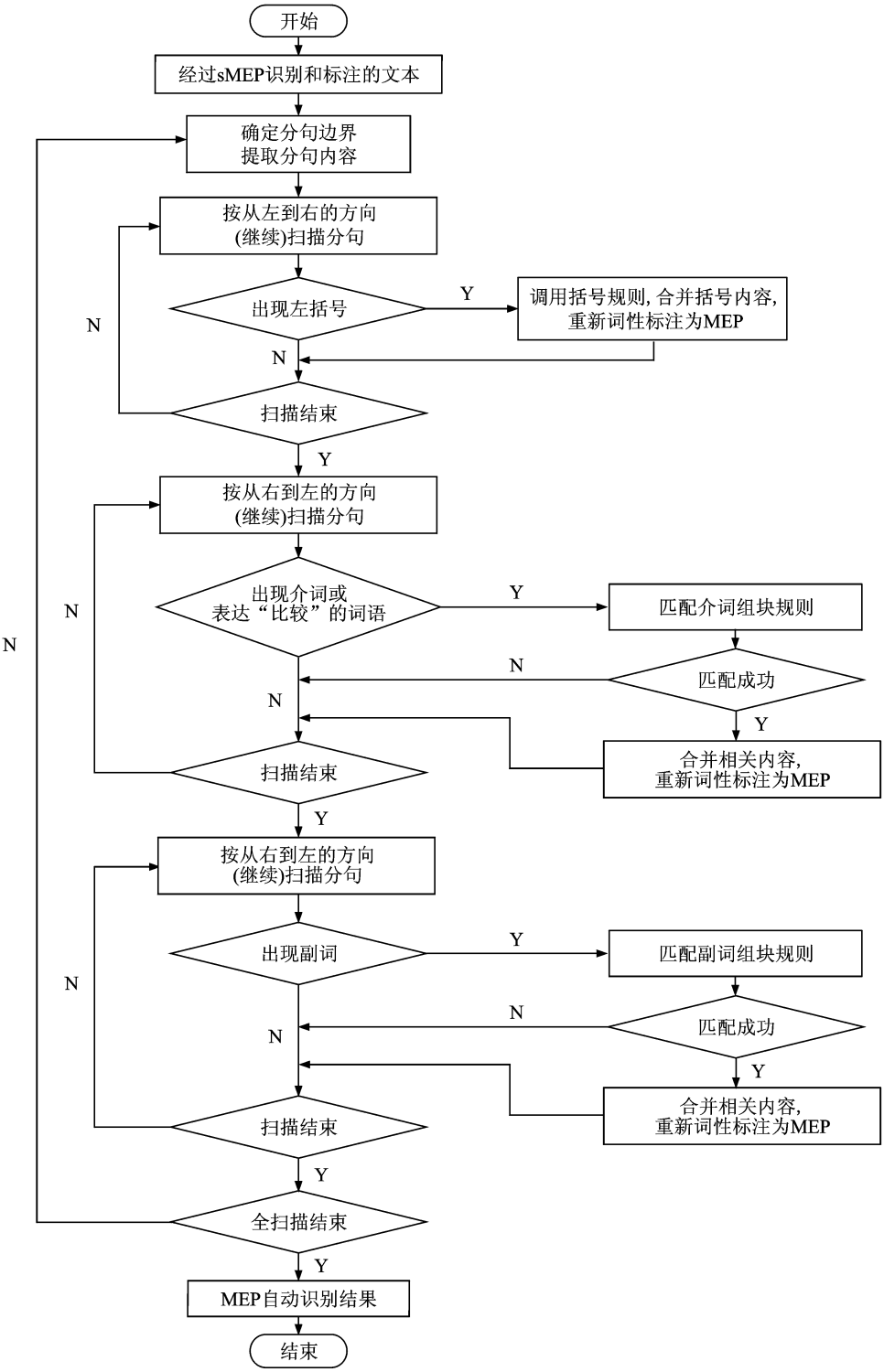


图 2 MEP 识别流程图

Fig. 2 Flowchart of MEP identification

2 规则评价搭配自动抽取算法

以句号、问号、叹号、逗号、分号和冒号等符号分割的句子称为分句。本文基于规则的情感评价单元抽取在分句层次上进行,即每次研究和处理一个分句内部的词语信息,若该分句中存在 MEP,则依次研究这些 MEP,对每个 MEP 应用相关规则定位并形成其对应的评价搭配。

2.1 规则库

规则库构建主要分为用于构建评价对象的规则库和用于抽取情感评价单元的规则库。因为评价词使用的是基于条件随机场和规则识别出的 MEP,所以无需建立构建评价词的规则库。

2.1.1 评价对象抽取规则库

使用规则库进行情感评价单元抽取的前提是该区段中含有被标记为 MEP 的词语组合。把 MEP 作为评价词,以此 MEP 为中心,构建其潜在的评价对象。

经过研究大量语料发现,评价词所修饰的评价对象,大多数情况下是其左侧或者右侧距离其最近的名词性词语。所以在 MEP 所在区段内,分向左和向右两个方向搜索距离其最近的连续名词性词语组合。其中,连续名词性名词组合指词性标记为名词(n)、代词(r)、字符串(x)、前缀(h)和后缀(x)等连续的词语组合。包含名词并列的情况,并列的名词由词性标记为顿号(wn)的符号和词性标记为并列连词(cc)的词语相连接,在构建评价对象时,若遇到上述两种情况,需要进一步判断相应方向上下一个词是否依然属于名词性词语,若属于,则继续构建下去,若不属于,则停止构建。

2.1.2 情感评价单元抽取规则库

MEP 是一定范围内能够表达情感色彩的最长评价短语,本文将 MEP 选为情感评价单元的评价词。使用评价词构建规则库,该区段中存在潜在的评价对象。

(1)MEP 右面紧邻评价对象。若该 MEP 是在识别 MEP 时应用了括号模板,即该 MEP 是由括号括起来的内容,则不抽取该搭配;否则,抽取该 MEP 和该评价对象的搭配作为情感评价单元。

(2)MEP 左面紧邻评价对象。若评价对象左面没有 MEP,则直接抽取该 MEP 和该评价对象的搭配作为情感评价单元。若该评价对象左面还有 MEP,则合并评价对象和其左面的 MEP 作为最终评价对象,和该 MEP 一起抽取出来作为情感评价单元。若这种含有 MEP 的评价对象的左侧存在并列连词及其所连接的名词性词语(包含紧邻的 MEP),则继续向左抽取这些词语作为最终的评价对象,和该 MEP 一起抽取出来作为情感评价单元。规则描述如下:

规则 1 Noun+MEP

规则 2 MEP+Noun+MEP

规则 3 MEP+Noun+cc+MEP+Noun+MEP

(3)MEP 左面存在评价对象,但是不相邻。评价对象和评价词语之间,按照如下所述规则进行匹配和抽取,形成情感评价单元。部分规则描述如下:

规则 1 Noun+v+MEP

规则 2 MEP+Noun+MEP

规则 3 MEP+Noun+cc+MEP+Noun+MEP

(4)使用评价词构建规则库,该分句中不存在潜在的评价对象时,进行隐式搭配的抽取。定义最近评价对象为表示语料中最近评论被评论过的对象的名词性词语的组合,需要不断被更新。在句号、问号和叹号等符号分割的句子中,评价对象很有可能出现变化。而出现在所研究分句的第 1 个名词(或名词

组合)极有可能是该分句所评价的对象。在所处理的语料中出现的隐式搭配中,95%是以距离其最近的分句中第1个名词(或名词组合)作为整个分句所描述的评价对象。所以生成最近评价对象的规则如下:

(a)当MEP左侧没有任何词(即MEP位于分句的最左端)的时候,直接使用最近评价对象与该MEP进行搭配,构成情感评价单元。

(b)当MEP左侧有且只有动词性词语的组合(包括并列连词连接多个动词性词语的情况),MEP很有可能对这些动词性词语组合起到修饰作用,在处理的语料中,这个比例达到了94%,所以,抽取这些动词性词语的组合与该MEP进行搭配,构成情感评价单元。

(c)其他情况下,直接使用最近评价对象与该MEP进行搭配,构成情感评价单元。

2.2 基于规则的评价搭配自动抽取

在基于统计和规则的MEP识别结果基础上,对已经经过标注(以MEP标注)的文本,以标记为MEP的词语为中心,如果其含有情感词和评价词词典中的词语,则对其进行匹配规则处理,定位其对应的评价对象,最终形成评价对象和最长评价短语的搭配组合。情感评价组合识别流程如图3所示。

3 实验设置

实验采用某汽车门户网站的用户评论信息作为实验数据集,以网络爬虫爬取的特定主题(例如特定汽车品牌和型号)的相关文本为处理对象,对该网络文本进行最长评价短语自动识别和情感评价单元抽取处理。如图4所示,系统分为3个功能模块:预处理模块、统计和规则相结合的MEP自动识别模块和基于规则的情感评价单元抽取模块。

中文词法分析是中文信息处理的基础与关键,当今中文分词的开源工具已经有很多,而且大多数工具的效率和准确率都很高,其中中国科学院的ICTCLAS(http://ictclas.nlpir.org/)是其中比较优秀的中文分词系统。本文的分词和词性标注等任务均由ICTCLAS支持完成,并且手工编制汽车领域专业词典,该词典包含2 327个词条,主要应用于对ICTCLAS分词的支持,本文实现了情感评价组合识别系统。

本系统的核心思想是向用户提供带有情感色彩的信息和为后续的情感倾向分析任务奠定基础,所以本系统的最终结果中的最长评价短语需要包含一定的情感词语或者评价词语。为此,本系统使用了HOWNET(http://www.keenage.com/html/c_index.html)所提供的情感词和评价词词典作为判别依据。对sMEP,MEP以及情感评价单元抽取的评价指标采用正确率(Precision)、召回率(Recall)和F值(F-measure),进行性能评价,如下所示

$$\text{Precision} = \frac{\#\,\text{system\_correct}}{\#\,\text{system\_proposed}}$$

$$\text{Recall} = \frac{\#\,\text{system\_correct}}{\#\,\text{person\_correct}}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

式中:# system_correct为系统正确识别出的数目;# system_proposed为系统自动识别出的数量,包含了错误的识别结果;# person_correct为语料中人工标注的实际数目。

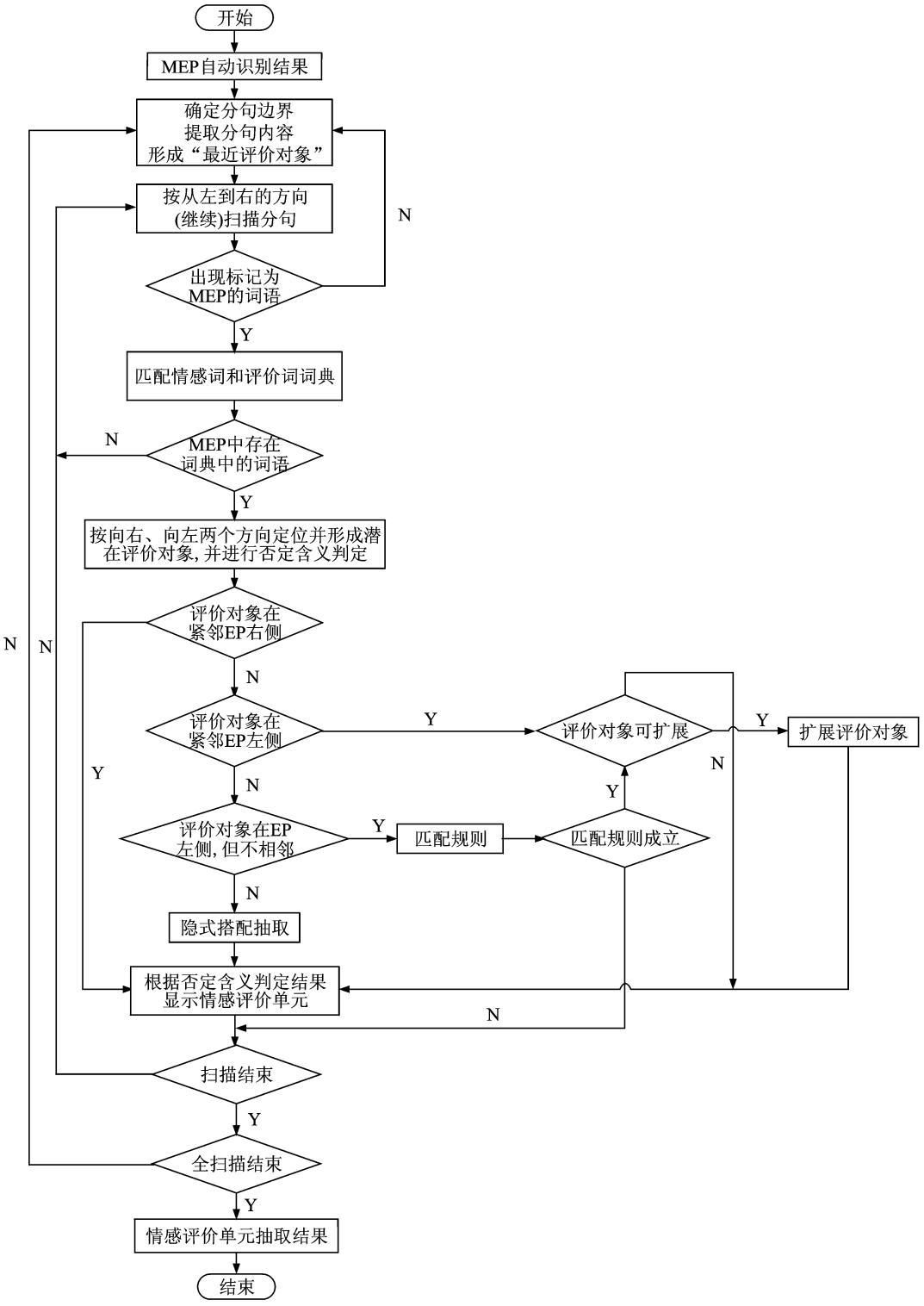


图 3 情感评价组合单元识别流程图

Fig. 3 Flowchart of extracting appraisal expression

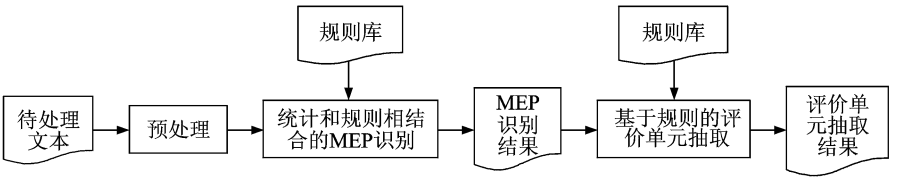


图 4 系统框架图

Fig. 4 Architecture of the system

4 实验结果与分析

sMEP, MEP 及情感评价单元的实验结果如表 3 所示，且将情感评价单元抽取与文献[9]的汽车评论意见挖掘系统的性能指标进行了对比，其结果如图 5 所示。

文献[9]的汽车评论意见挖掘系统需要汽车本体、极性词词典和匹配规则等基础资源，而这些基础资源的构建因人而异且费时费力，将本文的情感词词典和规则库应用到文献[9]的算法中而忽略汽车本体资源，其算法性能如表 3 所示。通过对比可以发现，在没有汽车本体资源且相同语料情况下，本文算法的 sMEP 识别准确率和召回率均比较高，这是由于 sMEP 构成比较有规律、语法成分相对固定，而 MEP 和情感评价单元抽取的性能指标同样达到了良好的效果；在有汽车本体资源的作用下，本文算法的性能同样接近于上海交大汉语汽车评论意见挖掘系统公布的平均召回率 80% 和平均准确率 60% 的指标，如图 5 所示，因此本文的算法合理、有效。同时从表 3 中可看出，汽车评论意见挖掘系统比较依赖于汽车本体资源。

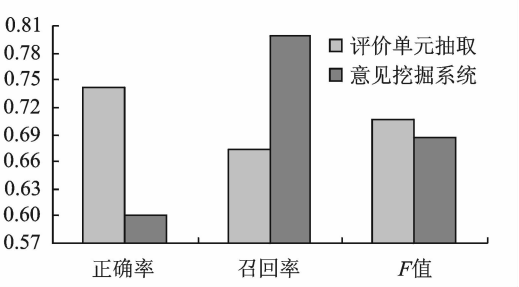


图 5 评价单元抽取与意见挖掘系统性能对比

Fig. 5 Comparison between appraisal expression extracting and opinion mining system

表 3 实验结果比较

Tab. 3 Comparison of experiment results %

识别种类	准确率	召回率	*F* 值
sMEP	89.53	95.73	92.52
MEP	76.01	69.09	72.38
情感评价单元	74.26	67.27	70.59
汽车评论意见挖掘系统[9]-汽车本体	56.08	74.07	63.83
汽车评论意见挖掘系统[9]	60.00	80.00	68.57

4.1 sMEP 自动识别过程中出现的错误

实验过程中，本文对 sMEP 的识别结果进行了进一步分析，发现 sMEP 的识别错误主要因为：

(1)部分 sMEP 以动词开头，特别是 sMEP 之前有动词，造成动词相连的情况，在基于统计标注时可能出现标注错误。(2)部分 sMEP 是名词性词语(或者词语组合)，当这种 sMEP 作为补语出现在名词性词语(或词语组合)的后面时，造成名词相连的情况，在基于统计标注时可能出现标注错误。(3)文本在语法中存在常识性错误，例如遗漏或错误使用标点符号、“的”及“地”错乱使用等情况，容易造成标注错误。(4)sMEP 最常见的形式是以结构助词“的”结尾的作为定语的短语，但是由于汉语语法的灵活性特

点,导致非定语的以“的”结尾的短语被错误地识别成为 sMEP。

4.2 MEP 自动识别过程中出现的错误

使用模板和规则进行结构相对复杂 MEP 的识别过程是基于正确的语法规则,所以如果文本中出现标点符号使用不当及词语省略等不符合正规语法现象时,很容易造成识别结果出现错误。例如:

“不仅仅/d 更/d 为/v 立体/b 圆润/a 最/d 关键/a 的/udel 是 /vshi 它/rr 的/udel 行车/nz 电脑显示屏/n 所/usuo 提供/v 的/udel 信息/n 非常/d 多样化/vi 并且/c 很/d 实用/a 。/wj”。

理论上“圆润”和“最”之间应该用分隔符号(如逗号)分隔,但是实际中并没有分隔成两个分句,导致这两个分句被视为一个分句进行处理,可能被匹配副词组块模板被识别成一个 MEP,造成 MEP 识别错误。

4.3 情感评价单元自动识别过程中出现的错误

通过对情感评价单元识别结果的统计分析,本文发现情感评价单元识别的原因主要有:(1)由于存在并列关系的两个评价对象分别位于两个分句当中,导致在定位评价对象的时候,会遗漏和 MEP 处在不同分句的评价对象,导致最终结果不够全面。(2)由于词性标注的错误,导致在使用规则和模板识别结构相对复杂的最长评价短语时出现错误。其中,“与”及“和”等兼有并列连词词性和介词词性的词语的词性标注错误占了很大比例。

5 结束语

基于规则的现有信息抽取算法中,规则构建的主观性较强且规则不宜扩展,过多的规则容易产生冲突。而本文提出一种统计和规则相结合的最长评价短语及其评价搭配抽取算法,避免了人工观测大量语料构建规则,简化了规则的构建过程,同时避免了由于过多规则而导致的冲突。该算法首先将结构简单的最长评价短语的抽取任务转换成为序列标注任务,利用统计的方法进行识别;然后创建并使用规则和模板,抽取结构相对复杂的最长评价短语。在完成最长评价短语抽取任务的基础上,提出一种基于规则的情感评价单元的抽取方法。本文将来的工作重点会放在以下 3 个方面:(1)最长评价短语自动识别规则库和评价搭配自动识别规则库的扩展算法;(2)实现数据集的自动标注算法,提高系统的智能化水平及其效率;(3)对 sMEP,MEP 以及情感评价单元识别过程中容易出现的错误深入分析,进一步完善本文的算法。

参考文献:

[1] Riloff E, Wiebe J. Learning extraction patterns for subjective expressions[C]//Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Sapporo, Japan: ACL, 2003:105-112.

[2] 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.
Zhao Yanyan, Qin Bing, Che Wanxiang, et al. Appraisal expression recognition based on syntactic path[J]. Journal of Software, 2011,22(5):887-898.

[3] Whitelaw C, Garg N, Argamon S. Using appraisal groups for sentiment analysis[C]//Proceedings of the 14th ACM International Conference on Information and Knowledge Management. Bremen, Germany: ACM, 2005:625-631.

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

[5] Bloom K, Garg N, Argamon S. Extracting appraisal expressions[C]//HLT-NAACL 2007. Rochester, NY, USA: ACL, 2007:308-315.

[6] Wilson T, Wiebe J. Annotating attributions and private states[C]//Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky. Ann Arbor, USA: ACL, 2005:53-60.

[7] Hu M, Liu B. Mining and summarizing customer reviews[C]// Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA, USA: ACM, 2004:168-177.

[8] Kim S M, Hovy E. Determining the sentiment of opinions[C]//Proceedings of the 20th International Conference on Computational Linguistics. Stroudsburg, PA, USA: ACL, 2004:1367-1373.

[9] 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[C]// 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. 北京：清华大学出版社，2006:260-281.
Yao Tianfang, Nie Qingyang, Li Jianchao, et al. An opinion mining system for Chinese automobile reviews[C]// The Advanced Progress in Chinese Information Processing—Proceedings of the Twenty-five Annual Conference of Chinese Information Processing Society. Beijing: Tsinghua University Press, 2006:260-281.

[10] Bloom K, Argamon S. Automated learning of appraisal extraction patterns[J]. Language and Computers, 2009,71(1):249-260.

[11] Popescu A M, Etzioni O. Natural language processing and text mining[M]. London: Springer, 2007.

[12] Abney S P. Parsing by chunks[M]. Netherlands: Springer, 1992:257-278.

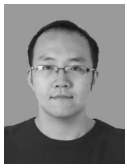
[13] 李素建，刘群. 汉语组块的定义和获取[C]// 语言计算与基于内容的文本处理——全国计算语言学联合学术会议论文集. 北京：清华大学出版社. 2003:110-115.
Li Sujian, Liu Qun. Research on definition and acquisition of Chunk[C]// Language Computing and Content-based Text Processing—Proceedings of the National Conference on Computational Linguistics. Beijing: Tsinghua University Press, 2003: 110-115.

[14] 干俊伟，黄德根. 汉语介词短语的自动识别[J]. 中文信息学报，2005,19(4):17-23.
Gan Junwei, Huang Degen. Automatic identification of Chinese prepositional phrase[J]. Journal of Chinese Information Processing, 2005,19(4):17-23.

作者简介：

刘全超(1982-)，男，博士研究生，研究方向：情感计算和机器学习，E-mail: liuquanchao@bit.edu.cn。

黄河燕(1963-)，女，教授，研究方向：自然语言处理和机器翻译。



王亚珅(1989-)，男，博士研究生，研究方向：机器学习。

冯冲(1977-)，男，副研究员，研究方向：自然语言处理和机器翻译。