基于模板的中文人物评论意见挖掘 *

李 娟[1,2], 张 全[2], 贾 宁[1,2], 臧翰芬[1,2]
(1. 中国科学院 研究生院, 北京 100039; 2. 中国科学院 声学研究所, 北京 100190)

摘 要: 使用基于模板的方法对中文人物评论语句进行意见元素挖掘,提取出句中所含的评价对象、评价词语,并分析出意见的倾向性。进行了中文人物评论语句的自动意见挖掘实验,实验中首先建立了一定数量的熟语料库,然后从语料库中生成意见模板,最后用生成的模板来提取语句的意见元素。实验获得了 72.55% 的 *F*-score,表明该算法是有效的。

关键词: 意见挖掘; 观点抽取; 基于模板
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2010)03-0833-04
doi:10.3969/j.issn.1001-3695.2010.03.007

Pattern-based opinion mining on figures comments in Chinese

LI Juan[1,2], ZHANG Quan[2], JIA Ning[1,2], ZANG Han-fen[1,2]
(1. *Graduate School, Chinese Academy of Sciences, Beijing* 100039, *China*; 2. *Institute of Acoustics, Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract: Using pattern-based method to mine opinions from Chinese figures comments, extract the object, opinion terms, and determine the orientation of the opinion. This paper carried out the experiment to automatically mine the opinions from Chinese figures comments. First established the training corpus, then extracted patterns from the corpus, then used the patterns to extract the opinion elements. The experiment result of 72.55% *F*-score shows that the method is effective.
Key words: opinion mining; opinion extracting; pattern-based

0 引言

意见挖掘是近年来自然语言处理领域研究中发展起来的一个新方向,意见挖掘研究的不是文档所谈论的话题,而是它所表达的倾向性观点,即肯定/否定或者褒扬/贬损性意见。意见挖掘的应用范围很广泛,针对产品的意见挖掘可以帮助用户作出理智的购买决定,针对人物的意见挖掘可以监测舆论倾向或民众意向等。

意见挖掘可以在三个层面上进行,即词语、语句和篇章,如图 1 所示。词语级的研究[1,2]可以判断出词语的语义倾向性,然后在此基础上可以挖掘语句中的意见元素,也可以判断篇章的情感倾向。语句级的研究既可以独立应用,也可以作为篇章级的研究基础,语句级的研究目标是提取出语句中的意见元素[3],如评价对象、评价词语、意见倾向等,研究的结果可以为人们提供具体而详细的意见内容,也可以提供宏观的结论,具有重要的价值。本文的研究就是提取出语句中的意见元素。

目前语句级的意见挖掘研究成果较少,已有的语句意见挖掘方法主要是对句子进行句法分析,然后在此基础上进一步挖掘意见信息。比较有代表性的是上海交通大学的娄德成等人[4,5]提出的方法:首先采用依存语法分析器对句子进行句法分析,然后再对分析出的主谓结构、动宾结构等进行分析处理,最后提取出句中的意见信息。从研究涉及的领域来看,现有的意见挖掘都是针对产品,如数码相机、汽车、笔记本电脑等,研究成果对于其他的领域则不适用。此外还有基于统计分析的意见挖掘方法[6]、统计分析与语义分析相结合的方法[7]等。

考虑到依存语法的分析能力有限,以此为基础进行的意见挖掘也会受到句法分析的限制和影响,因此本文尝试了基于模板的方法;同时,面向人物的意见挖掘也具有重大的价值,研究的结果将对舆情监测等社会生活产生较大影响,因此本文选择人物评论语句为对象进行意见挖掘。本文首先建立了人物评论语料库,并对其中的意见元素进行了标注,构成了熟语料库,再从熟语料库中提取意见模板,最后用模板提取出未标注的语句中的意见元素。

1 算法设计

1.1 整体设计

本文的基于模板的意见挖掘系统分为两大模块,即模板库生成模块和意见元素挖掘模块。系统模块如图 2 所示。

模板库生成模块的目的是将已标注的熟语料库转换为结构化的模板库。模块首先从熟语料中提取出包含模板所需信息的工作子串,然后将工作子串转换为候选模板,并对候选模板进行统计和过滤,最后形成结构化的模板并存储为模板库。

收稿日期: 2009-06-17; **修回日期**: 2009-08-31 **基金项目**: 国家"973"计划资助项目(2004CB318104);中国科学院声学所知识创新工程资助项目(0654091431);中国科学院声学研究所"所长择优"基金资助项目(GS13SJJ04);中国科学院青年人才领域前沿资助项目(0754021432)

作者简介:李娟(1981-),女,硕士研究生,主要研究方向为自然语言处理;张全,研究员,博士,主要研究方向为自然语言处理、HNC 理论;贾宁,博士,主要研究方向为自然语言处理;臧翰芬,博士研究生,主要研究方向为自然语言处理(zanghf@163.com).

意见元素挖掘模块的目的是对未标注的测试语句进行分析,得到句子中的人物评价的各个意见元素(也简称为意见元素)。模块首先获取句子的工作子串,提取子串中的部分特定词语作为检索关键字,并用关键字在模板库中检索模板。如果检索到相应的模板,则尝试将模板匹配到句子中,匹配不成功则挖掘失败;若匹配成功,通过对意见元素的具体分析可得到意见挖掘结果。

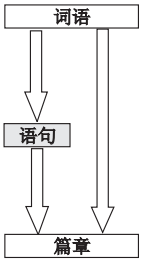

图1 意见挖掘层次图

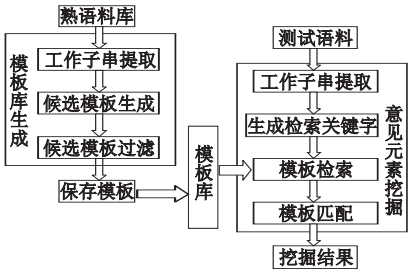

图2 基于模板的意见挖掘系统模块图

1.2 模板库生成

1.2.1 模板设计

虽然自然语言在具体的描述上表现得非常灵活,但是仍然有迹可循。这个规律可以是句子的语法结构,也可以是句子的语义结构,如主语—谓语—宾语的结构模式就是一种规律,或称为模板。这些模板与具体的句子之间形成了一种映射关系,熟语料库中的句子和未标注的生语料都通过某些规则映射到模板上,并通过这些模板形成相互的对应关系,从而达到对未标注语料进行意见挖掘的目的。因此,模板是整个系统中最重要的一环。

本文认为模板设计应遵循以下原则:a)模板必须包含熟语料的句子中实际出现的意见元素;b)通过模板应能较为方便准确地确定意见元素在句子中的位置;c)模板要有较好的适应性,能够适应句子在非主要内容上的变化。

这三点原则分别从不同方面对模板进行了限定。原则 a)指出了模板所要包含的基本内容,也就是熟语料句子中实际出现的意见元素。模板的目的就是提取意见元素,因此意见元素是模板最主要的成分。原则 b)强调的是模板的准确性,同样的意见元素在不同结构的句子中可能以不同的分布出现,要准确定位意见元素,模板中必须包括与句子结构相关的信息。原则 c)强调的是模板的适应性,要求模板能够适应结构相同而具体表达用词不同的句子。熟语料的规模是有限的,不可能覆盖所有的情况,因此需要将有限的模板尽可能多地应用到未知的情况中,以提高系统的适应性。

本文模板的构成分为以下两个方面:

a)模板的内容。模板包括的内容有意见元素、谓语动词、连词、助词、介词。根据上文提到的模板的设计原则 a),从熟语料中提取模板时,句中所包含的意见元素毫无疑问地会成为模板的内容。但是原则 b)要求模板能够有效地在生语料中定位意见元素,只包括意见元素的模板是无法达到这一目标的,因此必须在模板中加入句子结构的信息。

谓语动词是句子的核心,对句子的结构有至关重要的影响,因此谓语动词是一定要加入模板的。但是这样的条件仍然过于宽泛。连词、助词和介词这三类词在句子中起到了连接句子各部分和限定各部分之间关系的作用,因此这三类词也加入到模板中。

根据模板设计原则 c),模板需要有较好的适应性,因此将模板中的某些内容替换为符号的形式,主要包括表达意见元素所用的词,这样可以提高模板对于不同遣词的适应性。与句子结构相关的部分是不可以替换为符号的,这部分包括谓语动词、连词、助词和介词,这些部分必须以词的形式出现才能达到区分句子结构的作用。如:

例 1 其实,*张明海*本身就是爱岗敬业的标兵。

例 2 他的同事休· 维维安· 史密斯记得:“*老伊恩*肯定是世界上最*差劲*的证券经纪人之一。”

例 3 *巴克曼*是德军党卫军第 2“帝国”装甲师中最*优秀*的坦克指挥手。

三个例句中意见元素用斜体字标出,谓语动词用粗体字标出。例句是对不同人物不同方面的评价,评价意见有褒有贬。可以看到,三个例句的意见元素都不相同,表达的意见也不相同;但是三个例句在句子结构上都有共同之处,它们的谓语动词都是“是”,意见元素部分的句子结构可总结为被评价者 + 是 + 评价词。由此可见,模板中意见元素的部分应当替换为相应的符号。

综上所述,模板的内容包括意见元素、谓语动词、连词、助词、介词。其中意见元素以符号的形式出现,其他部分以词的形式出现。

b)模板的检索关键字。该关键字是用于在模板库中检索模板,检索关键字无须与模板一一对应,但需要将模板进行一定的区分以便有效地检索模板。本文在工作中选取谓语动词、评价内容和评价词三者的函数作为模板的检索关键字。函数如下:

$$key = sort(exist(C), word(V), exist(O))$$

其中:exist( ) 函数检测意见元素是否出现,如果出现则返回意见元素的符号,否则返回 null;word( ) 函数返回谓语动词的文本,如果没有谓语动词则返回 null;sort( ) 函数按照谓语动词、评价内容和评价词在熟语料中实际出现的顺序,对 exist( ) 函数和 word( ) 函数返回的结果进行排序,并返回排序结果。

例 4 来自大洋音尚的*叶世荣*、贾立怡、SOLER、陈尚实四组艺人去年在乐坛的杰出*表现*再次**得到**充分的*肯定*,而大洋音尚亦成为当晚本地乐坛的最大赢家。

该句出现了评价内容“表现”和评价词“肯定”,因此 exist(C) = C,exist(O) = O;该句的谓语动词为“得到”,因此 word(V) = 得到,最终提取出相应模板的检索关键字为 C 得到 O。

1.2.2 模板生成

模板生成的作用是从熟语料库中抽取出模板。熟语料库中的句子都已经做过人工标注,一个标注的例子如下:

例 5 爱德华的性情中,都是善良和 純。

```
〈tag〉
  〈sentence〉爱德华的性情中,都是善良和 純。〈/sentence〉
  〈seg〉爱德华/nr 的/u 性情/n 中/ng ,/w 都/d 是/v 善良/an
和/c 純/an 。/w 〈/seg〉
  〈P〉0〈/P〉
  〈V〉6〈/V〉
  〈C〉2〈/C〉
  〈O〉7〈/O〉
  〈ori〉正面〈/ori〉
〈/tag〉
```

例子中〈seg〉〈/seg〉标签的内容是例句的分词结果;〈P〉

〈/P〉、〈V〉〈/V〉、〈C〉〈/C〉、〈O〉〈/O〉分别是被评价者、评价的谓语动词、评价内容和评价词在分词结果中的词序号,序号以 0 开始,小于 0 的序号表示没有出现该项内容;〈ori〉〈/ori〉标签是评价的倾向性,正面表示褒义评价,负面表示贬义评价。

1)工作子串

根据 1.2.1 节的模板设计方案,句中所有的意见元素都转换为相应的符号,同时谓语动词 V、评价内容 C 和评价词 O 被提取出来,按照它们在句中的顺序生成模板的检索关键字。句中除意见元素、谓语动词、连词、助词、介词之外的词被去除,这样形成的结果称为工作子串。

2)模板的形式

工作子串过滤并结构化后形成模板,最后生成的模板如例 6 所示。

例 6

〈pattern〉
　〈V〉是〈/V〉
　〈key〉C 是 O〈/key〉
　〈ele〉PCVO〈/ele〉
　〈Pattern〉 P 的 C 是 O 〈/Pattern〉
　〈Ppos〉 P/nr 的/a C/n 是/v O/vn 〈/Ppos〉
　〈ori〉1〈/ori〉
〈/pattern〉

例子中〈pattern〉〈/pattern〉标签标记了一个模板,模板的第一个节点是模板中的谓语动词;〈key〉〈/key〉节点标记了模板中 VCO 元素的实际出现情况及其顺序;〈ele〉〈/ele〉节点标记了模板中实际的意见元素和谓语动词出现情况及其顺序;〈Pattern〉〈/Pattern〉节点标记了模板的内容;〈Ppos〉〈/Ppos〉节点给出了〈Pattern〉节点中每个词的词性;〈ori〉〈/ori〉节点标记了该模板的倾向性。

1.2.3　候选模板过滤

有些候选模板在熟语料库中出现的次数较多,这些候选模板相对比较可信,而那些出现次数较少的候选模板可信度相对较差,因此需要对候选模板进行过滤。过滤阈值为

$$d = \frac{\text{sum}}{\text{num}} \cdot \delta$$

其中:d 是一个计算得到的结果,它的含义是候选模板在熟语料库中出现的平均次数乘以一个比例系数,它的值与三个变量有关:sum 为所有候选模板在熟语料库中出现的次数之和;num 为候选模板的数量,δ 是比例系数,取值为(0, 1)。过滤后得到模板存入模板库中。在系统实际测试中,sum = 465;num = 198。

1.3　意见元素挖掘

意见元素的提取可以分为两步,即模板检索和模板匹配。其中,模板检索的作用是根据输入的未标注句子在模板库中检索相应的模板。显然,未标注的句子是无法直接用于检索模板的,这中间存在三个问题:用什么样的关键字来检索模板;如何获得这些关键字;怎样使用这些关键字来检索模板。模板匹配的作用是应用模板来提取句中的意见元素。

a)需要从未标注句中提取出检索关键字。检索关键字的内容包括谓语动词 V、评价内容 C 和评价词 O,候选谓语动词选取词性标注出的动词。评价内容一定是与人有关的名词,本文建立了内容词表(或称属性词表),共 103 词,然后将 VCO 组合以形成检索关键字。实际的系统中,各项元素需要按照其实际的出现顺序进行排列。组合的方式有四种,即 VCO、VO、CO 和 O。检索关键字采用这四种组合方式是由于倾向性的表达是通过评价词 O 来进行的,所以检索关键字中必须包含 O,谓语动词 V 和评价内容 C 可以有也可以没有,因此一共是四种组合。

b)使用检索关键字在模板库中检索可能与生语料匹配的模板。由于能够与模板中的意见元素匹配数量较多的检索关键字更为准确,模板检索关键字按照 VCO、VO、CO、O 的顺序在模板库中检索模板,当检索到模板并成功匹配时,停止继续检索,以匹配上的模板作为结果。

c)进行模板的匹配。将检索到的模板和从未标注句子中提取出的工作子串进行匹配,匹配方式采用寻找两者的最大公共子串的方式。当检索到的模板和工作子串的最大公共字串长度与模板长度相同时,表明工作子串和模板可以完全匹配。这时可以根据意见元素在模板中的分布获得意见元素在工作子串中的分布,并通过工作子串的还原得到原句中的意见元素。下面例 7 说明意见挖掘的过程。

例 7　北京时间 8 月 20 日,中国队 90 比 121 败于美国队,但是姚明的表现还是得到了足够的肯定。

北京/n　时间/n　8 月/t　20 日/t,/w　中国队/n　90/m　比/p　121/m　败/v　于/p　美国队/n,/w　但是/c　姚明/nr　的/u　表现/vn　还/d　是/v　得到/v　了/u　足够/a　的/u　肯定/v。/w

例子中是需要进行意见挖掘的句子,已经经过词性标注。首先在句子中查找候选的意见元素,查找得到评价内容 C:表现;评价词 O:肯定;人物 P:姚明,谓语动词 V:是、得到。根据候选意见元素提取工作子串:P 的 C 是得到的 O。候选意见元素组合成检索关键字:C 得到 O、得到 O、CO、O。检索关键字 C 得到 O 在模板库中检索到一个相应的模板:

〈pattern〉
　〈V〉得到〈/V〉
　〈key〉C 得到 O〈/key〉
　〈ele〉PCVO〈/ele〉
　〈Pattern〉P 的 C 得到 O 〈/Pattern〉
　〈Ppos〉P/nr 的/c C/n 得到/v O/vn 〈/Ppos〉
　〈ori〉1〈/ori〉
〈/pattern〉

模板的内容为 P 的 C 得到 O,工作子串为 P 的 C 是得到的 O,两者的最大公共字串为 P 的 C 得到 O,与模板相符,匹配成功。从模板中恢复意见元素的结果如表 1 所示。

表 1　例 7 意见挖掘结果

被评价者 P	谓语动词 V	评价内容 C	评价词 O
姚明	得到	表现	肯定

2　实验结果及分析

本章对系统进行了测试,首先从互联网收集了 600 条带有人物评价信息的句子,并对意见元素进行了人工标注。其中 400 句作为训练语料,另 200 句再加上 50 句没有意见的客观陈述句作为测试语料。先对测试语料进行分词,对分词的结果进行了人工校对,然后输入系统进行测试。测试结果如表 2 所示。

表 2　语句意见挖掘结果

语句总数	含有观点的句子数	提出观点的句子数	结果正确	结果错误	召回率/%	正确率/%	F-score
250	200	186	140	46	70	75.3	72.55

对 46 个错误的句子进行错误原因统计,结果如表 3 所示。

表 3　错误原因统计

错误总数	人物属性错	评价词错	意见倾向错	其他	对象人物错
46	5	5	4	12	20

以下举例给出实验的结果,如表 4 ~ 7 所示。

例 8　周娥皇的美,还是超乎了李煜的想象,双目流盼,明净澄澈;樱桃小嘴别致玲珑;如云乌发,高高挽起;如玉脖颈,颀长优雅。

表 4　例句 8 挖掘结果

被评价者	属性	评价词	情感倾向
李煜	嘴	别致玲珑	正面

例 9　张氏在其夫心里就是一个简单粗俗、常常醋意大发的人。

表 5　例句 9 挖掘结果

被评价者	属性	评价词	情感倾向
其夫	无	粗俗	负面

以上是对象判断错误的例子。例 8 中都是对周娥皇的评价意见,但是系统把第三小句及其后面的意见都判断为是对李煜的评价。这是因为系统没有对人物的句间省略进行详细的分析和处理,只取最近一个出现的人物进行恢复,因此句中靠后出现的“李煜”就作为第三小句的被评价者而提取出来,没有提取到正确的被评价者。

例 9 中出现了两个人物,而系统判断被评价者发生了错误。句中出现了人物张氏和其夫,被评价者应该是张氏,系统错误地判断成了其夫。这是在模板匹配时出现的错误。由于在对工作子串和模板进行匹配时采用的是 LCS 算法,当工作子串中有两个人物张氏和其夫,而模板中只有一个人物时,匹配上的是靠后的人物,即其夫。

例 10　冰冰做出似傻非傻的神情,很可爱。

表 6　例句 10 挖掘结果

被评价者	属性	评价词	情感倾向
冰冰	无	傻	负面
冰冰	神情	可爱	正面

例 10 是评价词判断错误的例子,句中的第一个“傻”被判断成了评价词,这是由于模板匹配的方法比较机械,没有对句子结构作详细分析。一旦出现了可能的意见元素并且有能够匹配的模板,就会提取出相应结果。

例 11　我从来就不曾觉得罗纳尔多是忠诚的。

表 7　例句 11 挖掘结果

被评价者	属性	评价词	情感倾向
罗纳尔多	无	忠诚	正面

例 11 是倾向性判断错误的例子,这是一个否定句,否定词和评价词距离较远,而系统只会对紧邻评价词的否定词作处理,因此没有判断出“忠诚”已经被否定。

通过上面的实验可以看出,本文所采用的基于模板的方法可以比较有效地从句中提取出意见元素。系统的问题主要在于:a)模板匹配时单纯采用 LCS 算法进行匹配,没有作进一步的分析;b)系统对于句间人物的省略没有作很好的处理;c)系统没有对句子的结构进行分析,因此当评价意见以较为复杂的形式出现时,提取会出现错误;d)极性词、属性词等的识别也会对提取效果产生影响。在各项意见元素的提取中,被评价者的错误较多,这主要是由于上述 a)b)这两个原因造成的;而第三点原因对其他意见元素的影响较大。在将来的工作中,将会对这些部分作有针对性的改进。

目前语句级的意见挖掘成型的系统不多,且都是针对产品的挖掘,正确率一般在 76% 以下。本文的工作达到了 75.3% 的准确率,与他人的工作相比,本文工作取得了较好的效果。召回率略低,主要是由于模板的覆盖面不够广,有些测试语料是模板没有覆盖到的,可以通过增加模板来改进系统的性能;由于模板依赖于熟语料库,本文中熟语料库规模较小,只有 400 个句子,影响了模板的覆盖能力,因此可以通过建立更为完备的熟语料库来增加模板的数量,从而提高系统的性能。本文选择的人物评论挖掘,相对于产品挖掘来说难度更大,主要表现在以下几个方面:a)对主题抽取。人物既可以发表评论,也可以被评论,而产品只能被评论,因此在被评价者抽取上更为困难。b)被评价对象的属性抽取。人物的属性复杂而且用语多样,如“眼”“目”“眼睛”“双眼”“双目”“眸”“眸子”“目光”“眼神”等都是指同一个评价属性。产品的属性相对较为简单,称谓较为固定,如数码相机的“像素”等。c)评价词提取。对人物的评价词灵活而丰富,如“杏眼”“蜂腰”“柳下惠”等,又如“婉顺”“宽和”等由多个词揉合而成,又如“风流”“风骚”等描述人物的词语常有二义性,因此增加了识别和判断的难度。d)意见的倾向分析。对人物的评论常常显得委婉含蓄,不够直接,表达方式也丰富异常,而产品的评价大多直截了当。因此,综合考虑上述原因,本文在中文人物评价挖掘方面采用基于模板方法取得了 75.3% 正确率,具有一定的价值。

3　结束语

本文提出了一种基于模板的中文人物评论语句意见挖掘算法,算法分为模板库生成和意见元素挖掘两个模块。在模板库生成模块中,首先在对意见元素及汉语词性特点的分析基础上,设计了评价挖掘结构化模板;根据模板设计,从已标注的语料中提取候选模板并过滤,生成模板库。在意见元素挖掘模块中,对输入的测试句子提取工作子串和检索关键字,使用检索关键字在模板库中检索相应的模板并将检索到的模板和工作子串进行匹配,根据匹配的结果挖掘出意见元素。实验表明,本文的方法能够较为准确地从测试语料中提取出意见元素,正确率达到 75.3%。

参考文献:

[1] 路斌,万小军,杨建武,等. 基于同义词词林的词汇褒贬计算[C]//中国计算技术与语言问题研究——第七届中文信息处理国际会议论文集. 北京:电子工业出版社,2007:17-23.

[2] 朱嫣岚,闵锦,周雅倩,等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报, 2006,20(1): 14-20.

[3] KIM S M, HOVY E. Determining the sentiment of opinions[C]// Proc of the 20th International Conference on Computational Linguistics. Morristown:Association for Computation Linguistics,2004:1367-1373.

[4] 娄德成,姚天　. 汉语句子语义极性分析和观点抽取方法的研究[J]. 计算机应用,2006,26(11):2622-2625.

[5] 娄德成. 基于 NLP 技术的中文网络评论观点抽取方法的研究[D]. 上海:上海交通大学,2007.

[6] 彭其伟. 基于统计方法的中文文本情感倾向分类研究[D]. 大连:大连理工大学, 2007.

[7] 徐琳宏,林鸿飞,杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报, 2007,21(1):96-100.