

基于关联特征词表的中文比较句识别

杜文韬^{1,2*}, 刘培玉^{1,2}, 费绍栋^{1,3}, 张 朕^{1,2}

(1. 山东师范大学 信息科学与工程学院, 济南 250014; 2. 山东省分布式计算机软件新技术重点实验室, 济南 250014;

3. 山东财经大学 图书馆, 济南 250014)

(* 通信作者电子邮箱 duwentao1949@126.com)

摘 要:中文比较句研究多集中于语言学领域,然而利用机器学习的方法识别比较句的研究才刚刚起步。根据关联规则挖掘算法的基本原理提出一种基于关联特征词表的比较句识别方法,该方法将词和词性作为一个基本元素,定义特征词表中核心词和依存词之间的关联方式,利用支持向量机(SVM)分类器进行比较句的识别。实验结果表明,该方法能够有效地识别出中文比较句,在准确率、召回率和F值上均取得不错的效果。

关键词:比较句识别;文本分类;中文比较模式库;类序列规则;关联特征词表

中图分类号:TP391 **文献标志码:**A

Chinese comparative sentences recognition based on associated feature vocabulary

DU Wentao^{1,2*}, LIU Peiyu^{1,2}, FEI Shaodong^{1,3}, ZHANG Zhen^{1,2}

(1. School of Information Science and Engineering, Shandong Normal University, Jinan Shandong 250014, China;

2. Shandong Provincial Key Laboratory for Normal Distributed Computer Software Technology, Jinan Shandong 250014, China;

3. Library, Shandong University of Finance and Economics, Jinan Shandong 250014, China)

Abstract: Chinese comparative sentences are more focused in the field of linguistics. Using machine learning methods to identify comparative sentences, however, has only just started. According to the basic principle of the association rules mining algorithm, a method of comparative sentences based on the associated feature vocabulary was proposed. This method regarded word and part of speech as basic elements, defined the connecting way between the table definition core words and interdependent relationship words, and used the Support Vector Machine (SVM) classifier for the identification of comparative sentences. The experimental results show that this method can effectively identify Chinese comparative sentences, and achieves good results in precision, recall and F-measure.

Key words: comparative sentences identification; text classification; Chinese comparative pattern database; class sequential rule; associated feature vocabulary

0 引言

比较是我们在日常生活中经常用到的一种表达方式,通过对两个事物的比较,可以判断出同类产品间的异同和优劣。尤其是随着网络技术的不断发展,许多博客、微博、日志、社会网络、论坛等新型网络元素迅速兴起,网络信息更加个性化和专业化。这些信息中不乏对各种新事物、新理论、新技术、新产品、新观点、新艺术等进行评论和比较的主观性信息,对这些主观信息的比较关系进行研究,分析同类产品的异同和优劣,可以对观点挖掘、信息推荐等应用提供重要的依据。从海量的评论信息中准确地识别比较句是研究比较关系的前提工作。

在国内对中文比较句的研究最初主要集中在语言学领域,包括比较的范畴、典型的比较句式、比较的语义以及比较的共时和历时研究等。刘焱^[1]在“除去特定语境影响的前提下,看一个句子在形式上是否具有比较句的结构特点、在功能上是否表达了比较意义”这一思想指导下,指出比较范畴应

该是一种“语义—句法”范畴。车竞^[2]从词汇短语角度研究了比较句的语义和句法的关系,讨论了比较句的分级和等级的度量。尚平^[3]认为汉语比较句的研究不仅要坚持语义同句法形式的结合,更要追求简洁明晰的分类结果。庞倩^[4]从结构和功能上分析了等比句和差比句两种句式的基本特点。语言学领域的研究工作对比较关系的挖掘有着指导意义,但并不能直接运用到计算机的自动挖掘上来。

最早用计算机进行比较句识别的是伊利诺伊大学芝加哥分校的Jindal等^[5],他们采用模式发现和监督学习的方法对英文比较句的识别进行了研究,达到了79%的准确率和81%的召回率。后来他们又用标签序列规则^[6]作为特征对比较句中的比较关系进行了进一步的研究。北京大学黄小江等^[7]使用支持向量机(Support Vector Machine, SVM)分类器和类序列规则挖掘的方法对中文比较句进行识别。Yang等^[8]从韩语文档中提取比较词汇作为关键词特征利用机器学习技术识别比较句。大连理工大学的宋锐等^[9]通过构建中文比较模式库和条件随机域模型结合的方法进行比较句识

收稿日期:2013-01-04;修回日期:2013-02-11。

基金项目:国家自然科学基金资助项目(60873247);国家社会科学基金资助项目(12BXW040);公安部科技创新计划项目(2011YYCXSDST057);山东省自然科学基金资助项目(ZR2012FM038, ZR2011FM030);山东省科技发展计划项目(2012GGB01194)。

作者简介:杜文韬(1987-),男,山东威海人,硕士研究生,CCF会员,主要研究方向:网络信息安全、网络舆情分析;刘培玉(1960-),男,山东潍坊人,教授,博士生导师,主要研究方向:计算机网络信息安全、网络系统规划、软件开发;费绍栋(1984-),男,浙江宁波人,博士研究生,主要研究方向:计算机网络信息安全、网络舆情分析、社会网络;张朕(1988-),男,山东烟台人,硕士研究生,主要研究方向:网络信息安全、网络舆情分析。

别和比较关系抽取;黄高辉等^[10]采用基于条件随机场(Conditional Random Field, CRF)的算法进行比较句的识别和关系抽取;李建军^[11]利用熵值平衡算法提取句中的统计特征和序列特征进行比较句识别。实验结果表明,这些方法都能有效地识别中文比较句,但是准确率还有待进一步提高。

本文通过分析比较句的结构特征,利用关联规则挖掘算法的原理建立关联特征词表,并结合规则中特征词的关联方向,建立了规则与特征词表之间的有向联系进行中文比较句的识别。实验结果表明,基于关联特征词表的方法可以更加有效地识别中文比较句。

1 比较句研究问题分析

比较句识别实际上就是判断一个句子是比较句还是非比较句的过程,从本质上讲属于一个文本分类问题。准确识别比较句的关键是找到一种能够区分比较句和非比较句特征的方法。从语言特征上看,比较句与非比较句之间在词汇和语序上存在一定的差异,这使得传统的文本分类计数和序列模式匹配方法用于比较句的识别成为可能。在英文比较句中,由于比较词特征比较明显,一般是形容词和副词的比较级或最高级形式,所以对英文比较句进行识别时,在语料预处理阶段,使用英文特有的词性标注工具,能够很好地分析出具有比较意义的形容词与副词等有效特征词。而中文比较句表达方式中不具有比较级或最高级的形态,往往普通的词语、成语以及谚语等信息就可以表达多个事物之间比较的含义,格式上也灵活多变,给准确地识别比较句带来了很大困难。

1.1 中文比较句式分析

从语义上分析,中文比较句描述的是同一类事物的两个或两个以上实体在同一个属性上的比较。在语言学领域认为一句完整的比较句通常包含四个基本要素^[2]:比较主体、比较客体、比较属性和比较结果。如例句:

- 1) 诺基亚 N8 的屏幕不如 iPhone 的好。
- 2) 相比我的森海 MX160 音质稍有不足。

在例句1)中“诺基亚 N8”为比较主体,“iPhone”为比较客体,“屏幕”为比较属性,“好”为比较结果。但通常我们遇到的比较句中并不一定包括完整的四要素,在例句2)中就缺少比较主体。由于比较元素不完整,使得在分析比较关系上增加了一定难度,然而比较句的识别只是判断该句是否为比较句,而比较关系的抽取则是对比较句四个基本要素的提取,所以在对比较句的识别中可以对四要素是否完整不作考虑。根据比较结果,又可以将比较句进一步划分为几个类别,夏群^[12]总结了马建忠^[13]对中文比较句的分类,包括三类:平比、差比和极比。目前这种分类方法在极比和差比的划分上还存在一定的争议,文献[9]根据语言学领域的分类分别构建了平比、差比、极比的比较模式,并将依靠模式无法判断类别的模式划分为“未定”类别。在本文中并没有将语句进行复杂的划分,只是根据匹配结果将句子划分为比较句和非比较句两大类。

1.2 现有技术问题分析

近些年,比较句识别研究在中文领域也相继展开,并取得一定得成果,其中主流的技术归纳起来有基于类序列规则的比较句识别和基于比较模式库的比较句识别。

文献[1]首先将类序列规则应用到比较句识别中,作为序列模式的一种与其他序列模式挖掘的思想一样,都是寻找满足用户定义好的最小支持度约束的模式,为后期的比较句识别提供特征输入。文献[7]在此基础上将该方法加以改进运用到中文比较句识别上,其主要改进是取消了文献[1]中

的滑动窗口策略,而是将句中每一个分句作为一个序列来提取与匹配规则。实验结果表明了该方法的可行性,但是问题在于,文献[7]中仍然延续了文献[1]的一个特点,就是仅将规则序列中的比较特征词做了词与词性的约束,而对序列的其他词仅有词性约束。由于中文句式的复杂性,表达方式灵活多样,使得很多非比较句也满足了序列规则。例如:

- 1) 此双核解决方案有利于提高系统性能。
- 2) 每次买到的东西差不多在下了订单之后的四天之内,这次居然不到三天!

在上述两个例句中,以序列规则为特征,例句分别满足序列规则“* / a 于 / p”和“* / n 差不多/a”,但例句并不是比较句。因此,如果采用类序列规则的方法,上述例句就会被误认为是比较句,这种类似句子结构在比较句和非比较句中都有很大比例,影响了分类的准确性。

在文献[9]中运用构建中文比较模式库,收集大量比较模式的方法进行比较句的识别,并对比较句进行了分类,取得很好的效果。但在该方法中仅用词作为特征进行模式匹配,而是否进行了比较往往并不是一个词能够表达清楚的,出现某个特征词的句子未必一定是比较句,而一些不是比较特征词的词语与其他词语搭配后往往能够形成比较句式。例如:

- 1) 诺基亚 5230 没有无线上网功能。
- 2) 诺基亚手机的屏幕色彩没有三星的好。

在例句1)中有一个关键词“没有”,但并没有表达出比较的含义。例句2)中是“没有……好”两个词组合在一起使句子构成一层比较的含义,这种灵活的比较方式不是单靠词特征就能够识别的。

2 基于关联特征词典的比较句识别

2.1 关联规则挖掘的基本原理

关联规则挖掘算法是由 Agrawal 等^[14]于 1993 年首先提出的,用于挖掘顾客交易数据库中项集间的关联规则问题,其核心方法是基于频集理论的递推方法,其形式化表示^[14]为:设 $I = \{i^1, i^2, \dots, i^m\}$ 是项的集合。设任务相关的数据 D 是数据库事务的集合,其中每个事务 T 是项的集合,使得 $T \subseteq I$ 。每个事务有一个标识符,称作 TID。设 A 是一个项集,事务 T 包含 A 当且仅当 $A \subseteq T$ 。关联规则是形如 $A \Rightarrow B$ 的蕴涵式,其中 $A \subset I, B \subset I$, 并且 $A \cap B = \emptyset$ 。关联规则 $A \Rightarrow B$ 可用如下参数描述。

支持度:

$$\text{Support}(A \Rightarrow B) = P(A \cup B) \quad (1)$$

置信度:

$$\text{Confidence}(A \Rightarrow B) = P(B | A) = \frac{\text{sup}(A \cup B)}{\text{sup}(A)} \quad (2)$$

提升度:

$$\text{Lift}(A \Rightarrow B) = \frac{\text{sup}(A \cup B)}{\text{sup}(A) * \text{sup}(B)} \quad (3)$$

支持度 *Support* 是指数据集 D 中事务同时包含 A 和 B 的百分比,是对关联规则重要性的衡量,支持度说明了该规则在事物集中有多大的代表性。如果项集大于最小支持度 min_sup ,则称它为频繁项集。置信度 *Confidence* 是指数据集 D 中包含事务 A 与同时也包含 B 的百分比,是对规则准确度的衡量。它反映在给定 A 的前提下, B 发生的条件概率。提升度 *Lift* 有时也称为 *Interest*,它是事物 A 和 B 同时发生的概率和在假定 A 和 B 独立的前提下 A 和 B 同时发生概率之间的比值。*Lift* 用来衡量 A 和 B 之间的关联与 A 和 B 相互独立偏离的程度。如果 *Lift* 接近于 1, A 和 B 就是相互独立的;如果 *Lift* 值小于 1,则 A

和 B 互为抑制; $Lift$ 值越大于 1, 则规则的实际意义就越好。

2.2 关联特征词表的生成

针对目前主流的比较句识别技术中存在的不足及它们之间的共性, 本文借鉴关联规则挖掘算法的基本原理, 提出了一种基于关联特征词表的识别方法, 利用 Apriori 算法从句中选取特征, 利用挖掘出来的特征中的前导词 (A) 和后继词 (B) 分别构建核心词表和依存词表, 并采用本体概念之间的语义相关性描述领域中的复杂关系, 通过语义相关度过滤掉领域中相关性较小的候选集, 以减少关联规则挖掘中候选集的数量^[15]。文中提升度 $Lift$ 设置为 1, 并删除特征度小于 1 的序列集, 以形成有效序列集。本文中使用的支持度和置信度约束分别为 $Support > 2/N$, $Confidence > 2/N_A$, 其中 N 是序列集中序列数目, N_A 是序列集中包含前导词 A 的序列数目。

该方法主要是针对现有主流方法的不足进行改进的一种方法, 文献[7]在利用类序列规则进行比较句识别时, 仅对特征词做了词和词性的绑定, 而对其他词只有词性的约束, 这就使得句子中的词只要满足序列规则中的词性约束即可满足规则。由于中文表达方式的复杂多样性, 仅用词性做约束会大大降低准确率, 所以在本文中, 将规则的前导和后继分为两部分, 对特征词和其他词都分别用词与词性做约束, 并且将关联规则中的特征词称作核心词, 能与核心词搭配构成比较句式的词语称作依存词, 不同的核心词与依存词之间的搭配具有方向性。核心词和依存词分别存放在核心词表和依存词表中, 两个词表有一个共同的标志字段用于建立两表之间的关联, 截止到目前为止, 关联词表中共收录了 64 个核心词和 238 个依存词。关于关联规则挖掘的基本原理参见文献[13]。存储样例如表 1 和表 2 所示。

表 1 核心词表

核心词	标志	核心词	标志
不如	0	与/p	1
相比之下	0	跟/p	1
相对于	0	和/p	1
比不上	0	比/p	2
无敌	0	没有/d	2
遥遥领先	0		

表 2 依存词表

依存词	标志	依存词	标志
一样/a	1	优秀/a	2
差不多/a	1	差/a	2
一模一样/a	1	高档/a	2
类似/a	1	清爽/a	2
相同/a	1	方便/a	2
好/a	2		

2.3 关联方法的运算

在关联算法中, 我们将每个例句形式化地表示为一个序列 $f = \langle s^1, s^2, \dots, s^m \rangle$, 将以逗号、分号、感叹号、句号作为分句标志的子句作为一个子序列, 记作 $s_k = \langle a^1 a^2 \dots a^n \rangle$, 其中 $1 \leq k \leq m$, a 表示序列中的元素。识别的实现过程中包括两大部分: 核心词的识别和依存词的识别, 分别用式 (4) 和式 (5) 进行匹配运算:

$$s_k = \begin{cases} 1, & \text{依存词覆盖 } s_{kj}; 1 \leq k \leq m, i \leq j \leq n \\ 0, & \text{其他} \end{cases} \quad (4)$$

$$sent = \begin{cases} 1, & \exists (s_k \in f \wedge s_k = 1) \\ 0, & \text{其他} \end{cases}; 1 \leq k \leq m \quad (5)$$

运算流程如下所示:

1) 依次输入子序列: $s_k = \langle a^1 a^2 \dots a^n \rangle (1 \leq k \leq m)$ 。

2) 定义两个指针变量 i 和 j , 令 i 指向 a^1 , j 始终指向 i 后边的元素。

3) 先对指针 i 所指的元素运用式 (4) 与核心词表中的词进行匹配, 其中 $s_k = 1$ 表示核心词表包含该元素, $s_k = 0$ 表示匹配失败。如果匹配成功, 则执行步骤 5); 否则执行步骤 4)。

4) 判断 i 所指示的元素 a^{ki} 是否是序列的最后一个元素: 如果是最后一个元素, 则表明该子序列不含有比较特征, 不属于比较句, 则从步骤 1) 开始运算下一个子序列; 否则指针 $i = i + 1$, 执行步骤 3)。

5) 记录指针 i 所指元素的位置和元素信息, 指针 j 指向 i 的下一个元素, 并将 j 所指的元素再次运用式 (4) 与依存词表中的依存词进行匹配。如果 $s_k = 1$, 则子序列符合序列规则, 执行步骤 6); 否则继续执行步骤 7)。

6) 执行式 (5), 并定义变量 $C: C += sent$ 。变量 $sent$ 表示满足序列规则的子序列, 变量 C 表示满足序列规则的子序列的个数。如果 C 为大于零的数, 那说明序列为比较句; 否则为非比较句。

7) 判断指针 j 是否指向子序列的最后一个元素, 如果是最后一个元素, 则该子序列不是比较句; 否则 $j = j + 1$, 重复执行步骤 5)。

2.4 观点句的分类

本文将比较句识别作为一个二分类问题来研究, 利用支持向量机 (SVM) 分类算法将经关联方法运算后的观点句进行分类, 整体流程如图 1 所示。

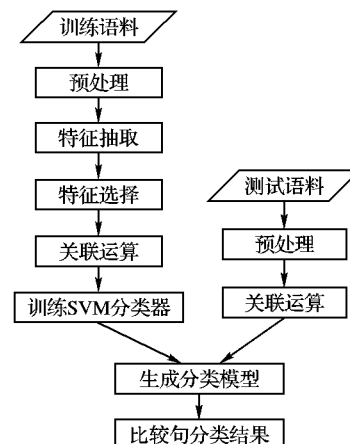


图 1 SVM 分类器识别比较句流程

具体流程可以分为以下几步:

1) 预处理。先用 LjParser 切词工具将文本进行切词处理, 将词和词性作为一个元素。

2) 特征提取、特征选择及关联运算已在 2.2 节和 2.3 节中具体介绍。从关联方法运算后的输出结果 C 中提取两种特征作为 SVM 分类器的候选特征: 一类为 C 等于零; 另一类为 C 为大于零的实数。

3) 训练 SVM 分类器。实验中使用的 SVM 分类器为 Libsvm 工具包。用序列化的训练语料训练 SVM 分类器, 生成分类模型, 再用此模型对序列化的测试语料进行分类。

最后根据分类特征将观点句分为比较句和非比较句两大类。

3 实验与结果

3.1 数据集

本文采用 COAE2012 提供的测试语料, 包含汽车产品和电子产品两个领域的产品评论信息, 其中训练语料各 1 200

条,测试语料各3600条,对训练语料进行人工提取核心词和依存词,提取结果进行去重处理,配合哈尔滨工业大学同义词词林对核心词和依存词进行同义词扩充,构建核心词词表和依存词词表,对能够构成比较关系的核心词和依存词建立关联规则。然后用任务中提供的7200句测试语料进行测试。数据集的样本情况如表3所示。

表3 数据集样本

数据	汽车产品	电子产品	总数
训练语料	1200	1200	2400
测试语料	3600	3600	7200

3.2 评价方法

本文采用 COAE2012 提供的评测方法。以准确率(Precision)、召回率(Recall)和F值(F-measure)作为评价标准,并借鉴文献[7]的评价方法,定义如表4所示。

表4 分类结果

句子类型	实际比较句	实际非比较句
判定比较句	tp	fp
判定非比较句	fn	tn

计算公式如下:

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$F-measure = \frac{2tp^2}{2tp^2 + tp \times fp + tp \times fn}$$

3.3 结果分析

本文选取了SVM分类器对结果进行分类,在文献[5,7,9]研究的基础上,使用关联特征词表进行模式匹配的方法识别比较句,该方法在COAE2012评测会议上取得了优异成绩。另外,在比较句识别实验中,本文还分别对比了以比较词、比较模式和类序列规则为特征的比较句识别方法。结果如表5所示。

表5 比较句识别实验结果

识别方法	准确率	召回率	F值
比较词	66.56	99.50	79.76
比较模式	77.00	80.00	78.56
类序列规则	91.40	79.60	85.00
关联特征词表	94.98	89.50	83.77

从表5可以看出,在基于四种不同特征的比较句识别方法中,关联特征词表的识别方法取得了不错的效果。基于比较词的识别方法中,虽然有很高的召回率,但准确率比较低,F值也处于平均水平;基于比较模式库的方法,在准确率上比基于比较词的方法有所提高,但召回率和F值均有所下降,而且该方法的各项评价指标对比较模式库的规模有很大依赖性;基于类序列规则的方法准确率和F值都达到了不错的结果,尤其是在准确率上较前两种方法有了很大的提高,但在召回率上还有待提高;基于关联特征词表的方法是在类序列规则的基础上将除了特征词以外的词也以词和词性作为特征进行关联匹配,并依据关联规则挖掘算法的基本原理定义了核心词与依存词之间的关联方式,从待识别的句子中先识别出是否存在核心词,然后依据核心词与依存词之间的规则查找是否与之匹配的依存词,从而判定是否为比较句。该方法与

其他特征识别方法相比在准确率上进一步得到了提高,但在召回率和F值上与基于比较词特征和类序列规则特征的识别方法相比还有所差距,有待进一步改进提高。

4 结语

目前,对中文比较句自动识别的研究在国内信息学领域还处于起步阶段。本文将语句中的词与词性作为元素构建关联特征词表进行规则匹配,克服了类序列规则中除特征词以外仅以词性做约束的缺点,并且规定了核心词与依存词之间关联的方向性。实验表明,该方法能够有效地识别中文比较句。但是,由于本文仅对汽车产品和电子产品两个领域的训练语料进行特征提取,导致特征词表的构建具有一定的领域局限性。另外,由于训练语料数量有限,提取的核心词和依存词不够全面,使得准确率还有待提高。下一步的工作,继续挖掘不同领域中比较句的特征,搜集更多的能够成为比较规则的核心词和依存词来扩充特征词表。

参考文献:

- [1] 刘焱. 现代汉语比较范畴的语义认知基础[M]. 上海: 学林出版社, 2004.
- [2] 车竞. 现代汉语比较句论略[J]. 湖北师范学院学报, 2005, 25(3): 60-63.
- [3] 尚平. 比较句系统研究综述[J]. 语言文字应用, 2006(S2): 77-80.
- [4] 庞倩. 略论现代汉语比较句之结构和功能特点[J]. 北方文学, 2012(2): 97-98.
- [5] JINDAL B, LIU B. Identifying comparative sentences in text documents[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2006: 244-251.
- [6] JINDAL N, LIU B. Mining comparative sentences and relations[C]// Proceedings of the 21st National Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2006: 1331-1336.
- [7] 黄小江, 万小军, 杨建武. 汉语比较句识别研究[J]. 中文信息学报, 2008, 22(5): 30-37.
- [8] YANG S, KO Y J. Extracting comparative sentences from Korean text documents using comparative lexical patterns and machine learning techniques[C]// Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Stroudsburg, PA: Association for Computational Linguistics, 2009: 153-156.
- [9] 宋锐, 林鸿飞, 常富洋. 中文比较句识别及比较关系抽取[J]. 中文信息学报, 2009, 23(2): 102-107.
- [10] 黄高辉, 姚天昉, 刘全升. 基于CRF算法的汉语比较句识别和关系抽取[J]. 计算机应用研究, 2010, 27(6): 2061-2064.
- [11] 李建军. 比较句与比较关系识别研究及其应用[D]. 重庆: 重庆大学, 2011.
- [12] 夏群. 汉语比较句研究综述[J]. 汉语学习, 2009(2): 58-64.
- [13] 马建忠. 马氏文通[M]. 上海: 商务印书馆, 1898.
- [14] AGRAWAL R, IMIELINSKI T, SWAMI A. Mining association rules between sets of items in large databases[C]// Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. New York: ACM, 1993: 207-216.
- [15] 武建华, 宋擒豹, 沈均毅, 等. 基于关联规则的特征选择算法[J]. 模式识别与人工智能, 2009, 22(2): 256-262.
- [16] 张磊, 夏士雄. 基于语义相关性的关联规则挖掘研究[J]. 东南大学学报, 2008, 24(3): 358-360.