基于 CRF 算法的汉语比较句识别和关系抽取 *

黄高辉, 姚天 , 刘全升
(上海交通大学 计算机科学与工程系, 上海 200240)

摘 要: 比较句是表明事物之间关系的常见表达方式,对于文本挖掘,特别是情感分析,具有重要的价值。目前汉语比较句的研究还是一个新颖的课题,包括汉语比较句的识别和比较关系的抽取。对于汉语比较句的识别,在前人研究的基础上,以 SVM 为分类器,以特征词和 CSR 序列规则为特征,同时利用 CRF 算法抽取实体对象,并增加以实体对象的信息作为特征,显著提高了比较句识别的准确率、召回率和 F-度量,最高分别达到96.55%、88.63%和 92.43%。对于汉语比较关系的抽取,在 CRF 算法抽取实体对象的基础上,通过定义一些规则,抽取比较主体和比较基准,也取得了较好的效果,其中比较主体的抽取效果要好于比较基准。

关键词: 比较句; 比较关系; CRF 模型; 比较主体; 比较基准
中图分类号: TP309 **文献标志码**: A **文章编号**: 1001-3695(2010)06-2061-04
doi:10.3969/j.issn.1001-3695.2010.06.018

Mining Chinese comparative sentences and relations based on CRF algorithm

HUANG Gao-hui, YAO Tian-fang, LIU Quan-sheng
(*Dept. of Computer Science & Engineering, Shanghai Jiaotong University, Shanghai* 200240, *China*)

Abstract: Comparative sentences are a common kind of expression to indicate the relations of different objects. They are valuable for text mining, especially for opinion mining. It is a novel research to identify Chinese comparative sentences and extract comparative relations. To identify Chinese comparative sentences, this paper took SVM as classifier and regarded keywords and class sequential rule as feature based on the previous research, and then used CRF algorithm to identify entity and also took the entity's information as feature. Finally, remarkably improve the precision, recall and F-measure for identifying comparative sentences and got the result up to 96.55%, 88.63% and 92.43% respectively. To mine comparative relations, extracted comparative subject and objected by defining some rules together with the result of CRF algorithm for identifying entity, and obtained good result. And the result to extract comparative subject is better than comparative object.

Key words: comparative sentence; comparative relation; CRF model; comparative subject; comparative object

0 引言

比较句是日常生活中常见的一种表达方式。通过比较,人们能够很方便地表达自己的想法和情感,从而得知不同事物之间孰优孰劣。比较句的研究对于文本挖掘,特别是情感分析,具有非常重要的价值。比如在对论坛和博客上的产品评论进行研究时,可以利用比较句分析同类产品之间的异同和优劣,从而正确指导用户购买产品。同时,在情感问答系统中存在着一类与比较相关的问题,如问句"宝马和奔驰相比,谁的性价比更高?",回答这个情感问题需要对宝马和奔驰进行比较,而直接对语料库中与宝马和奔驰相关的比较句进行分析,将是最直接最有效的方式。

目前,有关比较句的研究还不是很多。在英语方面,Jindal 等人[1]首先利用支持向量机(SVM)和 CSR(class sequential rules)算法对比较句进行识别,最高取得了 84% 的准确率和 83% 的召回率;然后他们又利用 LSR(label sequential rules)算法对比较元素进行抽取,取得了较好的效果[2]。Feldman 等人[3]利用规则的方法从论坛上抽取产品的名称和属性,并对不同产品进行比较。Sun 等人[4,5]则通过 Web 搜索来寻找两个对象之间的关系。

在汉语方面,比较句的研究主要集中在语言学领域,包括比较的范畴[6]、典型的比较句式[7]、比较的语义[8]等。但是从计算语言学和数据挖掘等领域对汉语比较句的研究较少,北京大学的黄小江等人[9]在 Nitin 研究的基础上,利用 SVM 和 CSR 算法对汉语比较句进行识别,取得了较高的准确率,但是召回率不高,平均只有 70% 左右。

由于汉语的句式更加灵活多样,汉语比较句的研究相对更加困难。比较句的研究包括两个任务:首先需要正确识别比较句;然后抽取比较关系。本文利用条件随机场(CRF)模型,对文献[1,9]中的比较句识别分类算法进行了改进,显著提高了准确率、召回率和 F-度量;然后结合 CRF 算法对比较关系进行分析,取得了较好的效果。

1 汉语比较句介绍

汉语比较句描述了两个或多个对象在某个属性上的位置关系,结合文献[1,9]的定义,一个典型的汉语比较句包含四

收稿日期: 2009-09-19; **修回日期**: 2009-10-26 **基金项目**: 国家自然科学基金资助项目(60773087)

作者简介:黄高辉(1985-),男,福建宁德人,硕士,主要研究方向为自然语言处理、情感倾向分析等(fangao2620@163.com); 姚天 (1957-),男,山西太原人,副教授,博士,主要研究方向为自然语言处理、信息抽取、机器学习等;刘全升(1986-),男,江西赣江人,硕士,主要研究方向为自然语言处理、情感倾向分析等.

个基本比较元素,即比较主体、比较基准、比较点和比较结果。本文把比较主体和比较基准称为比较实体对象,比较点称为比较属性。同时,这个四元组也构成了比较关系,如例句“诺基亚的性能比三星好”,这个比较句可以表示成四元组〈诺基亚,三星,性能,好〉。当然,并不要求一个比较句都具备这四个比较元素。比如,在比较句“诺基亚好于三星”中,比较点是缺省的;而在比较句“诺基亚的质量是最好的”中,比较基准则是缺省的。要抽取比较关系,就是要抽取这四个比较元素。

语义上,根据比较所描述的关系,可以进一步对比较句进行细分。目前,具体的划分方法在语言学界还存在争议,本文采用文献[9]中的划分方法,如表 1 所示。由于差比中的不同并没有直接反映比较主体与比较基准之间的关系,将只研究平比、高下和极比三种比较句类型。

表 1　比较的次范畴

次范畴		语义	例句
平比		被比较对象的属性相同或近似	法拉利和保时捷一样高贵
差比		两者之间有顺序上的差异	法拉利的速度比保时捷更快
	高下	多者之间的极值	保时捷是最时尚的
	极比	两者存在差异,但没有高低之分	法拉利和保时捷的生产商不同
	不同		

在英文中,比较句的特征往往比较明显,如形容词和副词的比较级和最高级形态等。汉语比较句的特征则复杂多样,这也使得准确识别汉语比较句比较困难。参考文献[9]所作的归纳,汉语比较句的特征如下:a)相对程度副词,包括最、极、顶、更、更加、越发、愈发等词,如例句“奥迪车是性价比最高的”;b)比较虚词,包括比、于等词,是比较句的重要标志,如例句“法拉利跑车的流线感好于保时捷”;c)比较实词,包括超过、一样、差不多等词;d)句法模式,主要是一些比较固定的比较句句法模式,如“X 像 Y 一样(R)”等。

2　汉语比较句的识别和关系抽取

2.1　汉语比较句的识别

比较句的识别任务就是对一个任意的句子判断其是否属于比较句,这是一个二分类问题。文献[1,9]均采用了基于特征词的分类算法和基于 CSR 的分类算法,分别对英语和汉语比较句进行研究。所谓 CSR 就是指类序列规则,主要是从句法上抽取一些序列模式作为特征。文献[9]列举了大量的汉语比较句的类序列规则,如“和/p … 相近/a”“和/p …比/v”“ */n… 比/p”“ */a… 于/p”“ */v… 于/p”“相比/v … */n”“ */n 比/p */n”“没有/d */v”等。其中“ *”表示任意满足条件的词。通过定义规则的窗口大小,可以抽取句子中满足条件的序列规则。采用 SVM 算法,分别以特征词和 CSR 序列规则为特征,文献[1]对于英语比较句最高取得了 84% 的准确率和83% 的召回率,文献[9]对于汉语比较句最高取得了 90% 以上的准确率,但是召回率平均只达到 70%。由此可见,不论以特征词为特征还是以 CSR 序列规则为特征,召回率都较低。通过观察语料,发现这是因为在比较句中有一大类特征词和 CSR 序列规则,在非比较句中也很常见,同时由于数据稀疏,从而造成了分类错误。如例句:

(1)诺基亚手机的质量强于三星。

(2)诺基亚的总部位于芬兰。

(3)三星没有诺基亚好用。

(4)三星手机没有抗震功能。

在上述例句中,以特征词为特征,例句(1)和(2)具有特征词“于”,例句(3)和(4)具有特征词“没有”;以 CSR 序列规则为特征,例句(1)和(2)都满足序列规则“ */v… 于/p”,例句(3)和(4)都满足序列规则“没有/d */v”。但是只有例句(1)和(3)是比较句,例句(2)和(4)则是非比较句。若只采用比较词或 CSR 序列规则作为特征,上述例句将很容易被误分类,而这类句子在比较句中占有很大的比例。

一个完整的比较句具有比较主体和比较基准两个比较对象。在比较句中,比较主体和比较基准往往位于比较特征词和 CSR 序列规则的两侧,如例句(1)中,比较主体“诺基亚手机”位于比较词“于”之前,而比较基准“三星”则位于比较词“于”之后;例句(3)中,比较主体“三星”和比较基准“诺基亚”分别位于 CSR 序列规则“没有/d */v”前后。而非比较句则往往不具备这样的特征,它们往往只有一个描述实体对象。因此,本文将在文献[1,9]的研究基础上,以特征词、CSR 序列规则和实体对象信息(主要是对象的位置和数量)作为特征,对比较句进行识别。而实体对象本质上是句子中的主题,具体针对某个领域则是领域主题,如例句(1)中的实体对象诺“基亚手机”和“三星”都是数码产品领域主题。目前,在主题的识别算法中,CRF 算法取得了较好的效果,因此本文将采用 CRF 算法识别实体对象。而且在抽取实体对象的信息时,只需要抽取实体对象的位置和数量,对于主题识别的准确率要求并不高,因此 CRF 算法识别主题的误差对于比较句分类的影响不大。同时,利用 CRF 算法抽取的实体对象,也可以用于后面的比较关系抽取。

2.2　条件随机场模型

条件随机场(CRF)由 Lafferty 等人[10]于 2001 年提出,结合了最大熵模型和隐马尔可夫模型的特点,是一种无向图模型,近年来在分词、词性标注和命名实体识别等序列标注任务中取得了很好的效果。条件随机场是一个典型的判别式模型,其联合概率可以写成若干势函数联乘的形式,其中最常用的是线性链条件随机场。若让 $x=(x_1,x_2,\cdots,x_n)$ 表示被观察的输入数据序列,$y=(y_1,y_2,\cdots,y_n)$ 表示一个状态序列,在给定一个输入序列的情况下,线性链的 CRF 模型定义状态序列的联合条件概率为

$$p(y|x)=\frac{1}{Z(x)}\exp\left\{\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(y_{i-1},y_i,x,i)\right\}$$

$$Z(x)=\sum_{y}\exp\left\{\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(y_{i-1},y_i,x,i)\right\}$$

其中:Z 是以观察序列 x 为条件的概率归一化因子;$f_j(y_{i-1},y_i,x,i)$ 是一个任意的特征函数;λ_j 是每个特征函数的权值。

对于条件随机场模型的训练,一般是利用最大似然法则获得定义特征的权值;而在预测实例中,则采用 Viterbi 算法来寻找在观测序列 x 条件下,具有最大条件概率的状态序列,即 $\arg\max P(y|x)$。

在本文中实体对象本质上属于句子中的主题,特别是领域主题,因此实体对象的识别实际上是一个序列标注问题,可以采用 CRF 模型。在 CRF 模型中,特征的选择至关重要。在这里选取的特征为词(W)、词性(P)、领域词典(D)和停用词词典(S),以及由它们组合而成的复杂特征。特征模板如表 2 所示。

表 2　CRF 模型特征模板

特征模板	原子特征	复合特征
模板 1	W,P,S	WnPnSn,PnPn+1　SnSn+1, WnWn+1　n=-2,-1,0,1
模板 2	W,P,S,D	WnPnSn,PnPn+1　SnSn+1, WnWn+1　DnDn+1　n=-2,-1,0,1

2.3　汉语比较句的关系抽取

在识别出比较句之后,可以进一步研究比较句的关系。根据第 1 章的定义,汉语比较句可以表示为一个四元组〈比较主体,比较基准,比较点,比较结果〉,抽取比较句的关系,就是抽取这四个比较元素。

对于比较主体和比较基准,Jindal 等人在文献[2]中采用 LSR 算法进行抽取,主要是利用一些英语短语的关系来抽取规则。由于汉语中的短语更加复杂,而且汉语分词和词性标注的准确率与英语相比还存在较大差距,这种基于短语规则的方法不能简单应用于汉语比较句。

本文将采用 CRF 算法与规则相结合的方法抽取比较主体和比较基准这两个对象。在比较句的识别过程中,已经利用 CRF 算法对实体对象进行了识别。在此基础上,针对比较句的每个特征词和序列句式定义了一系列规则。例如,在“X 相比/v Y”“X 比/p Y”“X 与/p Y 相比/v”等句式中,X 为比较主体,Y 为比较基准;而在“相比/v X,Y”“相对/p X, Y”等句式中,X 为比较基准,Y 为比较主体。通过定义这些规则,结合 CRF 算法识别出的实体对象,能够很方便地利用比较词与比较对象之间的位置关系抽取出比较主体和比较基准。

由于比较点往往是一些比较对象的属性,如在汽车领域,比较点往往是价格、质量、外观、性价比等属性词,暂时只利用一些领域属性词典进行抽取。

对于比较结果,首先对比较句进一步分类,如表 1 所示,包括“平比”“高下”和“极比”。由于这三类比较句的特征都很明显,将只采用关键字进行分类,然后利用情感词典识别比较句中的情感词,结合前面的分类结果最终得到比较结果。

3　实验

3.1　数据集

目前对于汉语比较句的研究还很少见,没有公开的评测数据集。因此,笔者收集了一些评测数据,并进行人工标注,数据来源于“百度知道”,包括汽车领域和数码产品领域。本文共有两个数据集,第一个数据集用于 CRF 算法的训练和测试,第二个数据集用于比较句的识别和关系抽取。数据集的样本情况如表 3、4 所示。

表 3　数据集一样本情况

数据	汽车领域/句	数码领域/句	总计/句
CRF 训练	700	700	1 400
CRF 测试	200	200	400

表 4　数据集二样本情况

数据	句子	汽车领域/句	数码领域/句	总计/句
训练	比较句	200	200	400
	非比较句	700	700	1 400
测试	比较句	150	150	300
	非比较句	300	300	600

3.2　实验结果

3.2.1　实体对象识别

由于将实体对象相对于比较词的位置和数量作为比较句识别的特征,首先需要采用 CRF 算法识别句子中的实体对象。实验数据采用数据集一,CRF 算法所选取的特征模板见表 2,实验结果如表 5 所示。

表 5　CRF 算法识别实体对象实验结果

模板	准确率/%	召回率/%	F-度量/%	位置准确率/%
1	82.64	90.09	86.20	91.66
2	87.30	94.82	90.90	95.45

在表 5 中,位置准确率是指 CRF 算法所识别出的实体对象与实际实体对象的位置是否一致,如在例句“奔驰的 E 级车性能比宝马 3 系车好”中,前一个实体对象为“奔驰的 E 级车”,如果 CRF 算法所识别出的实体对象为“奔驰”,那么识别结果是错误的,但是识别结果的位置是正确的。实体对象的位置信息将作为比较句识别分类算法的一个特征。从实验结果可知,模板 2 的各项实验评测指标都好于模板 1,这是由于实体对象往往是与领域相关的,而模板 2 增加了领域词典作为特征;同时模板 1 和 2 识别实体对象的位置准确率都达到了 90% 以上。其中模板 2 达到 95.45%,这也保证了将实体对象的信息作为比较句识别分类算法的一个特征时所造成的二次错误可以忽略。

3.2.2　比较句识别

在比较句识别中,本文选取了 SVM 算法作为分类器,在文献[1,9]的研究基础上,增加实体对象信息(位置和数量)作为特征。获取实体对象信息的 CRF 算法采用模板 2 作为特征模板,实验数据采用数据集二,实验结果如表 6 所示。其中,Keyword 表示以比较特征词作为特征,CSR 表示以 CSR 序列规则作为特征,Entity 表示以实体对象的位置和数量信息作为特征。

表 6　比较句识别实验结果

特征词	准确率/%	召回率/%	F-度量/%
Keyword	88.45	67.04	76.27
Keyword,Entity	95.05	87. 5	91.12
CSR	92.05	73.86	81.96
CSR,Entity	94.3	82.95	88.26
Keyword,CSR	93.85	80.68	86.77
Keyword,CSR,Entity	96.55	88.63	92.43

由表 6 可以看出,增加实体对象信息作为特征后,准确率、召回率和 F-度量都明显提高了,尤其是显著提高了召回率,有效克服了文献[1,9]中召回率低的缺点。其中以召回率为例,相比以 Keyword 为特征,召回率提高了 20.46%;相比以 CSR 为特征,召回率提高了 9.09%;相比以 Keyword 和 CSR 为特征,召回率提高了 7.95%。

3.2.3　比较关系抽取

比较关系由四元组〈比较主体,比较基准,比较点,比较结果〉构成。比较主体和比较基准的抽取采用 CRF 算法与规则相结合的方法。其中 CRF 算法采用模板 2 作为特征模板,实验数据采用数据集二,实验结果如表 7 所示。

表 7　比较主体和比较基准抽取结果

比较关系	准确率/%	召回率/%	F-度量/%
比较主体	80	92	85
比较基准	70	80	75

由表 7 可知,比较主体抽取的准确率、召回率和 F-度量分别达到 80%、92% 和 85%;而比较基准抽取的准确率、召回率和 F-度量则只达到 70%、80% 和 75%,均都要低于表 5 所示的各项指标。这是因为比较对象的抽取不仅要识别实体对象,还需要利用规则判断实体对象是属于比较主体还是比较基准,必然造成误差。同时,比较基准的准确率、召回率和 F-度量也均低于比较主体。通过观察语料,发现比较基准更倾向于以缩写或者指代的形式出现。比如例句"诺基亚 N97 的功能比 5630 更加丰富"和"奔驰的性能明显比其他车好"。前一个例句中,比较基准为"5630",是"诺基亚 5630"的简写;后一个例句中,比较基准为"其他车",是一种指代形式,这些因素都将对比较对象的正确识别带来更大的困难。

对于比较结果的抽取,首先需要进一步对比较句进行分类,包括"平比""高下"和"极比"。本文以关键字为特征进行分类,取得了很好的效果。其中平比的关键字如"一样""差不多"等;高下的关键字为"更""超过"等;极比的关键字如"最"等。实验结果如表 8 所示。然后利用情感词典进一步识别情感词,最终得到比较结果。

表 8 比较句类型的分类实验结果

比较句类别	准确率/%	召回率/%
平比	95	89
高下	92	90
极比	96	90

对于比较点的抽取,暂时只是简单采用领域属性词典进行识别,更加深入的工作有待于以后进一步研究。

4 结束语

目前,汉语比较句的研究还是一个新颖的课题。对于汉语比较句的识别,本文在 Jindal 和黄小江等人的研究基础上,采用 SVM 分类器,利用 CRF 算法识别出实体对象,然后增加实体对象的信息作为特征,显著提高了比较句识别的准确率、召回率和 F-度量。对于比较关系的抽取,本文利用 CRF 算法所识别出的实体对象,结合定义的规则,识别比较主体和比较基准,对比较主体的识别取得了较好的效果,对比较基准的识别效果比较一般。这是由于比较基准更倾向于采用缩写和指代的形式,这对比较基准的正确识别带来了困难,也是将来进一步研究的一个方向。对比较点和比较结果的抽取,本文只是简单利用词典进行识别,这也需要将来进行更加深入的研究。另外,在实验数据方面,建立一个更大规模的比较句语料库也是非常重要的。

参考文献:

[1] JINDAL N, LIU Bing. Identifying comparative sentences in text documents[C]//Proc of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:ACM Press,2006:244-251.

[2] JINDAL N, LIU Bing. Mining comparative sentences and relations [C]//Proc of the 21st National Conference on Artificial Intelligence. Boston:AAAI Press,2006:1331-1336.

[3] FELDMAN R, FRESKO M, GOLDENBERG J. Extracting product comparisons from discussion boards[C]//Proc of the 7th IEEE International Conference on Data Mining. Washington DC:IEEE Computer Society,2007:469-474.

[4] SUN Jian-tao, WANG Xuan-hui, SHEN Dou, *et al*. CWS: a comparative Web search system[C]//Proc of the 15th International Conference on World Wide Web. New York:ACM Press:2006:467-476.

[5] LUO Gang, TANG Chun-qiang, TIAN Ying-li. Answering relationship queries on the Web[C]//Proc of the 16th International Conference on World Wide Web. New York:ACM Press,2007:561-570.

[6] 许国萍. 现代汉语差比范畴研究[M]. 上海:学林出版社,2007.

[7] 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.

[8] 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.

[9] 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.

[10] LAFFERTY J D, McCALLUM A, PEREIRA F C N. Conditional random fields:probabilistic models for segmenting and labeling sequence data[C]//Proc of the 18th International Conference on Machine Learning. San Framcisce CA:Morgan Kaufmann Publishers Inc,2001.

(上接第 2060 页)双二次高通滤波器 15 种故障模式的诊断,取得了良好的效果,故障分类的正确率达到了 99%。实验结果表明,与经典的两种支持向量机的故障诊断方法相比,本文方法在简化分类器结构和计算复杂度的同时,也有效地提高了模拟电路故障诊断的正确率,对故障类型较多的故障诊断十分有效。

参考文献:

[1] LITOVSKI V, ANDERJEVIC M, ZWOLINSKI M. Analogue electronic circuit diagnosis based on ANN[J]. Microelectronics and Reliability,2006,46(8):1382-1391.

[2] AMINIAN M, AMINIAN F. Neural-network based analog-circuit fault diagnosis using wavelet transform as processor[J]. IEEE Trans on Circuits and System II: Analog and Digital Signal Processing, 2000,47(2):151-156.

[3] SONG Guo-ming, WANG Hou-jun, LIU Hong, *et al*. Fault diagnosis of analog circuits based on wavelet neural network[C]//Proc of International Symposium on Distributed Computing and Applications to Business, Engineering and Science. 2006:803-807.

[4] 王承,陈光禹,谢永乐. 基于小波—神经网络的模拟电路 IDDT 故障诊断[J]. 仪器仪表学报, 2005,26(11):1106-1108.

[5] 谭阳红,何怡刚. 模拟电路故障诊断的小波方法[J]. 电工技术学报,2005,20(8):89-93.

[6] LONG Bing, HUANG Jian-guo, TIAN Shu-lin. Least squares support vector machine based analog-circuit fault diagnosis using wavelet transform as preprocessor[C]//Proc of International Conference on Communications, Circuits and Systems Proceedings. 2008: 1026-1029.

[7] 孙永奎,陈光禹,李辉. 支持向量机在模拟电路故障诊断中的应用[J]. 电子测量与仪器学报,2008,22(2):72-75.

[8] SIWEK K, OSOWSKI S, MARKIEWICZ T. Support vector machine for fault diagnosis in electrical circuits[C]//Proc of the 7th Nordic Signal Processing Symposium. 2006:342-345.

[9] MARCELLIN S, ZIGHED D A, RITSCHARD G. Evaluating decision trees grown with asymmetric entropies[C]//Proc of the 17th International Conference on Foundations of Intelligent Systems. Berlin, Herdelberg:Springer-Verlag,2008:58-67.