文章编号: 1000-8608(2007)03-0433-06

基于支持向量机的中文文本中地名识别

李 丽 双*, 黄 德 根, 陈 春 荣, 杨 元 生

(大连理工大学 计算机科学与工程系, 辽宁 大连 116024)

摘要: 提出并实现了一种基于支持向量机(SVM)的中文文本中地名的自动识别方法.结合地名的特点,抽取单字本身、基于字的词性、是否在地名特征词表中及其上下文的信息作为向量的特性,并将其转化为二进制表示,在此基础上建立了训练集,并通过对多项式 Kernel函数的测试,得到了用支持向量机进行地名识别的机器学习模型.实验表明,所建立的 SVM地名识别模型是有效的,系统开式召回率和精确率分别达 86.69% 和 93.82%, F-值为 90.12%.

关键词: 支持向量机; 中文文本; 地名识别; 机器学习
中图分类号: TP301.6 文献标识码: A

0 引 言

汉语的自动分词是汉语信息处理领域中的"瓶颈"问题,是自动文摘、文本挖掘、语音处理和机器翻译等语言工程中的基础课题之一,而专有名词的自动识别是提高汉语分词系统正确率的关键技术.中文地名识别是专有名词识别的难点之一,主要表现为(1)中文地名数量大,没有明确规范的地名定义,并且新的地名不断涌现;(2)地名长度没有一定的限制,不像中文姓名那样,长度在 2~4个汉字;(3)地名内部相互成词,这些多字词已被核心词典收录;(4)切分错误,如"兴/城市";(5)地名结尾带有特征词,但地名特征词出现情况比较复杂,既可作为普通用词出现,又可出现在地名结尾甚至出现在地名前部.目前中文地名识别的方法主要是采用统计模型,以及利用统计与规则相结合的方法[1-3],效果均较好.文献[1]采用统计模型,利用属性矩阵和频级筛选,达到了较高的召回率,但精确率偏低;文献[2]采用基于语料库的方法,根据地名词典统计分析地名用字信息以及这些字在真实文本中使用程度信息进行地名初筛选,再结合上下文信息的规则来确定地名;文献[3]利用统计与规则相结合的方法对含特征词的中文地名进行识别,较好地解决了召回率和精确率之间的关系;文献[4]采用基于变换的错误驱动的机器学习方法,有效地提高了系统的召回率和精确率.总之,目前的研究基本上都是采取规则与统计相结合的方法,不同之处仅在于规则与统计的侧重不同.

本文提出基于支持向量机[5](SVM, support vector machine)的中国地名自动识别方法.SVM是 Vapnik等在统计学习理论的基础上发展起来的一种新的通用学习方法,它已表现出很多优于已有机器学习方法的性能.在自然语言处理领域,SVM应用于文本分类[6]、日语实体名词识别[7]、未登录词的识别[8]等,都取得了较好的效果.本文结合中文文本中地名的特点,对训练语料中的每个字进行分类标注及词性标注,然后抽取单字本身、词性、该字是否在地名特征词表中及其上下文的信息作为特征向量的属性,建立训练集,并通过对不同阶数多项式 Kernel函数的测

收稿日期: 2005-05-20; 修回日期: 2006-03-17.
基金项目: 国家自然科学基金资助项目(60373095; 60373096).
作者简介: 李丽双*(1967-),女,副教授, E-mail: lils@dlut.edu.cn; 黄德根(1965-),男,教授; 杨元生(1946-),男,教授,博士生导师.

试,得到用支持向量机进行地名识别的机器学习模型.

1 支持向量机 SVM

1.1 最优分类超平面

设原始输入空间 $\boldsymbol{X}\subseteq \mathbf{R}^n$ (n为输入空间的维数),训练集 $S=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_l,y_l)\}$, $\boldsymbol{x}_i\in \boldsymbol{X}$, $y_i\in\{-1,1\}$是 $\boldsymbol{x}_i$的标记,若 $\boldsymbol{x}_i$属于正类, $y_i=1$,若 $\boldsymbol{x}_i$属于负类, $y_i=-1$, l为样本的个数. SVM就是寻找能够将训练数据划分为两类的最优超平面,该超平面可以通过求下面凸二次规划问题的解得到[5]:

$$\max \sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)$$
$$\text{s.t.}\ \sum_{i=1}^{l}y_i\alpha_i=0,0\leqslant\alpha_i\leqslant c,i=1,2,\cdots,l \tag{1}$$

其中 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\Phi(\boldsymbol{x}_i)\cdot\Phi(\boldsymbol{x}_j)$,为 Kernel函数,其满足 Mercer条件[5], $\Phi(\boldsymbol{x})$为原始输入空间到高维特征空间的非线性映射; α_i为与每个样本对应的 Lagrange乘子; $c>0$是自定义的惩罚系数. 给定一个测试实例 $\boldsymbol{x}$,它的类别由下面的决策函数决定:

$$f(\boldsymbol{x})=\operatorname{sgn}\left[\sum_{\boldsymbol{x}_i\in \boldsymbol{sv}}\alpha_i y_i K(\boldsymbol{x}_i\cdot\boldsymbol{x})+b\right] \tag{2}$$

其中 $\boldsymbol{sv}$为支持向量, b是分类阈值,可用任一支持向量或通过两类中任一对支持向量取中值求得.

1.2 多类划分

SVM本身是解决两类分类问题的,对于多类(k类)划分问题可将其转化为两类划分问题加以处理,目前主要有两种方法: (1) pairwise方法[9],在任意两个类别之间构造一个二值分类器,从而生成 $k(k-1)/2$个二值分类器,每个分类器训练两种不同类别的数据,在分类中使用投票策略,即对于一个未知样本每个分类器都有一个选票,其结果是选票最多的类别. (2) one vs. others方法[9],构造 k个分类器,第 i个分类器的训练数据是第 i类的数据作为正例,其他类的数据作为负例,为每个类构造一个分类器,第 i个分类器在第 i类和其他类之间构造一个超平面,在多个两类分类器中具有最大输出的类别即是测试数据所属的类别.

2 用支持向量机识别中文文本中的地名

首先对训练语料进行自动分词和词性标注,再按字抽取特性建立训练集,通过选取合适的 Kernel函数,建立用 SVM识别地名的机器学习模型.

2.1 BIO分类标记

本文采用 IBO2的组块(chunk)表达方法来标识地名,即将每个字分为三类: B—— 地名首字、I—— 地名中部、O—— 地名外部,这里一个组块(BI或 B)视为一个地名. 对训练文本中的每个字进行 IBO2标注,即 $y_i\in\{B,I,O\}$,这样,用 SVM识别中文文本中的地名就是对文本中的每个字进行 B I O分类,如图 1所示.

	i					
位置	-2	-1	0	+1	+2	+3
单字	在	重	庆	市	举	行
词性	p-S	v-S	vg-S	n-S	v-B	v-E
是否为特征词	N	N	N	Y	N	N
IBO2标注	O	B	I	I	O	O

图 1 地名特性抽取实例

Fig. 1 An example of features' extraction

2.2 根据地名特点抽取向量特性

基于 SVM的中文地名识别,关键在于抽取地名的合适特性. 通过对中文文本中地名的特点进行分析得到地名向量的特征.

由于地名识别是对自动分词结果进行的,分词错误可能会影响地名的正确识别,如:

兴 /城市 /原 /种 /场 /种子公司 /还 /拖欠 /

为了解决分词错误导致地名的错误识别,这里,必须将每一个词分解为一个一个的字,按字抽取特性,最后对每一个字进行分类识别. 表 1给出了所选取的特性类型及相应值.

表 1 特性的类型及相应值

Tab. 1 Summary of features and their values

特性类型	值
单字	字本身
单字词性	n-B, v-I, p-S,…
是否在特征词表中	Y or N
前字的 IBO2标记	B, I, O

(1)“单字” 特性指该字本身. 中文地名一方面比较自由分散,地名录中共享汉字 3 685个,同时中文地名用词又有相对集中的覆盖能力,出现次数在 [1, 10]的地名用字占绝大多数,出现频次最多的前 100个占地名总数的 50% 左右,前 600个占地名总数的 90% 左右,前 2 000个占地名总数的 99% 左右. 由以上分析,地名用字可以充分反映地名的特点,由此将“单字本身” 作为地名的特性.

(2)“基于字的词性” 为该字所属词的词性加上其位置属性,标注方法见表 2. 例如: 若一个词包含三个字,第一、二、三个字的词性标注分别为词性 -B 词性 -I 词性 -E,单字词的词性标注为词性 -S. 其中“词性” 为该词 (多字词或单字词) 的词性,这里采用北大词性标注规范.

表 2　基于字的词性标注方法

Tab. 2　POS tags in a word

词性标注	字类型
词性 -S	单字词
词性 -B	多字词首字
词性 -I	多字词 (至少三字词) 中间字
词性 -E	多字词尾字

(3) 因为地名结尾经常有地名特征词出现,所以“是否在特征词表中” 是地名的一个很重要特性,如果该字为地名特征词,如“省”、“市” 等,则该特性的值为 Y,否则为 N.

(4)“前字的 IBO2标记” 是一种动态特性,一个地名往往由若干个字构成,如“重庆市”,若已识别出“重” 字在地名中,则“庆” 字很可能也在地名中,所以“重” 的 IBO2标记作为“庆” 的特性. 取前两个字的 IBO2标记作为当前字的特征.

(5) 一个字是否在地名中还依赖于该字的上下文信息,如“在重庆市” 中的“在” 是判别“重庆市” 是否为地名的很重要的上下文信息. 这里选定前后各两个字作为当前字的上下文信息窗口.

这样,每个样本的特性可用 3× 5+ 2= 17个三元组来表示,每个三元组定义为 (位置,特性类型,特性值). 图 1给出了特性抽取的实例. 第 i 个字“庆” 的特性可表示为下列三元组的集合: {(- 2,单字,在), (- 1,单字,重), (0,单字,庆), (+ 1,单字,市), (+ 2,单字,举), (- 2,词性,p-S), (- 1, 词性,v-S), (0, 词性,vg-S), (+ 1,词性, n-S), (+ 2,词性,v-B), (- 2,特征词,N), (- 1, 特征词, N), (0, 特征词, N), (+ 1, 特征词,Y), (+ 2, 特征词, N), (- 2, IBO2标注,O), (- 1, IBO2标注, B)}. 设训练集中有 n 种三元组 c_k ($k \leqslant n$),则第 i 个字的特性向量表示为 $\boldsymbol{x}_i = (b(c_1) \cdots b(c_k) \cdots b(c_n))$,若第 i 个字的上下文信息包含 c_k,则 $b(c_k) = 1$,否则 $b(c_k) = 0$.

综上, $(\boldsymbol{x}_i, y_i)$ 构成了一个训练实例.

学习时,单字“庆” 的特性为被框架包围的全部特性. 若同样的句子作为测试数据,则单字“庆” 的特性与学习时候一样,用框架内的全部特性.

2.3　Kernel函数的选取

常用的 Kernel函数有:

(1) 多项式 Kernel函数

$$K(\boldsymbol{x}, \boldsymbol{x}_i) = [(\boldsymbol{x} \cdot \boldsymbol{x}_i) + 1]^d, d \text{是自然数} \quad (3)$$

(2) 径向基 Kernel函数

$$K(\boldsymbol{x}, \boldsymbol{x}_i) = \exp\left\{-\frac{|\boldsymbol{x} - \boldsymbol{x}_i|^2}{\sigma^2}\right\}, \sigma > 0 \quad (4)$$

(3) Sigmoid Kernel函数

$\tanh(a(\boldsymbol{x} \cdot \boldsymbol{x}_i) + t)$, a, t 是常数,

tanh是 Sigmoid函数　　(5)

考虑到通过选取适当的参数, 3种类型 SVM的性能大致相同[10],而多项式函数形式简单且可以直观地比较各种特性不同组合时的分类效果,本文采用 d 次多项式作为 Kernel函数.

2.4　训练及测试

(1) 对训练语料及测试语料进行细标注 (基于字的标注)

① 对人工标注好的训练语料重新标注: 去掉人工标注结果还原到原始文本,并记录地名标注位置,然后用 NIHAO自动分词和标注系统 (不包含未登录词识别系统) 进行自动分词和词性标注 (未经人工修改),并进行基于字的词性标注 (词性 -S B I E),再根据记录地名的位置对语料中的每个字进行 IBO自动标注;

② 对测试语料同样进行自动分词和基于字的词性标注.

(2) 建立训练集和测试集

根据 2.2所述,对训练语料和测试语料抽取特性后转化为二进制向量表示,建立训练集和测试集.

(3) 用 SVM 算法进行训练及测试

选取多项式函数作为 Kernel函数,然后利用 SVM 算法进行训练,再分别用两种多类划分算法(pairwise和 one vs. others)判别每个字是 B I或 O.

3 实验结果及分析

为了评测地名识别效果,本文选取了与文献[3]相同的语料,即从 1998年《人民日报》上随机选取了 150万字的语料,对其进行人工标注后生成学习语料,其中含有 31 416个地名(包含重复地名);然后从 2000年《人民日报》上选取了含有 2 919个地名的句子作为测试语料进行了开放测试.

3.1 对不同阶数多项式的实验

首先用 pairwise多类划分算法对多项式不同阶数(d = 1, 2, 3)分别进行了实验,结果如表 3所示.

表 3 多项式 Kernel函数不同阶数的实验结果

Tab. 3 Results with different numbers of degree of polynomial Kernel function

阶数	召回率 /%	精确率 /%	F-值 /%
d = 1	84.66	91.95	88.16
d = 2	86.69	93.82	90.12
d = 3	86.27	94.23	90.07

$$\text{识别地名召回率} = \frac{\text{正确识别出地名数}}{\text{文本中地名总数}} \times 100\%$$

$$\text{识别地名精确率} = \frac{\text{正确识别出地名数}}{\text{所识别出地名总数}} \times 100\%$$

$$F\text{-值} = \frac{2 \times \text{召回率} \times \text{精确率}}{\text{召回率} + \text{精确率}} \times 100\%$$

从表 3可以看出,d = 2 d = 3时的测试结果明显好于 d = 1的结果,这说明进行地名识别时考虑各个特性的组合是十分必要的.

3.2 对不同多类划分算法的实验

本文分别用 pairwise和 one vs. others算法对多项式 Kernel函数(d = 2)进行了测试,如表 4所示. 结果表明: pairwise算法结果略好于 one vs. others算法.

表 4 不同多类划分算法的实验结果(d = 2)

Tab. 4 Results with different multi-class classifiers (d = 2)

方法	召回率 /%	精确率 /%	F-值 /%
pairwise	86.69	93.82	90.12
one vs. others	86.26	93.79	89.87

3.3 训练集规模对识别结果的影响

从图 2可以看出,当训练集很小时,识别效果较差,随着训练集规模的增加,召回率和精确率逐渐增加,F-值也增加,当训练语料(15万字)中地名数为 10 000时,召回率、精确率和 F-值都较高,并且此后随着语料规模的增大,F-值增大的幅度很小. 这说明采用本文所选取的特征,当训练集规模达到一定程度后用较少的训练样本就可以得到较好的识别效果,而采用概率统计的方法一般需要很多的训练语料,如文献[3]需要 200万字学习语料.

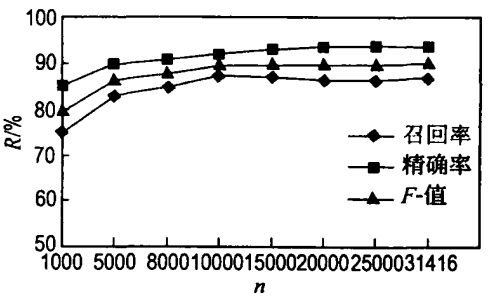

图 2 训练集规模对识别结果的影响

Fig. 2 Influence of different sizes of training sets on results

3.4 结果分析

通过对实验结果的分析,发现地名识别的错误主要如下.

(1) 训练语料规模较小或地名现象不明显: 很多地名实例只出现一次甚至没有出现,如"淳化是全国 /有 /名 /的 /苹果 /基地 /县 /"中的地名"淳化"未在训练语料中出现过,又不含特征词,所以未能召回;同一个地名,在有的上下文中能够正确识别,而在另外的上下文中却不能识别,这说明后者上下文信息在训练集中出现次数较

少. 以上类型的错误可通过优化训练集而得到改善.

(2) 单字词与地名相连后又具有地名的其他特征而导致该单字被作为地名首字错误识别: 如"/如新乡/某/安居/小区/一/住宅楼/"中的"如"字是地名前的单字,"如新乡"位于句首,训练样本中地名位于句首情况很多,又含有特征词,所以"如新乡"被作为地名错误识别.

(3) 人名中含有地名特征词时被当做地名错误召回: 如"/河北省/王贺亭/李/秀/花/一家/"中的"王贺亭",这可以利用规则的方法把错误剔除,若地名前面出现姓用字,并且地名后面出现称谓词的话,则可以认为是姓名. 另外,由于名词实体识别是先识别人名后识别地名,这类错误也可以通过提高人名识别系统的能力而排除.

3.5　与文献 [3]的比较

文献 [3]使用从大规模地名词典和真实文本语料库得到的统计信息以及针对地名特点总结出来的规则,通过计算地名的构词可信度和接续可信度对含特征词的中文地名进行识别. 为了衡量本文方法的识别效果,分别对仅含特征词和所有地名两种情况进行测试,表 5中列出了 $d = 2$时的开式测试结果与文献 [3]的比较.

表 5　SVM 方法与文献 [3]的比较

Tab. 5　Comparison of SVM method with Lit. [3]

方法	召回率 /%	精确率 /%	F-值 /%
文献 [3](含特征词的地名)	86.86	91.48	89.11
SVM(含特征词的地名)	88.93	94.06	91.42
SVM(所有地名)	86.69	93.82	90.12

从表 5可以看出,本文采用的 SVM方法较文献 [3]各项指标均有所提高. 文献 [3]只对含特征词的中文地名进行识别,而本文可以对中文文本中的所有地名(包含外国地名)进行识别,地名识别范围扩大,具有更广泛的适用性,如: /东榆林/已/是/城区/的/小康村/,/阿根廷/首都/布宜诺斯艾利斯/的/港口/,/甘孜/州/阿坝/州/凉山/等/少数民族/聚居/地区/,等等,文献 [3]不能识别,但用 SVM方法可以识别; 从测试的结果来看,即使对于后一种情况(识别所有地名),本文的方法在总体上依然显示出很好的识别效果: 与文献 [3]相比,召回率与之基本相当(仅下降 0.17%),精确率和 F-值都明显提高(分别提高 2.34% 和 1.01%). 很多文献 [3]不能识别或错误识别的情况,本文可以正确识别,如: "阎子庆同志是陕西省延长县人","图为兼庄村群众在投票点参加评议"中的"延长县"、"兼庄村"都可正确识别; 又如"调整产业结构以工兴镇","一两条沟里湾里的小鱼儿就足以让它快乐",其中的"以工兴镇"、"沟里湾"等未被错误召回. 可见,本文所采用的方法,不仅能更好地识别含特征词的地名,而且对外国地名以及不含特征词的中国地名也能较好地识别.

在所用资源上,文献 [3]计算地名构词可信度和接续可信度时需要建立大规模的地名字典及地名前部词表、特征词表,并且对 200万字的语料进行人工标注后建立单词频度词典和双词接续词典,而本文用到的资源只有地名特征词表和 150万字的学习语料; 从识别过程看,文献 [3]加入了很多规则进行后处理,而本文的结果未进行任何后处理.

4　结　语

本文针对中文文本中地名的特点抽取单字本身、词性、是否为地名特征词及其上下文的信息等特征用 SVM进行地名识别. 实验结果表明本方法选取较少的样本就可以得到较高的 F-值. 这说明所选取的特征充分反映了地名的特点并且充分利用了上下文信息.

文中向量的特性只选取了 3种及左右各两个字的上下文信息就能够得到较高的召回率和精确率,根据 SVM的学习误差不依赖于特性空间维数的特性,若增加地名的其他有用信息(如地名用字信息)、扩大上下文窗口并与一定的规则相结合,有望使系统的召回率和精确率得到进一步提高. 另外,由于训练的效果不在于训练集内样本的多寡,而在于集内的句子是否反映了各种典型的地名现象,可以通过优化训练样本而改善地名识别的正确率.

参考文献:

[1]沈达阳,孙茂松,黄昌宁. 中文地名的自动辨识 [C]//

计算语言学进展与应用. 北京: 清华大学出版社, 1995

[2]谭红叶,郑家恒,刘开瑛. 中国地名自动识别系统的设计与实现[J]. 计算机工程, 2002(8): 128–129

[3]黄德根,岳广玲,杨元生. 基于统计的中文地名识别[J]. 中文信息学报, 2003(2): 36–41

[4] TAN Hong-ye, ZHENG Jia-heng, LIU Kai-ying. Research on method of automatic recognition of Chinese place names based on transformation [J]. **J Software**, 2001, **12**(11): 1608–1631

[5] VAPNIK V N. **Statistical Learning Theory** [M]. New York: John Wiley & Sons, 1998

[6] JOACHIMS T. Text categorization with support vector machines: learning with many relevant features [C]// **Proceedings of the 10th European Conference on Machine Learning**. **Lecture Notes in Computer Science**. Heidelberg: Springer-Verlag, 1998

[7] YAMADA H, KUDO T, MATSUMOTO Y. Japanese named entity extraction using support vector machine [J]. **Trans of IPSJ**, 2002, **43**(1): 44–53

[8] GOH C L, ASHARA M, MATSUMOTO Y. Chinese unknown word identification using character-based tagging and chunking [C]// **Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics**. Sapporo: Interactive Poster/Demo Sessions, 2003: 33-54

[9] HSU C W, LIN C J. A comparison of methods for multi-class support vector machines [J]. **IEEE Trans on Neural Networks**, 2002, **13**(2): 415–425

Identification of location names from Chinese texts based on support vector machine

LI Li-shuang*, HUANG De-gen, CHEN Chun-rong, YANG Yuan-sheng

(Dept. of Comput. Sci. and Eng., Dalian Univ. of Technol., Dalian 116024, China)

Abstract Based on the characteristics of location names in Chinese texts, a method of automatic identification of Chinese location names using support vector machine (SVM) is proposed. The character itself, character-based part of speech (POS) tag, the information whether a character appears in a location name characteristic word table and context information are extracted as the features of the vectors. Each sample is represented by a long binary vector, and thus a training set is established. The machine learning models of automatic identification of location names are obtained by testing polynomial kernel functions. The results show that the models are efficient in identifying location names from Chinese texts. The recall, precision and F-measure are up to 86.69%, 93.82% and 90.12% respectively in open test.

Key words support vector machine; Chinese texts; identification of location names; machine learning