文章编号:1001 - 9081(2004)01 - 0014 - 03

# 基于最大熵方法的汉语词性标注

林　红[1] ,苑春法[2] ,郭树军[1]
(1. 河北省气象局 省气象台 ,河北 石家庄 050021;　2. 清华大学 计算机科学与技术系 ,北京 100084)
(linhong78426 @sina. com)

**摘　要**:最大熵模型的应用研究在自然语言处理领域中受到关注 ,文中利用语料库中词性标注的上下文信息建立基于最大熵方法的汉语词性系统。研究的重点在于其特征的选取 ,因为汉语不同于其它语言 ,有其特殊性 ,所以特征的选取上与英语有差别。实验结果证明该模型是有效的 ,词性标注正确率达到 97. 34 %。

**关键词**:语言模型;最大熵模型;词性标注
**中图分类号**: TP182;TP391. 1　　**文献标识码**:A

# A Chinese Part of Speech Tagging Method Based on Maximum Entropy Principle

LIN Hong[1] ,YUAN Chun-fa[2] ,GUO Shu-jun
(1. *Hebei Meteorological Observatory , Hebei Meteorological Bureau , Shijiazhuang Hebei* 050021 , *China;*
2. *Department of Computer Science and Technology , Tsinghua University , Beijing* 100084 , *China)*

**Abstract**: A lot of researches have been made on the application of the maximum entropy modeling in the natural language processing during recent years. This paper presents a new Chinese part of speech tagging method based on maximum entropy principle because Chinese is quite different from many other languages. The feature selection is the key point in this system which is distinct from the one used in English. Experiment results have shown that the part of speech tagging accuracy ratio of this system is up to 97. 34 %.

**Key words**: language model; maximum entropy; part of speech tagging



## 1　引言

目前汉语的词性标注基本上采用二元语法模型[1]和三元语法的隐马尔可夫模型 ,它们虽然也都有较好的标注效果 ,但由于预测信息的不足 ,对词性标注 ,特别是未登录词的词性标注精度影响很大 ,在上述模型中一般对未登录词的词性采用猜测(如猜测为名词)的方法。而最大熵模型使用特征的形式 ,有效的利用了上下文信息 ,在一定的约束条件下可以得到与训练数据一致的概率分布 ,即使是未登录词 ,由于其丰富的上下文信息 ,对它的词性标注也起到了很好的预测作用。实验结果证明最大熵方法取得了比上述模型较好的标注效果。

## 2　最大熵原理

假设 $p(x)$ 是在训练集$\{x_1, x_2, \cdots, x_n\}$上分布 ,通常的想法就是要求 $p(x)$ 尽量的和数据集的先验分布 $\tilde{p}(x)$ 相吻合, $\tilde{p}(x)$ 等于随机事件的出现概率。但是如果一味的追求使两者的值相同,会出现训练过适应的问题,以二元语言(bigram)模型 $x = (w_1, w_2)$ 为例,在训练集很可能是 $\tilde{p}(languge, paper) = 0$,但是不排除它在实际情况中可能出现,所以在数据分布的先验的限制和条件中,只要求求得的分布与我们认为重要和可靠的已知条件相符合,不对为未知的数据做任何的可能的先验假设引入了最大的不确定性,这也是最大熵模型成功的因素之一。

### 2. 1　问题描述

设随机过程所有的输出值构成集合 $T$(在本文中 $T$ 为词性标记集合),而已知语料库中的所有上下文信息为集合 $X$,对于一个词来讲,由于其输出 $t \in T$ 受其在文中上下文信息的影响。构造随机模型的任务是精确的描述该随机过程的特性。对于词性标记而言,其中 $x$ 为与待标记词有关的上下文信息的组合, $t$ 为该词可能出现的词性标记。对于给定的 $x \in X$,精确估计输出为 $t \in T$的条件概率,即对 $P(t \mid x)$ 进行精确估计。

### 2. 2　经验概率分布

经过人工标注和校对的训练语料库中有大量的词性标注实例(词的标记,词的上下文信息),即 $(t, x)$ 样本。经过统计可以得到训练语料中在特定的上下文中一个词出现某一词性的频率,从而得到训练语料中上下文信息与输出的经验概率分布:

$$p(x, t) = \frac{freq(x, t)}{\sum_{x \in X, t \in T} freq(x, t)}$$

### 2. 3　特征与约束

随机过程的输出受上下文信息的影响。如在词性标注系

**收稿日期**:2003 - 07 - 16;**修订日期**:2003 - 09 - 23
**基金项目**:国家自然科学基金资助项目(69975008);国家 973 规划资助项目(G1998030507)
**作者简介**:林红(1966 - ),女,福建仙游人,工程师,主要研究方向:中文信息处理、数据分析; 苑春法(1946 - ),男,河北曲阳人,教授,主要研究方向:中文信息处理、信息抽取; 郭树军(1968 - ),男,河北邯郸人,副高级工程师,主要研究方向:数据分析和挖掘.

统中,文本中待标记词可能出现的词性标记与其上下文有关,而上下文信息可以用特征来表示,下面举例来说明它们的定义。

如有这样一句话:我/rn 想/va 再/d 跟/pg 你/rn 聊/vgi 一/d 聊/vgi 。/。假设考虑第四个词"跟",它的词性标记为:"pg",而其上下文信息有:

(*a*) 中心词为:"跟",表示为:$w_4$ = "跟"

(*b*) 前一个词为:"再",表示为:$w_3$ = "再"

(*c*) 前一个词的词性为:"*d*",表示为:$T_3$ = "*d*"

(*d*) 下一个词为:"你",表示为:$w_5$ = "你"

还有许多对预测"跟"的词性标记有用的上下文信息,在此不再一一列举,而选取哪些上下文信息作为预测信息属于特征选择的范围,在第四部分再做详细分析。以上(*a*) ~ (*d*)所示的上下文信息对"跟"的词性标记为"pg"均有预测作用,为了建立一个预测模型,可以引入了特征函数来表述数据集中$(x,t)$的特性,它被定义为$\{0,1\}$域上的二值函数,例如可用如下所示的二值函数来表示:

$$f(x,t) = \begin{cases} 1(\text{如果 } w = \text{我 } \& \ t = pron) \\ 0(\text{其它}) \end{cases}$$

进一步可以引入一系列的特征函数$\vec{f} = \{f_1(x),\cdots,f_n(x)\}$来表示训练数据集的限制,并在此基础上通过对特征函数的期望值施加一定的约束来表述存在在原数据集中的上下文依赖关系,即要求特征函数在$p(x)$分布上的期望值和在先验模型$\widehat{p(x)}$上的相同。

$$\sum_{t\in T,x\in X} p(t,x)f_i(x) = \sum_{t\in T,x\in X} \hat{p}(t,x)f_i(x), i = 1,2,\cdots,N$$

### 2.4 最大熵原理

假设存在 $n$ 个特征,所求的模型是在满足约束集合$C \equiv \{p \in P \mid p(f_i) = p(\tilde{f}_i)\}, i \in \{1,2,\cdots,n\}$的条件下生成的模型,而满足约束集合的模型不只一个,需要的是具有最均匀分布的概率模型,而什么样的模型才是最均匀的呢?熵的概念为我们提供了解答,熵可以作为条件概率$P(t \mid x)$是否均匀的一种数学测量方法,而熵可表示为:

$$H(p) = -\sum_{a,b} \tilde{p}(t \mid x)\log p(t \mid x)$$

由此可以得到最大熵原理:在满足约束条件集合的前提下,具有使$H(p)$值最大的模型即为具有最均匀分布的模型,也就是我们需要的模型。即:

$$p^* = \arg\max_{p\in P} H(p)$$

## 3 算法描述

### 3.1 训练模型

假设满足最大熵条件的概率$p(t \mid x)$具有 Gibbs 分布:

$$p(t \mid x) = \frac{1}{Z_\lambda(x)} \exp\sum_{i=1}^{n} \lambda_i f_i(t,x) \qquad (1)$$

其中:

$$Z_\lambda(x) = \sum_t \exp\sum_{i=1}^{n} \lambda_i f_i(t,x)$$

每个特征函数对应一个权重值$\lambda_i$。

设有 $n$ 个特征函数,每个特征函数对应一个权重值$\lambda_i$,我们的目的是在满足约束集合的模型集合内,求出其中熵最大模型的一组$\lambda_i$值,过程如下:

(1) 从训练语料中经过统计得到$\tilde{p}_{t\in T,x\in X}(t,x)$,并计算出:

$$\tilde{E}(f_i) = \sum_{t,x} \tilde{p}(t,x)f_i(t,x)$$

(2) 从训练语料中得到每个词的上下文信息及其当前的词性标记,每一条记为一个 *event*,然后合并相同的 *events*;

(3) 通过对 *events* 的统计得到建立模型所需的特征,并增加"*correct*" 作为校正信息,来满足算法对特征函数的限制条件。

(4) 对于每个 *event*,计算其特征个数,设为 *num_feature*,如果 *num_feature* 为常数,则 $C$ = *num_feature*,否则 $C = \max(num\_feature)$。

同时增加$k = C - num_feature$个校正"*correct*",以便使每个 *event* 的 $C$ 均为常数。

(5) 对于所有的$i \in (1,2,\cdots,n)$,设置$\lambda_i = 0$;

(6) 对于每个$i \in (1,2,\cdots,n)$:

(*a*) 设 $m$ 为迭代次数,计算:

$$p^{(m)}(t \mid x) = \frac{1}{Z_\lambda(x)} \exp\sum_{i=1}^{n} \lambda_i^{(m)} f_i(t,x)$$

(*b*) 计算:

$$E^{(m)} f_i = \sum_{t,x} p^{(m)}(t \mid x)f_i(t,x)$$

(*c*) 计算:

$$\Delta\lambda_i = \frac{1}{C}\log\frac{\tilde{E}(f_i)}{E^{(m)}(f_i)}$$

(*d*) 更新$\lambda_i = \lambda_i + \Delta\lambda_i$。

重复(6) 直到$\lambda_i$收敛或循环若干次,如 $m$ = 100 次。

在本文中,对于每个 *event*,当$f^*(t,x)$不为常数时,都增加 $k$ 个校正信息"*correct*",使得每个 *event* 都有相同的特征个数,而不是对所有的 *events* 统一增加一个校正特征,这样在迭代时就不存在计算校正参数与其它参数不一致的问题了,这使得在模型训练时,训练时间减少了一半(本文第四部分的模型训练时间不到 2 个小时)。同时校正信息与可能的标记形成了不只一个的校正特征函数,能更精确的进行预测,这使得在对文本进行词性标记时正确率较增加一个校正参数形成的模型有所提高。

### 3.2 测试模型

给定一句需要词性标注的句子$(w_1,w_2,\cdots,w_n)$,利用(1)式可以得到一个词可能出现的词性的概率,而这一句话的词性标注$(t_1,t_2,\cdots,t_n)$的条件概率可表示为:

$$p(t_1\cdots t_n \mid w_1\cdots w_n) = \prod_{i=1}^{n} p(t_i \mid x_i)$$

测试模型的目的是要求出使上式的值最大的一组词性标注,在本系统中是用动态规划法实现的。

## 4 实验结果及分析

在实验中采用从证券时报网上下载的经过自动标注及人工校对的金融新闻语料库作为训练语料,训练语料中总词量为 190 万汉字,出现的词的个数为 110 万词次。标记集采用清华大学制定的有 102 个标记的标记集,在标记集规模对词性标注的影响一节中对标记集进行了合并。

4.1 模板建立

为了取得一个有效的特征模板,在实验中设计了如下两个特征模板,并分别对它们进行了训练和测试。最大熵模型是在 P4/ 主频 1.6G/ 内存 256M 的微机上生成的。

(1) 模板 1:取中心词、中心词的前两个词及它们的词性和下两个词作为特征,当前词为生词时统一取"$w$ = ",其余与如表 1 相同。

(2) 模板 2:当出现的词为生词时,利用最大匹配的方法找出词尾,如词尾在后缀表中,则建立候选特征为:$h$ = 词尾,否则建立候选特征为:"$w$ = ",模板如表 1。

表 1

| 条 件 | 特 征 |
|---|---|
| $W_i$ 不是生词 | $W = w_i$ &$t = t_I$ $h$ = 最后一个词尾 &$t$ = $t_I$ |
| $W_i$ 是生词 | 或 $w$ = &$t = t_I$ |
| 对于所有 $w_I$ | $t_{i-1} = X$ &$t = t_i$ |
| | $t_{i-2}t_{I-1} = XY$ &$t = ti$ |
| | $w_{I-1} = X$ &$t = t_i$ |
| | $w_{I-2} = X$ &$t = t_i$ |
| | $w_{I+1} = X$ &$t = t_i$ |
| | $w_{I+2} = X$ &$t = t_i$ |

根据训练语料的大小和反复试验,确定了若一个词在语料库中出现 3 次以下,则认为是生词;若特征出现次数为 3 次以下,对模型的预测是不可靠的,去掉。

4.2 模型建立及测试

利用上述 2 个模板,将语料库分为两部分,其中 9/ 10 作为训练语料,1/ 10 作为开放测试语料,分别生成了 2 个模型,并对这 2 个模型进行了封闭测试和开放测试,结果如表 2。

表 2

| | 模板 1 | 模板 2 |
|---|---|---|
| 封闭测试 | 97.10 % | 97.34 % |
| 开放测试 | 90.76 % | 90.92 % |

4.3 结果分析

(1) 根据前人的经验,首先选取了模板 1 作为特征模板,取得了较好的标注效果。

表 3

| | 模板 1 | | 模板 2 | |
|---|---|---|---|---|
| | 熟词 | 生词 | 熟词 | 生词 |
| 封闭测试 | 97.38 % | 86.35 % | 97.43 % | 93.88 % |
| 开放测试 | 91.33 % | 65.46 % | 91.37 % | 70.88 % |

(2) 后缀信息:通过对测试语料及其标注结果进行分析,发现模板 1 对熟词的标注效果还可以,但生词的标注效果差一些,分析其原因主要是模板 1 出现生词时只是生成了一个"$w$ = "的特征,没有利用好当前词本身的一些特点,为此经过对大量汉语语料的分析,考虑在模板 1 的基础上增加后缀信息。首先根据汉语的构词特点,如:"山"、"河"等一般用在词尾构成地名,"所"、"院"等一般用在词尾构成机构名,"红"、"军"、"峰"等一般用在词尾构成人名,建立后缀信息库,然后加入后缀信息作为特征,因此建立了模板 2 形式的特征模板。对两个模板的标注结果进行进一步分析、比较如表 3 所示。

从表 3 可以看出,后缀信息对熟词标注的影响不大,而对生词的标注影响较大,有效的提高了生词的标注正确率。

4.4 结论

从以上的实验中确定了用特征模板 2 生成的模型为这次实验的应用模型,迭代次数取 100 次,它的汉语词性标注正确率达到了 97.34 %,取得了较好的标注效果。

4.5 标记集对标注的影响

由于采用的标注集较详细,导致了标注正确率较低。如:在以上采用的 102 个标记的标记集中将一般动词分为了 vg、vga、vgd、vgi、vgn、vgs、vgv,而在测试中,经常会有 vgi 标为 vgn 或 vg 等类似情况,影响了标注正确率。为了得到较好的标注结果,而又不影响后续句法分析等工作的进行,合并了一些标注类型,利用模板 2 建立了模型并进行了测试,结果如表 4。

表 4

| 标记的个数 | 102 | 87 | 66 | 20 |
|---|---|---|---|---|
| 封闭测试 | 97.34 % | 98.18 % | 98.19 % | 98.11 % |
| 开放测试 | 90.92 | 93.56 % | 93.96 % | 95 % |

从上表可以看出,87 个标记的标记集能够达到实用的标注效果,同时对后续工作的影响很小,值得采纳。

## 5 展望

在汉语词性标注方面,与其它词性标注方法相比,最大熵模型因其提供了灵活的特征机制,有效的利用了上下文信息,得到了较好的标注效果。另外,由于其特征没有特殊的约束,只需发现什么特征是可能有用的,一旦发现了新的有用特征,只需重新生成模型,而无需修改程序,为今后的研究工作带来了极大的方便,该方法很容易移植到其它研究领域。

为了使标注正确率得到进一步的提高,在今后的工作中,应对汉语生词特征的选取进行深入研究。

参考文献

[1] 清华大学计算机科学与技术系. 汉语词性自动标注系统技术报告[R]. 1992.

[2] Berger AL, Pitra BAD, Pietra VJD. A Maximum Entropy Approach to Natrual Language Processing[J]. Computational Linguistics, 1996, 22(1): 39 - 71.

[3] Berger AL. The Improved Iterative Scaling Algorithm: A Gentle Introduction[R]. School of Computer Science Carnegia Mellon University, 1997.

[4] Ratnaparkhi A. A Maximum Entropy Part of Speech Tagger[A]. Conference on Empirical Methods in Natural Language Processing[C]. University of Pennsylvania, May 1996.

[5] Ratnaparkhi A. A Simple Introduction to Maximum Entropy Models for Natural Language processing[R]. Technical Report 97 - 08, Institute for Research in Cognitive Science, University of Pennsylvania, 1997.

[6] Darroch JN, Ratcliff D. Generalized Iterative Scaling for Log-Linear Models[J]. The Annals of Mathematical Statistics, 1972, 43(5): 1470 - 1480.

[7] 郭玲,周献中. 基于模糊最大熵原则的地图图像分割[J]. 计算机应用, 2002, 22(11).

---

**阅读此文的还阅读了：**

26. ICA算法及最大熵方法在闸瓦钎丢失故障自动识别中的应用

27. 基于最大熵方法的评论信息抽取方法

28. 基于递归奇异熵方法的波纹管压浆超声检测

29. 图像分割的最大熵方法的改进

30. 计算不变密度的一种二次样条最大熵方法

31. 最大熵方法在组合期权定价中的应用

32. 最大熵方法下的信度估计

33. 研究水库地震的熵方法

34. 期权定价的最大熵方法

35. 最大熵方法下的信度估计（英文）

36. 用最大熵方法估计海流旋转谱

37. 一个改进的基于最大熵原理的汉语词性标注系统

38. 反问题中的最大熵方法

39. 结构动力优化设计的最大熵方法

40. 基于最大熵方法对测量数据估计的改进方法研究

41. 基于最大熵方法的统计语言模型

42. 最大熵方法下的纯稳健信度估计

43. 基于最大熵方法的水下航行体结构动力响应概率建模

44. 基于最大熵模型的汉语词性标注研究

45. 一种基于密度核估计的最大熵方法

46. 用最大熵方法改善图像质量

47. 基于核密度最大熵方法的杂系混合信号盲分离

48. 基于最大熵方法的汉语词性标注

49. 基于最大熵方法面向零售业的数据挖掘

50. 解第一类算子方程的最大熵方法