

# 基于最大熵模型结合 CRFs 的中文词性标注

李泽中, 黄德根

(大连理工大学 计算机科学与工程系, 辽宁 大连 116024)

([huangdg@dlut.edu.cn](mailto:huangdg@dlut.edu.cn))

**摘要:** 中文词性标注是中文信息处理领域的一项基础工作。文中构建了一个基于最大熵结合 CRFs 的词性标注器。首先利用最大熵模型进行标注, 除保留最优标注路径外, 对于每个兼类词从其余的 N-Best 标注路径中根据适当的策略选择第二个候选词性, 然后利用 CRFs 在这两个候选词性中进行第二次选择作为最终的标注结果, 最后还尝试了采用模板的方式进行标注的修正。实验证明, 与单独基于最大熵的方法相比有所提高。在 2009 年 CIPS-ParsEval 的中文简体语料上进行了测试, 词性标注的精确率达到了 93.21%。

**关键词:** 中文词性标注; CRFs; 最大熵

## Chinese POS Tagging Based on Maximum Entropy Model with CRFs

LI Ze-zhong, HUANG De-gen

(Department of Computer Science and Engineering, Dalian University of Technology, Liaoning 116024, China)

**Abstract:** Chinese POS Tagging is a basic task in the field of Chinese information processing. This paper presents a Chinese POS tagger by combining Maximum Entropy Model and Conditional Random Fields. First, have a tagging by Maximum Entropy Model, besides that retaining the best POS tagging path as usual, also select the second best tag from the other N-Best paths through some proper strategy for every Multi-category word. Secondly, have another selection between the two best candidates for the final tagging result. At last, we also try a modification with templates. According to our experiments, the method achieves a better result with a tagging precision of 93.21% by using test data from CIPS-ParsEval-2009.

**Key words:** Chinese Part-of-Speech tagging ; CRFs ; Maximum Entropy

## 1 引言

词性标注是自然语言处理领域里一项非常基础且重要的工作。到目前为止, 词性标注的方法有基于规则的方法和基于统计的方法。当前, 基于统计的方法是比较主流的, 将统计模型应用到词性标注中取得了较好的效果, 如隐马尔科夫模型 (HMM)<sup>[1]</sup>、最大熵模型 (ME)<sup>[2]</sup>、条件随机域 (CRFs)<sup>[3]</sup>等。文献[4]采用了基于规则的方法, 这种方法的局限性在于自然语言的复杂性, 建立规则库需要很高的人工成本。文献[1]采用了隐马尔科夫模型取得了很好的效果, 它的缺点在于把求解一个条件概率的问题通过贝叶斯定理转化为求解联合概率的问题, 使得标注时不能有效地利用丰富的上下文特征。CRFs 有效地克服了独立性和“标记偏置”的问题但由于效率的问题并

不适合于词性标注。文献[3]采用了CRFs进行标注, 它充分利用了丰富的局部特征和长距离触发对特征, 但其漫长的训练时间难以被接受。最大熵的好处在于它可以方便地利用各种丰富的上下文特征, 而这些特征可以重叠, 它们之间没有任何独立性假设。在本文的标注方法中, 首先采用最大熵进行初次标注, 和正常的最大熵词性标注不同, 本方法不仅仅保留最优路径, 而且在其他几条比较好的路径中为每个兼类词挑选第二个候选词性, 然后利用CRFs对兼类词的两个候选词性进行一次选择, 作为最终的标记结果。

## 2 基于最大熵的词性标注

### 2.1 最大熵模型的基本原理

最大熵原理原本是热力学中一个非常重要的

原理，后来被广泛应用于自然语言处理方面。其基本原理很简单：对所有的已知事实建模，对未知不做任何假设。也就是建模时选择这样一个统计概率模型，在满足约束的模型中选择熵最大的概率模型<sup>[9]</sup>。

若将词性标注或者其他自然语言处理任务看作一个随机过程，最大熵模型就是从所有符合条件的分布中，选择最均匀的分布，此时熵值最大。

求解最大熵模型，可以采用拉格朗日乘法，其计算公式为：

$$p_{\lambda}(y|x) = \frac{1}{Z_{\lambda}(x)} \exp \left[ \sum_i \lambda_i f_i(x, y) \right] \quad (1)$$

其中， $Z_{\lambda}(x) = \sum_y \exp \left[ \sum_i \lambda_i f_i(x, y) \right]$ ， $\lambda_i$  是对应特征的权重， $f_i$  表示一个特征。每个特征对词性选择的影响大小由特征权重  $\lambda_i$  决定，而这些权重可由GIS学习算法自动得到。

## 2.2 最大熵模型的标注过程

最大熵模型的关键在于特征选取，特征选取的恰当与否会对结果有直接影响。从直观上来看，丰富的特征对于词性标注精确率的提高具有重要的作用。一个词的词性不仅仅与前一个词的词性有关，更与词本身的前后缀特征、前词和后词等特征有关，最大熵相对于隐马尔可夫模型的优势正在于此。最大熵模型可以综合地、比较随意地选择各种上下文特征和字特征，而这些特征之间并不需要任何独立性假设。

表 1 最大熵特征模板

特征编号	特征模板
1	$w_i = X \ \& \ t_i = T$
2	prefix = $w_i$ 的首字前缀 & $t_i = T$
3	suffix1 = $w_i$ 的最末字 & $t_i = T$
4	suffix2 = $w_i$ 的最后两个字 & $t_i = T$
5	suffix3 = $w_i$ 的最后三个字 & $t_i = T$
6	$t_{i-1} = X \ \& \ t_i = T$
7	$t_{i-2} \ t_{i-1} = XY \ \& \ t_i = T$
8	$w_{i-1} = X \ \& \ t_i = T$
9	$w_i = X \ \& \ t_i = T$
10	$w_{i+1} = X \ \& \ t_i = T$

最大熵特征模板如表 1 所示，其中  $t_{i-2}$ ， $t_{i-1}$ ， $t_i$ ， $w_{i-1}$ ， $w_i$ ， $w_{i+1}$  分别表示当前词前两个位

置处的词性，前词词性，当前词词性，前词，当前词和后词。对于低频词（训练语料中出现频数低于 5 的词）和命名实体抽取字特征（包含前缀特征和后缀特征），即特征 2 ~ 5。这对于未登录词的识别具有重要的意义。前缀特征对姓名的识别作用显而易见，后缀特征如“研究所”、“乡”、“村”、“山”等对于机构名和地名的识别具有重要的作用，而后缀特征如“事件”、“战争”等对于专有名词的识别作用也很明显。特征 1 和 6 ~ 10 是一般特征，对所有词都采用这些特征。

为了降低数据稀疏和语料中的一些噪音带来的预测不可靠性，需要对特征进行筛选。特征的筛选方法包括基于频数阈值的方法和增量式特征选择方法。基于频数的特征选择方法主要基于这样一个假设：不常出现的特征是噪音或不相关的，只有那些出现频数比较大的特征才真正代表了数据的特性。这里采用的就是基于频数的方法。不同的特征类型应该采用不同的频数阈值，这里对于一般特征采用阈值为 3，前后缀特征采用阈值为 10，这个阈值的大小和训练语料的大小有关，需要从实验中得知。

标注时的解码算法采用束搜索（beam search）算法，该算法实质上是一个近似的动态规划求解最优值的过程。束搜索的宽度为  $N = 5$ ，再大的宽度并不能有效提高标注的精确度<sup>[2]</sup>。

## 3 结合 CRFs 的标注过程

最大熵模型实际上是一个有向图结构的概率模型，基于最大熵的词性标注过程是一个单向标注过程，尽管它可以比较随意地组合利用各种特征，但是，它并不能很好地利用右边的特征（这里的标注过程为自左向右，如果是在自右向左的标注过程中反之）<sup>[7]</sup>。CRFs 则很好地解决了这一问题。CRFs 是一个无向图模型，可以更好地利用全局特征和右边的特征，但由于其训练的效率问题并不能直接用于词性标注。这里先用最大熵为每个兼类词选择两个候选词性，然后再用 CRFs 在这两个词性中进行选择，即仅仅需要标注 1 或 2 的过程（1 表示第一个候选词性正确，2 表示第二个正确），既能大大缩短 CRFs 的训练时间，又能很好地利用原先最大熵中所难以利用的右边的标记特征和全局特征。算法的整体流程如图 1 所示。

与传统的基于最大熵的词性标注不同，这里不仅仅保留最优路径，也就是为每个词选择最优的词性标注，而且将从其余几条N-Best路径中为每个兼类词选择出第二个最合适的候选词性（以后如果不加说明，均称这种词性为次优词性）。这里N-Best为前五条最优的路径，即束搜索算法保存的所有路径。考虑了四种选择次优词性的方案：

(1) 直接选择第二优路径中的词性为次优词性，如果和最优词性一样则不需选择次优词性。

(2) 在其余 4 条N-Best路径中选择出现次数最多的词性为次优词性，如果和最优词性一样则不需选择次优词性。

(3) 同方案(1)，只是当次优词性和最优一样时，改为在第三优路径中选择，如果还一样，则第四、第五依此类推。如果最后仍和最优词性一样则不需选择次优词性。

(4) 在其余 4 条N-Best路径中选择局部概率最大的词性作为次优词性，如果和最优词性一样则不需选择次优词性。

从原先丢弃的非最优标记中挑选出次优标记，这种最大熵结合CRFs的方法其实是通过提高词性标注的召回率的方法来提高词性标注的精确率的。为了分析这种方法的有效性，提出潜在精确率的概念。潜在精确率实际上是召回率的另一种说法，其公式为：

$$\text{潜在精确率} = \frac{m+n}{c} \quad (2)$$

其中  $m$  是最优标记为正确标记的个数； $n$  是次优标记为正确标记的个数； $c$  为测试语料的总词数。

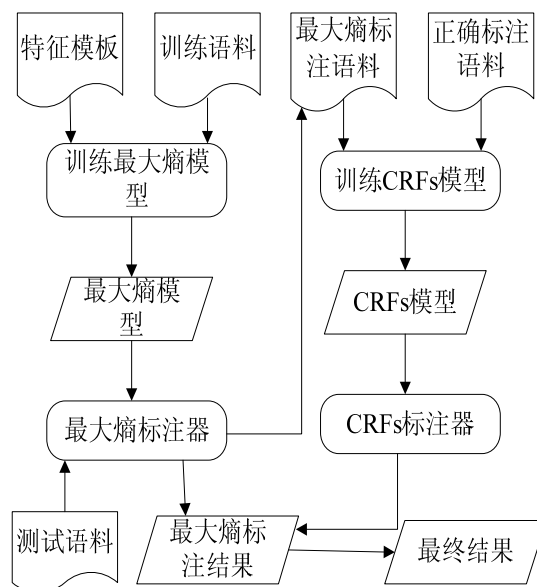


图 1 最大熵结合CRFs总体结构图

## 4 实验结果

### 4.1 最大熵词性标注

表 2 是基于最大熵词性标注的结果，最好结果为92.68%。训练语料为评测（CIPS-ParsEval-2009）发布的 3 M 语料，标记有66种，测试语料为715K语料，92687个词。

表 2 最大熵词性标注结果

一般特征阈值	前后缀特征阈值	精确率
3	5	92.49%
0	10	92.52%
3	10	92.68%
5	10	92.50%
4	10	92.60%
3	5	92.58%
3	15	92.54%

### 4.2 结合 CRFs 的标注过程

表 3 为四种不同的选择次优词性方案的实验结果。表 4 为CRFs在采用上述方案 4 时的修正效果。

表 3 四种方案的标注结果

方案	精确率	潜在精确率
方案 1	92.80%	94.27%
方案 2	92.89%	95.63%
方案 3	92.99%	95.94%
方案 4	93.11%	96.61%

表 4 CRFs对于最大熵模型的修正作用

统计项	结果
修正前精确率	92.68%
修正后精确率	93.11%
测试语料总词数	92687
原正确修正错误	512
原错误修正正确	911
维持错误	4877

从表 3 实验结果可以看出，方案 1 效果最差，对于精确率的提高作用比较微弱。方案 2、3 接近，方案 4 的提高作用最为明显。方案 1 效果差的原因在于最优路径和次优路径的差别很小，常常出现的情况是两者的区别仅仅在于这句话的最后一个标注，很显然这样并不能较大提高标注的潜在精确率，只有很少的一部分不同词性被召回回来，CRFs的作用也就微乎其微了。方案 2、3 和 4 都比较高地保证了兼类词候选词性的多样性，提高了潜在精确率。对于方案 4 效果最好的一个比较合理的解释是：原先所谓的最优标注不过是整体意义上（即整个路径上）的最优，是数学意义上的最优。而局部最优值常常有可能更接近自然语言的实际情况，这也说明了自然语言的复杂性。

### 4.3 利用模板进行修正

在 3 M 训练语料中，共有 389171 词次，共包含 26682 个词。其中频数超过 200 的称为超高频词，有 238 个，但这些词在全部训练语料中竟占 33.79%，起到“骨架”的作用。由于出现的频数非常高，所以一般也表现出多种词性，这些词也就成为词性标注中最难以标注的词。利用模板修正标注的方法来源于这样一个假设：直接保存这些超高频词的上下文作为测试时词性标注的实例，将对这类词的消歧具有重要意义。

模板实际上是一种固定格式的规则。同规则的选取一样，模板应该既要考虑到精确率又要照顾到覆盖率。一个低于统计标注精确率的模板是没有意义的，同样，一个覆盖率极低的模板也是无关大局的。以词性作为模板覆盖率很高，但精确率较低，词为模板则正好反之。考虑到这两个特点，采用了一种折中的三元模板： $(t_{i-1}, w_i, w_{i+1})$ ，即前词词性，当前词和后词。

除了上述三元模板之外，另外采用了一个二元模板  $(w_i, w_{i+1})$ ，用于扩大模板的覆盖率。

三元模板拥有比二元模板更高的优先级。

表 5 模板的修正作用

统计项	结果
修正前精确率	93.11%
修正后精确率	93.21%
模板匹配次数	16609
原正确修正错误	617
原错误修正正确	701

用模板扫描整个训练语料，当  $w_i$  为超高频时激发模板抽取模块。为了提高模板的精确率，对于抽取出的模板实例进行裁剪，凡模板实例精确率在 90% 以上的保留下来，否则去除。共抽出 106376 个三元模板实例，17400 个二元模板实例。两种模板实例的覆盖率为 17.92%，其中三元模板的精确率为 96.20%，二元模板为 95.11%。由于模板主要是针对高频词修正标注的，而高频词的精确率为 95.04%，因此如表 5 所示，模板的作用效果不是太明显。

## 5 结束语

本文根据最大熵和 CRFs 模型的特点，采用了结合两种模型进行汉语词性标注的尝试，并在实验中取得了较好的效果，最终的评测结果如表 6 所示。为了进一步提高标注的精确率，分析了评测结果中标记的错误类型。表 7 为易混淆的词性，其中横向为正确的标注，纵向为错标注成的词性，表中元素为这种错误所占总体错误的比例。如第二行三列的 7.82% 表示应该标记成 v 却错标记成 vN 占所有标注错误的 7.82%。表 8 显示了出错次数最多的 10 种词性，占了总错误数的 82.13%。因此，由于错误类型比较集中，针对易混淆的词性和最容易出现错误的词性，制定特殊的分类器，是未来研究工作中的重点。

表 6 标注的最终评测结果

统计项	精确率
非 OOV	94.43%
OOV	77.18%
总体	93.21%

表 7 易混淆的词性

	v	vN	n
v	-	7.82%	5.47%
vN	11.38%	-	3.78%

n	5.83%	3.78%	-
---	-------	-------	---

表 8 错误数最多的几种词性

词性	错误数	精确率
v	1620	86.68%
n	1291	94.48%
vN	769	75.15%
a	376	87.78%
d	241	93.01%
p	235	93.42%
b	205	81.68%
nP	185	81.92%
nS	140	89.34%
vC	104	92.21%

### 参考文献：

- [1] 梁以敏,黄德根. 基于完全二阶隐马尔可夫模型的汉语词性标注.计算机工程,2005,31(10):177-179
- [2] Adwait Ratnaparkhi. A Maximum Entropy Model for part-of-speech tagging. EMNLP 1,1996,21(4):133-142.
- [3] 姜维,关毅,王晓龙. 基于条件随机场的词性标注模型. 计算机工程与应用,2006,21:13-16.
- [4] 王广正,王喜凤. 一种基于规则优先级的词性标注方法.安徽工业大学学报,2008,第25卷第4期:426-429
- [5] Dan Klein,Christopher,D.Manning. Conditional Structure versus Conditional Estimation in NLP Models. EMNLP, 2002, 21(4):9-16.
- [6] Kristina Toutanova,Christopher D.Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger.EMNLP/VLC , 1996, 22(1):63-71.
- [7] Kristina Toutanova,Dan Klein,Christopher D. Manning, Yoram Singer. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. Proceedings of HLT-NAACL, 2003, 252-259.
- [8] Yoshimasa Tsuruoka,Jun'ichi Tsujii. Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data.Proceedings of HLT/EMNLP, 2005, 467-474.
- [9] 周雅倩.最大熵方法及其在自然语言处理当中的应用: (博士学位论文) .上海,复旦大学.2005