第 36 卷 第 4 期 计 算 机 工 程 2010 年 2 月
Vol.36 No.4 Computer Engineering February 2010
·博士论文· 文章编号：1000—3428(2010)04—0017—03 文献标识码：A 中图分类号：TP18

# 中文分词和词性标注模型

刘遥峰，王志良，王传经
(北京科技大学信息工程学院，北京 100083)

**摘 要：**构造一种中文分词和词性标注的模型，在分词阶段确定 $N$ 个最佳结果作为候选集，通过未登录词识别和词性标注，从候选结果集中选优得到最终结果，并基于该模型实现一个中文自动分词和词性自动标注的中文词法分析器。经不同大小训练集下的测试证明，该分析器的分词准确率和词性标注准确率分别达到 98.34%和 96.07%，证明了该方法的有效性。

**关键词：**分词；词性标注；最短路径

# Model of Chinese Words Segmentation and Part-of-Word Tagging

LIU Yao-feng, WANG Zhi-liang, WANG Chuan-jing
(School of Information Engineering, University of Science & Technology Beijing, Beijing 100083)

**【Abstract】** This paper proposes a model of Chinese words segmentation and part-of-word tagging. In the words segmentation stage, the top $N$ segmentation results are confirmed as the candidate. The final result among these candidates is gotten after unknown words recognition and part-of-word tagging. A Chinese lexical analyzer is developed. This model with different size of training set is tested. The lexical analyzer's accuracy of words segmentation and part-of-word is 98.34% and 96.07%. This proves the effectiveness of the method.

**【Key words】** words segmentation; part-of-word tagging; shortest path

## 1 概述

词是最小的、能够独立活动的、有意义的语言成分，但汉语中词语之间没有明显的区分标记，因此，中文词语分析是中文信息处理的基础与关键。分词的准确度和词性标注的准确度密切相关，有机地将分词过程和词性标注过程融合在一起，有利于消除歧义和提高整体效率。

本文应用 $N$-最短路径法，构造了一种中文自动分词和词性自动标注处理模型，实现了一个中文自动分词和词性自动标注处理的中文词法分析器。对最有潜力的粗分结果，分别进行未登录词识别和词性自动标注，形成候选集，再通过评估判优选择最优结果，有效地将词形信息和词性信息结合在一起[1]。

## 2 中文分词和词性标注处理模型

### 2.1 中文分词和词性标注模型描述

可以将中文分词及词性标注问题描述为一个已经被标注了词性的词串 $<W,T>=<w_1,t_1><w_2,t_2>\cdots<w_n,t_n>$。其中，$<w_i,t_i>$ 表示具有词性 $t_i$ 的词 $w_i$，输出端输出序列 $C=c_1c_2\cdots c_n$，通过找出跟 $C$ 对应的 $<W,T>$，比较得出具有最大概率的结果 $<W,T>^o$。

$$<W,T>^o=\arg\max_{W,T}P(<W,T>|C) \tag{1}$$

### 2.2 中文分词和词性标注统计模型

式(1)从概率统计角度描述了分词及词性标注的一般模型。为将中文自动分词和词性自动标注模型化处理，实现将词语信息和词性信息融合在一起作为最终结果的评价依据，引入 $N$-最短路径法，用以产生包含 $N$ 个最大结果的候选集[2]。

对于字序列 $C$，经过分词处理后，得到 $N$ 个概率最大的分词结果 $(W_1,W_2,\cdots,W_n)$。对每个分词结果进行未登录词的识别，并且保留了未登录词的词性，产生具有少量词性信息的词序列 $(<W_1,T_1>',<W_2,T_2>',\cdots,<W_n,T_n>')$。然后分别进行词性自动识别，得到具有完备词性信息的词序列 $(<W_1,T_1>,<W_2,T_2>,\cdots,<W_n,T_n>)$。最后，用评估函数进行判优，得到概率最大的结果 $<W,T>^o$。其处理过程如图 1 所示。

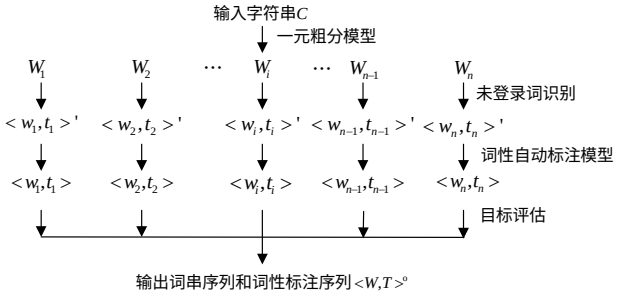


**图 1 分词和词性标注处理过程**

### 2.3 一元粗分模型

自动分词就是已知一个汉字串，求跟它对应的、有最大概率的词串。即

$$W'=\arg\max_{W}P(W|C) \tag{2}$$

由贝叶斯公式

$$W'=\arg\max_{W}P(W|C)=\arg\max_{W}\frac{P(W)P(C|W)}{P(C)} \tag{3}$$

**基金项目：**国家“863”计划基金资助项目“智能感知与先进计算技术”(2007AA01Z160)；北京市自然科学基金资助重点项目“基于情绪认知模型的个性化数字教育关键技术研究”(KZ200810028016)

**作者简介：**刘遥峰(1982－)，男，博士研究生，主研方向：自然语言处理，个人机器人技术；王志良，教授、博士生导师；王传经，硕士研究生

**收稿日期：**2009-09-30 **E-mail：**yaofeng-liu@163.com

其中，$P(C)$ 是汉字串的概率，它是一个常数，不必考虑；$P(C|W)$ 是词串到汉字串的条件概率。在已知词串的条件下，出现相应的汉字串的概率是 1，也不必考虑。仅仅需要考虑的是 $P(W)$，即词的概率。

上述公式可简化为

$$W' = \arg\max_{W} P(W) \tag{4}$$

词串概率采用一元语法进行求解，则

$$P(W) = \prod_{i=1}^{n} p(w_i) \tag{5}$$

这就是说，概率最大的词串便是最佳的词串。

### 2.4 词的自动标注模型

在词性自动标注过程中，在已知词串 $W$，寻求最大概率的词性标注列 $T'$。

$$T' = \arg\max P(T|W) \tag{6}$$

根据贝叶斯公式：

$$P(T|W) = \frac{P(T)P(W|T)}{P(W)} \tag{7}$$

其中，分母 $P(W)$ 是词串 $W$ 的概率，是一个常量，上式可以简化为

$$P(T|W) = P(T)P(W|T) \tag{8}$$

假定词语之间是相互独立的，并且词语的出现只依赖于它本身的标注，则已知标记 $T$ 的条件下词串的概率可近似地用每个词在已知标记时的条件概率的乘积来表示。

$$P(T) \approx P(w_1|t_1)P(w_2|t_2)\cdots P(w_n|t_n) \tag{9}$$

假定一个标记的概率取决于出现在它前面的那个标记，那就可以用 $T$ 中每个标记的概率的乘积来表示：

$$P(T) \approx P(t_1|t_0)P(t_2|t_1)\cdots P(t_n|t_{n-1}) \tag{10}$$

因为 $t_0$ 是虚设的标记，所以 $P(t_1|t_0)$ 实际上是 $P(t_1)$，综合起来，词性标记的统计模型可以表示为

$$T' = \arg\max \prod_{i=1}^{n} P(t_i|t_{i-1})P(w_i|t_i) \tag{11}$$

用隐马尔科夫模型来描述词性标注问题：以词串 $W = w_1w_2\cdots w_n$ 作为观察到的输出序列，以对应的词性标注串 $T = t_1t_2\cdots t_n$ 作为隐藏的状态转移序列，将词语以某词性出现的概率变相作为状态转移的发射概率。在求解过程中，应用维特比算法，算法有 3 步：初始化，推导，终止和路径读出[3]。

(1)初始化

第 1 个词处于状态 $j$ (标注为 $j$)的概率：

$$\delta_1(t^j) = P(t^j|w_1) \tag{12}$$

(2)推导

词 $i+1$ 处于状态 $j$ (标注为 $j$)的概率：

$$\delta_{i+1}(t^j) = \max_{1\le k\le T}[\delta_i(t^k)\times P(w_{i+1}|t^k)\times P(t^j|t^k)], 1\le j\le T \tag{13}$$

词 $i+1$ 处于状态 $j$ (标注为 $j$)时，词 $i$ 所处的最有可能的状态(标记)为

$$\Psi_{i+1}(t^j) = \arg\max_{1\le k\le T}[\delta_i(t^k)\times P(w_{i+1}|t^j)\times P(t^j|t^k)],\ 1\le j\le T \tag{14}$$

(3)终止和路径读出

$t_1, t_2, \cdots, t_n$ 是词语序列 $w_1, w_2, \cdots, w_n$ 选择的标记，有

$$t_n = \arg\max_{1\le j\le T} \delta_n(t^j) \tag{15}$$

$$t_i = \Psi_{i+1}(t_{i+1}), 1\le i\le n-1 \tag{16}$$

$$P(t_1, t_2, \cdots, t_n) = \max \delta_{n+1}(t^j), 1\le j\le T \tag{17}$$

### 2.5 目标评估函数

一个词串对应的汉字串是唯一的，即

$$P(C|W) = 1 - P(CW) = P(W) \tag{18}$$

$$P(W,T|C) = P(T|CW)P(W|C) = P(T|W)P(W|C) = P(T)P(W|T) \tag{19}$$

利用隐马尔科夫展开 $P(T)P(W|T)$，并引入共现概率

$$P(<W,T>|C) = \prod P(t_i|t_{i-1})P(w_i|w_t) \tag{20}$$

$$P^0(W,T) = \ln P(W,T) = \sum \ln P(t_i|t_{i-1}) + \sum \ln P(w_t|t_i) \tag{21}$$

评估函数如下：

$$R^{\#} = \arg\max_{W,T}[\sum \ln P(t_i|t_{i-1}) + \sum \ln P(w_t|t_i)] \tag{22}$$

## 3 分词及处理

### 3.1 预处理阶段

在预处理过程中，需要进行 2 次扫描。第 1 次，扫描待处理文本，根据标点符号进行断句。第 2 次，扫描通过断句得出的每一句话，区分汉字和非汉字字符，并对非汉字字符用自动机识别数字和英文字串。汉字串作为自动分词模块的输人序列。

### 3.2 分词及未登录词识别

本文采取的分词方法是基于字符串匹配的分词方法中的正向最大匹配分词和逆向最大匹配分词相结合的分词方法。

正向最大匹配算法的核心是，对于待切分的一段字符串，用户首先以该段语句的首字母起点进行搜索，直到找到以该首字母为起点，在字符串中出现的最长的词，并以此为标志切出第一个词，并将剩余字符串作为另一待切分字符串进行相同处理[4]。

逆向最大匹配算法的基本思想与正向最大匹配算法相同，唯一的区别是最大匹配的顺序不是从首字母开始而是从末尾开始。正向最大匹配算法流程如图 2 所示。

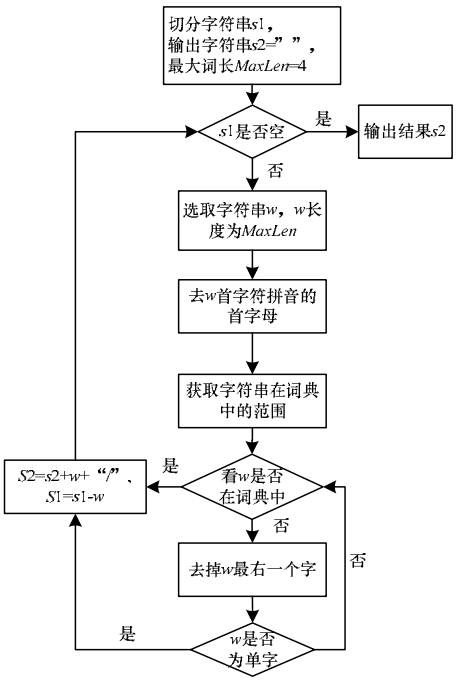


**图 2 分词算法流程**

采用 2 种分词算法相结合的方式，主要是为了消除组合歧义，比如如下的 2 个句子：

他的确切地址在这里。

这块肉的确切得不错。

这种歧义称为组合歧义。单一使用正向最大匹配算法和逆向最大匹配算法都会产生错误，为了减少这种歧义的影响，

使用 2 种分词方法相结合的方式。通过统计方法消除部分歧义。描述如下：$C = c_1c_2\cdots c_m$ 表示输入的由 $m$ 个汉字组成的歧义切分字段。$W = w_1w_2\cdots w_n$ 表示把 $C$ 切分后得到的由 $n$ 个词组成的词串。$V = v_1v_2\cdots v_k$ 是另一种切分结果。用 $frq(w)$ 表示 $w$ 的频率。

如果有

$$frq(w_1)\times frq(w_2)\times\cdots\times frq(w_n) > frq(v_1)\times frq(v_2)\times\cdots\times frq(v_k) \quad (23)$$

则选择切分结果 $W$ 。

使用这种算法的弊端是对于频率较少的不能得到正确的分词结果，如果输入的是“ 这块肉的确切得不错 ”，则得到的就是错误的分次效果。对于这种歧义的消除就要建立分词知识库，根据分词规则处理歧义现象。

未登录词的识别包括姓名识别和地名识别两部分，识别过程如图 3 所示。姓名的识别采用基于概率统计的方法。地名的识别采用基于规则的方式，主要通过检查其后级来识别。识别出的未登录词保留其词性信息。

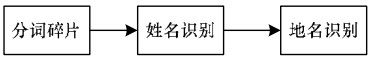


**图 3　未登录词识别模块流程**

### 3.3　词性标注及评估判优阶段

词性自动标注的对象是一句话的词串，目标是寻找一条概率乘积最大的词性序列。将词串中的每个词的所有词性及其费用添人到该节点中(未登录词的词性信息由前面的处理获得，其余从词典获得)。从左到右扫描词串，计算该词在某一词性的费用和前一词最有可能的词性。当前词为终点词时，进行回退，得出该词串的词性序列。

对每个分词结果分别进行未登录词识别和词性标注之后，形成了 $N$ 个具有词性信息的词串。根据式(22)对每个路径进行打分，费用最小者为最终结果。在这个模块中，还进行叠词识别和合并部分未登录词的操作[5]。

## 4　实验与分析

词性标注方法的好坏最终还是要通过在大规模数据上的实验来进行评价。根据上文方法构建了一个汉语词性自动标注器。训练语料训练和封闭开放测试均采用北京大学的人民日报 1998 年 1 月的语料，以下是实验的结果与分析。

### 4.1　实验语料与标记集

选取了 2.05 MB 共 306 930 词的语料作为训练语料，训练过程是从 20 万词次逐步递加到 30 万词次。又另外从中选取了 35 万语料作为测试语料，将测试语料处理经过分词但未标注的文本，再重新进行词性标注。本文所有的实验都是在 26 个词类组成的小标注集上进行的。

### 4.2　评价标准

本文采用 3 个评估函数：中文分词(标注)准确率，中文分词(标注)召回率和 $F$-值，分别如式(24)~式(26)所示。为了方便，用 $a$ 表示正确分词(标注)的词的个数，$b$ 表示所识别的词的个数，$c$ 表示文本中词的总数。

$$\text{准确率 } p = \frac{a}{b}\times 100\% \quad (24)$$

$$\text{召回率 } r = \frac{a}{c}\times 100\% \quad (25)$$

$$F\text{-值} = \frac{2\times p\times r}{p+r}\times 100\% \quad (26)$$

### 4.3　实验结果

实验结果如表 1 所示，其中总词数为 7 351。

**表 1　系统实验结果**

| 训练集大小 | $N$ 值 | 自动分词 | | 词性标注 | |
|---|---|---|---|---|---|
| | | 正确切分词数 | 准确率/(%) | 正确标注词数 | 准确率/(%) |
| 20 万 | 1 | 7 129 | 96.98 | 6 902 | 93.89 |
| | 2 | 7 203 | 97.98 | 6 919 | 94.13 |
| | 3 | 7 168 | 97.51 | 6 935 | 94.34 |
| | 4 | 7 172 | 97.56 | 6 907 | 93.97 |
| 25 万 | 1 | 7 160 | 97.40 | 6 940 | 94.41 |
| | 2 | 7 168 | 97.51 | 6 975 | 94.88 |
| | 3 | 7 234 | 98.40 | 6 988 | 95.06 |
| | 4 | 7 217 | 98.17 | 6 966 | 94.76 |
| 30 万 | 1 | 7 219 | 98.21 | 7 046 | 95.85 |
| | 2 | 7 226 | 98.29 | 7 051 | 95.92 |
| | 3 | 7 229 | 98.34 | 7 062 | 96.07 |
| | 4 | 7 186 | 97.76 | 7 038 | 95.74 |

从表 1 中的数据可看出：(1)该系统的性能良好，分词准确率和词性标注准确率高于使用分词和马尔科夫模型标注的准确率 96%和 94%。(2)引用最短路径法后，分词准确率和词性标注准确率都得到提高，证明前期保留多个粗分结果是合适的。(3)词性信息的引人提高了分词的准确率。(4)随着 $N$ 值的增大，系统性能提升变得不明显。针对这种情况，并且考虑到系统的运行效率，选取 $N$ 值为 3。(5)随着训练集大小的增加，分词和标注效果会越来越好，正确率逐渐增大，但是增大的趋势是减小的。训练集的大小与标注正确率的提高是呈非线性分布的。

### 4.4　错误分析

首先，未登录词的识别是影响系统识别准确度的一个重要方面。因为使用隐马尔科夫模型，后一个词的词性判断和前一个词的词性密切相关，容易引起识别错误的未登录词会在系统中形成累积效应，严重影响后续文本的分词及词性标注准确性。

另外，连续的兼词也会对系统产生比较严重的影响。兼词中前一个词的词性不确定性，对系统后续的词性标注也会产生叠加的影响。

最后，分词产生的歧义也是不可忽略的重要因素。在本系统中对分词歧义的考虑不多，但是现实字符串中存在的各种各样的分次歧义都会对分词结果产生比较大的影响，对中文分词歧义的消除也是自然语言处理的一个重要议题，分词的效果直接影响词性标注的效果。

## 5　结束语

本文应用最短路径法构造了一种中文自动分词和词性自动标注的模型，并实现了一个中文自动分词和词性自动标注的中文词法分析器。经测试，该系统具有较高的分词准确率和词性标注准确率，初步证明该方法是有效的。

未登录词及分词歧义是影响本系统性能的重要因素，需要在后续的工作中给予进一步的研究，以进一步提高本系统的性能。

## 参考文献

[1] 张华平, 刘　群. 基于 $N$-最短路径的中文词语粗分模型[J]. 中文信息学报, 2002, 16(5): 1-7.

[2] Manning C D, Schutze H. 统计自然语言处理基础[M]. 苑春法, 李庆中, 王　昀, 等, 译. 北京: 电子工业出版社, 2005.

[3] 梁以敏, 黄德根. 基于完全二阶隐马尔科夫模型的汉语词性标注[J]. 计算机工程, 2005, 31(10): 177-179.

[4] 苗夺谦, 卫志华. 中文文本信息处理的原理与应用[M]. 北京: 清华大学出版社, 2007.

[5] 张素智, 刘放美. 基于矩阵约束法的中文分词研究[J]. 计算机工程, 2007, 33(15): 98-100.

编辑　顾逸斐