一种基于统计的分词标注一体化方法①

Integrated Chinese Words Segmentation and Labeling Based on Statistic Method

褚颖娜 廖 敏 宋继华 (北京师范大学 信息科学与技术学院 北京 100875)

摘 要: 分词标注是中文信息处理的基础。传统方法的处理步骤大都是首先对文本进行预处理，得到文本的粗分模型，在此基础上对词语进行词性标注。粗分模型集合的大小取决于采用的分词方法，粗分模型的准确性直接影响着后续处理结果的准确性。提出一种基于统计的分词标注一体化方法即概率全切分标注模型，该方法的特点是将分词、标注两部分工作融为一体同时进行，在利用全切分获得所有可能分词结果的过程中，计算出每种词串的联合概率，同时利用马尔可夫模型计算出每种词串所有可能标记序列的概率，由此得到最可能的处理结果。该方法提高了结果的召回率和准确率，由于在查询词典时采用的是单次查询双数组 Trie 树索引，因此效率也很高。

关键词: 分词标注 粗分模型 双数组 Trie 树索引 马尔可夫标注模型 全切分

中文词法分析是进行中文信息处理的基础，中文词法分析一般包括 3 个过程：预处理过程的词语粗切分、切分排歧与未登录词识别、词性标注[1]。预处理过程的粗分结果对后续处理起着关键作用，较为熟悉的最大匹配、最短路径、最大概率等分词方法均只保留了一种切分结果，较差的包容性导致有可能失去正确的切分结果，为了防止预处理阶段的错误对后续工作产生错误累加，我们采用全切分算法对文本进行切分，获得所有可能的切分结果，并在切分的同时获得候选词的属性用来计算词串的标记序列概率模型。最终结果由切分词串的联合概率和词串的标记序列概率共同决定。

1 训练词典和词性标记序列规则

1.1 建立词典

本文对人民日报一个月的标注语料库进行了词条统计，得到了词条的属性，其中包括词条数目、词条以不同标记出现的计数。每一词条在词典中的组织形式为：

word $a_1|c_1$ $a_2|c_2$……$a_i|c_i$ C，其中 $a_1,a_2\cdots a_n \in \{A\}$({A}为词典的词类标注集)，$c_1,c_2...c$ 为 word 以不同词类在文本中出现的次数，C 为 word 在文本中出现的总次数，因此有：$c_1 + c_2 + \ldots + c_n = C$。例如：“包 nr|19 q|12 v|61 n|8 100” 为“包”在词典中的表现形式。

1.2 双数组 Trie 树词典索引

在对文本进行分词的过程中大部分的操作都在对词典进行搜索，词典的索引效率是分词效率的一个重要影响因素。目前一般常用的索引包括线性索引表、倒排表、散列表以及各种搜索树。本文采用的是双数组 Trie 树索引，Trie 树属于搜索树，前缀相同的词在 Trie 树索引中处在同一分支，它的实质是一个确定的有限状态自动机，每个节点代表一种状态，根据输入变量的不同，进行状态转移。

所谓双数组 Trie 树是用两个整型数组来表示 Trie 树， 该种表示方法很大程度上弥补了 Trie 树空间浪费的问题[2]。Trie 树可以看成是一个有限状态自动机。定义函数 g(s, c) = t 为状态之间的转移函数，

① 基金项目:国家社科基金(05BYY022)
收稿时间:2009-03-06

其中 s 表示一个结点，c 表示转移条件，t 表示下一个可接受状态。当 Trie 树用两个一维数组(设为 base 和 check 数组)表示时，函数 g(s, c)=t 拆分成如下表示：

① t = base[s] + c

② check[t] = s

在这两个式子中，base[s]作为状态转移的基值，c 为状态转移变量，check[t]为校验值，只有当 check[t]=s 成立时，t 才被判定为可接收状态。若 t 为可终止状态，则 base[s] 设定为负值，若 t 无孩子节点，则 base[t]=-∞。第一个式子表示 t 的位置可由 s 的 base 值和 c 的序列码来计算，第二个式子限定了 t 的前一个状态为 s。已知一个状态和转移条件，便可以唯一确定下一个可接受状态。此外本文在创建双数组的过程中采用了前人的优化算法，即在创建数组时优先考虑孩子多的节点。这一方法进一步提高了空间利用率，减少了双数组数据稀疏的问题。

由双数组Trie树以上性质可知查询时间只取决于输入字串的长度，因此效率很高。

2 分词标注概述

中文词法分析包括分词、切分排歧与未登录词识别、词性标注三个部分[3]。大部分的处理办法均是将三个部分分开进行，逐步处理。首先是分词技术，现有的分词方法可分为三大类：基于规则的字符串匹配分词方法、基于理解的分词方法和基于统计的分词方法。常用的分词方法有正向和逆向最大匹配、最短路径、全切分、最大概率、N—最短路径等方法。

中文在切分的过程中经常出现具有多种切分结果的字段，被称为歧义字段[4]。歧义字段可分为交叉型歧义字段(又称交集型歧义字段)和覆盖型歧义字段(又称组合型、多义型或包孕型歧义字段)。假设在待切分文本中有一连续字段 ABC，A、B、C 均代表由一个或多个字组成的字串，其中 AB、BC 均可被切分为词，则该字段为交叉型歧义字段。假设在被处理文本中有一连续字段 EF，E、F 分别代表由一个或多个字组成的字串，如果 EF、E、F 都是词典中的词，则称 EF 为覆盖型歧义字段。

最大匹配分词方法简单、容易实现，但是无法解决上面提到的歧义问题。因此分词结果的正确率不是很高，导致最终的标注结果的准确率较低。最短路径方法是使切分出来的词数最少，但是最短路径经常不只一条，不科学的舍弃原则也影响了分词结果。前人在此基础上引入了 N—最短路径方法，该方法保留了 N 条较短路径，即分词结果有 N 条，体现了很好的包容性，可以最大限度的包容正确结果。此外最大概率分词方法也是一个较好的分词方法，它的理论依据是联合概率最大的刺穿就是最终的切分结果。而全切分方法与以上方法的不同之处在于它切分出了所有可能的切分结果，不在分词阶段做排除工作。

词性标注的基本方法有两种：基于规则的方法和基于统计的方法。基于规则的方法需要采用人工的方法构建大量的语法规则，该方法不易保证规则的完备性和在真实文本处理中的有效性。基于统计的方法主要有基于阴马尔可夫模型、基于最大熵的方法和决策树等方法。其中基于马尔可夫模型的方法是词性标注领域应用最广泛、最成熟的方法。

3 概率全切分标注模型

本文改进的分词标注一体化方法，是基于概率全切分分词模型和马尔可夫模型。该方法的特点是将分词、标注两部分工作融为一体同时进行，在利用全切分得到所有可能分词结果的过程中，同时计算每种可能词串的联合概率，并利用马尔可夫模型标注器计算出每种词串所有可能标记序列的概率，由此得到最可能的处理结果。该方法在最大程度上提高了结果的召回率和准确率。

3.1 全切分的定义

设 $S=c_1c_2...c_n$ (c_i ∈汉字集合，1<=i<=n)为待切分的汉字串；

$R=w_1w_2...w_m$ (w_j ∈基于词典的汉词集合，1<=j<=m)为 C 的一种切分形式，设 K 为所有可能的切分形式的个数，为 S 的所有可能的切分形式集合，则 R(S)是对 S 的全切分集合，对 S 的全切分过程就是求解 R(S)的过程[5]。

3.2 全切分集合的求解公式

定义 1：字符串集合的串接运算·：设字符串集合 A、B，则 A·B={a-b|a∈A，b∈B}(“-”表示两个字符串的连接) 例如：{乒乓}·{球，球拍}= {乒乓-球，乒乓-球拍}。

定义 2：串首词集合 FW (S)：设汉字串 $S=w_1c_1c_2...c_n$ (c_i ∈汉字集合，1<=i<=n)，则 S 的串首词集合定义为：

FW (S)={FW_j | FW_j = $C_1C_2\cdots\cdots C_i$, $C_1C_2\cdots\cdots C_i$ ∈汉字集合，1<=i<=n，1<=j<=n }

设 R(S)是关于 S 的全切分集合，那么关于输入字符串 S 的全切分集合的求解公式为：

$$R(s)=\bigcup_{k}(\{FWj\}\bullet R(substr(S,strlen(FWj)+1))) \quad (1)$$

其中 $FW_j \in$ FW (S),k 为 FW (S)的元素个数，substr()是取子串函数，strlen()是串长度函数。根据全切分的求解共识可以看出，对一个汉字串的全切分过程就是首先求得串首词，然后对剩余的字串递归全切分的过程。

3.3 马尔可夫模型

马尔可夫模型本身是一个马尔可夫过程的概率函数[6]。马尔可夫过程最早由 Andrei A .Markov 提出，它最原始的目的也是用于语言处理。马尔可夫过程的特性如下：若当某随机过程{X(t),t ∈ T}在某时刻 t_k 所处的状态已知的条件下：过程在时刻 t(t>t_k)处的状态只会与过程在 t_k 时刻的状态有关，而与过程在 t_k 以前所处的 状态无关。若{X(t)，t ∈ T}为离散型随机变量，这种特性可表示为：

$$P(X_{i+1}=t^j \mid X_i,\ldots,X_1)=P(X_{i+1}=t^j \mid X_i) \quad (2)$$

在马尔可夫模型标注器中，将文本中的标记序列看成是一条离散型马尔可夫链，根据马尔可夫的特性，文本中任意一个词语的标记只依赖于前面的标记。这种属性可表示为：$P(t_{i+1}|\ t_{1,i})= P(t_{i+1}|\ t_i)(\ t_1,\cdots..\ t_{i+1}$ ∈T, t_i 为句子位置 i 处的词语 w_i 的标记，T 为标注集)。因此对文本进行标记可转化为获得一个词序列最佳的标记序列，即使 $P(t_{1,n}|w_{1,n})$最大的 $t_{1,n}$($w_{1,n}$ 表示位置 1 与 n 之间出现的词语，$t_{1,n}$ 表示 $w_{1,n}$ 的标记 $t_1,\cdots..\ t_n$)。根据贝叶斯原理，可将此内容表示如下：

$$\begin{aligned}&\arg\max_{t_{1,n}} P(t_{1,n}\mid w_{1,n})\\&=\arg\max_{t_{1,n}}\frac{P(w_{1,n}\mid t_{1,n})P(t_{1,n})}{P(w_{1,n})}\frac{n!}{r!(n-r)!}\\&=\arg\max_{t_{1,n}}P(w_{1,n}\mid t_{1,n})P(t_{1,n})\end{aligned} \quad (3)$$

在假设词语之间是独立的和词语的出现只依赖于它本身的标注的前提下，可得到下式：

$$\begin{aligned}&P(w_{1,n}\mid t_{1,n})P(t_{1,n})=\prod_{i=1}^{n}P(w_i\mid t_{1,n})\times P(t_n\mid t_{1,n-1})\times P(t_{n-1}\mid t_{1,n-2})\times\ldots\times P(t_2\mid t_1)\\&=\prod_{i=1}^{n}P(w_i\mid t_i)\times P(t_n\mid t_{1,n-1})\times P(t_{n-1}\mid t_{1,n-2})\times\ldots\times P(t_2\mid t_1)\\&=\prod_{i=1}^{n}[P(w_i\mid t_i)\times P(t_i\mid t_{i-1})]\end{aligned} \quad (4)$$

3.3 概率全切分标注模型

为了提高文本切分标注的结果，并减少索引词典的次数，概率全切分标注模型采用全切分的分词方法，以及分词标注同时进行的处理策略。该方法对待处理的词进行单次索引词典操作来获得词的所有信息，计算所有分词结果的联合概率，以及每种结果的最可能的标记序列。

对于字符串 S,设 R(S)={R^m=|1<=m<=K}为全切分结果集，在假设上下文无关，词与词之间是相互独立的前提下设 $R^m=w_1w_2\cdots\cdots w_n$ 可得 Rm 的联合概率公式：

$$P(R^m)=P(w_1..w_n)=\prod_{i=1}^{n}P(w_i) \quad (5)$$

则求文本分词标注的处理结果可转化为求使 $P(R^m|\ t_{1,n})\ P(t_{1,n})P(R^m\ )$值最大的 R^m 和 $t_{1,n}$。可将此内容表示为下面的公式：

$$\arg\max_{t_{1,n},R^m}P(R^m\mid t_{1,n})P(t_{1,n})P(R^m) \quad (6)$$

$$\begin{aligned}&P(R^m\mid t_{1,n})P(t_{1,n})P(R^m)=P(w_{1,n}\mid t_{1,n})P(t_{1,n})P(w_{1,n})\\&=\prod_{i=1}^{n}[P(w_i\mid t_i)\times P(t_i\mid t_{i-1})]\times\prod_{i=1}^{n}P(w_i)\\&=\prod_{i=1}^{n}[P(w_i\mid t_i)\times P(t_i\mid t_{i-1})\times P(w_i)]\end{aligned} \quad (7)$$

因此，确定一个句子最优分词标注的等式是：

$$\begin{aligned}&\arg\max_{t_{1,n},R^m}P(R^m\mid t_{1,n})P(t_{1,n})P(R^m)=\\&\arg\max_{t_{1,n},w_{1,n}}\prod_{i=1}^{n}[P(w_i\mid t_i)\times P(t_i\mid t_{i-1})\times P(w_i)]\end{aligned} \quad (8)$$

3.4 求解与实现

如果按照上式计算的话，仅在标注部分其复杂度就是句子长度的指数倍，因此要采用有效的 Viterbi 算法进行优化。Viterbi 算法有三步：(1)初始化；(2)推导；(3)终止和路径读出计算两个函数：$\delta_i(j)$给出词 i 处于状态 j 时的概率；$\varphi_{i+1}(j)$给出了词 i+1 处于状态 j 时词 i 所处的最可能状态。

① 初始化步骤把概率 1.0 分配给标记 PERIOD (句点标记)：δ_1(PERIOD)=1.0； δ_1(t)=0.0(t≠PERIOD)

② 句子从句点开始，推导步骤公式如下：

$$\delta_{i+1}(t^j)=\max_{1\le k\le T}\begin{bmatrix}\delta_i(t^k)\times P(w_{i+1}\mid t^i)\\\times P(t^i\mid t^k)\times P(w_{i+1})\end{bmatrix} \quad (9)$$

$$\varphi_{i+1}(t^j)=\arg\max_{1\le k\le T}\begin{bmatrix}\delta_i(t^k)\times P(w_{i+1}\mid t^j)\\ \times P(t^j\mid t^k)\times P(w_{i+1})\end{bmatrix} \tag{10}$$

为了处理方便可将上式转化为求下面的最小值

$$-\ln\delta_{i+1}(t^j)=-\begin{bmatrix}\ln\delta_i(t^k)+\ln P(w_{i+1}\mid t^j)\\ +\ln P(t^j\mid t^k)+\ln P(w_{i+1})\end{bmatrix}=$$
$$-\begin{bmatrix}\ln\delta_i(t^k)+\ln\dfrac{C(w_{i+1},t^j)}{C(t^j)}+\ln\dfrac{C(t^k,t^j)}{C(t^k)}\\ +\ln\dfrac{C(t^k,t^j)}{C(t^k)}+\ln\dfrac{C(w_{i+1})}{\sum_{i=0}^{n-1}C(w_{i+1})}\end{bmatrix} \tag{11}$$
$$=\left[\ln C(t^j)+\ln C(t^k)+\ln\sum_{i=0}^{n-1}C(w_{i+1})\right]-\begin{bmatrix}\ln\delta_i(t^k)+\ln C(w_{i+1},t^j)\\ +\ln C(t^k,t^j)+\ln C(w_{i+1})\end{bmatrix}$$

将此推导过程融合在串行全切分算法的过程中，得到下面切分标注一体化的求解算法：首先要考虑一下串行全切分算法的切分分支数据需要保存的数据项:

$$x=\left[\mathrm{W},\ \mathrm{pos},\ \min\left(-\mathrm{In}\delta_{i+1}(t^j)\right)\right]$$

其中 w 表示该切分分支上的汉词标记串，第二项为当前的切分位置，第三项保存的是当前词的 min(-In $\delta_{i+1}(t^j)$)值。设待切分标记的字串为 S,则可以归结该方法为如下步骤：

① 置起始搜索位置 pos=1,将

$x=\left[\text{""},\ \mathrm{pos},\ \min\left(-\mathrm{In}\delta_1(t^j)\right)\right]$压栈；

② 若栈为空，则退出；否则转③；

③ 弹出栈顶元素 x,设:

$x=\left[\mathrm{W},\ \mathrm{pos},\ \min\left(-\mathrm{In}\delta_1(t^j)\right)\right]$;

④ 若 pos>=strlen(S)，则输出 w,转②；否则，转⑤；

⑤ 调用搜索算法 search(pos)，得串首词集合 fwc，设 n 为集合 fwc 的元素个数；

⑥ 若 n>0；则对 fwc 的任意一个元素 fwi+1,，按照公式(11)计算 x 中第三项的值，即 min(-in $\delta_{i+1}(t^j)$)，并计算使得该值 最小的前一词所处的最可能状态(t_i),将 x=[w+fwi+1,pos+strlen(fwi+1),min(-in $\delta_{i+1}(t^j)$)]压栈，转②。

3.5 实验结果与分析

本文在同等条件下，针对训练的语料库全文句子总数为 201,103。将本文的分词标注一体化方法与常用的方法如最大匹配、最大概率、N-最短路径做了分词标注测试比较，将切分标注的句子结果进行了统计，其数据如表 1 所示。由数据分析可见，本文提出的分词标注方法明显高于最大匹配、最大概率、最短路径+隐马尔可夫三种方法，由于本文考虑了所有分词结果，因此在切分上要高于 N-最短路径方法，最终在句子召回率上高于其他方法。

表 1　与常用分词标注方法的对比实验结果

方法	正确切分的句子	正确标注的句子	句子召回率
最大匹配规则分词标注	172,798	124,155	61.43%
最大概率规则标注	186,885	134,557	66.58%
最短路径分词+隐马尔可夫标注	182,297	175,370	86.77%
2-最短路径分词+隐马尔可夫标注	201,436	194,385	96.18%
分词标注一体化方法	202,103	195,232	96.60%

4　讨论

本文从最大限度的获得正确结果的角度出发，提出了基于统计的分词标注方法，首先介绍了基于训练语料 的词典结构模型，并介绍了本研究过程中使用的改进的词典索引模型双数组 Trie 树，在综合了各种分词算法和标注模型的基础上，提出了一种基于概率全切分和马尔可夫模型的分词标注算法。该算法的优点是获得比以往的算法更高的准确率。并为了减少同一词为了获得词条属性频繁索引词典的次数，采用了分词、标注两部分工作同时进行的策略。

参考文献

1 张华平,刘群.基于 N-最短路径方法的中文词语粗分模型.中文信息学报, 2002,16(5):24－30.

2 王思力,张华平,王斌.双数组 Trie 树算法优化及其应用研究.中文信息学报, 2006,20(5):24－30.

3 冯志伟.中文信息处理与汉语研究.北京:商务印书馆, 1992.18－24.

4 张会鹏.中文词法分析技术的研究与实现[硕士学位论文].哈尔滨:哈尔滨工业大学, 2006.

5 于源,衣袭.中文全切分快速分词方法.大连铁道学院学报, 2005,26(2):84－85.

6 苑春法,李庆中,王昀,李伟,曹德芳.统计自然语言处理基础.北京:电子工业出版社, 2007.216－225.