中 文 信 息 学 报
第 16 卷 第 5 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 16 No. 5

基于 N－最短路径方法的中文词语粗分模型

张华平 刘 群

（中国科学院计算技术研究所软件实验室 北京 100080）

摘要：预处理过程的词语粗切分，是整个中文词语分析的基础环节，对最终的召回率、准确率、运行效率起着重要的作用。词语粗分必须能为后续的过程提供少量的、高召回率的、中间结果。本文提出了一种基于 N－最短路径方法的粗分模型，旨在兼顾高召回率和高效率。在此基础上，引入了词频的统计数据，对原有模型进行改进，建立了更实用的统计模型。针对人民日报一个月的语料库（共计 185,192 个句子），作者进行了粗分实验。按句子进行统计，2－最短路径非统计粗分模型的召回率为 99.73%；在 10－最短路径统计粗分模型中，平均 6.12 个粗分结果得到的召回率高达 99.94%，比最大匹配方法高出 15%，比以前最好的切词方法至少高出 6.4%。而粗分结果数的平均值较全切分减少了 64 倍。实验结果表明：N－最短路径方法是一种预处理过程中实用、有效的的词语粗分手段。

关键词：N－最短路径方法；粗分；中文词语分析

中图分类号：TP391.2

Model of Chinese Words Rough Segmentation Based on N-Shortest-Paths Method

ZHANG Hua-ping LIU Qun

(Software division Institute of Computing Technology The Chinese Academy of Sciences Beijing 100080 China)

Abstract: As the very first step of Chinese word segmentation, rough segmentation tries to cover the correct segmentation with as few candidates as possible. This paper presents a model of rough segmentation, which is based on the N-shortest-paths method, to achieve the goal. In parallel, a statistical model can easily be obtained by attaching frequencies to the edges of the word-graphs. Experiments have been made on a one-month news corpus of 185,192 sentences from the People's Daily. By sentence, the recalling rate of the non-statistical model based on 2-shortest-paths method is 99.73%. When the statistical model is applied, a recalling rate as high as 99.94%, nearly 6.4% higher than known best approach and 15% higher than the maximum matching segmentation, can be reached with 6.12 candidates on average. In addition, the average number of segmentation candidates is reduced by 64 times as compared to the approach of full segmentation. The result shows that the N-shortest-paths method is effective for the task of rough segmentation.

Key words: N-shortest paths method; words rough segmentation; Chinese lexical analysis

一、引言

“词是最小的能够独立活动的有意义的语言成分”[1]，但汉语是以字为基本的书写单位，

* **收稿日期**：2001－12－18

基金项目：国家重点基础研究项目（G1998030507－4、G1998030510）.

作者**张华平**，男，1978 年生，硕士生，主要研究方向为自然语言处理与中文词语分析. **刘群**，男，1966 年生，在职博士生，副研究员，主要研究方向为机器翻译，自然语言处理与中文信息处理.

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。而中文词语分析一般包括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。目前中文词语分析采取的主要步骤是:先采取最大匹配、最短路径、概率统计方法、全切分等方法,得到一个相对最好的粗分结果,然后进行排歧、未登录词识别,最后标注词性。例如:北大计算语言所分词系统采用了统计方法进行词语粗分[2,3,5],北航1983年的CDWS系统则采用了正向或逆向最大匹配方法[4,5],而清华大学的SEGTAG系统采用的是全切分方法[5]。在实际的系统中,这三个过程可能相互交叉、反复融合,也可能不存在明显的先后次序。

预处理过程产生的粗切分结果是后续过程的处理对象,粗分结果的准确性与包容性(即必须涵盖正确结果),直接影响系统最终的准确率、召回率。预处理得到的粗分结果一旦不能成功召回正确的结果,后续处理一般很难补救,最终的词语分析结果必然会导致错误,粗分结果的召回率往往是整个词语分析召回率的上限。同时,粗分结果集的大小也决定了后续处理的搜索空间与时间效率,最终也会影响系统的运行效率。因此,词语粗分是后续处理的基础和前提,其关键在于如何以尽量高的效率寻找数量极少、涵盖最终结果的粗分结果集。

我们采取当前常用的粗分方法,对大规模真实语料库的进行测试实验,词语粗切分的召回率均不足93.50%。因此,改进预处理过程中的汉语词语粗分方法,是提高排歧、未登录词识别、词性标注最终效果的基础性措施,也是提高中文词语分析质量的重要途径。

本文提出了一种旨在提高召回率同时兼顾准确率的词语粗分模型——基于N-最短路径方法的中文词语粗分模型。根据我们针对大规模真实语料库的对比测试,粗分结果的召回率有较大提高,模型的运行效率也令人满意,该方法是行之有效的。本文第二节将系统描述非统计模型的基本思想与实现,然后加入词频信息,得到N-最短路径的一元统计模型,最后给出对比实验的结果及分析。

二、基于N-最短路径的非统计粗分模型

粗切分的目标是快速(粗分结果集尽量少)、高召回率(即可能的涵盖最终结果)。一个很直接的研究思路是先快速的找出包含正确结果在内的N(N≥1)种粗分结果。然后综合考虑速度和召回率,通过试验,确定N的最佳值,最终得到涵盖最终结果在内的尽量小的粗分结果集。

2.1 基本思想

我们采取的是最短路径的改进方法——N-最短路径方法。其基本思想是根据词典,找出字串中所有可能的词,构造词语切分有向无环图。每个词对应图中的一条有向边,并赋给相应的边长(权值)。然后针对该切分图,在起点到终点的所有路径中,求出长度值按严格升序排列(任何两个不同位置上的值一定不等,下同)依次为第1,第2,…,第i,…,第N的路径集合作为相应的粗分结果集。如果两条或两条以上路径长度相等,那么他们的长度并列第i,都要列入粗分结果集,而且不影响其他路径的排列序号,最后的粗分结果集合大小大于或等于N。

2.2 模型求解

设待分字串 $S=c_1,c_2,\cdots,c_n$,其中 $c_i(i=1,2,\cdots,n)$ 为单个的字,n 为串的长度,$n\geqslant 1$。建立一个节点数为 $n+1$ 的切分有向无环图 G,各节点编号依次为 $V_0,V_1,V_2,\cdots,V_n$。

通过以下两种方法建立 G 所有可能的词边。

(1)相邻节点 V_{k-1},V_k 之间建立有向边 $<V_{k-1},V_k>$,边的长度值为 L_k,边对应的词默认为 $c_k(k=1,2,\cdots,n)$

(2)若 $w=c_i,c_{i+1},\cdots,c_j$ 是一个词,则节点 V_{i-1},V_j 之间建立有向边$<V_{i-1},V_j>$,边的长度值为 L_w,边对应的词为 $w(0<i<j\leqslant n)$

这样,待分字串 S 中包含的所有词与切分有向无环图 G 中的边一一对应。如图 1 所示:

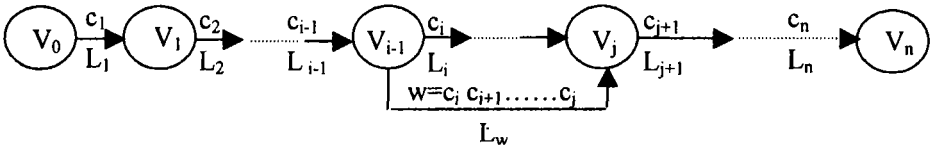

图 1　切分有向无环图

在非统计粗分模型中,我们假定所有的词都是对等的,为了计算方便,不妨将词的对应边长的边长均设为 1。

设 NSP 为 V_0 到 V_n 的 N－最短路径集合;而 RS 是最终的 N－最短路径粗分结果集,则 RS 是 NSP 对应的分词结果,即我们所求的粗分结果集。因此,N－最短路径方法词语粗切问题转化为如何求解有向无环图 G 的集合 NSP。

2.3　N－最短路径求解与复杂度分析

求解有向无环图 G 的集合 NSP,可以采取贪心方法[6]。我们使用的算法是基于 Dijkstra[3]的一种简单扩展。改进的地方在于每个节点处记录 N 个最短路径值,并记录相应路径上当前节点的前驱。如果同一长度对应多条路径,必须同时记录这些路径上当前节点的前驱。最后通过回溯即可求出 NSP。

我们以“他说的确实在理”为例,给出了 3－最短路径的求解过程。如图 2 所示。

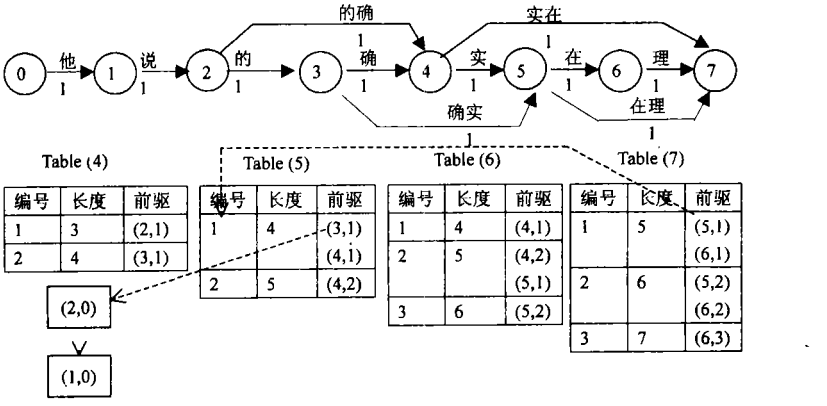

图 2　“他说的确实在理”的求解过程示例(N=3)

(其中虚线是回溯出的是第一条最短路径,对应的粗分结果为:“他/说/的/确实/在理/”)

Dijkstra 算法的时间复杂度为 $O(n^2)$,它求的是图中所有点到单源点的最短路径,而在切分有向图中的应用,有 2 个本质区别:1)有向边的源点编号均小于终点编号,即所有边的方向一致;2)我们求的是有向图首尾节点的 N－最短路径。所以我们使用的算法中,运行时间与 n(字串长度)、N(最短路径数)以及某个字作为末字的平均次数 k(等于总词数除以末字总数,对应的是切分图中结点入度的平均值)成正比。整个算法的时间复杂度是 $O(n*N*k)$。

三、基于 N－最短路径的统计粗分模型

在非统计模型构建粗切有向无环图有向边的过程中,我们给每个词的对应边长度赋值为 1。随着字串长度 n 和最短路径数 N 的增大,长度相同的路径数急剧增加,同时粗切分结果数

量必然上升。例如,N=2 时,句子"江泽民在北京人民大会堂会见参加全国法院工作会议和全国法院系统打击经济犯罪先进集体表彰大会代表时要求大家要充分认识打击经济犯罪工作的艰巨性和长期性"的粗切结果居然有 138 种之多。大量的切分结果对后期的处理,以及整个性能的提高是非常不利的。

3.1 基本原理

假定一个词串 W 经过信道传送,由于噪声干扰而丢失了词界的切分标志,到输出端便成了汉字串 C,这是一个典型的噪声-信道问题[7]。N-最短路径方法词语粗分模型可以相应的改进为求取 W,使得概率 $P(W|C)$的值是最大 N 个。为了简化 $P(W|C)$的计算,我们采用的是一元统计模型,即只引入词频并假定词与词之间是相互独立。基于以上分析,我们引入词 w_i 的词频信息 $P(w_i)$,对模型进行了改进,得到一个基于 N-最短路径的一元统计模型。

3.2 一元统计粗分模型的求解与实现

$P(W|C)$的求解如: $P(W|C)=P(W)P(C|W)/P(C)$ (1)

其中,$P(C)$是汉字串的概率,它是一个常数,不必考虑。从词串恢复到汉字串的概率 $P(C|W)=1$(只有唯一的一种方式)。因此,我们的目标就是确定 $P(W)$最大 N 种的切分结果集合。

$W=w_1,w_2,\cdots,w_m$ 是字串 $S=c_1,c_2,\cdots,c_n$ 的一种切分结果。w_i 是一个词,$P(w_i)$表示 w_i 的出现的概率。在大规模语料库训练的基础上,根据大数定理[8],即:在大样本统计的前提下,样本的频率接近于其概率值。所以 $P(w_i)$的极大似然估计值[9]等于词频,有:

$$P(w_i)\approx k_i/\sum_{j=0}^{m}k_j \quad (\text{其中 } k_i \text{ 为 } w_i \text{ 在训练样本中出现的次数}) \tag{2}$$

粗切分阶段,为了简单处理,我们仅仅采取了概率上下文无关文法[10],即假设上下文无关,词与词之间相互独立。因此,根据(1)、(2),我们可以得到:

$$\text{则 } W \text{ 的联合概率} P(W)=\prod_{i=1}^{m}P(w_i)\approx\prod_{i=1}^{m}(k_i/\sum_{j=0}^{m}k_j) \tag{3}$$

为了处理的方便,令 $P*(W)=-\ln P(W)=\sum_{i=0}^{m}[-\ln P(w_i)]\approx\sum_{i=0}^{m}[-\ln(k_i/\sum_{j=0}^{m}k_j)]$

$$=\sum_{i=0}^{m}[\ln(\sum_{j=0}^{m}k_j)-\ln k_i] \tag{4}$$

那么就可以将(3)极大值的问题转化为求解(4)极小值的问题。适当修改切分有向无环图 G 边的长度(加 1 主要是为了数据的简单平滑处理):

1)* $<V_{k-1},V_k>$的长度值 $L_k=-\ln(0+1),(k=1,2,\cdots,n)$

2)* $w=c_i,c_{i+1},\cdots,c_j$ 对应的有向边为$<V_{i-1},V_j>$,

其长度值 $L_w=\ln(\sum_{j=0}^{m}k_j+m)-\ln(k_i+1)$

针对修改边长后的切分有向无环图 G*,使用 2.3 中的算法,就可实现问题的最终求解。

四、与常用方法对比分析

分词问题的研究已经比较成熟,常用的方法主要有:最大匹配(包括向前、向后以及前后相结合)、最短路径方法(切分出来的词数最少)、全切分方法(列出所有可能的分词结果)、以及最大概率方法(训练一个一元语言模型,通过计算,得到一个概率最大的分词结果。原理和 3.1 类似)。下面针对各自的优缺点,对比分析如下:

1.最大匹配分词是一种纯粹基于规则的方法,简单有效。在没有大规模预先切分好的熟语料的情况下,是唯一行之有效的方法。但是该方法仅仅是从最大匹配的角度出发,很多问题无法解决,如交叉歧义、组合歧义。最终的准确率不会太高,预处理的粗分过程一旦采用最大匹配方法,后期处理必须做很多补救措施,才能保证最终的分词质量。另外一个不足在于它缺少合理的评分机制,我们就很难再选出一个次优的切分结果。

2.最短路径方法采取的规则是使切分出来的词数最少,符合汉语自身的语言规律。可以取得较好的效果,但是同样不能正确切分许多不完全符合规则的句子。如果最短路径有多条,往往只保留其中一个结果,这对其他同样符合要求的路径时不公平的,也缺乏理论根据。

3.全切分方法列举出所有可能的切分结果,避免在粗分过程中就出现切分错误,将优选排错的任务交给后续过程,有一定合理性。但是,全切分产生的切分结果数随着句子长度的增大而成指数级增大,其中大多数是无效结果,对正确结果的生成没有太大帮助。无论是求取所有切分结果,还是后续过程对大量结果的分析处理都是非常困难而且费时。因此该方法和实际需求还有一定差距,实用性不强。

4.最大概率分词方法[7]的根据是:联合概率(各个词的词频相乘)最大的词串就是最终的切分结果,是一种效果较好的分词方法。最大概率分词方法实质上是一种简单变形的最短路径方法,改进的地方在于它的切分有向图的边长等于词频。

5.N-最短路径方法实际上是最短路径方法和全切分的有机结合。该方法的出发点是尽量减少切分出来的词数,这和最短路径分词方法是完全一致的;同时又要尽可能的包含最终结果,这和全切分的思想是共通的。通过这种综合,一方面避免了最短路径分词方法大量舍弃正确结果的可能,另一方面又大大解决了全切分搜索空间过大,运行效率差的弊端。同时我们还可以看到最短路径方法和全切分方法分别是N-最短路径方法在N=1(而且只能选择唯一的路径)和N=∞时的退化。N-最短路径一元统计方法在N=1(而且只能选择唯一的路径)时就退化为最大概率分词方法。

作为预处理阶段的一种粗分方法,N-最短路径方法的优势还体现在它的包容性,该方法通过保留少量大概率的粗分结果,可以最大限度地包容正确结果,粗分结果仅仅是解决粗分阶段能解决的一些问题,而将歧义字段、未登录词等问题,尽量保留给下一阶段专门处理。常用方法共同的弊端就在于保留一个自己认为最优的结果,而这一结果往往会因为歧义或未登录词问题而丢失正确结果。例如,用最大匹配或最短路径方法粗分"结合成分子时",得到的结果均为"结合/成分/子时"。显然,由于交叉歧义的存在,这两种方法在粗分阶段就抹杀了事实上的交叉歧义,导致最终的错误,而采取2-最短路径方法就完全可以将正确的结果"结合/成/分子/时"成功召回。

N-最短路径方法相对的不足就是粗分结果不唯一,后续过程需要处理多个粗分结果。但是,对于预处理过程来讲,粗分结果的高召回率至关重要。因为低召回率就意味着没有办法再作后续的补救措施。预处理一旦出错,后续处理只能是一错再错,基本上得不到正确的最终结果。而少量的粗分结果对后续过程的运行效率影响不会太大,后续处理可以进一步优选排错,如词性标注、句法分析等。

五、实验及结果分析

采用80812个词条的词典,针对已经切分标注好的人民日报一个月语料库(共计185,192个句子),我们作了三组实验。实验一、实验二分别采用N-最短路径非统计粗分模型和统计

粗分模型，变换 N，进行粗分实验，实验三是 N－最短路径粗分模型与常用方法的对比实验。

5.1 句子正确粗分的评价标准

1.如果句子中不存在未登录词，粗分结果与语料库中给出的参考结果完全匹配才认为是正确的。

2.存在未登录词的情况，正确的粗分结果必须满足 2 个条件：

1)粗分结果中除未登录词外的其它部分与参考结果必须完全一致；

2)未登录词部分的字串必须可以组合还原成参考结果中对应的未登录词。这样可以保证不影响后续处理过程识别出正确的结果，保证可以召回。

例如："安徽省合肥市长江路"被粗切成"安徽省／合肥市／长江／路／"我们认为是正确的，因为并不会影响后续过程将正确结果("安徽省／合肥市／长江路")识别出来。但是"尉健行李岚清"被粗切成"尉／健／行李／岚／清／"就不对，因为正确结果("尉／健行／李／岚清／")无法由粗切结果召回。

5.2 N－最短路径方法的粗分实验

实验测试结果(被测试的句子总数为 185,192)见表 1、表 2。

表 1 非统计粗分模型的实验结果

N	粗分结果数	正确粗分的句子数	句子召回率
1	1	169,992	91.80%
1	2	175,283	94.65%
1	4	175,598	94.28%
1	8	175,612	94.83%
2	2	175,873	96.59%
2	4	182,299	94.44%
2	16	183,991	99.35%
2	48	184,684	99.73%

表 2 表统计粗分模型的实验结果

N	粗分结果数	正确粗分的句子数	句子召回率
1	1	173,156	93.50%
2	2	182,251	98.42%
3	3	183,819	99.26%
4	4	184,463	99.61%
5	5	184,742	99.75%
6	6	184,876	99.83%
8	8	184,997	99.89%
10	10	185,078	99.94%

其中，粗分召回率＝正确粗分的句子数/句子总数＊100%。(下同)

实验一、二表明了 N－最短路径方法可以取得较好的句子召回率，都超过了 99.50%。由于统计模型加入了词频信息，能更快速的得到数量更少、召回率更高的粗分结果集。比非统计模型更加实用，但是它需要事先对语料库进行训练，得到带有词频的概率词典。对于缺少大规模熟语料库的使用者来说，非统计粗分模型也不失为一种非常理想的粗分方法。

5.3 与常用方法的对比测试

我们利用同一核心词典，分别采取最大匹配方法、最短路径方法、最大概率方法、全切分方法、2－最短路径非统计粗分模型、10－最短路径统计粗分模型针对同一熟语料库(句子总数为 185,192)进行粗分实验。对比结果如表 3 所示。对比实验的数据表明：

1.N－最短路径方法的句子召回率比目前最好的方法至少高出 6.4%，与最大匹配比较，句子召回率提高了将近 15%。

原因在于，N－最短路径模型粗切结果的数量一般大于 N 个(统计模型一般为 N 个，非统计模型的粗切结果数更大)，而且一旦第 1 个结果没有召回，后续的结果可以高概率的将正确结果召回，例如：在第 1 个粗分结果没有成功召回正确结果的前提下，非统计模型的第 2 个结果将其召回的概率为 58.43%，而此时统计模型为 75.57%。

2. 运行效率大大提高

与全切分方法相比,2－最短路径非统计粗分模型的召回率低了0.27%,但是每句的切分结果平均数仅仅是全切分方法的1/90。10－最短路径非统计粗分模型的召回率损失仅仅是0.06%。但粗分结果平均数是全切分的1/64。

表3 与常用粗分方法的对比实验结果

方 法	一个句子粗分结果数的最大值	每句粗分结果数的平均值	正确粗分的句子数	句子召回率
最大匹配	1	1	158,263	85.46%
最短路径	1	1	169,992	91.80%
最大概率	1	1	173,156	93.50%
全切分	>3,424,507	>391.79	185,192	100.00%
2－最短路径非统计粗分	164	4.40	184,684	99.73%
10－最短路径统计粗分	20	6.12	185,078	99.94%

六、结论

汉语预处理过程的粗切分,是整个词法分析过程的基础环节,对系统最终的召回率、准确率、运行效率起着重要的作用。本文综合最短路径方法与全切分方法,提出了一种基于N－最短路径方法的粗切分方法,N=2时,非统计粗切模型句子召回率达到了99.73%。在原有模型的基础上,进一步加入了词频信息,建立统计模型,使整个系统搜索空间和最终的结果数量锐减,运行效率得到了进一步提高,粗切召回率也有较好的改善。统计模型在N=8时,句子粗切召回率已经达到了99.90%。我们下一步的工作将引入规则以及优化函数,对数量不多的、高召回率的粗分结果进行未登录词的识别和其他的相应处理,集中地提高准确率。最终采取一体化方法,利用隐马模型选取切分及标注的最佳结果,实现中文词语的系统分析。

致谢 感谢北京大学计算语言所提供的2个月《人民日报》语料库!中国科学院计算技术研究所软件室的程学旗主任对本文的工作给予了细心的指导,张浩、李继锋、李素建、王长胜、邹纲对本文的完成提出了很多有益的建议,在此一并表示感谢。

参 考 文 献

[1] 朱德熙．语法讲义．北京:商务印书馆,1982

[2] 周强．规则与统计相结合的汉语词类标注方法．中文信息学报,1995,9(2):1-10

[3] 周强,俞士汶．一种切分与词性标注相融合的汉语语料库多级处理方法．计算语言学研究与应用,北京:北京语言学院出版社,1993

[4] 梁南元．书面汉语自动分词系统-CDWS. 中文信息学报,1987,1(2);44－52

[5] 孙斌．切分歧义字段的综合性分级处理方法--北京大学计算语言学研究所讨论班;99.4.13;

[6] 余祥宣,崔国华,邹海明．计算机算法基础．武汉:华中理工大学出版社,2000,67-87

[7] 陈小荷．现代汉语自动分析．北京:北京语言文化大学出版社,2000,97-98

[8] Yuan S. C, Henry T., Probability Theory, Springer-Verlag New York Inc., 1978, 324-338

[9] Christopher D. Manning, Hinrich S., Foundations of statistical natural language processing, MIT press, 1999, 197-202

[10] 翁富良,王野翊．计算语言学导论．北京:中国社会科学出版社,1998,136-145