基于层叠隐马模型的汉语词法分析

刘 群[1,3] 张华平[1,2] 俞鸿魁[1] 程学旗[1]
[1](中国科学院计算技术研究所 北京 100080)
[2](中国科学院研究生院 北京 100039)
[3](北京大学计算语言学研究所 北京 100871)
(liuqun@ ict. ac. cn)

摘 要 提出了一种基于层叠隐马模型的汉语词法分析方法, 旨在将汉语分词、词性标注、切分排歧和未登录词识别集成到一个完整的理论框架中. 在分词方面, 采取的是基于类的隐马模型, 在这层隐马模型中, 未登录词和词典中收录的普通词一样处理. 未登录词识别引入了角色 HMM: Viterbi 算法标注出全局最优的角色序列, 然后在角色序列的基础上, 识别出未登录词, 并计算出真实的可信度. 在切分排歧方面, 提出了一种基于 N-最短路径的策略, 即: 在早期阶段召回 N 个最佳结果作为候选集, 目的是覆盖尽可能多的歧义字段, 最终的结果会在未登录词识别和词性标注之后, 从 N 个最有潜力的候选结果中选优得到. 不同层面的实验表明, 层叠隐马模型的各个层面对汉语词法分析都发挥了积极的作用. 实现了基于层叠隐马模型的汉语词法分析系统 ICTCLAS, 该系统在 2002 年的"九七三"专家组评测中获得第 1 名, 在 2003 年汉语特别兴趣研究组(ACL Special Interest Group on Chinese Language Processing, SIGHAN)组织的第 1 届国际汉语分词大赛中综合得分获得两项第 1 名、一项第 2 名. 这表明: ICTCLAS 是目前最好的汉语词法分析系统之一, 层叠隐马模型能够解决好汉语词法问题.

关键词 汉语词法分析; 分词; 词性标注; 未登录词识别; 层叠隐马模型; ICTCLAS

中图法分类号 TP391. 1; TP391. 2

Chinese Lexical Analysis Using Cascaded Hidden Markov Model

LIU Qun[1,3], ZHANG Hua-Ping[1,2], YU Hong-Kui[1], and CHENG Xue-Qi[1]
[1]*(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)*
[2]*(Graduate School of the Chinese Academy of Sciences, Beijing 100039)*
[3]*(Institute of Computational Linguistics, Peking University, Beijing 100871)*

Abstract This paper presents an approach for Chinese lexical analysis using cascaded hidden Markov model (CHMM), which aims to incorporate Chinese word segmentation, part-of-speech tagging, disambiguation and unknown words recognition into an integrated theoretical frame. A class-based HMM is applied in word segmentation, and in this model, unknown words are treated in the same way as common words listed in the lexicon. Unknown words are recognized with reliability on roles sequence tagged using Viterbi algorithm in roles HMM. As for disambiguation, the authors bring forth an n-shortest-path strategy that, in the early stage, reserves the top N segmentation results as candidates and covers more ambiguity. Various experiments show that each level in the CHMM contributes to Chinese lexical analysis. A CHMM-based system ICTCLAS is accomplished. The system ranked top in the official open evaluation, which was held by the "973" project in 2002. And ICTCLAS achieved 2 first ranks and 1 second rank in the first international word segmentation bakeoff held by SIGHAN (the ACL Special Interest Group on Chinese Language Processing) in 2003. It indicates that ICTCLAS is one of the best Chinese lexical analyzers. In a word,

收稿日期: 2003-03-04; 修回日期: 2003-08-26
基金项目: 国家"九七三"重点基础研究发展规划项目(G1998030507-4, G1998030510); 中国科学院计算技术研究所领域前沿青年基金项目(20026180-23)

CHMM is effective for Chinese lexical analysis.

Key words Chinese lexical analysis; word segmentation; POS tagging; unknown words recognition; cascaded hidden Markov model; ICTCLAS

1 引 言

词是最小的能够独立活动的有意义的语言成分[1]. 在汉语中, 词与词之间不存在分隔符, 词本身也缺乏明显的形态标记, 因此, 中文信息处理的特有问题就是如何将汉语的字串分割为合理的词语序列, 即汉语分词. 汉语分词是句法分析等深层处理的基础, 也是机器翻译、信息检索和信息抽取等应用的重要环节.

从 1983 年第一个实用分词系统 CDWS[2] 的诞生到现在, 国内外的研究者在汉语分词方面进行了广泛的研究, 提出了很多有效的算法. 我们可以粗略地将这些方法分为两大类: 第 1 类是基于语言学知识的规则方法, 如: 各种形态的最大匹配、最少切分方法、以及综合了最大匹配和最少切分的 *N*-最短路径方法[3], 还有的研究者引入了错误驱动机制[4], 甚至是深层的句法分析[5]. 第 2 类是基于大规模语料库的机器学习方法[6~8], 这是目前应用比较广泛、效果较好的解决方案. 用到的统计模型有 *N* 元语言模型、信道-噪声模型、最大期望[9]、隐马模型等. 在实际的分词系统中, 往往是规则与统计等多类方法的综合. 一方面, 规则方法结合使用频率, 形成了可训练的规则方法[6]; 另一方面, 统计方法往往会自觉不自觉地采用一些规则排除歧义、识别数词、时间及其他未登录词. 同时, 我们也注意到国外同行提出的一些探索性算法, 如: 基于压缩的方法[10], 分类器的方法[11], 无词典的自监督学习方法[12]. 目前, 对不含歧义和未登录词[1] 的文本进行分词, 已有方法和系统表现相当出色, 其准确性已经达到相当高的水平.

实际上, 汉语分词的主要瓶颈是切分排歧和未登录词识别. 切分歧义和未登录词降低了自身正确切分的可能性, 同时还干扰了其相邻词的正确处理. 更糟糕的是, 未登录词往往和切分歧义交织在一起, 进一步增加了分词的难度. 如: 在“克林顿对内塔尼亚胡说”中,“内塔尼亚胡”是一个词典中没有收录的译名, 实际切分的时候,“对”与“内”,“胡”与“说”往往会粘在一起, 最终导致错误的切分结果:“克林顿/ 对内/ 塔尼亚/ 胡说/ ”.

文献[13] 对切分歧义进行了较好的形式化描述, 并引入了嘈杂度的概念对歧义的程度进行了定量的描述. 目前切分排岐的研究路线基本上以规则为主[14~16], 还有的针对某一类歧义, 引入了一些成熟的模型作专门处理, 如引入向量空间解决组合歧义问题[17]. 在未登录词识别方面, 主要的出发点是综合利用未登录词内部构成规律及其上下文信息. 未登录词识别处理的对象主要是人名、地名、译名和机构名等命名实体. 在语料库不足的条件下, 未登录词识别惟一的出路是采取精细的规则[18~21], 规则一般来源于观察到的语言现象或者是大规模的专名库. 目前比较成功的解决方案大都是从大规模的真实语料库中进行机器学习, 解决方案有隐马模型[22,23]、基于 Agent 的方法[24]、基于类的三元语言模型[25]等.

经过 20 余年的努力, 研究者们在分词算法、切分排歧和未登录词识别方面均取得了较大的进展. 然而, 现有方法和系统往往缺乏一个相对统一的模型框架将 3 者进行有机地融合. 排歧、未登录词识别往往和分词相对独立, 排歧的结果和识别出的未登录词缺乏科学的可信度计算, 即使量化, 往往流于经验, 很难在量值上与普通词作真正意义上的比较. 一般都倾向于假定排歧和未登录词结果正确无误, 忽略具体的分词算法, 直接修正分词结果. 现有的分词方法更大程度上是专门切分出词典中收录的词, 基本上没有将未登录词和歧义纳入到同一个算法体系当中, 一旦遇到歧义或者未登录词就作为特例进行专门处理, 因此使用的模型和方法都没有贯彻到底, 缺乏统一的处理算法, 对切分结果也缺乏统一的评估体系. 最终导致分词的准确率在开放测试的条件下并不像宣称的那样理想, 处理含有未登录词、歧义字段的真实文本时, 效果更是不尽人意.

本文提出了一种基于层叠隐马模型的方法, 旨在将汉语分词、切分排歧、未登录词识别、词性标注等词法分析任务融合到一个相对统一的理论模型中. 首先, 在预处理的阶段, 采取 *N*-最短路径粗分方

[1] 我们这里所说的未登录词指的是核心词典中没有收录而又不能用正则表达式描述的词, 如没有被收录的人名、地名.

法, 快速地得到能覆盖歧义的最佳 N 个粗切分结果; 随后, 在粗分结果集上, 采用低层隐马模型识别出普通无嵌套的人名、地名, 并依次采取高层隐马模型识别出嵌套了人名、地名的复杂地名和机构名; 然后将识别出的未登录词以科学计算出来的概率加入到基于类的切分隐马模型中, 未登录词与歧义均不作为特例, 与普通词一起参与各种候选结果的竞争. 最后在全局最优的分词结果上进行词性的隐马标注. 该方法已经应用到了中科院计算所汉语词法分析系统 ICTCLAS 中, 取得了较好的分词和标注效果. ICTCLAS 在"九七三"专家组机器翻译第 2 阶段的评测和 2003 年 5 月 SIGHAN 举办的第 1 届汉语分词大赛中, 取得了不俗的成绩, 是目前最好的汉语词法分析系统之一.

本文第 2 节将概述层叠隐马模型和汉语词法分析的总体框架, 随后介绍基于类的切分隐马模型; 然后分别叙述基于角色隐马模型的未登录词识别方法, 以及切分排歧的 N-最短路径粗切分策略, 最后给出各种条件下的对比测试结果, 以及国家"九七三"开放评测和国际分词大赛的测试结果, 并给出简单分析.

2　层叠隐马模型和汉语词法分析

2.1　层叠隐马模型概述

隐马模型(hidden Markov model; HMM)[26] 是经典的描述随机过程的统计方法, 在自然语言处理中得到了广泛的应用. 然而, 相对于复杂的自然语言现象来说, 传统的 HMM 仍然略显简单, 为此, 我们需要采用多个层次的隐马模型对汉语词法分析中遇到的不同情况进行分别处理. 文献[27]提出的层次隐马模型(hierarchical hidden Markov model, HHMM)的思想. 在 HHMM 中, 有多个状态层和一个输出层. 每一个上一层状态都对应于若干个下一层的子状态, 而每个状态的子状态的分布都是不同的, 由一个隶属于该状态的初始子状态概率矩阵和子状态转移概率矩阵所决定. 最底层状态通过一个输出概率矩阵输出到观察值. HHMM 实际上是一种不同于 HMM 的更复杂的数学模型, 并且具有比 HMM 更强的表达能力, 不过使用起来时空开销也比较大. HHMM 的解码问题求解的时间复杂度是 $O(NT^3)$, 而 HMM 的解码问题求解的时间复杂度只有 $O(NT)$. 本文采用的也是一种多层隐马尔可夫模型, 称为层叠隐马尔可夫模型(cascaded hidden Markov model, CHMM). 不同于 HHMM 的是, CHMM 实际上是若干个层次的简单 HMM 的组合, 各层隐马尔可夫模型之间以下几种方式互相关联, 形成一种紧密的耦合关系: 各层 HMM 之间共享一个切分词图作为公共数据结构; 每一层隐马尔可夫模型都采用 N-Best 策略, 将产生的最好的若干个结果送到词图中供更高层次的模型使用; 低层的 HMM 在向高层的 HMM 提供数据的同时, 也为这些数据的参数估计提供支持. 整个系统的时间复杂度与 HMM 相同, 仍然是 $O(NT)$, 与句子的长度成线性关系, 速度非常快. 所有各层隐马模型都采用《人民日报》标注语料库作为训练语料库, 通过对该语料库进行不同形式的改造以适应各层隐马尔可夫模型的使用, 而这种改造绝大部分都是自动进行的, 只需要介入很少量的人工校对.

2.2　基于 CHMM 的汉语词法分析框架

针对汉语词法分析各个层面的处理对象及问题特点, 我们引入 CHMM 统一建模, 该模型包含原子切分、普通未登录词识别、嵌套的复杂未登录词识别、基于类的隐马切分、词类标注共 5 个层面的隐马模型, 如图 1 所示. 其中, N-最短路径粗切分可以快速产生 N 个最好的粗切分结果, 粗切分结果集能覆盖尽可能多的歧义. 在整个词法分析架构中, 二元

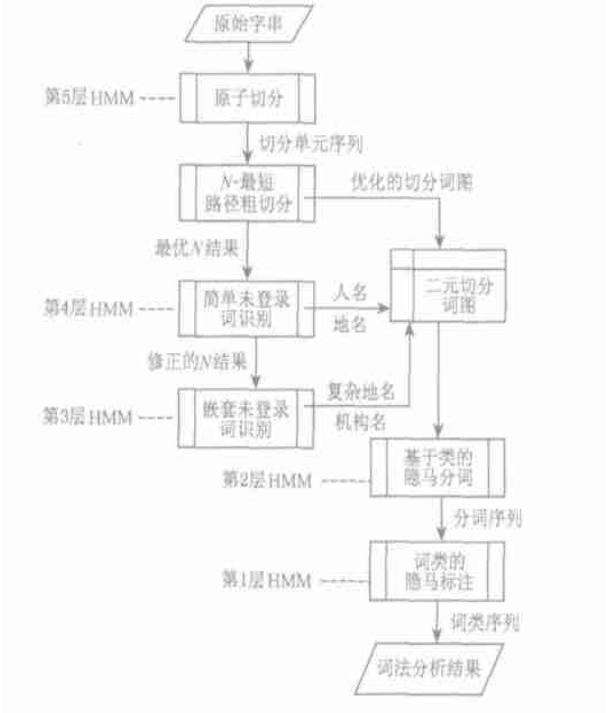


图 1　基于 CHMM 的汉语词法分析框架

切分词图是个关键的中间数据结构, 它将未登录词识别、排歧、分词等过程有机地进行了融合, 在分词模型中会详细地介绍.

原子切分是词法分析的预处理过程, 主要任务是将原始字符串切分为分词原子序列. 分词原子指的是分词的最小处理单元, 在分词过程中, 可以组合成词, 但内部不能做进一步拆分. 分词原子包括单个汉字, 标点以及由单字节、字符、数字等组成的非汉字串. 如"2002. 9, ICTCLAS 的自由源码开始发布"对应的分词原子序列为"2002. 9/, /ICTCLAS/的/自/由/源/码/开/始/发/布/". 在这层 HMM 中, 终结符是书面语中所有的字符, 状态集合为分词原子, 模型的训练和求解都比较简单, 本文就不赘述. 词法分析的最高层隐马模型是词性标注过程, 和第 5 节中的角色标注过程本质一样, 在这里不重复论述. 下面主要介绍汉语词法分析中其他层面的隐马过程.

3 基于类的隐马分词算法

本算法处于 CHMM 的第 2 层, 也就是在所有的未登录词识别完成后进行. 首先, 我们可以把所有的词分类:

$$c_i=\begin{cases} w_i \text{ iff } w_i \text{ 在核心词典中收录;}\\ \text{PER iff } w_i \text{ 是人名 and } w_i \text{ 是未登录词;}\\ \text{LOC iff } w_i \text{ 是地名 and } w_i \text{ 是未登录词;}\\ \text{ORG iff } w_i \text{ 是机构名 and } w_i \text{ 是未登录词;}\\ \text{NUM iff } w_i \text{ 是数词 and } w_i \text{ 是未登录词;}\\ \text{TIME iff } w_i \text{ 是时间词 and } w_i \text{ 是未登录词;}\\ \text{BEG iff } w_i \text{ 是句子的开始标记;}\\ \text{END iff } w_i \text{ 是句子的结束标记;}\\ \text{OTHER iff 其他} \end{cases}$$

其中, 核心词典中已有的每个词对应的类就是该词本身. 这样假定核心词典中收入的词数为 $|Dict|$, 则我们定义的词类总数有: $|Dict|+6$.

给定一个分词原子序列 S, S 的某个可能的分词结果记为 $W=(w_1, \cdots w_n)$, W 对应的类别序列记为 $C=(c_1, \cdots c_n)$, 同时, 我们取概率最大的分词结果 $W^{\#}$ 作为最终的分词结果. 则

$$W^{\#}=\arg\max_W P(W).$$

利用贝叶斯公式进行展开, 得到

$$W^{\#}=\arg\max_W P(W\mid C)P(C).$$

将词类看做状态, 词语作为观测值, 利用一阶 HMM 展开, 得

$$W^{\#}=\arg\max_W \prod_{i=1}^{n} p(w_i\mid c_i)p(c_i\mid c_{i-1}).$$

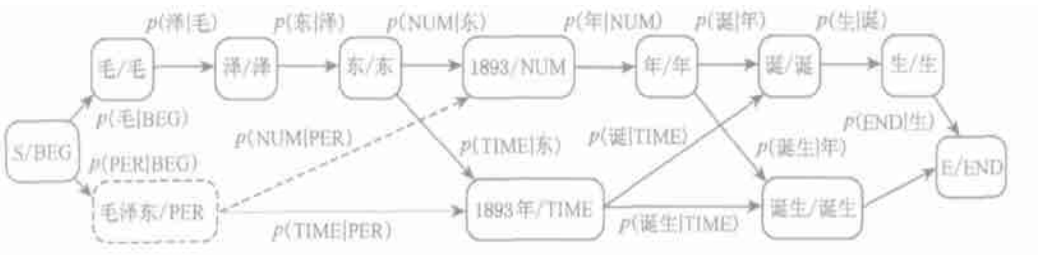
(其中, c_0 为句子的开始标记 BEG, 下同)

为计算方便, 常用负对数来运算, 则

$$W^{\#}=\arg\max_W \sum_{i=1}^{n}[-\ln p(w_i\mid c_i)-\ln p(c_i\mid c_{i-1})].$$

根据类 c_i 的定义, 如果 w_i 在核心词典收录, 可以得到 $c_i=w_i$, 因此, $p(w_i\mid c_i)=1$. 在分词过程中, 我们只需要考虑未登录词的 $p(w_i\mid c_i)$. 在图 2 中, 我们给出了"毛泽东 1893 年诞生"的二元切分词图. 最终所求的分词结果就是从初始结点 S 到结束结点 E 的最短路径, 这是个典型的最短路径问题, 可以采取贪心算法, 如 Dijkstra 算法快速求解.

在实际应用基于类的分词 HMM 时, 切分歧义能否在这一模型内进行融合并排解是一个难题; 另外一个关键问题还在于如何确定未登录词 w_i, 识别其类别 c_i 并计算出可信的 $p'(w_i\mid c_i)$; 本文的随后两节将依次阐述这两类问题的解决思路.



说明: [1] 结点中表示的是"词语/类"(即 w_i/c_i), 结点的权值为类到词语的概率 $p(w_i\mid c_i)$.

②有向边的权值为相邻类的转移概率 $p(c_i\mid c_{i-1})$; S 为初始结点; E 为结束结点.

③"毛泽东/PER"相关的虚线部分是人名识别 HMM 作用过之后产生的.

图 2 基于类的二元切分词图(原始字串为"毛泽东 1893 年诞生")

4　N-最短路径的切分排歧策略

从构成形态上划分, 切分歧义一般分为交叉歧义和组合歧义. “结合/成/分子/时”是个典型的交叉歧义,“这/个/人/手/上/有/痣”中的“人/手”构成了一个组合歧义字段.

从排歧的角度看, 切分歧义可以分为全局歧义和局部歧义. 全局歧义指的是必须结合当前句子的上下文才能准确排除的歧义. 如“乒乓球拍卖完了”, 在缺乏语境的情况下, 可以合理地切分为“乒乓球/拍卖/完/了”和“乒乓球拍/卖/完/了”. 与此相反, 局部歧义完全可以在句子内部进行排除, 本节开头列举的两个歧义句子均属于局部歧义. 根据我们对大规模语料的统计发现, 局部歧义占绝大多数, 全局歧义几乎可以忽略不计.

我们采取的是 *N*-最短路径的切分排歧策略[3]. 其基本思想是在初始阶段保留切分概率 $P(W)$ 最大的 N 个结果, 作为分词结果的候选集合. 在未登录词识别、词性标注等词法分析之后, 再通过最终的评价函数, 计算出真正最优结果. 实际上, *N*-最短路径方法是最少切分方法和全切分的泛化和综合. 一方面避免了最少切分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.

该方法通过保留少量大概率的粗分结果, 可以最大限度地保留歧义字段和未登录词. 常用切分算法往往过于武断, 过早地在初始阶段做出是否切分的判断, 只保留一个自己认为最优的结果, 而这一结果往往会因为存在歧义或未登录词而出错, 这时候, 后期补救措施往往费时费力, 效果也不会很好. 表 1 给出了 8-最短路径与常用算法在切分结果包容歧义方面的对比测试结果.

表 1　*N*-最短路径与常用算法对比

方法	切分最大数	切分平均数	正确切分覆盖率/%
最大匹配	1	1	85.46
最少切分	1	1	91.80
最大概率	1	1	93.50
全切分	> 3424507	> 391.79	100.00
8-最短路径	8	5.82	99.92

说明: [1] 切分最大数指的是句子可能的最大切分结果数.
[2]切分平均数指的是单个句子平均的切分结果数.
[3]正确切分覆盖率= 正确切分被覆盖的句子数/句子总数.
[4]测试语料大小为 200 万汉字.

同时, 我们对最终选择出的惟一切分标注结果进行了开放歧义测试, 测试集合是北大计算语言所收集的 120 对常见组合歧义、99 对常见交叉歧义, 最终组合歧义和交叉歧义排除的成功率分别为 80.00% 和 92.93%.

5　未登录词的隐马识别方法

未登录词识别的任务有: [1] 确定未登录词 w_i 的边界和类别 c_i; [2]计算 $p(w_i | c_i)$. 我们在 N 个候选切分结果的词类序列基础上, 引入了高层 HMM 来实现未登录词的识别.

5.1　未登录词识别角色表

和基于类的隐马分词模型类似, 我们对初始切分得到的各个词按照其在未登录词识别中的作用进行分类, 并将词所起的不同作用称为角色. 表 2 是人名识别的角色表. 与隐马分词中定义的类相比, 角色不同的是: 类和词是一对多的关系, 而角色与词是多对多的关系, 即: 一个词可以充当多个角色, 而一个角色也可以对应多个词.

表 2　人名识别角色表

角色	意义	示例
A	人名的上文	又/来到/于/洪/洋/的/家
B	人名的下文	新华社/记者/黄/文/摄
C	中国人名的姓	张/华/平/先生; 欧阳/修
D	双名的首字	张/华/平/先生
E	双名的末字	张/华/平/先生
F	单名	张/浩
G	人名的前缀	老/刘、小/李
H	人名的后缀	王/总、刘/老、肖/氏
L	译名的首部	蒙/帕/蒂/·/梅/拉/费
M	译名的中部	蒙/帕/蒂/·/梅/拉/费
N	译名的末部	蒙/帕/蒂/·/梅/拉/费
O	日本人名末部	小泉/纯/一/郎
X	连接词	邵/钧/林/和/稽/道/青/说
Z	其他	人民/深切/缅怀/邓/小/平

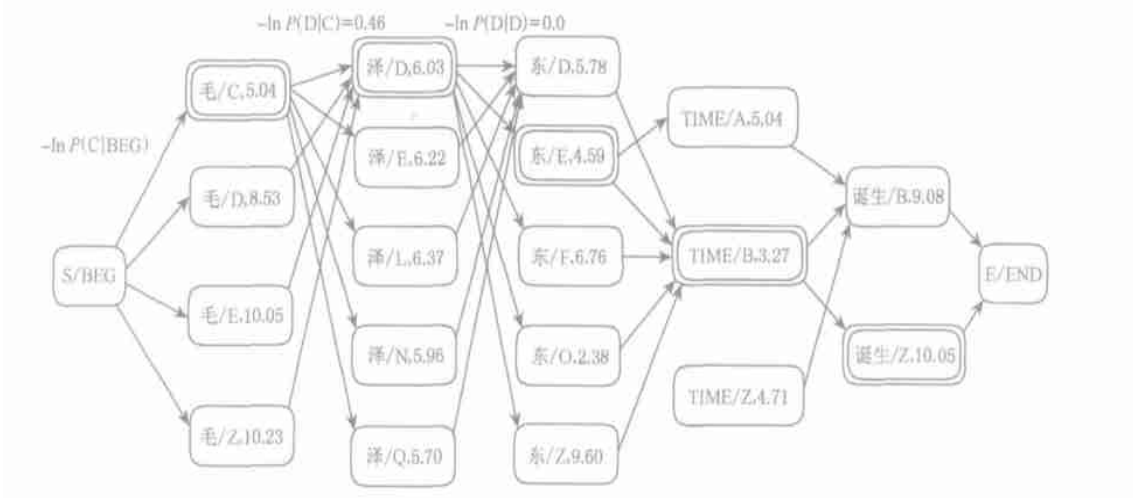
5.2　角色标注与未登录词识别

对于一个给定的初始切分结果 $W=(w_1, \cdots, w_n)$, 在一个角色集合的范畴内, 假定 $R=(r_1, \cdots, r_n)$ 为 C 的某个角色序列. 我们取概率最大的角色序列 $R^{\#}$ 作为最终的角色标注结果. 和第 3 节隐马分词的推导过程类似, 我们最终可以得到:

$$R^{\#} = \arg\min_{R} \sum_{i=1}^{n} [-\ln p'(w_i \mid r_i) - \ln p(r_i \mid r_{i-1})].$$

其中, r_0 为句子的开始标记 BEG, 下同.

$R^{\#}$ 可以通过 Viterbi 算法[26]选优得到; 图 3 给出了在词类序列"毛/泽/东/TIME/诞生"的 Viterbi 算法标注人名角色的过程. (这里 TIME 是在原子切分阶段通过简单的有限状态自动机识别出来的.)



说明: ① 图中结点的格式为: 词类c_i/角色r_i, $-\log P(c_i \mid r_i)$, 双线结点 Viterbi 选优结果.

④图中的有向边权值为相邻角色的转移概率$-\log P(r_i \mid r_{i-1})$, 这里没有全部列出.

图 3 角色标注的 Viterbi 算法选优过程

在最大概率角色序列的基础上, 我们可以简单地通过模板匹配实现特定类型未登录词的识别. 在图 3 中, 我们可以求解出最优的角色标注结果: "毛/C 泽/D 东/E TIME/B 诞生/Z", 而 CDE 正好构成一个典型的汉语人名, 因此"毛泽东"被识别为人名 PER.

识别出来的未登录词为 w, 类别为 c, 利用隐马过程可以得到:

$$p(w \mid c) = \prod_{j=0}^{k} p(w_{p+j} \mid r_{p+j}) p(r_{p+j} \mid r_{p+j-1}),$$

其中, w_i 由第 $p, p+1, \cdots, p+k-1$ 个初始切分单元组成.

最后, 识别结果及其概率加入到二元 HMM 切分图中, 和普通词一样处理, 竞争出最佳结果, 如图 2 中的虚线部分所示.

5.3 嵌套未登录词的识别

复杂地名和机构名往往嵌套了普通无嵌套的人名、地名等未登录词, 如"张自忠路"、"周恩来和邓颖超纪念馆". 对于这种嵌套的未登录词, 我们的做法是: 在低层的 HMM 识别过程中, 先识别出普通不嵌套的未登录词, 然后在此基础上, 通过相同的方法采取高层隐马模型, 通过角色标注计算出最优的角色序列, 在此基础上, 进一步识别出嵌套的未登录词. 以切分序列片断"周/恩/来/和/邓/颖/超/纪念馆"为例, 我们先识别出"周恩来"和"邓颖超"为人名 PER, 得到新的词类序列"PER/和/PER/纪念馆", 最终就可以识别出该片段为机构名. 这样的处理优点在于能够利用已经分析的结果, 并降低数据的稀疏程度.

我们用来训练 HMM 角色参数的语料库是在北大计算语言所切分标注语料库的基础上, 甄别出各种类型的未登录词之后, 自动转换得到的.

6 实验与分析

采取 CHMM 的方法, 我们研制出了计算所汉语词法分析系统 ICTCLAS (Institute of Computing Technology, Chinese lexical analysis system, 该系统全部的源码和文档, 均可在中文自然语言处理开放平台 www.nlp.org.cn 中自由下载, 免费使用). 下面我们给出 ICTCLAS 在不同条件下的测试结果, 提供 ICTCLAS 在国家"九七三"英汉机器翻译第 2 阶段评测中开放测试的结果, 并介绍我们在第 1 届国际分词大赛中的比赛情况. 在这里, 我们按照惯例引入如下评测指标: 切分正确率 SEG, 上位词性标对率 $TAG1$, 下位词性标对率 $TAG2$, 命名实体(NE)识别的准确率 P 和召回率 R, 以及 F 值. 它们的定义分别如下:

SEG= 切分正确的词数/总词数×100%;

$TAG1$= 上位词性标注正确数/总词数×100%;

$TAG2$= 下位词性标注正确数/总词数× 100%;

P= 正确识别该类 NE 数/识别出该类 NE 总数× 100%;

R= 正确识别该类 NE 数/该类 NE 总数× 100%;

$F=\frac{R\times P\times(1+\beta^2)}{R+P\times\beta^2}$, 这里我们取 $\beta=1$, 称为 F-1 值.

6.1　词法分析与 CHMM

我们使用北京大学计算语言学研究所加工的《人民日报》语料库进行了训练和测试. 在《人民日报》1998 年 1 月份共计 1108049 词的新闻语料库上, 我们进行了如下 4 种条件下的性能测试:

(1) BASE, 基准测试, 即仅仅做隐马分词和词性标注, 不引入其他层面的 HMM;

(2) + PER, 在 BASE 的基础上引入人名识别 HMM.

(3) + LOC, 在+ PER 的基础上引入地名识别 HMM.

(4) + ORG, 在+ LOC 的基础上引入机构名识别 HMM.

图 4 给出了 4 种条件下, 词法分析的分词正确率 SEG、上位词性标对率 $TAG1$、下位词性标对率 $TAG2$, 人名识别的 F-1 值 FP、地名识别的 F-1 值 FL 以及机构名识别的 F-1 值 FO.

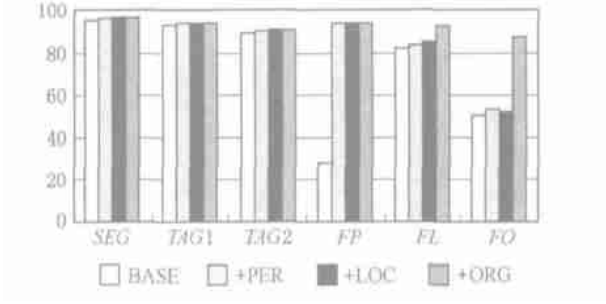


图 4　四种条件下的词法分析的性能指标

从图 4 中我们可以发现:

(1) 随着各层隐马模型的逐层加入, 词法分析的效果逐步提升. 其中, 人名识别引入后, 切分正确率 SEG 在 96.55%的基础上, 增加到97.96%, 增幅最大. 人名、地名、机构名等识别 HMM 均加入后, 切分正确率 SEG、上位词性标对率 $TAG1$、下位词性标对率 $TAG2$ 分别达到了 98.38%, 95.76%, 93.52%. 这表明: 各层 HMM 对最终词法分析均发挥了积极作用.

(2) 同时, 随着各层 HMM 的加入, 不仅极大地提高了本层 HMM 的最终性能, 还改进了低层 HMM 处理精度. 人名识别 HMM 加入后, 人名识别的 F-1 值, 立即从 27.86% 提升到 95.40%, 低层的分词 HMM 的正确率提高了 1.41%; 机构名识别 HMM 引入后, 机构名识别的 F-1 值提高了 35.59%, 同时低层的地名识别也提高了 8.49%, 人名识别的 F-1 值也达到了最高点 95.58%. 其原因在于: 高层 HMM 的成功应用在解决当前问题的同时, 也消除了低层 HMM 部分的歧义, 排除了低层 HMM 的错误结果. 例如: 在人名识别 HMM 中很容易将"刘庄的水很甜"中的"刘庄"错误识别为人名, 然而, 高层的地名识别 HMM 会正确地将"刘庄"作为地名召回, 因此达到了排歧的作用.

6.2　ICTCLAS 在"九七三"评测中的测试结果

2002 年 7 月 6 日, ICTCLAS 参加了国家"九七三"英汉机器翻译第 2 阶段的开放评测, 测试结果如表 3:

表 3　ICTCLAS 在"九七三"评测中的测试结果

领域	词数	SEG/%	$TAG1$/%	$RTAG$/%
体育	33348	97.01	86.77	89.31
国际	59683	97.51	88.55	90.78
文艺	20524	96.40	87.47	90.59
法制	14668	98.44	85.26	86.59
理论	55225	98.12	87.29	88.91
经济	24765	97.80	86.25	88.16
总计	208213	97.58	87.32	89.42

说明: ① 数据来源: 国家"九七三"英汉机器翻译第 2 阶段评测的评测总结报告;

④标注相对正确率 $RTAG=TAG1/SEG\times100\%$.

㈣由于我们采取的词性标注集和"九七三"专家组的标注集有较大出入, 所以词性标注的正确率不具可比性.

专家组的开放评测结果表明: 基于 CHMM 的 ICTCLAS 能实际地解决汉语词法分析问题, 和兄弟单位的类似系统对比, ICTCLAS 的分词结果表现出色.

6.3　第 1 届国际分词大赛的评测结果

为了比较和评价不同方法和系统的性能, 第 41 届国际计算语言联合会(The 41st Annual Meeting of the Association for Computational Linguistics, 41th ACL) 下设的汉语特别兴趣研究组(the ACL Special Interest Group on Chinese Language Processing, SIGHAN; www.sighan.org) 于 2003 年 4 月 22 日至 25 日举办了第 1 届国际汉语分词评测大赛(First International Chinese Word Segmentation Bakeoff)[28].

报名参赛的分别是来自于大陆、台湾、美国等 6 个国家和地区, 共计 19 家研究机构, 最终提交结果的是 12 家参赛队伍.

大赛采取大规模语料库测试, 进行综合打分的方法, 语料库和标准分别来自北京大学(简体版)、宾州树库(简体版)、香港城市大学(繁体版), 台湾"中央院"(繁体版). 每家标准分两个任务(track): 受限训练任务(close track)和非受限训练任务(open track).

ICTCLAS 分别参加了简体的所有 4 项任务, 和繁体的受限训练任务. 其中在宾州树库受限训练任务中综合得分 0.881[28], 名列第 1; 北京大学受限训练任务中综合得分 0.951[28], 名列第 1; 北京大学非受限训练任务中综合得分 0.953[28], 名列第 2. 值得注意的是, 我们在短短的两天之内, 采取 ICTCLAS 简体版的内核代码, 将层叠隐马模型推广到繁体分词当中, 同样取得了 0.938[28] 的综合得分.

7 结 论

本项研究的贡献在于: 针对汉语词法分析的具体问题和目前方法的不足, 引入了 CHMM, 将分词、词性标注、未登录词识别、切分排歧有机地融合到一个统一的理论模型当中. 提出并实现了歧义排除的 *N*-最短路径策略, 采取了两层 HMM, 能够识别出普通的人名地名和复杂嵌套的地名和机构名, 并计算出未登录词的概率, 将未登录词的计算结果加入到二元切分词图中, 运用基于类的 HMM 切分方法, 实现了未登录词和普通词的统一竞争和筛选. 在此理论框架的基础上, 最后实现了中科院计算所汉语词法分析系统 ICTCLAS, 经过不同条件下的测试表明, 各层 HMM 对最终的词法分析起到了积极的作用, 在国家"九七三"专家组大规模不同领域语料的开放评测中, 以及 2003 年第 1 届国际分词大赛中, ICTCLAS 均取得不错的分析效果, 是目前最好的汉语词法分析系统之一. 在现有研究的基础上, 我们将进一步开展能够自动适应各种不同领域的词法分析系统, 同时会针对信息安全的需要, 开展面向字符流的实时分词算法研究.

致谢 感谢王斌博士、孙健博士、张浩、李继锋、周栋等中国科学院计算技术研究所知识挖掘组的同事们在 ICTCLAS 研制过程中所提供的诸多建议和帮助. 特别感谢邹纲同学开发了词法分析评测程序.

参 考 文 献

1 朱德熙. 语法讲义. 北京: 商务印书馆, 1982
(Zhu Dexi. Lectures on Grammar(in Chinese). Beijing: Commerce Publishing House, 1982)

2 梁南元. 书面汉语自动分词系统——CDWS. 中文信息学报, 1987, 1(2): 101~106
(Liang Nanyuan. CDWS: A word segmentation system for written Chinese texts. Journal of Chinese Information Processing(in Chinese), 1987, 1(2): 101~106)

3 张华平, 刘群. 基于 *N*-最短路径的中文词语粗分模型. 中文信息学报, 2002, 16(5): 1~7
(Zhang Huaping, Liu Qun. Model of Chinese words rough segmentation based on N-shortest-paths method. Journal of Chinese Information Processing(in Chinese), 2002, 16(5): 1~7)

4 J Hockenmaier, C Brew. Error-driven learning of Chinese word segmentation. In: J Guo, K T Lua, J Xu, eds. The 12th Pacific Conf on Language and Information, Singapore, 1998

5 Andi Wu, Zixin Jiang. Word segmentation in sentence analysis. 1998 Int'l Conf on Chinese Information Processing, Beijing, 1998

6 D Palmer. A trainable rule-based algorithm for word segmentation. The 35th Annual Meeting of the Association for Computational Linguistics (ACL'97), Madrid, 1997

7 Y Dai, C S G Khoo, T E Loh. A new statistical formula for Chinese text segmentation incorporating contextual information. ACM SIGIR99, Berkeley, 1999

8 高山, 张艳, 等. 基于三元统计模型的汉语分词及标注一体化研究. 见: 自然语言理解与机器翻译. 北京: 清华大学出版社, 2001. 116~122
(Gao Shan, Zhang Yan, *et al*. The research on integrated Chinese word segmentation and labeling based on trigram statistical model. In: Natural Language Understanding and Machine Translation(in Chinese). Beijing: Tsinghua University Press, 2001. 116~122)

9 F Peng, D Schuurmans. A hierarchical EM approach to word segmentation. The 6th Natural Language Processing Pacific Rim Symposium (NLPRS-2001), Tokyo, 2001

10 W J Teahan, Y Wen, R McNabI, *et al*. A Compression-based algorithm for Chinese word segmentation. Computational Linguistics, 2001, 26(3): 375~393

11 Nianwen Xue, Susan P Converse. Combining classifiers for Chinese word segmentation. First SIGHAN Workshop Attached with the 19th COLING, Taipei, 2002

12 F Peng, D Schuurmans. Self-supervised Chinese word segmentation. The 4th Int'l Symp on Intelligent Data Analysis (IDA-2001), Lisbon, 2001

13 Jiangsheng Yu, Shiwen Yu. Some problems of Chinese segmentation. The 1st Int'l Workshop on Multimedia Annotation (MMA2001), Tokyo, 2001

14 张仕仁. 利用语素词规则消除切分歧义. 见: 1998 年中文信息处理国际会议论文集. 北京: 清华大学出版社, 1998. 157~

162
(Zhang Shiren. Disambiguous segmentation based on rules of word-constructing of morphemes. In: Proc of 1998 Int'l Conf on Chinese Information Processing(in Chinese). Beijing: Tsinghua University Press, 1998. 157~ 162)

15 Chunyu Kit, Haihua Pan, Hongbiao Chen. Learning case-based knowledge for disambiguating Chinese word segmentation: A preliminary study. First SIGHAN Workshop attached with the 19th COLING, Taipei, 2002

16 J H Zheng, F F Wu. Study on segmentation of ambiguous phrases with the combinatorial type. In: Collections of Papers on Computational Linguistics. Beijing: Tsinghua University Press, 1999. 129~ 134

17 Xiao Luo, Maosong Sun, Benjamin K Tsou. Covering ambiguity resolution in Chinese word segmentation based on contextual information. The 19th COLING, Taipei, 2002

18 H Luo, Z Ji. Inverse name frequency model and rules based on Chinese name identifying. In: C N Huang, P Zhang, eds. Natural Language Understanding and Machine Translation. Beijing: Tsinghua University Press, 2001. 123~ 128

19 Z Luo, R Song. Integrated and fast recognition of proper noun in modern Chinese word segmentation. Int'l Conf on Chinese Computing 2001, Singapore, 2001

20 M S Sun. English transliteration automatic recognition. In: L W Chen, Q Yuan, eds. Computational Language Research and Development. Beijng: Beijing Institute of Linguistic Press, 1993

21 H Y Tan. Chinese place automatic recognition research. In: C N Huang, Z D Dong, eds. Proc of Computational Language. Beijing: Tsinghua University Press, 1999

22 张华平, 刘群. 基于角色标注的中国人名识别. 计算机学报, 2004, 27(1): 85~ 91
(Zhang Huaping, Liu Qun. Automatic recognition of Chinese person based on roles tagging. Chinese Journal of Computers(in Chinese), 2004, 27(1): 85~ 91)

23 Zhang Huaping, Liu Qun, Zhang Hao, *et al*. Automatic recognition of Chinese unknown words recognition. First SIGHAN Workshop Attached with the 19th COLING, Taipei, 2002

24 S R Ye, T S Chua, J M Liu. An agent-based approach to Chinese named entity recognition. The 19th Int'l Conf on Computational Linguistics, Taipei, 2002

25 J Sun, J F Gao, L Zhang, *et al*. Chinese named entity identification using class-based language model. The 19th Int'l Conf on Computational Linguistics, Taipei, 2002

26 Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. Proc of IEEE, 1989, 77(2): 257~ 286

27 Shai Fine, Yoram Singer, Naftali Tishby. The hierarchical hidden Markov model: Analysis and applications. Machine Learning, 1998, 32(1): 41~ 62

28 Richard Sproat, Thomas Emerson. The first international Chinese word segmentation bakeoff. The First SIGHAN Workshop Attached with the ACL2003, Sapporo, Japan, 2003. 133~ 143

刘 群 男, 1966 年生, 博士, 副研究员, 主要研究方向为机器翻译和自然语言处理.

张华平 男, 1978 年生, 博士研究生, 主要研究方向为计算语言学、中文信息处理与信息抽取.

俞鸿魁 男, 1978 年生, 硕士, 主要研究方向为计算语言学.

程学旗 男, 1973 年生, 研究员, 主要研究方向为信息检索与网络安全.