基于遗传算法的隐马尔可夫模型在名词短语识别中的应用研究

李 荣[1] 郑家恒[2] 郭梅英[1]

(忻州师范学院计算机系 忻州 034000)[1] (山西大学计算机与信息技术学院 太原 030006)[2]

摘 要 为了进一步提高名词短语的识别精度,针对遗传算法和隐马尔可夫模型各自的特点,提出一种基于遗传算法的隐马尔可夫模型识别方法。该方法是在高准确率词性标注的基础上实现的。在训练阶段,用遗传算法获取HMM参数;识别阶段先用一种改进的Viterbi算法进行动态规划,识别同层名词短语,然后用逐层扫描算法和改进Viterbi算法相结合来识别嵌套名词短语。实验结果表明,此联合算法达到了94.78%的准确率和94.29%的召回率,充分融合了遗传算法和隐马尔可夫模型的优点,证明它较单一的隐马尔可夫模型识别法具有更好的识别效果。

关键词 短语识别,遗传算法,隐马尔可夫模型,Viterbi算法,层次分析

中图法分类号 TP391 **文献标识码** A

Application Study of Hidden Markov Model Based on Genetic Algorithm in Noun Phrase Identification

LI Rong[1] ZHENG Jia-heng[2] GUO Mei-ying[1]

(Computer Department, Xinzhou Teachers' College, Xinzhou 034000, China)[1]

(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)[2]

Abstract To increase further the accuracy of noun phrase(NP) identification and utilize features of the genetic algorithm(GA) and the hidden markov model(HMM), a novel HMM identification method based on GA was proposed. The method was based on a high-performance POS(parts of speech) tagging. During the training phase, model parameters were gained by the genetic algorithm. And during the identifying phase, an improved Viterbi algorithm for dynamic programming was first presented to identify the same hierarchy noun phrase, then the combination method of hierarchical syntax parsing and Viterbi algorithm was brought forward to identify those recursive noun phrases. Experimental results show that this combined algorithm has achieved a high precision and recall rate of 94.78% and 94.29%, respectively, fully inosculating the strength of genetic algorithms and hidden markov model. This proves that the combination method has much better identification effect than the unitary hidden markov model identification approach.

Keywords Phrase recognition, Genetic algorithm, Hidden markov model, Viterbi algorithm, Hierarchical analysis

1 引言

名词短语(NP)是非常重要的短语类型,在句子中占有很大比例。识别名词短语是自然语言处理领域中非常重要的子任务,它直接关系到文本分析和文本处理的正确性。作为自然语言处理的一种预处理手段,它的分析结果可简化句子结构,并可解决歧义问题,从而降低句法分析的难度和复杂度,为进一步的短语划分、句法分析等提供基础,同时对机器翻译、信息检索、信息抽取以及语料库的研究等应用领域都具有非常重要的意义。目前,国内外关于NP识别的方法很多,如Church[1]采用边界统计的方法识别英语BaseNP;Brill[2]提出了基于转换的学习方法;Cardie & Pierce[3]使用词性串规则,并应用剪枝(pruning)技术进行修正,用匹配的办法识别NP;Endong Xun[4]建立了识别英语BaseNP的统计模型,首先用N-best方法进行词性标注,然后进行英语BaseNP的识别。Tjong Kim Sang分别用多数投票法(Majority voting)[5]和基于记忆的学习方法[6]进行了BaseNP识别的浅层结构分析。另外,支持向量机法[7]、最大熵法[8]等也是目前组块识别的代表方法。

本文提出了一种基于遗传算法(GA)和隐马尔可夫模型(HMM)的名词短语识别方法。在训练阶段,用遗传算法获取HMM参数;识别阶段用逐层扫描算法和一种改进的Viterbi算法相结合的方法来识别名词短语。结果表明,此联合方法可以在短语定界问题上发挥优势。

2 基于GA和HMM的NP识别

2.1 HMM模型的建立

短语识别的HMM模型[9,10]可描述为:对给定经过分词和词性标注的文本 $T=w_1/p_1, w_2/p_2, \cdots, w_n/p_n$,要找出一个短语边界状态序列 $Q=Q_1Q_2Q_3\cdots Q_n$,使得 $P(Q|W)$ 能最

到稿日期:2008-11-11 返修日期:2009-02-02 本文受国家自然科学基金(60775041),山西省忻州师范学院科研基金(200623)资助。

李 荣(1974—),女,硕士,副教授,主要研究方向为人工智能和中文信息处理等,E-mail:lirong_1217@126.com;郑家恒 女,教授,主要研究方向为人工智能和中文信息处理;郭梅英 女,副教授,主要研究方向为人工智能。

好地说明文本 T 的短语构成情况，这正是 HMM 的第二个问题：解码问题。该模型中可观察序列是一对对词性标记对，隐藏状态是词性标记对两词性间的边界状态。所以，短语识别的目标就是由词性对序列反推出最合理的边界状态序列。用于短语识别的 HMM 模型是一个 5 元组 (Q_N, O_M, A, B, Π)[11]。其中 $Q_N=\{q_0, q_1, \cdots, q_{N-1}\}$ 为 NP 边界状态的有限集合；$O_M=\{v_1, v_2, \cdots, v_M\}$ 为观察值词性标记对的有限集合，M 表示不同的观察值数目，不同句子中词性对数 M 是不同的；$A=\{a_{ij}\}$，$a_{ij}=P(X_{t+1}=q_j \mid X_t=q_i)$ 为状态转移矩阵；$B=\{b_j(k)\}=P(O_t=V_k \mid X_t=q_j)$ 为观察值输出矩阵；$\Pi=\{\pi_i\}$，$\pi_i=P(X_1=q_i)$ 为初始状态分布。

图 1 表示 NP 识别的 HMM 模型中边界状态的转移情况。

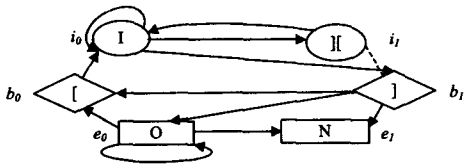

图 1　状态转移拓扑

图 1 中 3 种几何图形表示模型的不同状态，各状态间的箭头表示状态间的转换途径。正方形表示 NP 边界外部的可能状态："O"表示在 NP 外部，"N"是一种不存在的状态，引入的目的是为了使 $\sum a_{ij}=1$，$\sum b_j(k)=1$；菱形表示 NP 可能的边界状态："["表示 NP 左边界，"]"表示 NP 右边界；圆表示 NP 内部的可能状态："I"表示在 NP 内部，"]["表示两个 NP 相邻，本文也把它视为内部状态。

2.2　基于遗传算法的 HMM 参数估计

HMM 的参数估计算法一般用 Baum-Welch 算法，但是由于该算法易收敛于局部最优，本文设计了一种遗传算法[12]来训练 HMM，即寻找最优的参数 (A, B, Π)[13]。

2.2.1　编码方案

将一个状态序列看作一个 NP 边界状态序列。由于模型中有边界外部、边界处、边界内部 3 种状态，因此本文采用如下的编码方案：e_k，b_k，i_k 表示为十进制 $3k+2$，$3k+1$，$3k$，对于状态序列"[I][I]ON"，即 $Q=$"$b_0 i_0 i_1 i_0 b_1 e_0 e_1$"，对应的编码串为 $g=$"1,0,3,0,4,2,5"。

2.2.2　适应度函数

对于给定的一个序列 O，将编码的适应度函数定义为后验概率 $P(O, Q_k \mid \lambda)$，即

$$f(g_k, O)=P(O, Q_k \mid \lambda)\ (k=1,2,\cdots,n) \tag{1}$$

式中，Q_k 为 g_k 描述的 NP 边界状态序列，n 为编码个数。

2.2.3　杂交算子和变异算子

杂交算子(Crossover)：本文采用的是单点杂交算子，杂交点 C 的选择是随机的，$C=\text{random}(n)$。

变异算子(Mutation)：若编码的长度为 T，随机选择 g_k 中的第 i 个位置，$i=\text{random}(T)$，此位置的编码若为 $3k$，则可能变为 $3(k+1)$，$3(k+1)+1$，$3(k-1)$；若为 $3k+1$，则可能变为 $3k$，$(k-1)+1$，$3(k-1)+2$，$3k+2$；若为 $3k+2$，则可能变为 $3k+1$，$3(k+1)+2$。

2.2.4　选择算子

本算法采用转盘式选择策略。若编码总数为 n，有 M 个词性对序列 $O^{(k)}(k=1,\cdots,M)$，且 $n=lM$，$(l\neq 0$ 为整数)，对每个序列 $O^{(k)}$ 按转盘式选择 l 个编码，那么编码 g 被选中的概率为

$$P_c=f(g_c, O^{(k)}) / \sum_{k=l,n} f(g_k, O^{(k)}) \tag{2}$$

2.2.5　训练 HMM 模型的算法

将模型初始化为 λ_0，并由其生成初始边界状态序列；

DO{

根据模型参数估计值 $\hat{\lambda}$ 对编码进行选择、杂交和变异；

FOR 每个序列 $O^{(k)}(k=1,\cdots,M)$ DO

{选取最优编码 g_k^*，使 $g(g_k^*, O^{(k)})$ 最大；

计算 $\hat{\lambda}$

$$\hat{\pi}_i=\frac{\sum_{k=1}^{M}(g_k \text{ 中 } i \text{ 的出现次数})}{M} \tag{3}$$

$$\hat{a}_{ij}=\frac{\sum_{k=1}^{M}(g_k \text{ 中由 } i \text{ 转换到 } j \text{ 的次数})}{\sum_{k=1}^{M}(g_k \text{ 中 } i \text{ 的出现次数})} \tag{4}$$

$$b_j(O_K)=\frac{\sum_{k=1}^{M}(g_k \text{ 中由 } j \text{ 产生 } O_j \in O^{(k)} \text{ 的次数})}{\sum_{k=1}^{M}(g_k \text{ 中 } j \text{ 的出现次数})} \tag{5}$$

}

UNTIL 最大迭代次数

算法终止并输出 $\hat{\lambda}$

本文对具有 6 个状态的 HMM 模型进行了训练。

2.3　NP 识别

NP 的识别过程主要由训练和识别两个阶段构成，整个识别过程如图 2 所示。

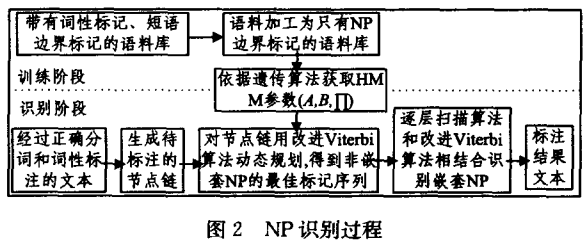

图 2　NP 识别过程

2.3.1　非嵌套 NP 的识别

NP 的识别过程可转化为在词性对之间插入不同边界状态符 Q^* 的过程，本文用一种改进的 Viterbi 算法来解决 HMM 的解码问题，以求取使 $P(Q \mid O, \lambda)$ 最大的 Q^*。

改进的 Viterbi 算法：

输入：参数模型 $\lambda=(A, B, \Pi)$，字符串 $W=w_1 w_2 \cdots w_n$，词性标记序列 $P=p_1 p_2 \cdots p_n$

输出：Q^*(最佳边界标记状态序列)

(1)初始化，计算句首词性对的 δ 和 ψ 的值。

(a) i 从 1 到 6，计算句首词性对的 6 种可能边界状态概率值

$$\delta_1(i)=\log\pi_i+\log[b_i(O_1)]+4 \tag{6}$$

(b) 句首词性对前无词性对，其前一状态记为"N"，即

$$\psi_1(i)=0 \tag{7}$$

(2)从第二个词性对开始，递归计算通向句中每个词性对的每个边界标记的最佳路径。

$$\text{(a)}\delta_t(j)=\max_{1\leqslant i\leqslant 6}[\delta_{t-1}(i)+\log(a_{ij})]+\log[b_j(O_t)]+4 \tag{8}$$

//计算到各时刻最佳路径的概率权值

$$\text{(b)用 }\psi_t(j)=\operatorname{argmax}_{1\leqslant i\leqslant 6}[\delta_{t-1}(i)+\log(a_{ij})] \tag{9}$$

//保存通向当前边界状态的最佳的上一个边界状态

(3)到达句中最后一个词性对时，计算这个词性对的最佳边界

标记

$$(a)\ \log p^* = \max_{1\leqslant i\leqslant 6}[\delta_t(i)] \tag{10}$$

//计算所选最佳路径上最后一词性对的累计最大概率权值

$$(b)\ q_t^* = \arg\max_{1\leqslant i\leqslant 6}[\delta_t(i)] \tag{11}$$

//保存最后所选择的最佳标记路径的边界状态

(4)回溯查找最佳路径上的边界状态标记,即求取最佳状态序列

$$q_t^* = \psi_{t+1}(q_{t+1}^*) \tag{12}$$

则 $q_1^*, q_2^*, \cdots, q_t^*$ 为所求的词性序列串中第一词性对、第二词对…最后一个词性对的边界标记状态。

Viterbi 算法的改进之处:

A)在对 Viterbi 公式中概率值都乘以比例因子 102 的基础上,再对公式两边取对数,得到改进的 Viterbi 公式,见上述(1. a)、(2. a)、(3. a)。

B) 许多情况下,概率值为 0,为了避免取对数无意义,对以上改进公式做如下规定:

1) if($\prod_I = 0$) or($b_i(O_t) = 0$) then $\delta_t(i) = 0$ (13)

2) 令 $temp = {}_{t-1}(i) + \log(a_{ij})$, if($a_{ij} = 0$) or $\delta({}_{t-1}(i) = 0)$ then $temp = 0$ (14)

3) if($\max_{1\leqslant i\leqslant 6}(\delta_{t-1}(i) + \log(a_{ij})) = 0$) or($b_j(O_t) = 0$) then $\delta_t(j) = 0$ (15)

2.3.2 嵌套 NP 的识别

一个完整的汉语句子都是由各类短语层层嵌套而成的。把短语中的嵌套关系化解为层次关系来解决,用改进的 Viterbi 算法同样可识别每一层面上的嵌套短语。根据现代语法中的层次分析法和上下文相关思想设计了一个逐层扫描算法。

逐层扫描算法:

(1)初始化,$I=1$。

(2)确定当前观察值矩阵行 $O=O_I$。

(3)从字串 O_I 起始值开始,利用改进的 Viterbi 算法找出当前 O_I 词性间最佳状态序列 Q_I。

(4)$I=I+1$。

(5)分析结束了吗? 是,则结束,否则转 2。

例:Managers/NNS tell/VBP newcomers/NNS the/DT tale/NN of/IN their/PP countrymen/NNS. /.

观察值矩阵:

$$\begin{bmatrix} \$ & NNS & VBP & NNS & DT & NN & IN & PP & NNS & \cdot & \# \\ \$ & NNS & VBP & NNS & DT & NN & IN & * & & \cdot & \# \\ \$ & NNS & VBP & NNS & * & & & & & \cdot & \# \end{bmatrix}$$

边界状态矩阵:

$$\begin{bmatrix} O & O & O & O & O & O & [& I &] & O \\ O & O & O & [& I & I & I & &] & O \\ O & O & O & O & & & & & O & O \end{bmatrix}$$

最终识别结果:

Managers/NNS tell/VBP newcomers/NNS [the/DT tale/NN of/IN [their/PP countrymen/NNS]]. /.

3 实验结果及分析

这里采用 Penn Treebank 语料库的 WSJ15-18(共 8937 个句子,211727 个词汇和标点符号,约 2.18MB)作为训练集,WSJ20(共 2103 个句子,47377 个词汇和标点符号)作为测试集。实验分别进行了单纯 HMM 方法和联合方法的 NP 识别,测试结果如表 1 所列。

表 1 两种方案的结果对照表

测试指标	HMM	GA+HMM
Closed test precision	93.23%	94.78%
Closed test recall	90.41%	94.29%
Closed test $F_{\beta=1}$	91.80%	94.53%
open test precision	82.76%	86.22%
open test recall	76.06%	77.59%
open test $F_{\beta=1}$	79.27%	81.68%

由表 1 结果可计算得到联合方法的综合指标 $F_{\beta=1}$ 值比 HMM 方法的 $F_{\beta=1}$ 值平均提高了 2.57 个百分点,可见第二种方案的结果优于第一种方案。这主要是因为,在 HMM 训练阶段,用遗传算法进行了参数的估计,单纯 HMM 方法中所用的传统 Baum-Welch 估计算法易陷入局部最优,而遗传算法收敛于全局最优的概率要大大高于 Baum-Welch 算法,遗传算法具有较好的收敛速度和估计精度。实验结果表明,把遗传算法用于 HMM 的参数估计进行 NP 识别,是行之有效的。

结束语 本文提出了一种基于遗传算法的隐马尔可夫模型识别名词短语的方法。本方法在 HMM 模型的训练中,充分利用遗传算法全局最优的特点进行 HMM 参数估计,并提出一种改进 Viterbi 算法和逐层扫描算法相结合的方法来识别名词短语。实验结果较令人满意,证明此联合方法确实是可行的。下一步的工作是通过导入更多的上下文信息,包括加入语义、搭配和共现等信息,进一步提高识别的准确率和召回率。

参 考 文 献

[1] Church K W. A stochastic parts program and noun phrase parser for unrestricted text[C]// Proceedings of the 2nd Conference on Applied Natural Language Processing. Texas, USA, 1988: 136-143

[2] Brill E. Transform - based Error - driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging [J]. Computational Linguistics, 1995, 21(4): 543-565

[3] Cardie C, Pierce D. Error-driven pruning of the treebank grammars for base noun phrase identification[A] // Proceedings of the 36th International Conference on Computational Linguistics [C]. Montreal, Canada, 1998: 218-224

[4] Xun Endong. A unified statistical model for the identification of English BaseNP[A]// Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics[C]. 2000: 104-111

[5] Erik F, Sang Tjong Kim. Noun phrase representation by system combination[A]// Proceedings of ANLP-NAACL 2000[C]. Seattle, WA, USA, 2000: 50-55

[6] Erik F, Sang Tjong Kim. Memory - based shallow parsing [J]. Machine Learning Research, 2002, 2(3): 559-594

[7] Kudo T, Matsumoto Y. Chunking with Support Vector Machines[A] // Proceedings of NAACL 2001 [C]. Pittsburgh, USA: Morgan Kaufman Publishers, 2001

[8] 周雅倩,郭以昆,黄萱菁,等. 基于最大熵方法的中英文基本名词短语识别[J]. 计算机研究与发展, 2003, 40(3): 440-446

[9] Mak B, Bocchieri E. Direct training of subspace distribution clustering hidden Markov model[J]. IEEE Transactions on Speech and Audio Processing, 2001, 9(4): 378-387

(下转第 261 页)

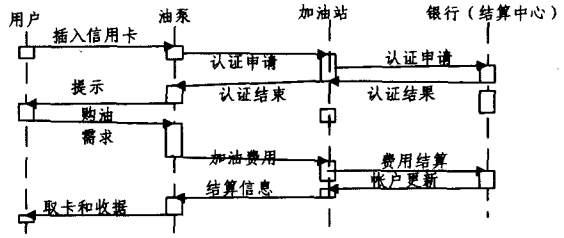

图 7　FastPass 系统的业务时序跟踪图

根据时序关系进一步细化 OV-5 模型，可以建立 OV-5 的 IDEF0 模型的顶层模型，如图 8 所示。

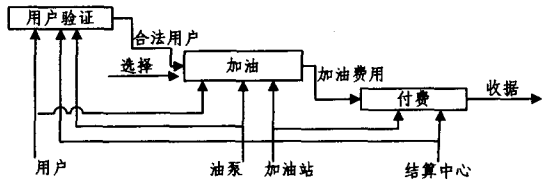

图 8　FastPass 的 OV-5 的顶层模型

根据规范分析得到的规范说明，可以开发 OV-6a 产品。针对 N1 和 N2，建立对应的业务规则 Rule1 和 Rule2 模型，如下所述。

```
Rule1:
        if Credit_Valid = true then
        pump_state= on
        else
            pump_state=off
        endif
Rule2:
        if Fuel_State= success then
        count_charge;
        update_credit;
        endif
```

其中变量 Credit_Valid 和 Fuel_State 在其他体系结构产品中定义。

结束语　MEASUR 方法是基于组织符号学理论提出的一套用于问题和业务域分析的有效方法。MEASUR 方法用于体系结构开发前期的分析阶段，不仅可以帮助设计人员更加清楚地了解问题，全面、清楚理解与体系结构设计相关的业务问题，而且各步骤形成的结果可以直接用于体系结构产品的开发，为体系结构产品提供基础的数据和词汇，使得产品设计可以结合问题的实际背景，从具体设计内容上为产品开发提供保证。与以往设计前期分析不充分，单纯依靠设计人员的经验设计相比，在体系结构设计中结合 MEASUR 方法进行前期分析可以保证体系结构设计更科学和合理。虽然，本文针对 DoDAF1.0 开展研究，但是本文的思路和方法可以推广到其他信息系统体系结构框架中使用。

参 考 文 献

[1] Zachman J A. A Framework for Information Systems Architecture[J]. Systems Journal, 1987, 26(3): 454-470

[2] Department of Defense Architecture Framework Working Group. DoD Architecture Framework Version 1.0, Volume I. Definitions and Guidelines[R]. USA, 2004

[3] Department of Defense Architecture Framework Working Group. DoD Architecture Framework Version 1.0, Volume II. Product Description[R]. USA, 2004

[4] The Open Group. The Open Group Architecture Framework, version 8 Enterprise Edition. The Open Group, 2005. http://www.opengroup.org/togaf/

[5] Levis A H, Wagenhals L. C4ISR Architectures I: Developing a Process for C4ISR Architecture Design[J]. Systems Engineering, 2000, 3(4): 225-247

[6] Wagenhals L W, Kim I S D, Levis A H. C4ISR Architecture II. A Structured Analysis Approach for Architecture Design[J]. Systems Engineering, 2000, 3(4): 248-287

[7] Bienvenu M P, Shin I, Levis A H. C4ISR Architectures III: An Object-Oriented approach to architecture design[J]. Systems Engineering, 2000, 3(4): 288-312

[8] 修胜龙，罗雪山，罗爱民. C4ISR 体系结构产品开发工具研究[J]. 计算机工程与科学，2004，26(9)：26-29

[9] Sun L, Liu K. A method for interactive articulation of information requirements for strategic decision support[J]. Information and Software Technology, 2001, 43(4): 247-267

[10] Liu K. Requirements Reengineering from Legacy Information Systems Using Semiotic Techniques, Systems, Signs and Actions [J]. The International Journal on Communication, Information Technology and Work, 2005, 1(1): 36-61

[11] Liu K. Semiotics in Information System Engineering[M]. Cambridge Cambridge: University Press, 2000

[12] Stampe P K, Althaus K, Backhouse J. MEASUR: Method for Eliciting, Analyzing and Specifying User Requirement, Computerized Assistance During the Information System Life Cycle. Elsevier Science, Amsterdam, 1988

[13] Filipe J, Liu K. Organisational Semiotics[C]// Proceedings of the 7th International Workkshop on Organisation Semiotics. Portugal: INSTICC Press, 2004

（上接第 246 页）

[10] 郑家恒，徐健. 基于隐马尔可夫模型的现代汉语句法分析[J]. 计算机工程与应用，2003，39(27)：109-112

[11] 张晓燕，王挺，陈火旺. 命名实体识别研究[J]. 计算机科学，2005，32(4)：44-48

[12] Kong S, Chau C W, Man K F, et al. Optimisation of HMM topology and its model parameters by genetic algorithms[J]. Pattern Recognition, 2001, 34(2): 509-522

[13] 杨国亮，王志良，等. 一种改进的 HMM 训练算法及其在面部表情识别中的应用[J]. 计算机科学，2006，33(11)：200-204