第 14 卷 第 2 期
1988 年 3 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 14, No. 2
Mar., 1988

句法-词义模式识别中递归结构的文法推断

路 浩 如
（浙 江 大 学）

摘 要

本文在文献[1]的基础上进一步研究递归结构的句法-词义文法的推断问题。描述体系采取以基本递归结构为子模式形成分层文法的形式。推断策略是从样本的词义化表述入手，以形式自相关搜索递归结构，通过模式结构分割建立模式的分层模型，从而构成主干文法和子文法。这一推断策略具有相当普遍的意义。

一、推断路线勾划

句法-词义方法是模式结构分析的有力手段，但其文法推断则尚研究甚罕。

模式递归结构的句法-词义描述、文献[1]已推荐之，原则归纳为：以基本递归结构作为子模式的规范形式；以分层文法作为模式描述的规范形式；以单独生成线性独立递归语言的句法描述作为文法句法部分的规范形式；而以程序文法、属性文法、递归条件文法和递归属性文法作为描述的适用形式。

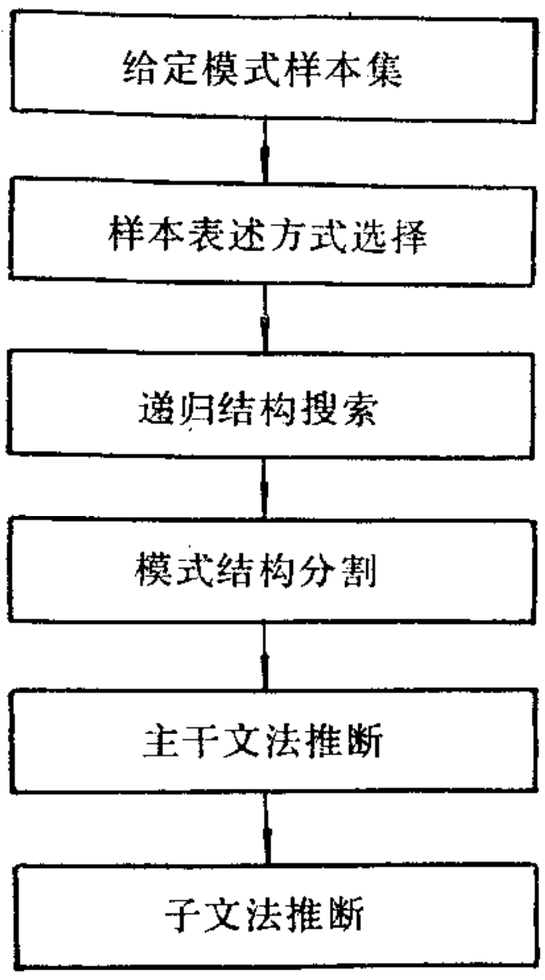


图 1 推断路线

本文根据以上原则，对递归结构的句法-词义文法推断问题提出了相应的策略，其推断路线归结如图 1 所示。它包括如下环节：1) 给定足够大的模式样本集；2)选择模式样本的表述方式；3)在样本集中搜索一切递归结构；4)进行模式结构分割，形成主干和子模式分层结构；5)为主干结构推断其主干文法；6)为子模式结构推断它们各自的子文法。

二、样本集和样本表述

定义1. 一个模式类 S 按其结构形式而有若干个子类 S_i

$$S = \bigcup_{i=1}^{n} S_i,\ n \geqslant 1. \tag{1}$$

本文于 1985 年 6 月 22 日收到。

不可再分的子类为模式的基本子类.

定义 2. 模式类 S 的有限样本集 S^+ 应有样本数大于最小必需样本数 $N_{\min}$，以使每一基本子类 S_i 的样本数大于某一 θ_r 之值. 在按递归结构划分模式的基本子类时，θ_r 称为递归阈值.

最小必需样本数 $N_{\min}$ 决定于概率 $P(S_i|S)$，

$$N_{\min}=\frac{\theta_r}{\mathrm{Min}(P(S_i|S)|\,i=1,2,\cdots,n)}. \tag{2}$$

若在已有的样本总数 N 中，基本子类 S_i 的样本数为 N_i，则可近似估计 $N_{\min}$ 为

$$N_{\min}=\frac{N\theta_r}{\mathrm{Min}(N_i|\,i=1,2,\cdots,n)}. \tag{3}$$

关于样本表述问题，文献[1]已把相对特征描述时的词义化串表述定义为 $a_1\Omega_1a_2\Omega_2\cdots\Omega_{t-1}a_t$，或者 $a_1a_2\cdots a_t|\,\Omega_1(a_1,a_2),\ \Omega_2(a_2,a_3),\ \cdots,\ \Omega_{t-1}(a_{t-1},a_t)$；而递归结构的词义化串表述定义为 $\{\beta^m|\,\Omega(\beta,\beta)\}$. 在绝对特征描述时则所有连接关系 $\Omega=CAT$，并可忽略不记. 应该指出，绝对特征描述下的递归结构变换为相对特征描述时，仍保持为递归结构，如图 2(a) 及 (b)；但相对特征描述下的递归结构并不都能变换成绝对特征描述下的递归结构，如图 2(c) 及 (d).

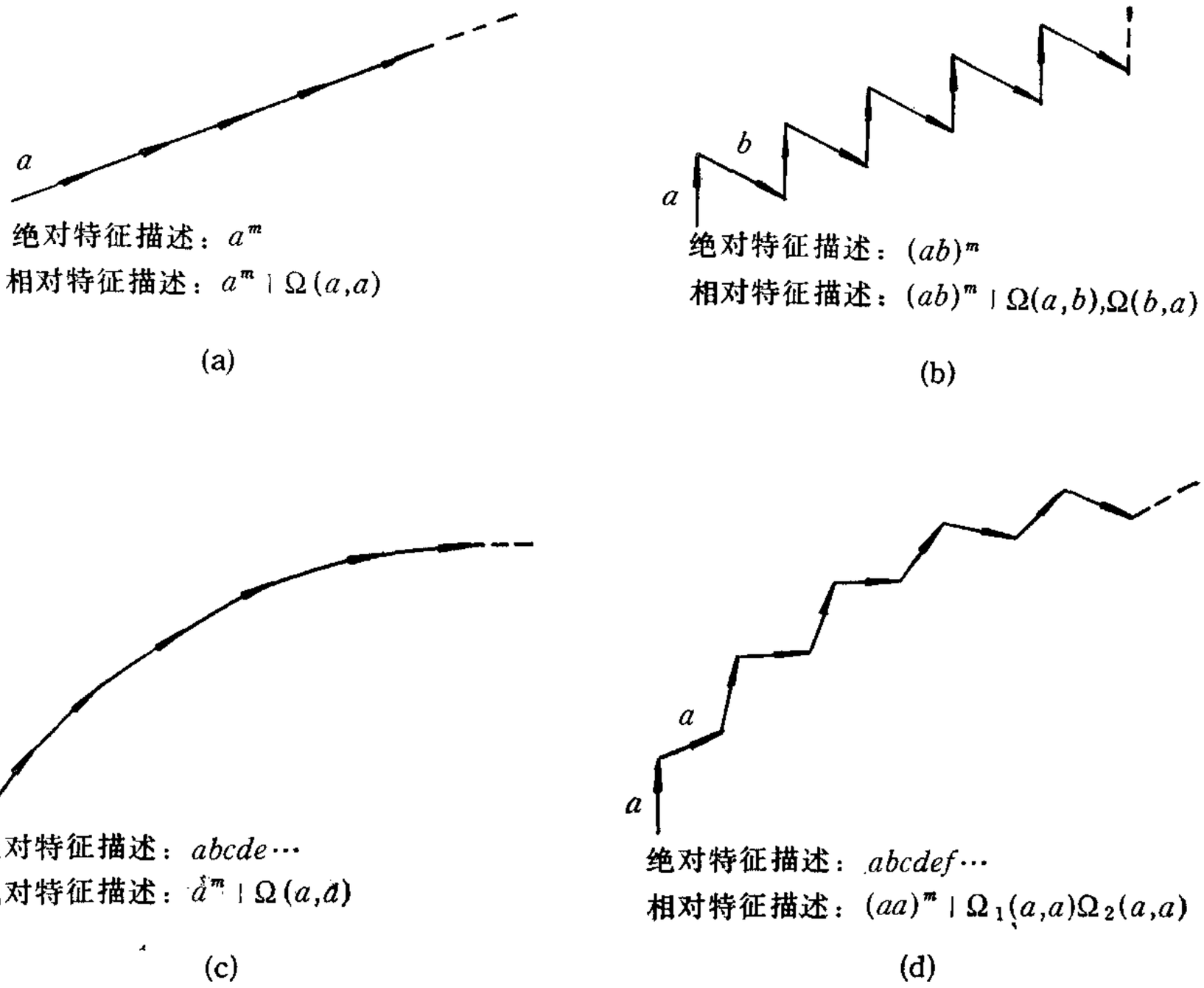


图 2　绝对特征与相对特征描述下的递归结构

三、递归结构搜索

定义 3. 模式的同构样本指模式有限样本集之中各个以若干同样的重复子串和各自

的重复次数，按同一连接关系和连接次序组成的诸样本。由同构样本归纳而成的归一化表述形式称为归一化样本，它代表模式类的一个基本子类。由样本集归纳而得的全部归一化样本组成模式的归一化样本集 $\hat{S}^+$.

规则 1. 模式递归结构的搜索和归一如下：

1）对给定样本集 $S^+ = \{s_k | k = 1, 2, \cdots, q\}$ 中的每一样本 s_k，搜索其一切重复子串，并应注意搜索嵌套于重复子串中的重复子串；

2）任一同构样本子集中的样本数超过选定的递归阈值 θ，时，即可确认其为一递归结构基本子类；

3）将每一递归结构基本子类中的递归参数归纳成独立变量或简单函数，并注意递归参数之间相互关连的可能；

4）每一基本子类都归纳成归一化样本，并组成归一化样本集 $\hat{S}^+ = \{\hat{s}_k | k = 1, 2, \cdots, p\}$.

为了从样本集中搜索递归结构，必须首先从样本串中搜索一切重复子串。文献[2,3]提出的形式自相关概念对此十分有效。并已证明绝对特征描述下串语句 s 的第 n 次位移为 i 的形式自相关 $K^{(n)}(s, i)$ 中若有无 0 子串 $\eta \neq \gamma^l$，且 $i \leqslant |\eta| < 2i$，则 s 中必存在子串 $\alpha = \beta^{n+1}$，$|\beta| = i$. 这时若 $|\eta| = i$，有 $\beta = \eta$；否则 β 为 η 中任一长度为 i 的子串 ρ. 现将形式自相关推广适用于相对特征描述下的词义化串表述，可以证明以下定理成立：

定理 1. 形式自相关析出相对特征描述下串语句 s 所含有的各重复子串。设 s 的词义化串表述中包含全部连接符，成为基元与连接符相间的符号串。位移 i 为偶数时，若第 n 次形式自相关 $K^{(n)}(s, i)$ 之中有无 0 子串 $\eta \neq (\gamma^l | \Omega(\gamma, \gamma))$，且 $i-1 \leqslant |\eta| < 2i$，则 s 中必存在子串 $\alpha = (\beta^{n+1} | \Omega(\beta, \beta))$，$|\beta| = i - 1$（连接符计入长度中）。可以确定：1)若 $|\eta| = i - 1$，且 η 不以连接符为其首尾，则 $\beta = \eta$，$\Omega(\beta, \beta)$ 为紧接 η 前的连接符；2)若 $|\eta| = i$，有 $\beta\Omega$ 或 $\Omega\beta = \eta$；3)若 $i < |\eta| < 2i$，则 $\beta\Omega$ 或 $\Omega\beta$ 为 η 中任一长度为 i 的子串 ρ.

四、模式结构分割

定义 4. 模式结构分割是将模式的各基本子类按结构特点以一定方式划分为子模式；子模式又划分为更小的子模式。本文所取基本子类以递归结构同构样本组成，即以归一化样本 $\hat{s}_k$ 表示，并以基本递归结构作为子模式的规范形式。由此，$\hat{s}_k$ 将代之以基本递归结构组成的归一化分层样本 $\tilde{s}_k$. 它们的集合为归一化分层样本集 $\tilde{S}^+$。如：

归一化分层样本集 $\tilde{S}^+ = \{\tilde{s}_k | k = 1, 2, \cdots\}$，
$\tilde{s}_k \in \{V_T \cup V_{N1}\}^+$.

一级子模式集 $V_{N1} = \{X_i | i = 1, 2, \cdots\} \subset V_N$，
$X_i = (x^{m_i} | \Omega_i)$.

二级子模式集 $V_{N2} = \{Y_j | j = 1, 2, \cdots\} \subset V_N$，
$Y_j = (\gamma^{m_j} | \Omega_j)$.

… …

V_N 和 V_T 各为非终止符集和终止符集.

若模式可分割成 n 级子模式,其中只有基元(最小的子模式)在模式描述中是不可或缺的. 因其它各级子模式或采用,或不采用,故模式的表述经路将有 $N = 2^n$ 种可能. 图 3 示出 $n = 4$ 的情况. 于是不同的模式分割将形成不同的句法复杂度和词义复杂度. 我们以基本递归结构作为子模式的规范形式,避免了此种不确定性.

在模式结构复杂的实际问题中,往往必须在按基本递归结构划分子模式之前,先按模式具体性质进行某种结构预分割. 例如图 4, 先分割区域 1—5 如(*a*);再分割线段 *A*—*E* 如(*b*). 此种以表述方便为目的的非规范分割,本文不拟详论,而把它们考虑在模式的主干结构中.

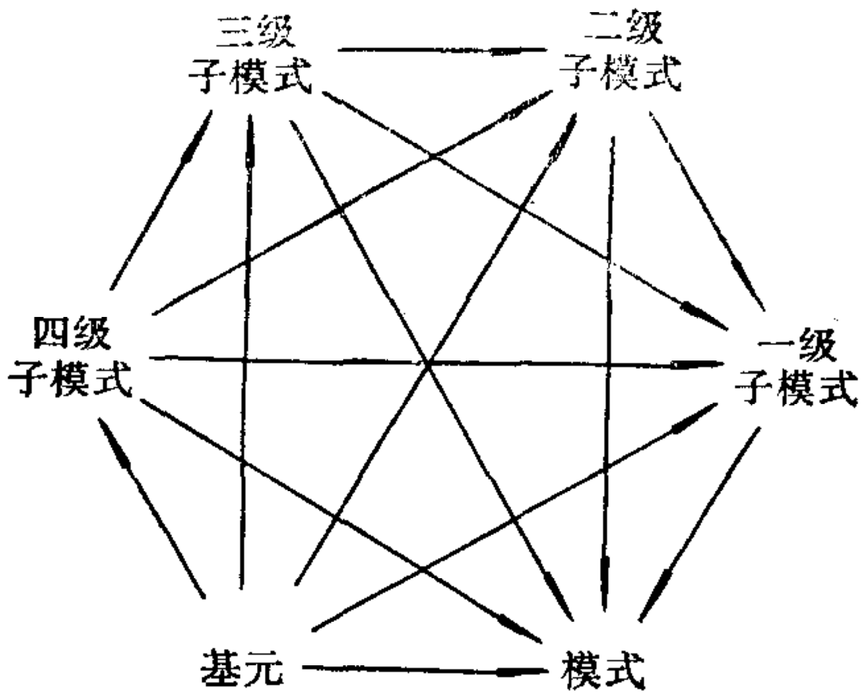
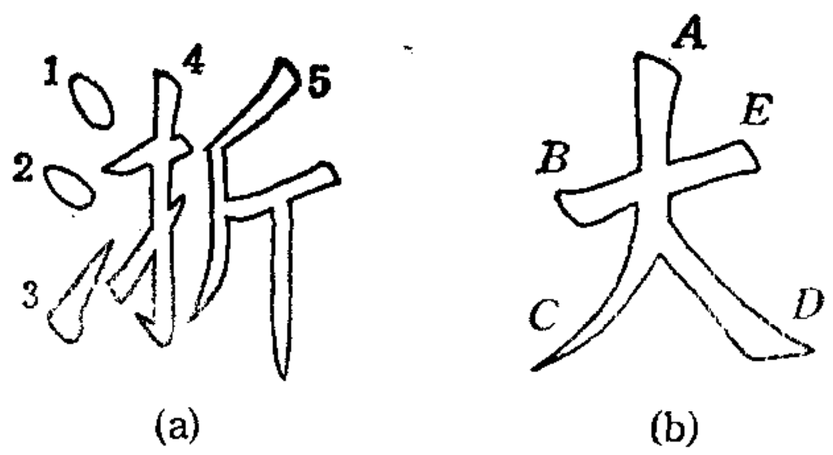


图 3　$n = 4$ 时的表述经路



(a)　　(b)

图 4　模式结构预分割

规则 2. 归一化样本的结构分割规则:

1) 对每一归一化样本 $\hat{s}_k$,取其非递归子串及基本递归结构作为子模式;

2) 保留非递归结构,而将基本递归结构代之以一级子模式 X_i, 二级子模式 Y_j, $\cdots$ 等;

3) 不论同一归一化样本之中或不同归一化样本之中(程序文法情况只限于后者),相同的基本递归结构,以相同的子模式符号表示;

4) 组成归一化分层样本集 $\tilde{S}^+$, 其主干结构及子模式结构均用词义化串表述表示(绝对特征描述情况除外);

5) 根据基元特征及连接关系,确定属性转换函数[1]:

$$\boldsymbol{A}(X\Omega Y) = \Phi(\boldsymbol{A}(X),\ \boldsymbol{A}(\Omega),\ \boldsymbol{A}(Y)). \tag{4}$$

6) 分别组成独立的和相关的子模式集:

$$W_1 = \{X_g, X_h, \cdots, Y_g, Y_h, \cdots\}, \tag{5}$$

$$W_2 = \{(X_r, X_s), \cdots (Y_r, Y_s), \cdots\}. \tag{6}$$

五、主干文法推断及子文法推断

1. 主干文法推断

归一化分层样本集 $\tilde{S}^+$ 给出了模式的树状分层结构，据此可构成其句法-词义分层文

法.

定义 5. 模式的主干结构指归一化分层样本集的主干表述:

$$\tilde{S}^+ = \{\tilde{s}_k | k = 1, 2, \cdots\}. \tag{7}$$

$$\tilde{s}_k = \alpha_i \Omega_{i1} X_i \Omega_i \alpha_j \Omega_{j1} X_j \Omega_j \cdots \Omega_p \alpha_q. \tag{8}$$

$\alpha \in V_T^*$ 为非递归子串; X 为子模式. 模式的主干结构为非递归结构. 生成主干结构的主干文法为非递归文法.

规则 3. 主干文法推断规则:

1) 将各个归一化分层样本 $\tilde{s}_k$ 的表述式(8)中各 $\alpha\Omega$ 及 $X\Omega$ 暂时考虑为复合的终止符

$$\tilde{s}_k = (\alpha_i \Omega_{i1})(X_i \Omega_i)(\alpha_j \Omega_{j1})(X_j \Omega_j) \cdots \alpha_q. \tag{9}$$

2) 先不考虑词义描述部分, 为模式主干结构建立一有限状态文法或线性文法, 其生成式都有如下数种形式, 其中 M, $N \in V_N$:

$$M \to \alpha\Omega N, \quad M \to \alpha, \quad M \to X\Omega N, \quad M \to X.$$

3) 按照各生成式中连接符所表示的连接关系以及式 (4) 形式定义的属性转换函数 Φ, 一一加入词义描述, 然后删去句法规则中的各个 Ω. 得到如下几种形式的生成式:

$$\begin{aligned} M \to \alpha N \qquad & \Omega(\alpha, N), \\ & \boldsymbol{A}(M) = \Phi(\boldsymbol{A}(\alpha), \boldsymbol{A}(\Omega), \dot{\boldsymbol{A}}(N)). \\ M \to XN \qquad & \Omega(X, N), \\ & \boldsymbol{A}(M) = \Phi(\boldsymbol{A}(X), \boldsymbol{A}(\Omega), \boldsymbol{A}(N)). \\ M \to \alpha \qquad & \boldsymbol{A}(M) = \boldsymbol{A}(\alpha). \\ M \to X \qquad & \boldsymbol{A}(M) = \boldsymbol{A}(X). \end{aligned}$$

4) 主干文法的终止符集为基元和一级子模式的集合; 非终止符集为 $\{S, M, N, \cdots\}$.

2. 子文法推断

子模式的规范形式既已定为基本递归结构, 子文法自然就是生成基本递归结构的文法.

规则 4. 属性文法的子文法推断规则:

1) 设子模式 $X = (\beta^m | \Omega(\beta, \beta))$. 已经约定子文法句法描述部分的规范形式生成线性独立递归语言. 子文法生成式的句法规则取为: $X \to \beta X$, $X \to \beta$.

2) 为各生成式建立词义规则, 成如下规范形式:

$$\begin{aligned} X^{(i)} \to \beta X^{(i+1)} \qquad & \Omega(\beta, X) = \Omega(\beta, \beta), \\ & \boldsymbol{A}(X^{(i)}) = \Phi(\boldsymbol{A}(\beta), \boldsymbol{A}(\Omega), \boldsymbol{A}(X^{(i+1)})), \ i < m. \\ X^{(i)} \to \beta \qquad & \boldsymbol{A}(X^{(i)}) = \boldsymbol{A}(\beta), \ i = m. \end{aligned}$$

其中 i 为 X 的递归序号.

3) 子文法的终止符集由 β 所包含基元及下一级子模式组成; 非终止符集中则为 X.

规则 5. 递归条件文法的子文法推断规则不同于属性文法者, 只在于按如下形式组成词义描述:

$$\begin{aligned} X \to \beta X \qquad & \text{for } N_X < m, \\ & \Omega(\beta, X) = \Omega(\beta, \beta), \\ & \boldsymbol{A}(X) = \Phi(\boldsymbol{A}(\beta), \boldsymbol{A}(\Omega), \boldsymbol{A}(X)), \end{aligned}$$

$$N_X \Leftarrow N_X + 1.$$

$$X \to \beta \qquad \text{for } N_X = m,$$

$$\boldsymbol{A}(X) = \boldsymbol{A}(\beta).$$

递归次数 N_X 的初始值为 1.

规则 6. 递归属性文法的子文法推断规则:

1) 子模式 $X = (\beta^m | \Omega(\beta, \beta))$ 的子文法句法规则按 $X \to \beta^{m_X}$ 形式列入,递归参数 m_X 由词义规则加以制约;

2) 子文法词义规则按如下形式组成:

$$X \to \beta^{m_X} \qquad m_X = m,$$

$$\Omega(\beta, \beta),$$

$$\boldsymbol{A}(X) = \Phi^{(m-1)}(\boldsymbol{A}(\beta), \boldsymbol{A}(\Omega), \boldsymbol{A}(\beta)).$$

其中 $(m-1)$ 阶属性转换函数 $\Phi^{(m-1)}$ 定义为

$$\left.\begin{aligned} &\Phi^{(m-1)} = \Phi(\Phi^{(m-2)}, \boldsymbol{A}(\Omega), \boldsymbol{A}(\beta)),\\ &\cdots\\ &\Phi^{(i)} = \Phi(\Phi^{(i-1)}, \boldsymbol{A}(\Omega), \boldsymbol{A}(\beta)),\\ &\cdots\\ &\Phi^{(1)} = \Phi(\boldsymbol{A}(\beta), \boldsymbol{A}(\Omega), \boldsymbol{A}(\beta)). \end{aligned}\right\} \tag{10}$$

规则 7. 对于相关基本递归结构,例如: (X_1, X_2), $X_1 = \beta_1^{f_{1_1}(m)}$, $X_2 = \beta_2^{f_{2_2}(m)}$, 它们的子文法分别推断后,应合并而成统一的子文法,以使变量 m 能取同一数值.

六、程序文法的子文法推断

作为一种句法-词义描述的特殊形式,程序文法一般只用于绝对特征描述情况. 文献[2,3]定义了基本递归结构的递归格式,论述了程序文法的可推断性. 已经证明: 基本递归结构的参数域递归格式有如下形式时,都能推断其程序文法:

$$\begin{bmatrix} f(n+1) \\ f_1(n+1) \\ f_2(n+1) \\ \vdots \\ f_i(n+1) \end{bmatrix} = \begin{bmatrix} k & h \cdots q \\ k_1 & h_1 \cdots q_1 \\ k_2 & h_2 \cdots q_2 \\ \vdots & \vdots \quad \vdots \\ k_i & h_i \cdots q_i \end{bmatrix} \begin{bmatrix} f(n) \\ f_1(n) \\ f_2(n) \\ \vdots \\ f_i(n) \\ 1 \end{bmatrix}. \tag{11}$$

其中 $f(n)$ 为递归参数; $f_1(n), \cdots, f_i(n)$ 为导出函数; 系数矩阵各元均为常值. 在满足式(11)的条件下,程序文法的子文法有表 1 所列的规范化普遍形式.

程序文法的推断问题在文献[4]中已详尽论述. 本文仅归纳成几条简单的规则,作为程序文法子文法推断的要义.

规则 8. 程序文法的子文法推断规则:

1) 根据基本递归结构的参数域递归格式,若满足式 (11) 的形式,按表 1 构成生成式集的核心,并仿照此规范形式标定生成式的标号和去向.子文法的入口恒为 $1X$ 和 $2X$, 出口则为去向 "Next".

2) 若有 $k=1$，$k_1=k_2=\cdots=k_i=0$。则核心中的 $\hat{X}$ 应代以 X，且删去生成式 $(i+3)X$，标号可顺移。此时 $1X$ 的成功去向改为 "Next"，$2X$ 的成功去向改为 "$3X$"，它们的失败去向则为 "ϕ"。

表 1 程序文法子文法的规范化普遍形式

标 号	核 心	成功去向	失败去向
$1X$	$X\to\beta^{f(1)}$	$1X$	Next
$(2i+4)X$	$X_1\to\beta^{f_1(1)}$	$(2i+4)X$	$(2i+5)X$
$(2i+5)X$	$X_2\to\beta^{f_2(1)}$	$(2i+5)X$	$(2i+6)X$
⋮	⋮	⋮	⋮
$(3i+3)X$	$X_i\to\beta^{f_i(1)}$	$(3i+3)X$	$1X$
$2X$	$X\to\hat{X}^k\hat{X}_1^h\hat{X}_2^g\cdots\hat{X}_i^p\beta^q$	$2X$	$3X$
$3X$	$X_1\to\hat{X}^{k_1}\hat{X}_1^{h_1}\hat{X}_2^{g_1}\cdots\hat{X}_i^{p_1}\beta^{q_1}$	$3X$	$4X$
$4X$	$X_2\to\hat{X}^{k_2}\hat{X}_1^{h_2}\hat{X}_2^{g_2}\cdots\hat{X}_i^{p_2}\beta^{q_2}$	$4X$	$5X$
⋮	⋮	⋮	⋮
$(i+2)X$	$X_i\to\hat{X}^{k_i}\hat{X}_1^{h_i}\hat{X}_2^{g_i}\cdots\hat{X}_i^{p_i}\beta^{q_i}$	$(i+2)X$	$(i+3)X$
$(i+3)X$	$\hat{X}\to X$	$(i+3)X$	$(i+4)X$
$(i+4)X$	$\hat{X}_1\to X_1$	$(i+4)X$	$(i+5)X$
$(i+5)X$	$\hat{X}_2\to X_2$	$(i+5)X$	$(i+6)X$
⋮	⋮	⋮	⋮
$(2i+3)X$	$\hat{X}_i\to X_i$	$(2i+3)X$	$2X$, $(2i+4)X$

3) 若参数域递归格式的形式为

$$f(n+1)=f(n)+q, \tag{12}$$

相应的子文法生成式集成为表 2 规范形式:

表 2

标号	核 心	成功去向	失败去向
$1X$	$X\to\beta^{f(1)}$	Next	ϕ
$2X$	$X\to X\beta^q$	$1X$, $2X$	ϕ

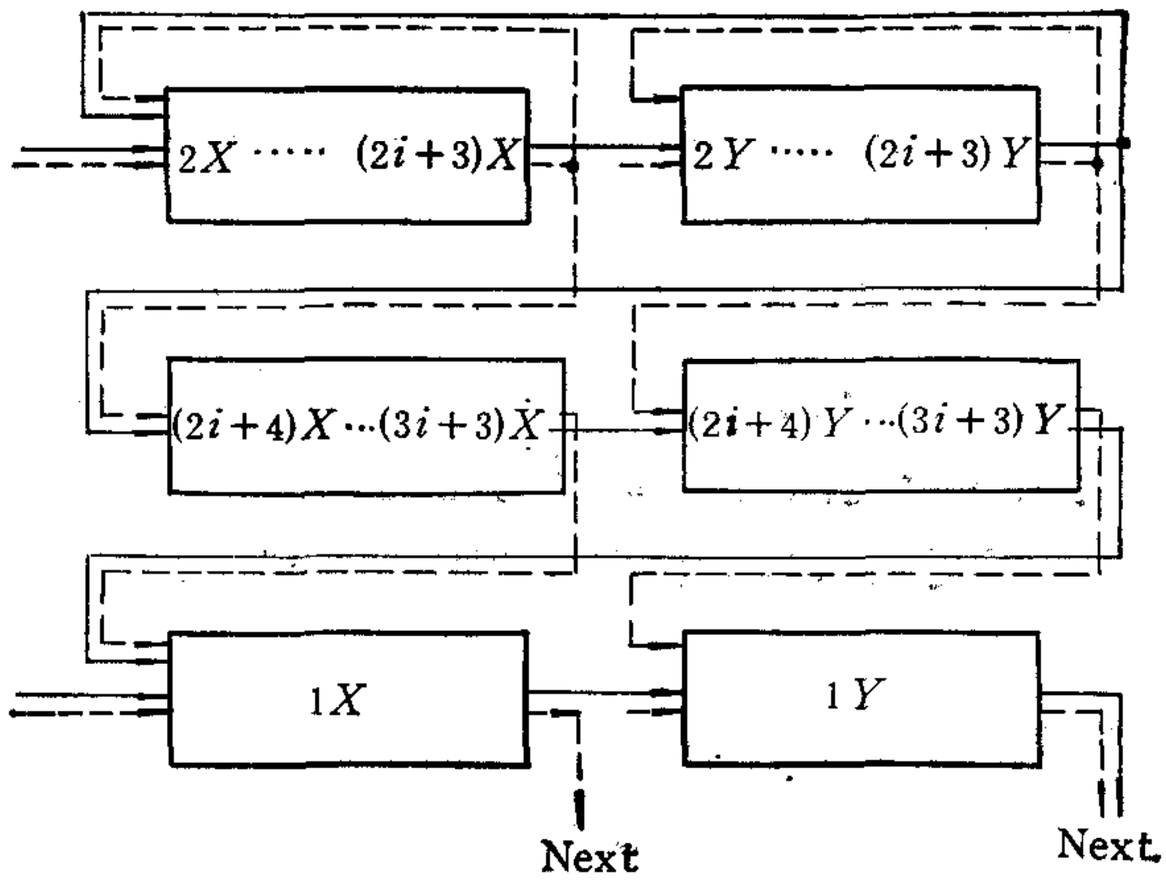


图 5 相关基本递归结构的子文法

⟶复合子文法流程; ⇢ 独立子文法流程

规则 9. 相关基本递归结构（X，Y）的子文法推断规则：

1）分别为 X 及 Y 按规则 8 各推断一独立的子文法；2）按图 5 的流程将 X，Y 两子文法合并成为复合子文法.

结　　语

本文论述了模式识别中递归结构的句法-词义文法的推断理论和方法，系统地提出了一种可行的推断策略和路线，适用于不同类型的句法-词义文法描述，从而证明了它具有普遍意义. 策略的特点是：对模式样本作词义化表述；按基本递归结构进行模式结构分割；以分层文法描述模式的分层结构；主干文法的非递归性质和子文法的规范化. 这些特点使推断过程极为简单.

本文的工作得到美国普渡大学傅京孙教授生前的热情支持，谨借本篇表示作者的沉痛悼念.

参　考　文　献

[1] 路浩如，模式识别中的递归结构及其句法-词义描述，自动化学报 **13**(1987)，87—93.

[2] 路浩如，傅京孙，上下文无关程序文法的可推断性理论，模式识别与机器智能，**1**(1987)，傅京孙教授纪念专辑.

[3] Lu, H. R. and Fu, K. S., Inferability of Context-free Programmed Grammars, *Int. J. Comput. Inf. Sci.* **13** (1984).

[4] Lu, H. R. and Fu, K. S., A General Approach to Inference of Context-free Programmed Grammars, *IEEE Trans. System, Man, Cybern.* SMC-**14**(1984), 191—202.

GRAMMATICAL INFERENCE FOR RECURSIVE STRUCTURES IN SYNTACTIC-SEMANTIC PATTERN RECOGNITION

Lu Haoru

(*Zhejiang University*)

Abstract

Based on [1], the paper further studies the inference problems of syntactic-semantic grammar in recursive structures. The inference strategy starts from the semantic representation of pattern samples, then searches all the existing recursive structures by means of formal autocorrelation, and establishes hierarchical pattern models by structural segmentation, finally constructs the trunk grammar and all the subgrammars. Such a strategy may be characterized by its simplicity and wide generality in practice.