数理解析研究所講究録
第353巻 1979年 220-227

線形文法の推定

北九州大　商　棚次奎介

任意の線形言語Lに対し、Lに属するか否かだけが知られている語のある有限系列を入力として $L=L(G)$ なる線形文法Gを推定するための手続きを示す。この方法はBiermannが提示した正則言語に対する学習方法[1]の自然な一般化である。

§1. 言語Lからの推定

アルファベット Σ 上の言語Lに対して、$u,v\in\Sigma^*$ によるLの等言語を $\{x;uxv\in L\}$ とし、$\overline{u}L\overline{v}$ で表わす。

Lに対していつでも構造図 $D(L)=(N,f)$ が一意に対応する[3]。ここにNは節の集合で $\{\overline{u}L\overline{v};u,v\in\Sigma^*\}$ を表わす。fは節からの節へのラベルつき矢印を与える写像で、$a\in\Sigma$ に対し

$$f(X,Y)=\begin{cases} a\cdot & (Y=\overline{a}X \text{ のとき}) \\ \cdot a & (Y=X\overline{a} \text{ のとき}) \\ \text{定義されない} & (\text{その他のとき}) \end{cases}$$

と定める。Lが正則であればD(L)はLを受理する有限オートマトンを与えることになる。そうでないときはD(L)は無限構造図となるが、その場合でもLの文法的構造を充分に内包しており、それを

どうとり出すかが ここでの中心的課題となる.

いま長さ $\leq k$ 以下の L の語全体を $(L)_k$ で表わす. $\Sigma = \{a_1, a_2, \cdots, a_n\}$ として L の深さ k の構造木 $T(L,k)$ を次のように帰納的に構成する:

ステップ0: 根節 $(\varepsilon, \varepsilon)$ に $(L)_k$ をラベルせよ.

ステップ n: 現時点で得られている木において, まだチェックされていない葉のうち, 最も早く生成された節 (u,v) を選べ. 同じラベルをもつ他のいかなる節からもまだ枝が張られていないなら, 節 (u,v) の下に左から右へ順に節 (ua_1, v), (u, a_1v), (ua_2, v), (u, a_2v), $\cdots$, (ua_n, v), (u, a_nv) を並べ, 節 (u,v) とそれぞれラベル $a_1\cdot$, $\cdot a_1$, $a_2\cdot$, $\cdot a_2$, $\cdots$, $a_n\cdot$, $\cdot a_n$ をもつ枝で結べ. 新しい各節 (u',v') には $(\overline{u}L\overline{v})_k$ をラベルせよ. 節 (u,v) と同じラベルをもつ他の節から既に枝が張られているなら節 (u,v) をチェックせよ.

$(\Sigma^*)_k$ が有限なので上の手続きはいつでも有限ステップで終了する.

いま, $T(L,k)$ における節の全体を $T^*(L,k)$ とし, $\{\overline{u}L\overline{v} ; (u,v) \in T^*(L,k)\}$ を $N(L,k)$ で表わす. $k \leq h$ として, $L \subset \Sigma^*$ の深さ (k,h) の構造図 $D(L,k,h) = (N, f)$ を次のように定める:

$$N = \{(X)_h ; \ X \in N(L,k)\}$$

$f(X,Y) = \alpha \iff T(L,k)$ において $(X)_h$ とラベルされた節から $(Y)_h$ とラベルされた節へラベル α をもつ枝が張られている.

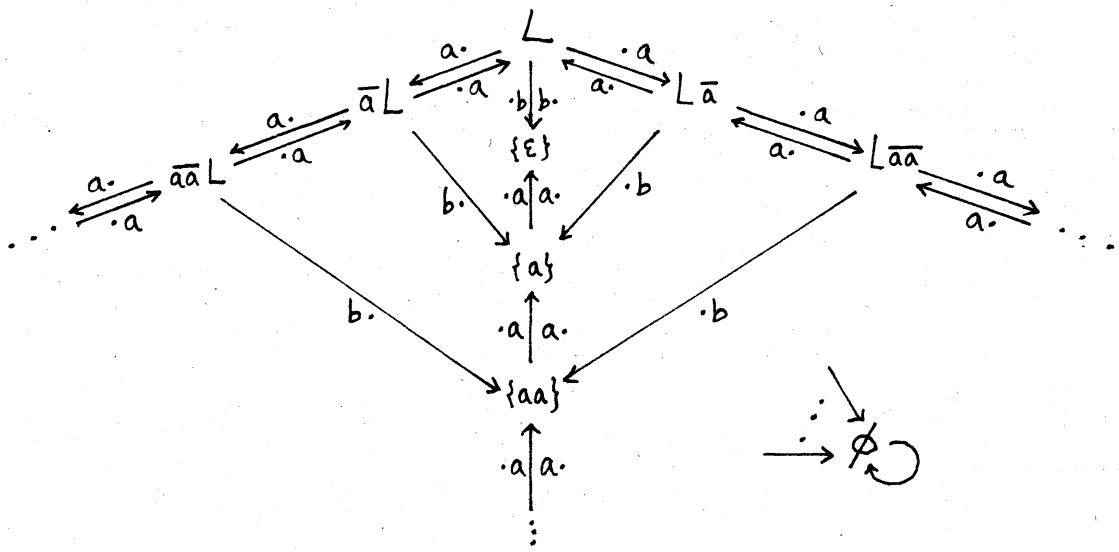
この構造図 $D(L, k, h) = (N, f)$ から次のような線形文法 $G(L, k, h) = (\Gamma, \Sigma, P, S)$ を対応させることができる：

$\Gamma = N - \{\phi\}$, Σ は L のアルファベット, $S = (L)_h$, P は $X, Y \in \Gamma$, $a \in \Sigma$ として $f(X, Y) = a\cdot$ なら $(X \to aY) \in P$, $f(X, Y) = \cdot a$ なら $(X \to Ya) \in P$, さらに $\varepsilon \in X$ なら $(X \to \varepsilon) \in P$ とする.

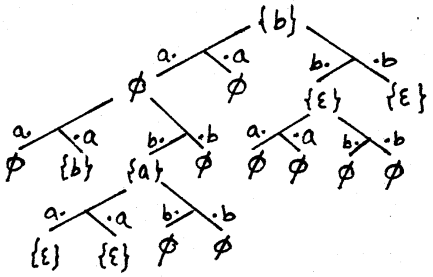
$L(G(L, k, h))$ を単に $L(L, k, h)$ と書く. 以上の手続きによって, 与えられた L から一つの線形文法を得る.

例 1. $L = \{ a^n b a^n ; n \geq 0 \}$ とする.

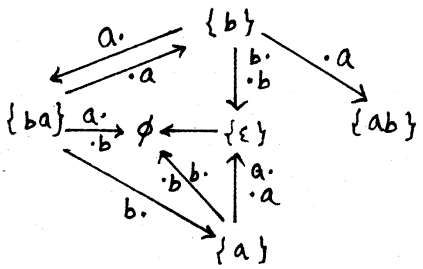
(1) 構造図 $D(L)$



(2) 構造木 $T(L, 1)$



(3) 構造図 $D(L, 1, 2)$



(4) 文法 $G(L,1,2)=(\Gamma,\Sigma,P,S)$: $\Gamma=\{\{b\},\{ba\},\{ab\},\{\varepsilon\},\{a\}\}$, $\Sigma=\{a,b\}$, $S=\{b\}$, $P=\{\{b\}\to a\{ba\}/\{ab\}a/b\{\varepsilon\}/\{\varepsilon\}b,\ \{ba\}\to\{b\}a/b\{a\},\ \{\varepsilon\}\to\varepsilon,\ \{a\}\to a\{\varepsilon\}/\{\varepsilon\}a\}$.

P を簡単化すると P と等価な規則 $P'=\{S\to aSa/b\}$ を得る.

明らかに $L(L,1,2)=L$ である.

§2. 手続きの完全性

任意の線形言語 L に対して適当な非負整数の組 (k,h) を選べば $L(L,k,h)=L$ とできることを示そう.

いま線形文法 $G=(\Gamma,\Sigma,P,S)$ に対して, $\Gamma\cup\Sigma=V$ として

$T_G=\{(u,v)$; 導出 $S\overset{*}{\Rightarrow}xAy\Rightarrow xuyv$ には各非端末記号はたかだか一度しか現われない, $x,y\in V^*$, $u,v\in\Sigma^*\}$

とし, 言語 L の次数を

$$d(L)=\min_{L=L(G)}\left(\min\{k;\ T^*(L,k)\supset T_G\}\right)$$

と定める. また

$$d(L,k)=\max_{\substack{L_1,L_2\in N(L,k)\\ L_1\neq L_2}}\left(\min\{i;\ (L_1)_i\neq(L_2)_i\}\right)$$

とおく.

<u>命題.</u> $L\subset\Sigma^*$, $X,Y\in N(L,k)$ とするとき, $G(L,k,h)$ において $(X)_k\overset{*}{\Rightarrow}u(Y)_kv$ なる導出が存在するための必要十分条件は $Y=\bar{u}X\bar{v}$ である. ここに $h=d(L,k)$ とする.

証明. $D(L,k,h)=(N,f)$, $G(L,k,h)=(\Gamma,\Sigma,P,S)$

とする. $N(L,k)$ の要素は互いにたかだか k の長さで異なるので $X, Y \in N(L,k)$ として

$$((X)_k \to a(Y)_k) \in P \iff f((X)_k, (Y)_k) = a\cdot \iff Y = \bar{a}X$$

$$((X)_k \to (Y)_k a) \in P \iff f((X)_k, (Y)_k) = \cdot a \iff Y = X\bar{a}$$

が成りたつ. しかも

$$(X)_k = (Y)_k \iff X = Y$$

なので $(X)_k \overset{*}{\underset{G}{\Rightarrow}} u(Y)_k v$ なる導出が存在することと $Y = \bar{u}X\bar{v}$ であることが同等になる. □

補題 1. 任意の $L \subset \Sigma^*$ に対し, $L(L,k,d(L,k)) \subset L$.

証明. $d(L,k) = h$, $G(L,k,h) = (I, \Sigma, P, S)$ とおく. $w \in L(L,k,h)$ とすると命題から

$$S = (L)_k \overset{*}{\Rightarrow} u(\bar{u}L\bar{v})_k v \Rightarrow uv = w$$

なる導出が存在する. 導出の最後には生成規則 $(\bar{u}L\bar{v})_k \to \varepsilon$ が適用されており, $\varepsilon \in \bar{u}L\bar{v}$ となる. ゆえに $w = uv \in L$. □

補題 2. G をある線形文法とし, $L(G) = L$ とおく. $T^*(L,k) \supset T_G$ ならば $L(L,k,d(L,k)) = L$ である.

証明. $G = (I, \Sigma, P, S)$ とする. $(X \to uxv) \in P$ ($u, v \in \Sigma^*$, $x \in V^*$) に対し, 各非端末記号がたかだか一度しか現われないような適当な導出 $S \overset{*}{\underset{G}{\Rightarrow}} u'Xv'$ ($u', v' \in \Sigma^*$) が存在し, $u'Xv' \overset{G}{\Rightarrow} u'uxvv'$ となるので T_G の定義から (u', v'), $(u'u, vv') \in T_G \subset T^*(L,k)$. したがって $\bar{u'}L\bar{v'}$, $\overline{u'u}L\overline{vv'}$

$\in N(L,k)$ である。 そのとき命題1によって

$$(\overline{w}L\overline{v})_k \overset{*}{\Rightarrow} u(\overline{uu}L\overline{vv'})_k v$$

なる導出が $G(L,k,h)$ において存在する。 ここに $h=d(L,k)$ とする。 また、$w\in\Sigma^*$ で $(X\to w)\in P$ に対しては、上の生成規則で $x=\varepsilon$, $uv=w$ とした場合に相当し、$\varepsilon\in(\overline{uu}L\overline{vv'})_k$ なので $G(L,k,h)$ において

$$(\overline{w}L\overline{v})_k \overset{*}{\Rightarrow} u(\overline{uu}L\overline{vv'})_k v \Rightarrow uv=w$$

なる導出が存在することになる。

以上から、G によって生成される語は $G(L,k,h)$ によっても生成される。一方、補題1から $L\supset L(L,k,h)$ である。 □

補題2と $d(L)$ の定義からただちに次の定理を得る。

定理1. 任意の線形言語 L に対して

$$L(L,\ d(L),\ d(L,d(L)))=L$$

§3. サンプルからの推定

$L\subset\Sigma^*$ から $L=L(G)$ なる文法 G を推定する場合、L 全体を入力とすることは非現実的である。 実は L の適当な有限部分集合によって G を推定することが可能である。 ここではそのような部分集合の条件とそれからの推定手続きを示す。

いま L の特徴集合を

$$C(L)=\{w\in u(\overline{u}L\overline{v})_{d(L,d(L))}v\ ;\ (u,v)\in T^*(L,d(L))\}$$

とする． $C(L)$ は L の有限部分集合である．

定理2． L を線形言語とする． $C(L) \subset C \subset L$ のとき，適当な非負整数 k, h が存在して $L(C, k, h) = L$ とできる．

証明． $k = d(L)$， $h = d(L, d(L))$ とする． $(u, v) \in T^*(L, k)$ に対して $C(L)$ の定義から $u(\overline{u}L\overline{v})_k v \subset C(L) \subset C$ を得る．そこで $(\overline{u}L\overline{v})_k \subset \overline{u}C\overline{v}$ となり， $(\overline{u}L\overline{v})_k \subset (\overline{u}C\overline{v})_k$ が導かれる． 一方， $C \subset L$ だから $(\overline{u}C\overline{v})_k \subset (\overline{u}L\overline{v})_k$． したがって $(\overline{u}L\overline{v})_k = (\overline{u}C\overline{v})_k$ を得る． $k \leq h$ だから $(\overline{u}L\overline{v})_k = (\overline{u}C\overline{v})_k$ となって $T(L, k)$ と $T(C, k)$ は全く一致し， $D(L, k, h) = D(C, k, h)$， すなわち

$$L(C, k, h) = L(L, k, h) = L.$$ □

上の定理の応用として線形文法 G を推定するためのアルゴリズムを図に示す． これは言語 L に属するか否かは知られているものとして語の系列 $w_1, w_2, \cdots, w_n$ を次々に入力とし， L の文法を逐次的に推定していく方法である． I_+ は L に属する語の集合を表わし，直接推定に用いられる． I_- は L に属さない語の集合を表わし推定された文法が妥当かどうかの判断にのみ用いられる．

例2． $L = \{a^n b a^n ; n \geq 0\}$ とする． L の特徴集合は $C(L) = \{b, aba\}$ である． L のサンプル入力として a, b, ab, aba をこの順に与えると，各段階で次のような文法を推定する：

(1) $I_+ = \{b\}$, $I_- = \{a\}$: $G = G(I_+, 0, 1)$, $L(G) = \{b\}$,

$I_+ \subset L(G)$, $I_- \cap L(G) = \phi$

(2) $I_+ = \{b, aba\}$, $I_- = \{a, ab\}$:

① $G = G(I_+, 0, 1)$, $L(G) = \{b\}$, $I_+ \not\subset L(G)$

② $G = G(I_+, 1, 1)$, $L(G) = \{b\}$, $I_+ \not\subset L(G)$

③ $G = G(I_+, 0, 2)$, $L(G) = \{b\}$, $I_+ \not\subset L(G)$

④ $G = G(I_+, 1, 2)$, $L(G) = L$, $I_+ \subset L(G)$, $I_- \cap L(G) = \phi$.

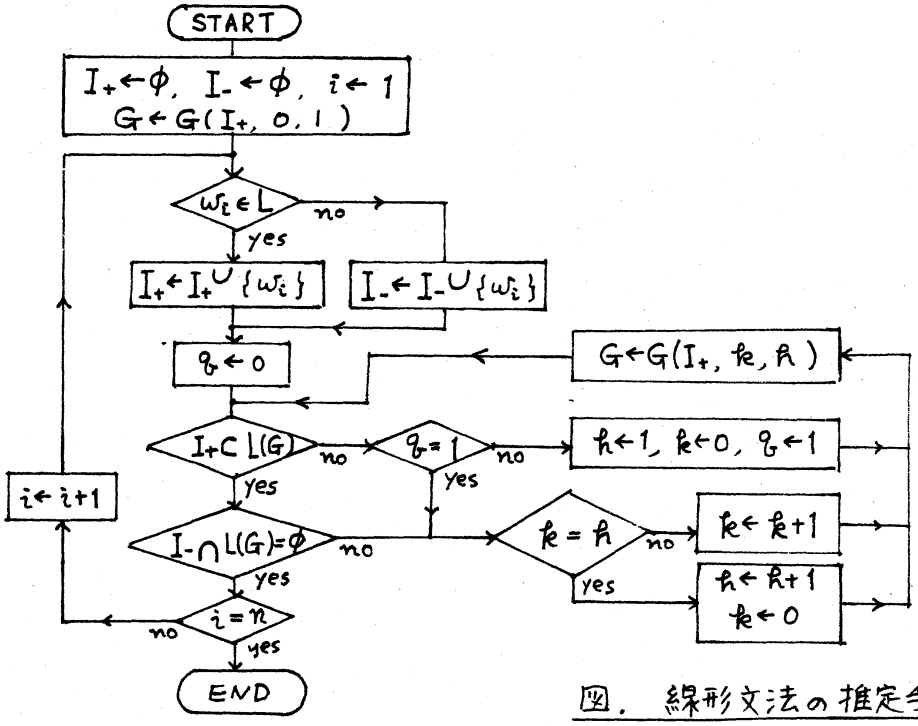


図. 線形文法の推定手続き

参考文献

[1] A.W. Biermann, An Interactive Finite-State Language Learner, Proc. 1-st USA-JAPAN Comput. Conf. (1972).

[2] K. Tanatsugu and S. Arikawa, On Characteristic Sets and Degrees of Finite Automata, Int. Jour. of Comput. and Inform. Sci., Vol.6, No.1 (1977).

[3] 飯盛末夫・藤野精一, 自由言語の構造図を利用したプッシュダウン・オートマトンの作成について, 佐賀大学教育学部研究論文集, 第24集 (II) (1976).