数理解析研究所講究録
229巻 1975年 79-88

系列パターン認識システムの考え方
── 系列パターンの誤り処理 ──

京大 工学部　堂下 修司
山崎 進

§1 まえがき

パターン認識の問題において，刺激として与えられた位置パターンと，最終的な識別結果の論理的パターンとの間に，通信チャネルとして系列パターンが存在する。位置パターンや論理的パターンに於いては，信号として，または論理的関係としての近さが存在し，それにもとずいて，パターン認識が行なわれるのであるが，系列パターンに関しては，そのような近さの尺度を導入することは，必ずしも自明ではない。一般的には，この系列パターンは言語の問題として理論的に扱うことが可能である。系列パターンの分類を言語の問題として扱う場合，文法あるいはオートマトンを用いて，系列パターンの各クラスに対して文法を構成し，その各々のクラスの識別を，文法による系列の受理非受理で行うことになるが

あるクラスに対して，それに属する系列を受理するとともに，そのクラスの系列パターンのもっている構造上の関係を正しく反映する文法あるいはオートマトンを構成することが一つの問題であり，オートマトンの構成に関し榎本氏ら[1]のグループによって詳細な研究が行なわれている。

一方，ある文法と言語が与えられているときに，その言語に属する記号列に対して，実際に観測される系列パターン入力に誤りを含む場合，それをもとの言語から生成されたものとして認識する（構文を求める）問題は類似の問題ではあるが上記の系列パターンの分類の問題とは異なった観点からの取扱いが必要となる．すなわち，ある言語を正しく受理する文法が与えられている場合，その言語に属する記号列が誤りを含んだパターンとして得られたとき，文法から，その誤りを検出し訂正する方法が実際的な課題となる．とくに，系列を構成する記号が独立に，記号間の置換，記号の挿入，記号の削除によって誤るという条件が最も簡単な場合として考えられるが，正規言語より上のクラスの言語に対しては，有限個の誤りに関して，その誤り検出が可能かどうかを，一般的には決定できないし，誤り系列の訂正に関してもあいまい性が残る[2]．

したがって，ある種の制限下での誤り訂正，例えば最小誤

りの誤り訂正を行うということが現実的な課題となり，そのためのアルゴリズム手法が提案されている[3]．

パターン認識の立場においては，系列パターンが，ある確率をもった誤りの構造の下で生起していると見るのが自然である．とくに記号の置換によって誤りが確率的に生じているとするとき，入力記号列に対し確率的に見て最も「確からしい」ものへの訂正方法が述べられている[4]．

その他，文法による系列パターンの分類において，系列パターンの生成および誤り変形を確率的に見て，系列パターンの識別に確率的文法を用い stochastic syntax analysis を適用し，染色体の識別に応用した例が論じられている[5]．

我々は，系列パターンを言語論的に扱う上で，言語に属するそれぞれの系列パターンの生起確率が予じめ提示され，入力系列パターンの誤り生成が確率的にわかっているとしたとき，系列パターン集合の系列の生起確率から見た特性を反映した文法を構成することおよび，入力記号列パターンのもとの言語への同定という問題に興味をもっている．ここでは，記号列集合Lと，それを特性づける確率的文脈自由形文法Gが，何らかの方法で与えられているとし，実際に観測される誤りを含んだ系列パターン集合$L'(\supset L)$とするとき，入力記号列パターン$w'(\in L')$の記号列$\{w\}\subset L$への同定を，記号列wの生

起確率およびwがw'へ誤る確率の積が最も大きいようなwを求めるいわゆる最尤推定の問題として考え，{w}を求める手法としては，文献(3)の手法を変形すれば，効率的なものが，得られることを示す。

§2. 系列パターンの同定

まず若干の準備を行う．文脈自由形文法(cfg) $G=(V_N, V_T, P, S)$ 等は通常の記法に従う。確率的自由形文法(scfg)[6] G_S $G_S=(G,\varphi)$である．Gはcfgであり，φはPから実数区間(0,1]への写像，$\varphi: P \to (0,1]$である．ただし$\varphi(P_i)$, $P_i \in P$ を同一の左辺をもつP_iについて加えたものは1に等しいとする．$w \in L(G)$に対し，$\psi(w) = \sum_{P_1 P_2 \cdots P_n \in E(w)} \varphi(P_1)\varphi(P_2)\varphi(P_3)\cdots\varphi(P_n)$ ($E(w)$はwの左生成過程の全体)によってwの生成確率を与える．$L(G_S)=\{(w, \psi(w)) \mid w \in L(G)\}$とする。

ここでは簡単な場合としてつぎのような誤りを考える．系列に対して，系列を構成する記号の誤りが独立に，記号間の置換，記号の削除，記号の挿入によって生ずるものとし，$a \in V_T \cup \{\varepsilon\}$が$b \in V_T \cup \{\varepsilon\}$に誤る確率を$r(b|a)$と書く．$a \in V_T$が誤らないときも$r(a|a)$なる確率での一種の誤りと見なす．いま$r(b|a)$が図1のような形で予め定義されているものとする．($r(\varepsilon|\varepsilon)$も形式的に定義される．)系列$x=a_1 a_2 \cdots a_n$が$a_1 \to b_1, a_2 \to b_2, \cdots, a_n \to b_n$と誤ることによって系列$y=b_1 b_2 \cdots b_n$が得られると

したときの誤り確率 $g(y|x)$ は $g(y|x)=r(b_1|a_1)r(b_2|a_2)\cdots r(b_n|a_w)$ で与えられるとする。

入力記号 ＼ 誤り記号	V_T	ε
V_T	置換 / 誤りなし / 置換	削除
ε	削除	誤りなし

$a\in V_T$ に対し

$$\sum_{b\in V_T\cup\{\varepsilon\}} r(b|a)=1$$

図1 入力系列パターンに関する誤りの構造

いま，系列パターンの誤り処理を次のような問題と考える。「$G_S=(G,\varphi)$，$G=(V_T,V_N,P,S)$ が定義されているとしたとき，$w\in V_T^*$ に対し $\max_{w\in L(G_S)} \varphi(w)g(w'|w)$ なる w の集合を求め，w の構文を得る手順を考えること」

以下，手順としては，文献(3)での手法を変形すれば適用できることを示す。

文献(3)ではEarleyのアルゴリズムで用いる，解析中のプロダクションの番号とその右辺の参照位置および入力系列上の参照位置の三つ組からなっている“状態”に誤りの回数を表わす変数を組み込んでいるが，その代りに $[0,1]$ の確率を組み込むことにする。概略を示すと次のようになる：

1) G_S に $Z\to S\dashv$ を付け加える。G_S の生成規則に番号付けを行ない，$Z\to S\dashv$ のそれを1とする。入力は $w\dashv=t(1)t(2)\cdots t(m)\dashv\in V_T^*\dashv$ とする。($t(i)\in V_T$)

2) (p,j,f,e) [p: 生成規則の番号，j: p の生成規則の右辺の位置を示す，f: $t(1)t(2)\cdots t(m)\dashv$ のある位置を示すバックポイン

ヲ，e：実数値］を状態とする．

3) 状態の集合 statesets {S(i)} を考え，入力の位置 t(i) に対応して S(i) とする．S(i) に後述するように4つ組の状態を付け加えてゆく．w'=t(1) t(2) … t(n) に対し n+2 の statesets がある．Statesets へ 状態 (P,j,f,e) を加えるとき，その中に (P,j,f,e') (e>e') があれば (P,f,j,e') は (P,f,j,e) におきかえられる．

4) 処理の手順 ① S(i) は (K,1,i,1) 1≦k≦#P と初期設定される．② SCAN が S(i) の state のために よばれる．③ SCAN の後 Completer は S(i) の全ての final state をおきかえる．ここで final state とは状態の4つ組において k番目の生成規則 Pk に対し，j が Pk の右辺の最後を示しているときの状態である．④ S(i) に②,③の処理が行なわれたら，S(i+1) への処理②,③を繰返し，S(n+2) で終了する．

SCAN ① ここでは入力 t(i) の check と推定の処理を行う．② 入力 t(i) に対してつぎのようにする．(i) t(i) が P番目の生成規則の右辺の j 番目 C(P,j) に (P,j,f,e) (j+1)∈S(i) で一致すれば (P,j+1, f, e·r(t(i)|t(i)))∈S(i+1)．(P,1,f,e)∈S(i) で t(i) が C(P,1) に一致すれば (P,2,f,e·r(t(i)|t(i))·φ(P))∈S(i+1) (φ(P) は P の適用確率)

(ii) (P,j,f,e)∈S(i) (j+1) で C(P,j)≠t(i)，C(P,j)∈VT ならば (P,j+1, f, e·r(t(i)|c(P,j)))∈S(i+1)．(P.1,f,e)∈S(i) で，C(P,1)≠t(i)，C(P,1) ∈VT ならば (P,2,f,e·r(t(i)|c(P,j))·φ(P))．(iii) (P,j,f,e)∈S(i)(j+1)

で，$C(P,j)\in V_T$ ならば $(P,j+1,f,e\cdot r(\varepsilon|C(P,j)))\in S(i)$. $(P,1,f,e)\in S(i)$ で $C(P,1)\in V_T$ ならば $(P,2,f,e\cdot r(\varepsilon|C(P,1))\varphi(p))\in S(i)$. (iv) $(P,j,f,e)\in S(i)$ ならば $(P,j+1,f,e\cdot r(t(i)|\varepsilon))\in S(i)$ とする。

Completer ① $S(i)$の上で SCAN が終了したとき，働く処理ルーチンである。② SCAN で $C(P,j)$ が V_N になっているときは，$C(P,j)$ を左辺とする生成規則をもつ状態を $S(i)$ に加える。③ final state を次の例のような手順でおきかえる。$(V_P\to W\cdot V,e)\in S(x)$ (j は・の位置によって表わされる。) $(V_x\to r\cdot,e')\in S(i)$ のとき $(V_P\to WV\cdot,a,ee')\in S(i)$ とする。すなわち部分系列の構文の完了を意味する。この構文の部分木は適当に記憶しておく。

最後に $(Z_1\to S\cdot,e)\in S(n+2)$ が得られたとき，$e=\max_{w\in L(G_S)}\varphi(w)\cdot p(w|w)$ であり，w の構文木も得られる。

次に例を示す。

$G_S=\{\{S\},\{a,b\},\{S\to aSb,S\to b\},S,\{\varphi(S\to aSb)=0.6,\varphi(S\to b)=0.4\}\}$ とし，誤りの構造を図2のように仮定する。一般的にいって対角線要素の値は，他の要素の値に比べて大きく，また一定以下の確率での誤り（例えば0.1以下）はゼロとみなして無視することができる。これにより，入力系列パターンの信頼性が高く，あいまいさが小さい場合，構文解析への負担が軽くなり，平等な全ての誤りを可能とみてeを誤り個数として計算する場

w \	a	b	ε
a	0.7	0.2	0.1
b	0.1	0.7	0.2
ε	0.1	0.1	0.8

図2.

合に比べて利点があると思われる。このとき，State sets は図3のようになる。図において，M,S,D,Iはそれぞれ SCAN によって，マッチング がとれたこと，置換，削除，挿入による誤りがあるとして処理することに対応している。4つ組の状態の第1番目の生成規則において，$(P_i, j) \in V_N$ から出ているポインタは Completer による。①から④までは，最終状態のおきかえで

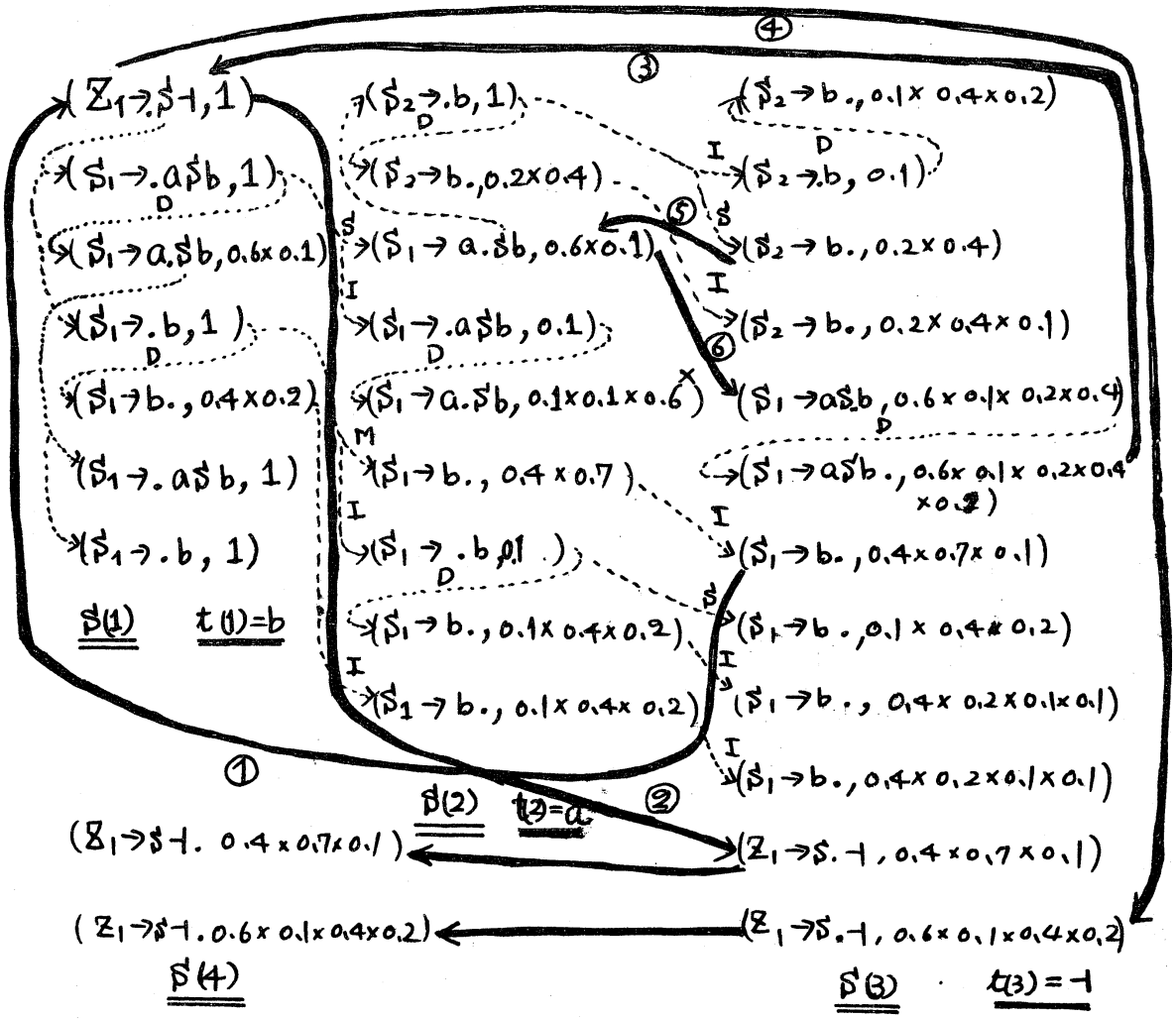

図3. 文法 G_8，入力：ba

ある。この単純な例では，$b \in L$ が $\max_{w \in L(G_S)} \psi(w) g(w'|w)$ を満たす w になっており a の挿入誤りが生じたと判断している。上記の手順においては，$P(w',w) = \max_{w \in L(G_S)} \psi(w) g(w'|w)$ を満足する w のみを考えた．しかしこの最良の判定における $P(w',w)$ の値が識別結果として十分大きく，$P(w',w)$ が他の $P(w',\tilde{w})$ $(\tilde{w} \in L, \tilde{w} \neq w)$ に対する値に比べて十分大きい場合には，これで良いが，$P(w,w)$ の値が不満足であったり，$P(w,w)$ と $P(w,\tilde{w})$ の比が接近している場合，一意的に w を w と見なすことは問題がある．実際には，$\psi(w) g(w'|w)$ の値が大きい方からいくつかの記号系列 $\{w\}$ を候補として列挙することが望ましい。特にこのような判定を用いて，さらに上位の識別の過程（たとえばセマンティックス）などを考える場合、最も確からしいいくつかの候補をその判定の値 P と共に求めておくことは必要であると思われる。

§3 むすび

系列パターンを文法によって扱うとき，系列パターンが記号の間である確率をもって独立に生ずる誤りを含んで生成される場合に，系列パターンを最も高い確率で同定するのに必要な手順は，言語の最小誤り訂正の手法において用いられたエラーカウンターを別の量で定義すれば，その手法と同等の記憶量，解析時間で可能であることを検討した。

今後の問題として，ある言語に属する記号列の生起頻度が

予じめ知られている場合，それを規則の確率または，規則の列に対する条件は，確率として近似的に文法に賦与する方法も考える必要がある。

文献

1) 穂坂，堂下，富田：サンプル記号列を識別する最簡オートマトンの構成，信学論 vol 55-D, No.3, pp 210-217, 1973年3月

2) 岩元，沢野：正則言語，文脈自由言語の誤り訂正，信学論 vol 56-D, No 12, pp 675-680, 1973年12月

3) G. Lyon: Syntax-directed least errors analysis for Context-free languages: A practical approach, CACM. vol 17, No1, pp3-14 Jan.'74

4) L.W. Fung and K.S. Fu: Stochastic syntactic classification of noisy paterns, Proceedings of the 2nd International Joint Conference on Patern Recognition, pp102-103, '74

5) H.C. Lee and K.S. Fu: A stochastic syntax analysis procedure and its application to pattern classification, IEEE Trans. on Computers, vol C-21, No.7, pp660-662, July, '72

6) 尾関：文脈自由形言語の情報論的性質，信学論，vol 55-D, No.11, pp753-760, 1972年11月

7) T. L. Booth: Probabilistic representation of formal languages, IEEE Tenth Annual Symposium on Switching and Automata Theory, pp74-81, Nov. '69