

言語空間における 2 つの測度

京大 エ 高岡 忠雄

1. はじめに

1 つの言語に 1 つの実数に対応させる試みとして、言語のエントロピーが、Benerji (1963), Kuich (1970) らによって計算されている。一方、言語の複雑さを表すため、その Turing Machine による認識時間を数論的関数で表すこと、所謂 $T(n)$ -認識可能性の問題が Hartmanis & Stearns (1965) らによって論じられている。

本稿では、言語空間から実数空間への写像として 2 つの測度を定義する。エントロピーは加法性をみたさないが、これらの測度は当然に加法性をみたす。一般に、ある Context-Free 言語が正則 (Regular) であるかどうかは決定不可能であるが、これらの測度のいくつかの性質を応用して、いくつかの言語の非正則性を示すことができる。これにより、完全平方数や素数を 2 進展開して得られた言語の非正則性を示すこ

となるが、完全平方数についての Ritchie (1963) の理論、及び素数の場合の Minsky & Papert (1966) の理論より簡潔な証明が与えられる。その結果、数論的関数の複雑さと、言語理論的階層とが比較される。

2. 諸定義

Σ を有限集合とし、 $|\Sigma| = r$ とする。 Σ^* も、空語入を含み、アルファベット Σ から生成されるすべての語の集合とする。これを Σ からつくられる自由モノイドともいう。

$R(\Sigma^*)$ を Σ^* の部分集合のすべてからなる集合とする。

Σ^* の濃度は可算、 $R(\Sigma^*)$ の濃度は連続である。 $R(\Sigma^*)$ の2つの元に対し、その和集合の演算を“+”で、連鎖 (Concatenation) の演算を“ $\cdot$ ”または juxtaposition で表すことにより、代数系として半環になる。また、当然に $R(\Sigma^*)$ は σ -完備ブール束である。要するに、台集合 $R(\Sigma^*)$ の上に種々の代数構造を定義することができるが、必要に応じて、 $R(\Sigma^*)$ の前に代数系の言葉を冠して用いる。なお半環 $R(\Sigma^*)$ は三根、高岡 (1969) で定義されており、正則集合はこの半環の上の不動点形式の線型方程式の解で与えられることが知られている。

定義 1. $A \subseteq R(\Sigma^*)$ を言語族と呼ぶ。写像 $\mu: A \rightarrow R$ (R : 実数集合) が任意の $A, B \in A$ に対して

$$A \cap B = \emptyset \Rightarrow \mu(A \cup B) = \mu(A) + \mu(B) \quad (\text{加法性}) \quad (1)$$

を満たすとき, μ は言語族 A に対する測度であるという。

定義 2. $A \in R(\Sigma^*)$ に対して数列 $\{p_n(A)\}$ を

$$p_n(A) = \frac{1}{n} \sum_{t=0}^{n-1} N(A, t) r^{-t} \quad (2)$$

で定義する。ここに, $N(A, t)$ は, A に含まれる長さ t の語の数を表す。数列 $\{p_n(A)\}$ が収束するとき, $p(A)$ を

$$p(A) = \lim_{n \rightarrow \infty} p_n(A)$$

で定義する。

定義 3. $A \in R(\Sigma^*)$ に対して, 数列 $\{\omega_n(A)\}$ を

$$\omega_n(A) = \sum_{t=0}^{n-1} N(A, t) r^{-t} \quad (3)$$

で定義する。数列 $\{\omega_n(A)\}$ が収束するとき、 $\omega(A)$ を

$$\omega(A) = \lim_{n \rightarrow \infty} \omega_n(A)$$

で定義する。

$p(A)$ の存在する A の全体を $\mathcal{A}_p$ で表し、 $\omega(A)$ の存在する全体を $\mathcal{A}_\omega$ で表す。明らかに p, ω は $\mathcal{A}_p, \mathcal{A}_\omega$ の上で加法性をみたすので、それぞれ p -測度、 ω -測度と呼ぶ。

また、明らかに、

$$p(\Sigma^*) = 1, \quad 0 \leq p(A) \leq 1, \quad p(A^c) = 1 - p(A)$$

が成り立つ。こゝに、 A^c は A の補集合である。それ故、 p -測度のことを確率測度とも呼ぶ。

補題 1.

$$\left. \begin{array}{l} \forall A \in \mathcal{A}_p \quad (p(A) > 0 \Rightarrow \omega(A) = \infty) \\ \forall A \in \mathcal{A}_\omega \quad (\omega(A) < \infty \Rightarrow p(A) = 0) \end{array} \right\} \quad (4)$$

すなわち, $A \in R(\Sigma^*)$ に対して, $\rho(A)$ と $\omega(A)$ が同時に non-zero finite の値をもつことはできない。(証明略)

定義 4. 有限オートマトン $S = \langle S, M, a, F \rangle$ は次で定義される 1 つの代数系である。

S : 内部状態の集合で有限集合

Σ : 前述のアルファベット

a : 初期状態

M : 写像 $S \times \Sigma \rightarrow S$

F : S の部分集合で, 最終状態の集合.

有限オートマトン S により, 受理される言語 $\beta(S)$ とは,

$$\beta(S) = \{x \in \Sigma^* \mid M(a, x) \in F\}$$

で定義される。 $A \in R(\Sigma^*)$ に対して, ある有限オートマトン S が存在して, $A = \beta(S)$ となれば A は正則であるという。正則な言語の族を R で表す。よく知られているように, R は集合の演算に関してブール代数を構成する。

$S = \{s_1, \dots, s_n\}$, $a = s_1$ とし, 有限オートマトン S に対して, (n, n) -行列 $Q_\sigma(S)$ を

$$\left. \begin{aligned} [Q_\sigma(S)]_{ij} &= 1, \text{ if } M(s_i, \sigma) = s_j \\ &= 0, \text{ otherwise} \end{aligned} \right\} \quad (5)$$

で定義し、行列 $Q(S)$ を

$$Q(S) = \sum_{\sigma \in \Sigma} Q_\sigma(S) \quad (6)$$

で定め、更に行列 $P(S)$ を

$$P(S) = Q(S) \cdot r^{-1} \quad (7)$$

で定める。 $P(S)$ は確率行列となる。

3. 正則集合族 $\mathcal{R}$ に対する測度

補題 2.

$$N(\beta(S), t) = (1, 0, \dots, 0) [Q(S)]^t \eta_F^T$$

ニニに、 η_F は $s_i \in F$ であるときのみ i 番目の要素が 1 であるような横ベクトルである。“ T ” は転置を示す。ベクトル $(1, 0, \dots, 0)$ を α で表す。また、

$$\begin{aligned}
 p_n(\beta(S)) &= \frac{1}{n} \sum_{t=0}^{n-1} \alpha [Q(S)]^t \eta_F^T \cdot r^{-t} \\
 &= \alpha \left\{ \frac{1}{n} \sum_{t=0}^{n-1} [P(S)]^t \right\} \eta_F^T
 \end{aligned}$$

が成り立つ。

定理 1. $A \in R(\Sigma^*)$ が正則であるとき, $p(A)$ の値が存在し, かつその値は有理数である。

証明. Doob (1952) によれば, Markov Chain の理論より, 確率行列 P のべき P^n の Césaro 和

$$\frac{1}{n} (P + \dots + P^n)$$

は確率行列 P^* に収束し, P^* は方程式

$$P^* P = P^*$$

と, P^* が確率行列であるという条件から求められる。よって, 我々の場合

$$p(\beta(S)) = \alpha [P(S)]^* \eta_F^T$$

となり, 従ってこの値は有理数である.

(証明終)

既に述べたように, 正則集合族 $\mathcal{R}$ はブール代数を構成しているから, $\langle \mathcal{R}, \mu \rangle$ は コルモゴロフ (1958) の定義による測度空間を構成することがわかる. なお, Recursive Set の範囲内で $\mu(A)$ の値が存在する A のような A を見出すことができる. それ故, μ -測度を $\mathcal{R}$ からどれ位拡張できるか, 例えば Context-Free 言語族に対して定義できるか, 等, 今後の研究課題である.

以後, 一般性を失うことなく, 有限オートマトン S は Connected でかつ最小化されているものとする.

定義5. 有限オートマトン S において, $E \subseteq S$ がエルゴード集合であるとは, 任意の $s, t \in E$ に対して

$$(\exists x \in \Sigma^* M(s, x) = t) \wedge (\exists y \in \Sigma^* M(t, y) = s)$$

が成り立ち, かつ

$$M(E, \Sigma) = E$$

が成り立つときをいう。 $d \in S$ が dead state であるとは

$$\forall x \in \Sigma^* \quad M(d, x) \notin F$$

なるときをいう。 S は最小化されていると考えられるから、
高々 1 個の dead state が存在し、上の条件は

$$\forall \sigma \in \Sigma \quad M(d, \sigma) = d, \quad d \notin F$$

と等価である。 dead state d に対して、 $\{d\}$ はエルゴード集合である。 どんなエルゴード集合にも属さない state を transient state といい。以上、Markov Chain の場合と Parallel につくった定義である。

定理 2. $A \in R(\Sigma^*)$ が正則であるとき、 $p(A)$ か $w(A)$ の中、どちらか一方のみが必ず non-zero finite の値をもつ。

証明. $s_i \in F$ があるエルゴード集合に属していれば、
 ~~$[p(S)]^*$~~ $[p(S)]^*$ は non-zero の値をとる。このとき $p(p(S)) > 0$ となる。 F に属する状態がすべて transient であるとき、
dead state が存在し、

$$\omega_n(\beta(S)) = \alpha \sum_{t=0}^{n-1} [P(S)]^t \eta_F^T$$

が有限確定値 $\omega(\beta(S))$ に収束する。これらの結果と補題1とを合せて定理が得られる。(証明終)

定理3. $A \in \mathcal{R}$ に対して, $\omega(A)$ が finite な値をもてば, それは有理数である。

証明. 確率行列を $\begin{bmatrix} P & O \\ R & Q \end{bmatrix}$ と書くとき, Q のべき級数

は収束し

$$\sum_{t=0}^{\infty} Q^t = (I - Q)^{-1}$$

となることを用いる。上記の記法については Kemeny & Snell (1960) 参照。(証明終)

例1. Hopcroft & Ullman (1968) による Half Weight Language

$$L = \{x \in \{a, b\}^* \mid \text{equal number of } a\text{'s and } b\text{'s occur in } x\}$$

に対して,

$$N(L, 2n) = {}_{2n}C_n = \frac{(2n)!}{(n!)^2} \sim \frac{1}{\sqrt{\pi n}} 2^{2n}$$

$$N(L, 2n) \cdot 2^{-2n} \sim \frac{1}{\sqrt{\pi n}}$$

これにより

$$N(L, 2n) \cdot 2^{-2n} \rightarrow 0 \quad (n \rightarrow \infty), \quad p(L) = 0$$

$$\sum_{t=0}^{n-1} N(L, 2t) \cdot 2^{-2t} \rightarrow \infty \quad (n \rightarrow \infty)$$

よって L は正則ではない。なお, L を生成する Unambiguous Context-Free Grammar が存在する。

例 2. 素数の 2 進展開により得られる言語を L で表す。
すなわち

$$L = \{10, 11, 101, \dots\}$$

素数定理により, x を大きな自然数として, 0 から x までに

含まれる素数の個数は大体 $x/\log x$ である。よって、

L に含まれる長さ n の語の個数 $N(L, n)$ は、 n が大きいとき、

$$\begin{aligned} N(L, n) &= \frac{2^{n-1}}{n-1} - \frac{2^{n-2}}{n-2} \\ &= \frac{(n-2)}{(n-1)(n-2)} \cdot 2^{n-2} \\ &= \frac{1}{n-1} \cdot 2^{n-2} \end{aligned}$$

これより、

$$N(L, n) \cdot 2^{-n} \rightarrow 0 \quad (n \rightarrow \infty), \quad p(L) = 0$$

$$\sum_{t=0}^{n-1} N(L, t) \cdot 2^{-t} \rightarrow \infty \quad (n \rightarrow \infty), \quad w(L) = \infty$$

よって L は正則でない。

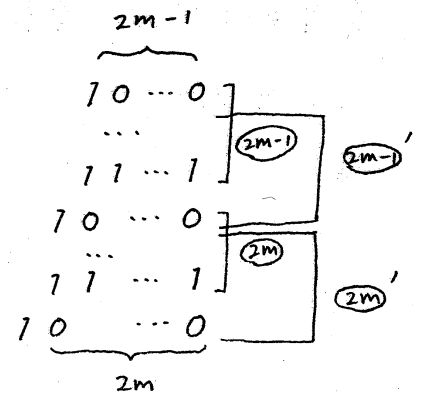
なお、Hartmanis & Shank (1968) によっても、この L が non-Context-Free であることまで示されている。

例3. 完全平方数を2進展開して得られた言語を L とする。

即ち、

$$L = \{1, 100, 1001, \dots\}$$

右の図で、 $(2m)$ の部分に含まれる
完全平方数の個数は $(2m)'$ の部分に
含まれるそれより 1 個少なく、
 $(2m-1)$ の部分に含まれる完全平方
数の個数は $(2m-1)'$ の部分に含まれ
るそれより 1 個多い。従って、



$$\omega_{2n}(L) = \sum_{m=1}^n \{ ([2^{m-\frac{1}{2}}] - 2^{m-1} + 1) \cdot 2^{-(2m-1)} + (2^m - [2^{m-\frac{1}{2}}] - 1) \cdot 2^{-2m} \}$$

$$= \sum_{m=1}^n \{ [2^{m-\frac{1}{2}}] \cdot 2^{-2m} + 2^{-2m} \}$$

$$\omega(L) = \sum_{m=1}^{\infty} [2^{m-\frac{1}{2}}] \cdot 2^{-2m} + \frac{1}{3}$$

ここで $[]$ はガウス記号を表し、自然数 n までに含まれる
完全平方数の個数は $[\sqrt{n}]$ であることを用いている。さて、
上式中 1 項は $\sqrt{2}/2$ を 2 進展開したものを $0.\alpha_1\alpha_2\alpha_3\dots$ と
すると、次の図から理解されるように、 $0.\alpha_1 0\alpha_2 0\alpha_3 0\dots$
となる。

$$\begin{array}{rcl}
 m & & \\
 1 & 0.0\alpha_1 & \\
 2 & 0.00\alpha_1\alpha_2 & \\
 3 & 0.000\alpha_1\alpha_2\alpha_3 & \\
 4 & 0.0000\alpha_1\alpha_2\alpha_3\alpha_4 & \\
 5 & 0.00000\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5 & \\
 6 & 0.000000\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5\alpha_6 & \\
 & \dots\dots & \\
 +) & \hline
 & 0.\alpha_10\alpha_20\alpha_30\dots &
 \end{array}$$

$0.\alpha_1\alpha_2\alpha_3\dots$ は Ultimately periodic ではないから, $0.\alpha_10\alpha_20\alpha_30\dots$ も, ultimately periodic ではない。よって, $\omega(L)$ は有理数ではない得ず, L は正則ではない。

例 4. 再び Half Weight Language L について,
収束半径 1 で

$$\frac{1}{\sqrt{1-x}} = 1 - \frac{1}{2}x + \dots + \frac{(2n)!}{(n!)^2} 2^{-2n} (-x)^n + \dots$$

と展開できる。言語 L が, c, d を dummy として, アルファベット $\Sigma = \{a, b, c, d\}$ の上で定義されていると考えると

$$\omega(L) = \sum_{n=0}^{\infty} \frac{(2n)!}{(n!)^2} 4^{-2n}$$

一方

$$\frac{1}{\sqrt{1+\frac{1}{4}}} = \frac{2}{\sqrt{5}} = 1 - \frac{1}{2} \cdot \frac{1}{4} + \dots + \frac{(2n)!}{(n!)^2} 4^{-2n} + \dots$$

であるから、

$$w(L) = 2/\sqrt{5}$$

これは無理数であり、この方法によっても、 L の non-regularity が証明されることがわかった。

3. Context-Free 言語に対する測度

N. Chomsky (1963) によれば Context-Free Grammar の defining equation は半環 $R(\Sigma^*)$ の上の不動点形式の非線型方程式として与えられ、CF 言語はその最小解として与えられる。その方程式を

$$x_i = g_i(x_1, \dots, x_n) \quad (i=1, \dots, n) \quad (8)$$

とする。与えられた Grammar は Unambiguous とする。

上式の関数 g_i の中におかれた各 $\alpha \in \Sigma$ を複素変数 z に置き換え、 x_i を複素変数 y_i に置き換えて得られる方程式を

$$y_i = M_i(y_1, \dots, y_n; z) \quad (i=1, \dots, n) \quad (9)$$

と記す。Kuich によれば、この方程式は $z=0$ のまわりの disk において, unique to analytical solution

$$y_i = f_i(z), \quad f_i(0) = 0 \quad (i=1, \dots, n) \quad (10)$$

をもつ,

$$y_1 = \sum_{t=0}^{\infty} N(L, t) z^t \quad (11)$$

となる。ここに、 x_1 は L を与える CF Grammar の初期シンボルである。(11) を structure generating function とする。(11) 式のべき級数の収束半径は、Cauchy-Hadamard の定理により、

$$\Lambda = \lim_{n \rightarrow \infty} \sqrt[n]{N(L, n)}$$

とするとき、 $1/\Lambda$ で与えられる。よって、Unambiguous な CF 言語 L に対し、 $r > \Lambda$ のとき、 $\omega(L)$ は値をもつ。
 $\omega(L)$ は方程式

$$y_i = f_i(r^{-1}) \quad (i=1, \dots, n) \quad (12)$$

を解いて、 y_i として得られる。よって、 $w(L)$ が値をもつとき、それは代数的な数である。従って、もし、 $w(L)$ が値をもって、それが超越数ならば、 L は unambiguous な CF 言語ではないといえる。なお、 $r \geq 1$ であるから、そして、dummy alphabet を加えても、grammar の ambiguity は不変であるから、適当に Σ を大きくすることによって、いつでも $w(L)$ が値をもつようにすることが出来る。

例 5. Simple Deterministic Grammar $\{\sigma \rightarrow a\sigma\sigma, \sigma \rightarrow b\}$ を与えよう。定義方程式は

$$\sigma = a\sigma\sigma + b$$

方程式 M_i は

$$y = zy^2 + z$$

$$y = \frac{1 - \sqrt{1 - 4z^2}}{2z}$$

$$\begin{aligned}
 N(L, k) &= \frac{(2n)!}{n!(n+1)!} & (k=2n+1) \\
 &= 0 & (\text{otherwise})
 \end{aligned}$$

$$w(L) = 1 \quad \text{for } \Sigma = \{a, b\}$$

$$w(L) = 2 - \sqrt{3} \quad \text{for } \Sigma = \{a, b, c, d\}$$

よって, L は non-regular である。

例 6. 言語 $L = \{a^n \mid n \geq 1\}$ を考えよう。明らかに L は Context-Sensitive である。 $\Sigma = \{a, b\}$ の上で, $w(L)$ を計算すると,

$$w(L) = \sum_{n=1}^{\infty} \frac{1}{2^n}$$

これは Liouville Number として知られている超越数である。

よって, L は unambiguous CF 言語ではない。

文 献

1. R. Banerji (1963), "Phrase structure languages, finite machines and channel capacity," Information and Control 6.
2. W. Kuich (1970), "On the entropy of context-free languages," Inf. and Cont. 16.
3. J. Hartmanis and R.E. Stearns (1964), "On the computational complexity of algorithms," Trans. Amer. Math. Soc.
4. M. Minsky and S. Papert (1966), "Unrecognizable set of numbers," JACM 13.
5. 三根, 高岡 (196), "有限オートマトンの状態特性方程式の一般解について," 信学会誌 S2-C.
6. J. L. Doob (1952), Stochastic Processes, John Wiley & Sons, Inc.
7. コルモゴロフ (1962), 関数解析の基礎, 岩波.
8. J. G. Kemeny and J. Snell (1960), Finite Markov Chains, Van Nostrand.
9. J. E. Hopcroft and J. D. Ullman (1969), Formal Languages and Their Relation to Automata, Addison-Wesley.

10. N. Chomsky (1963), " Formal properties of grammars," *Handbook of Math. Psych.*, John Wiley & Sons, Inc.
11. J. Hartmanis and H. Shank (1968), " On the recognition of primes by automata," *JACM* .