

文脈自由形言語の情報論的性質

NHK総合技術研究所 尾関 和彦

§ 1. まえがき

自然言語であれ人工言語であれ，言語の持つ最も重要な機能のひとつは情報の伝達であろう．それでは言語が伝達する情報量というものをどのように考えたらよいであろうか．

情報理論が創始されて以来，自然言語に対しても Markov 過程のモデルが考えられ，それにもとづいて一文字当りの平均情報量が計算されたこともある．Markov 過程のモデルを考えることは，その言語の文法として正規文法を考えることにほかならないが，近年の言語理論の進歩は正規文法が自然言語の文法として不十分であることを明らかにした．ここでは文法のクラスとして，自然言語の構造の基本的な部分を記述することができるといわれている文脈自由形文法をとりあげ，それを言語能力として持つ話手モデルとして，プロダクションに確率が付与された，いわゆる確率的文脈自由形

文法を考える．そしてその話し手の話す言語について情報論的議論を展開する．

§2. 言語が情報をにんう機構

ここには二人の人 A , B がいるとし, 彼らは共通の言語を話すとする．彼らの言語能力のモデルとしてつぎのような文法自由形文法 G を仮定しよう:

$$G = (V_T, V_N, P, S),$$

$$V_T = \{\text{花, 鳥, 娘, は, とても, 美しい}\},$$

$$V_N = \{S, \text{主部, 述部, 名詞}\},$$

$$P = \{S \rightarrow \text{主部述部}, \text{主部} \rightarrow \text{名詞は}, \text{名詞} \rightarrow \text{花}, \\ \text{名詞} \rightarrow \text{鳥}, \text{名詞} \rightarrow \text{娘}, \text{述部} \rightarrow \text{美しい}, \text{述部} \rightarrow \text{とても述部}\}.$$

A および B が話しかつ理解するここのでできる文の全体, すなわち彼らの言語は, $\{\text{花は美しい, 娘は美しい, ---, 花はとても美しい, ---}\}$ という可算無限集合である．いま A が B に

$$\text{‘娘はとても美しい’} \tag{1}$$

と言ったとせう．この文が A の伝えた情報にになり, かつ B がこの文を受けとって A の意思を知ることのできる機構はどのようなものであろうか．(1) の生成過程を考えてみる

と、

S ⇒ 主部述部 ⇒ 名詞は述部 ⇒ 娘は述部 ⇒ 娘はとても述部

⇒ 娘はとても美しい (2)

というステップ⁰を踏んで (1) が生成されることがわかる。この生成過程の第3ステップ⁰において名詞→娘というプロダクションが適用されているが、Aはここで名詞→花、あるいは名詞→鳥というプロダクションを選ぶこともできたはずである。つまり三つの可能なプロダクションの中からAは特に、名詞→娘というプロダクションを選択したわけである。このようにして (1) は‘花’でもなり‘鳥’でもなり‘娘’は…というAの意思をになうことができるのである。一方Bは(1)を受けとり、その生成過程の第3ステップ⁰において三つの可能なプロダクションの中から特に名詞→娘が選択され適用されていることを見出しAの意思を知ることができるであろう。すなわち文が情報をになうことができるのは、その生成過程の各ステップ⁰において適用するプロダクションを話し手が選択できたり、聞き手が文を受けとって情報を得ることができるのは、その文の生成過程の各ステップ⁰においていくつかの可能なプロダクションの中から実際にどのひとつが選択されたかを知るからにほかならない。このように文を生成過程に分析して考えれば、‘情報を得る’とはいくつかの可能

性の中からどれが選択されたかを知ることである。' という情報に付する Shannon の基本的な考え方がそのまま適用できることがわかる。

ここでは話し手-聞き手の言語能力のモデルとして文脈自由形文法を考え、話し手の意思を、各プロダクションにそれが適用される確率を付与するという形でモデル化する。つまり確率的文脈自由形文法を話し手のモデルとする。それをもとにして文脈自由形言語に情報論的考察を加えていこう。

§ 3. 確率的文脈自由形言語

記法 1. 文脈自由形文法 (cfg) G を通常の記法にしたがって $G = (V_T, V_N, P, S)$ と表わす。ここで V_T は末端記号の集合, V_N は非末端記号の集合, P はプロダクションの集合, $S \in V_N$ は出発記号である。

定義 1. 確率的文脈自由形文法 (scfg) G_s は順序対 (G, φ) である。ここで $G = (V_T, V_N, P, S)$ は既約な cfg であり, φ は P から実数区間 $(0, 1]$ への写像である。ただし $\varphi(P_i)$, $P_i \in P$ を同一の左辺を持つ P_j について加えたものは 1 に等しいとする。 G を G_s の台という。

記法 2. $G = (V_T, V_N, P, S)$ を cfg とするとき $\alpha \in (V_T \cup V_N)^* - V_T^*$ から出発する G による左生成過程の全体を D_α , D_α の元で文生成過程となすものの全体を E_α と書く. D_α , E_α の元で n ステップのもの全体の全体をそれぞれ $D_\alpha^{(n)}$, $E_\alpha^{(n)}$ と書く. 特に $D_S, E_S, D_S^{(n)}, E_S^{(n)}$ をそれぞれ $D, E, D^{(n)}, E^{(n)}$ と書く.

記法 3. $d \in D_\alpha^{(n)}$ の第 i ステップ ($i=1, \dots, n$) において α ロケーション P_i が適用されているとき, $d = P_1 \dots P_n$ と書く.

定義 2. $G_\beta = (G, \varphi)$ を acfg とする. そのとき写像 $\psi_\alpha: D_\alpha \rightarrow (0, 1]$ によって定義する: $d = P_1 \dots P_n \in D_\alpha$ に対して $\psi_\alpha(d) = \varphi(P_1) \dots \varphi(P_n)$.

定義 3. 写像 $\psi'_\alpha: 2^{D_\alpha} \rightarrow [0, \infty)$ によって定義する: $M \subset D_\alpha$ に対して $\psi'_\alpha(M) = \sum_{d \in M} \psi_\alpha(d)$, したがって $\psi'_\alpha(\emptyset) = 0$.

ψ'_α は ψ_α の拡張であるが, 両者ともに ψ_α で表わす. ψ_S を ψ と書く. ψ の定義域を E に制限したものも ψ で表わす.

命題 1. $(E, 2^E, \psi)$ は測度空間である.

定義 4. E の元は, それによつて生成される文を対応させる写像を μ とする.

記法 4. $\mu(E)$ を $L(G)$ と書く.

μ によつて E から $L(G)$ 上に引きおこされる測度を考えることができる. すなわち,

命題 2. $\tilde{\psi} : 2^{L(G)} \rightarrow [0, \infty)$ を $M \subset L(G)$ に対して $\tilde{\psi}(M) = \psi(\mu^{-1}(M))$ と定義すると $(L(G), 2^{L(G)}, \tilde{\psi})$ は測度空間である.

$(L(G), 2^{L(G)}, \tilde{\psi})$ を G_{μ} によつて生成される確率的文脈自由形言語 (scfl) という.

命題 3. $\tilde{\psi}(L(G)) = \psi(E)$.

命題 4. $G_{\mu} = (G, \mu)$ を scfl とする. G が "あり" でなければ $(E, 2^E, \psi)$ と $(L(G), 2^{L(G)}, \tilde{\psi})$ は測度空間として同型である.

定理 1. 任意の σ -f.g. に μ によって生成される σ -f.g. $(L(G), 2^{L(G)}, \tilde{\mu})$ について $0 < \tilde{\mu}(L(G)) \leq 1$ が成立する.

定理 1 を証明するために二つの補題を用意する.

補題 1. 任意の σ -f.g. に対して $\psi(D^{(n)} - E^{(n)}) = \psi(D^{(n+1)})$ が成り立つ.

補題 2. 任意の σ -f.g. に対して $\psi(D^{(n)} \cup \bigcup_{i=1}^{n-1} E^{(i)}) = 1$.

(定理 1 の証明) 命題 3 により $\tilde{\mu}(L(G)) = \psi(E)$ であるから $0 < \psi(E) \leq 1$ を証明する. $E = \lim_{n \rightarrow \infty} \bigcup_{i=1}^n E^{(i)}$ であり, また $\{\bigcup_{i=1}^n E^{(i)}\} (n=1, 2, \dots)$ は単調非減少な集合列であるから測度の一般論により

$$\psi(E) = \psi\left(\lim_{n \rightarrow \infty} \bigcup_{i=1}^n E^{(i)}\right) = \lim_{n \rightarrow \infty} \psi\left(\bigcup_{i=1}^n E^{(i)}\right).$$

$M_1 \subset M_2 \subset E$ ならば $\psi(M_1) \leq \psi(M_2)$ だから $\{\psi(\bigcup_{i=1}^n E^{(i)})\} (n=1, 2, \dots)$ は単調非減少な実数列であり, また補題 2 により $\psi(\bigcup_{i=1}^n E^{(i)}) = 1 - \psi(D^{(n+1)}) \leq 1$ であるから $\psi(\bigcup_{i=1}^n E^{(i)}) (n=1, 2, \dots)$ は有界である. したがって $\lim_{n \rightarrow \infty} \psi(\bigcup_{i=1}^n E^{(i)})$ が存在する. この値を α とすれば, $0 < \psi(\bigcup_{i=1}^n E^{(i)}) \leq 1$ であ

るから $0 < \ell \leq 1$.

(証明終).

例 1. $G = (\{a, b\}, \{S\}, \{S \rightarrow a, S \rightarrow bSS\}, S)$ を台として持つ ncfg $G_A = (G, \varphi)$, $\varphi(S \rightarrow a) = p_1$, $\varphi(S \rightarrow bSS) = p_2$, $p_1 + p_2 = 1$, $p_1 > 0$, $p_2 > 0$ を考える. この G_A により生成される ncfl について $\tilde{\psi}(L(G))$ をおめると

$$\tilde{\psi}(L(G)) = \begin{cases} 1; & p_1 \geq \frac{1}{2} \text{ のとき,} \\ \frac{p_1}{1-p_1}; & \frac{1}{2} > p_1 > 0 \text{ のとき.} \end{cases}$$

例 1 は ncfl が確率空間に存在する場合とそうでない場合のあることを示しているが, ncfl が確率空間に存在ということの意味についてもう少し考察を加えてみよう. いま ncfg を文を生成するオートマトンと考え, このオートマトンは文が生成されたときに停止するものとする. $D^{(n)} - E^{(n)}$ は n ステップの生成過程のうち文生成過程でないものの全体であり, $\bigcup_{i=1}^n E^{(i)}$ は n ステップ以下の文生成過程の全体であり. したがって $\psi(D^{(n)} - E^{(n)})$ はこのオートマトンが n ステップ以下では停止しない確率, $\psi(\bigcup_{i=1}^n E^{(i)})$ は n ステップ以下で停止する確率であり, さてすでに示されたように

$$\tilde{\psi}(L(G)) = \lim_{n \rightarrow \infty} \psi\left(\bigcup_{i=1}^n E^{(i)}\right)$$

であり, $\psi(\bigcup_{i=1}^n E^{(i)})$ は n に對して單調非減少でありから,
 ‘ $\tilde{\psi}(L(G)) = 1$ ’ と ‘任意の $\varepsilon > 0$ に対して自然数 n_0 が存
 在して $\psi(\bigcup_{i=1}^{n_0} E^{(i)}) > 1 - \varepsilon$ ’ とは同値であり. すなわち, つ
 ぎの命題が得られた.

命題 5. つぎの (i), (ii) は同値であり.

- (i) $\Delta C f g$ G_δ によって生成される $\Delta C f g$ が確率空間となる.
- (ii) 任意の $\varepsilon > 0$ に対して自然数 n_0 が存在し, オートマト
 ンとしての G_δ が n_0 ステップ以下で停止する確率が $1 - \varepsilon$ より大きい.

命題 5 の (ii) を直感的に言えば ‘ G_δ が殆んど確実に有限
 ステップで停止する’ となる. 話し手のモデルとしてはその
 ような G_δ のみを考えておけば十分であろうから, 以下の議
 論では $\Delta C f g$ はつねに確率空間であるとす.

§ 4. 言語の情報量

4.1 文生成過程および言語のエントロピー

定義 5. $\Delta C f g$ G_δ による文生成過程の全体 E のエントロ
 ピー $H(E)$ を

$$H(E) = - \sum_{d \in E} \psi(d) \log_2 \psi(d)$$

と定義する.

$H(E)$ は文生成過程 μ とつ当りの平均情報量, すなわち P -マーカーを一つ受けと, t ときに得られる情報量の期待値である.

定義 6. $G_\lambda = (G, \varphi)$ を scfg とする. G_λ によつて生成される scfl のエントロピー $H(L(G))$ を

$$H(L(G)) = - \sum_{w \in L(G)} \tilde{\varphi}(w) \log_2 \tilde{\varphi}(w)$$

と定義する.

$H(L(G))$ は G_λ によつて生成される文 μ とつ当りの平均情報量である.

あとでわかるように一般に $H(L(G)) \leq H(E)$ であるから $H(E)$ が有限ならば $H(L(G))$ も有限である. しかし $H(E)$ は常に有限とは限らない. P -マーカーを一つ受けと, t ときに, それから無限の情報を得ることが期待できるということは一見不合理のようであるが, 以下に述べることでこの疑問を解決するであろう.

定義 7. $N: E \rightarrow \mathbb{R}$ をつきのように定義する: $d \in E$ に対し $d \in E^{(n)}$ のとき $N(d) = n$.

N は $(E, 2^E, \psi)$ 上の確率変数であり, $N = n$ となる確率 $\psi\{N = n\}$ は $\psi(E^{(n)})$ で与えられる. N の平均値を $M(N)$ と書く: $M(N) = \int_E N d\psi = \sum_{n=1}^{\infty} n \psi(E^{(n)})$.

定理 2. $H(E)$ は $M(N)$ が有限なとき, かつそのときに限り有限である.

$M(N)$ は文生成過程の平均ステップ数であるから, G_0 を話し手のモデルと考える以上 $M(N)$, したがって $H(E)$ は有限と考えてよいであろう.

4.2 言語のあいまいさ

言語にはあいまいさのある場合がある. これは同一の文を生成する複数個の生成過程が存在する場合で, 文が与えられてもそれがどの生成過程によるものか聞き手が決定できず, あいまいさが残るわけである.

ある acfg によって生成される二つの測度空間 $(E, 2^E, \psi)$, $(L(G), 2^{L(G)}, \tilde{\psi})$ について考える. $w \in L(G)$ が生成されたと

この条件のもとでの $d \in E$ の確率 $P_w(d)$ は条件付確率の定義により $P_w(d) = \psi(\{d\} \cap h^{-1}(w)) / \psi(h^{-1}(w))$ で与えられる。

定義 8. $w \in L(G)$ のあいまい度 $A(w)$ を

$$A(w) = - \sum_{d \in E} P_w(d) \log_2(P_w(d))$$

で、また $L(G)$ の平均あいまい度 $A(L(G))$ を

$$A(L(G)) = \sum_{w \in L(G)} \tilde{\psi}(w) A(w)$$

で定義する。

命題 6. $A(L(G)) = H(E) - H(L(G))$ 。

命題 7. $A(L(G)) = 0$ となるのは G があいまいでないとき、かつそのときに限る。

謝辞 この研究を進めるに際しては、NHK総合技術研究所内 数理工学研究会の諸氏に討論をしていただいた。とくに例1の結果は上坂吾則氏および坂井徹男氏に負うところが大きいことを記して謝意を表したい。