北星短大紀 *Journal of Hokusei Jr. Col. Vol.29 : 67－73 , 1993*

知能作用過程のシュミレーション

その2　文字情報に関する理論的考察

寺岡　宏　*矢吹哲夫

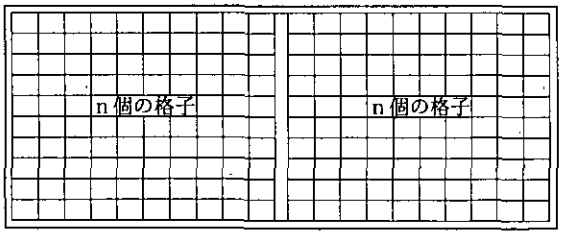
既述の論文その1[1]において、文字情報の読み取りのための文章作成をZipfの法則[2]に基いて行なった。本稿では、その入力文字情報の作製を人工言語系の設定とみなし、その人工言語系のもつ情報量、正確には1文字当たりの平均情報量を算定する。この1文字当たりの平均情報量は、Shannon[3]によってはじめて情報理論の中に導入され確立されたエントロピー（情報エントロピー）に他ならない。そこで、本稿においては、この情報エントロピーの理論的出発点として、物理学の一分野である古典統計力学におけるエントロピーの基本概念を概説し、次に情報理論への応用、その理論と具体的方法を述べる。最後に、その方法に基いて、我々の設定した人工言語系のもつ情報エンドロピーを算出する。

●情報量とエントロピー

一つの文字や一つの文字集団の担う平均情報量を測る物差しとして〝エントロピー〟という量が定義されている。この〝エントロピー〟は、まず熱力学の分野の中で考案され、その後の古典統計力学の発展に伴い厳密に定式化された。これによってより幅の広い応用性を獲得し、今日では、経済学や生物学そして本論文でとりあげる情報理論の分野にまで浸透する包括性の高い概念として定着している。本論文においては、はじめに出発点である古典統計力学におけるエントロピー概念の定式化を具体例を用いて明確にし、次に本題の〝情報エントロピー〟の概念とその計算方法を敷衍することにする。

A：古典統計力学とエントロピー

ある物理系の相異なる巨視状態の出現確率は、異なる巨視状態各々の中に含まれる微視状態の数に比例する。例えば巨視状態M_1、M_2があるとき、その中に含まれる微視状態の数を各々$\Omega(M_1)$、$\Omega(M_2)$とすると、M_1、M_2が出現する確率の比は$\Omega(M_1) : \Omega(M_2)$



＊北海道大学・理学部・物理学科・大学院博士課程

である。このことを具体的に、非可逆過程の例として広く引用される次の例で考察する。

一定の全容積Vの直方体の容器に気体を詰める。容器の真中より左の部分を Left、右の部分を Right として、気体のとり得る2つの異なる巨視状態として、すべて Left に含まれる場合をM_1、それ以外の場合即ち容器全体のどこに気体分子がいてもいい場合をM_2とする。このとき、この2つの巨視状態M_1、M_2の中に含まれる微視状態の数Ω（M_1）、Ω（M_2）を評価しよう。

ここで気体分子の総数をNとして気体分子1個の体積をvとする。このとき、気体分子1個を収容する〝格子〟を想定すると、Left に存在する格子数nは$n=\frac{V}{2}\div v=\frac{V}{2v}$で与えられる。そして容器全体には2n個の格子が存在することになる。N個の気体分子がすべて Left に収容され得るためには$n>N$でなければならないが，気体という前提が崩れない限りこれは成立している。このとき、Ω（M_1）は、N個の気体分子のすべてが Left の中のn個の格子のいずれかに収容される場合の数であるが、それはn個の格子の各々がN個の気体分子のどの分子にあてがわれるかの総数に他ならない。すなわち，Ω（M_1）$={}_nP_N$である。（ここで${}_nP_N$は順列を表わしている。）同様にして、Ω（M_2）は2n個の格子の各々がN個の気体分子のどの分子にあてがわれるかの総数であり、Ω（M_2）$={}_{2n}P_N$である。この両者の比Ω（M_1）：Ω（M_2）$={}_nP_N:{}_{2n}P_N$は、Nが大きくなるにつれて天文学的な数字の開きとなる。このことを下に式で表わそう。気体分子のすべてが Left に含まれる確率（巨視状態M_1の確率）をP（M_1）とすると、

$$P(M_1)=\frac{\Omega(M_1)}{\Omega(M_2)}$$

$$=\frac{{}_nP_N}{{}_{2n}P_N}\to 0\ (N\to\infty)$$

（上式で、N→∞という条件は、気体分子数Nの日常的な大きさであるアボガドロ数の規模で十分適用可能である。）

これは、仮に容器全体に分布する気体をしきりで人為的に左側に押しこめたとしても、しきりを開くや否や気体は速やかに容器全体に広がり、元の左側に押しこめられた状態に戻る逆過程の確率は殆ど0であるという非可逆過程を表わしている。

今の例で、Nが十分大きくなると微視状態の数Ω（M_1）、Ω（M_2）は膨大な数になる。一般に微視状態の数Ω（M）はそのままでは大き過ぎて扱いづらい。そこでその対数 log Ω（M）で新しい量が定義された。これが則ちエントロピーである。

〈定義〉

巨視状態MのエントロピーS_Mは次式で与えられる。

$$S_M=k\log\Omega(M)$$

（注）対数は通常自然対数である。またkは、古典統計力学に先行する熱力学で定義されていたエントロピーと結びつけるためのものでボルツマン定数と呼ばれる。ここではkは便宜上のものとみなしてよい。

上述の議論から分かるように、エントロピーSの大きい巨視状態ほど実現確率は高い。

B：情報量とエントロピー

今、m種類の文字または文字集団A_1、A_2、…、A_mが各々n_1、n_2、…、n_m個ある。このとき、これらのA_1、A_2、…、A_mを一列に並べる並べ方の総数を考える。異なる並べ方には異なる情報を対応させることができるから、この並べ方の総数が、与えられた素材

A_1、A_2、…、A_mから構成し得る情報の総数である。

	出現確率
A_1…n個	$\frac{n_1}{N}=P_1$
A_2…n個	$\frac{n_2}{N}=P_2$
⋮	⋮
A_m…n個	$\frac{n_m}{N}=P_m$

（$n_1+n_2\cdots+n_m=N$）全体N個の中から、任意にA_1、A_2、…、A_mをとり出して一列に並べる並べ方の総数は次式で与えられる。

$$\text{並べ方の総数}=\frac{N!}{n_1!\,n_2!\cdots n_m!}$$

（同種のものを含む順列の公式より）

次にこの対数をとりStirlingの近似(注)公式をあてはめる。

（注）Stirlingの近似方式
Nが十分大きいとき
$\log N!\fallingdotseq N(\log N-1)$

$$\log\frac{N!}{n_1!\,n_2!\cdots n_m!}$$

$$=\log N!-\sum_{i=1}^{m}\log n_i!$$

$$=N(\log N-1)-\sum_{i=1}^{m}n_i(\log n_i-1)$$

（∵Stirlingの近似公式より）

$$=N\log N-N-\sum_{i=1}^{m}n_i\log n_i+\sum_{i=1}^{m}n_i$$

$$=N\log N-N-\sum_{i=1}^{m}n_i\log n_i+N$$

（$\because\sum_{i=1}^{m}n_i=N$）

$$=N\log N-\sum_{i=1}^{m}n_i\log n_i$$

$$=N\left(\sum_{i=1}^{m}\frac{n_i}{N}\log N-\sum_{i=1}^{m}\frac{n_i}{N}\log n_i\right)$$

$$=N\sum_{i=1}^{m}\frac{n_i}{N}\log\frac{N}{n_i}$$

$$=-N\sum_{i=1}^{m}\frac{n_i}{N}\log\frac{n_i}{N}\cdots①$$

よって、m種類の素材A_1～A_mを用いたN個の文字列あるいはN個の文字集団の列により表現でき得る情報の総数に対数をとったものが①式となる。このとき、文字1個当たりの情報数の対数（あるいは文字集団1つ当たりの情報数の対数）は①式をNで割って得られ、更に$\frac{n_i}{N}$が素材A_iの出現確率P_iに等しいことを考慮すると、

文字（または文字集団）1個当たりの情報数の対数$=-\sum_{i=1}^{m}P_i\log P_i$……②（$P_i$は$A_i$の出現確率）となる。

②式は、情報理論におけるエントロピーの定義式であり、以後本論文では情報エントロピーと略称する。

換言すると、

情報エントロピー＝log（1個当たりの情報数）……③

となる。

③式は、その形から古典統計力学のエントロピー＝$k$log（状態の総数）の自然な拡張であることが分かる。

上式②で定義された情報エントロピーはシャノンによって初めて導入され、平均情報量とも呼ばれる。

C：自然言語とエントロピー

Bで与えられた情報エントロピーの定義式「情報エントロピー$=\sum_{i=1}^{m}P_i\log P_i$」のlogの底は通常2をとる。このとき情報エントロピー（＝平均情報量）の単位はビットになる。今後、本論文では情報エントロピーをSで表わすがそれは、次式で定義されるものとする。

$$S=-\sum_{i=1}^{m}P_i\log_2 P_i$$

この情報エントロピーSに対して、次の不等式が成り立つ。

$0 \leqq S \leqq \log_2 m$ (4)

（mは異なる文字または文字集団A_iの種類の数）

ここでSが、最小値S＝0になる場合と最大値$S=\log_2 m$になる場合について考察しておく。

（i）最小値S＝0になるとき

ある特定の1文字(または文字集団)A_kしか現われない文字列のときである。則ち式で書くと、

$$P_i = \begin{cases} 1 & (i=k) \\ 0 & (i \neq k) \end{cases}$$ のときである。

（ii）最大値$S=\log_2 m$になるとき

すべての文字（または文字集団）A_iの出現確率P_iが等しくなるときである。則ち式で表わすと、

$$P_i = \frac{1}{m} \quad (i=1、2、\cdots、m)$$

のときである。

この最小値S＝0、最大値$S=\log_2 m$は両極端の場合であり、一般の自然言語ではこの間の値をとる。その一例として英語の場合をみてみよう。

文字集団（＝単語）ではなく文字（＝アルファベット）に着目して情報エントロピーを考えると、素材A_i（i＝1、2、…、m）は、アルファベット26文字（a、b、c、…、z）に空白を加えたものになる。このとき素材数は、m＝27である。このアルファベットに空白を加えた27文字がすべて等確率に出現するとき、情報エントロピーSは最大値となり、

$$S=\log_2 27=4.75 \text{ビット}$$

になる。しかし実際の英語の文章では、これらの27文字がすべて等確率で現われることはなく、文字ごとに出現頻度が異なる。例えばイギリス、アメリカで刊行されている英字の新聞や雑誌、小説等の広範な出版物からとった27文字の出現頻度統計を用いてみよう。（表(1)を参照）

表1　英語の文字の出現確率

文　字	確　率	文　字	確　率
スペース	0.1859	N	0.0574
A	0.0642	O	0.0632
B	0.0127	P	0.0152
C	0.0218	Q	0.0008
D	0.0317	R	0.0484
E	0.1031	S	0.0514
F	0.0208	T	0.0796
G	0.0152	U	0.0228
H	0.0467	V	0.0083
I	0.0575	W	0.0175
J	0.0008	X	0.0013
K	0.0049	Y	0.0164
L	0.0321	Z	0.0005
M	0.0198		

F.M.Reza著、鶴見・大石訳「確率、情報、コード」（共立出版、1973）

この出現頻度に基いて情報エントロピーSを計算するとS＝4.03ビットとなり最大値4.75ビットより小さくなっているのが分かる。

因みに、表(1)の統計は我々の論文その1で議論したZipfの法則にほぼ従っている。実際の英語の文章においては、この各文字の出現頻度の差に加えて、各文字間の相関が無視できない。例えばアルファベットQの次にUが現われる確率は1であるが、Qの次にU以外の他の文字が現われる確率は0である。このように各素材が相関をもちながら出現するマルコフ連鎖も考慮して、英語の情報エントロピーを計算すると、S＝3.32ビットで最大値4.75ビットに比べて更に小さくなる。

英語の例で見られるように、実際の自然言語においては、情報エントロピーSはその最

大値を大きく下回っている。このことは、1文字当たりの平均情報量の減少を意味し、情報の能率の低下を表わしている。しかし、このことは同時に、スペルにささいな誤まりがあっても推測でそれを正していける柔軟性というべきものを言語が備えていることを意味している。情報の能率を表わす量として効率があり、情報の柔軟性を表わす量として冗長度（redundancy）がある。この2つは、

$$\begin{cases} \text{効率}\ \eta = S / S\max \\ \text{冗長度} = 1 - \eta = 1 - S / S\max \end{cases}$$

で定義される。

英語の例では、文字相関を無視すれば、

効率 $\eta = 4.03 / 4.75 = 0.85$

冗長度 $= 1 - 0.85 = 0.15$

となる。

D：まとめ、入力データの情報エントロピー

本論文で設定されている人工言語系の情報エントロピーについて以下考察し本論文のまとめとする。

本論文その1のD[1]で詳述したように、我々の用いた人工言語系においては、それぞれに定められた長さの文字集団を設定し、それらにZipfの法則に基く出現頻度を与えるという方法で情報(＝読み取りのための文章)を作製している。このとき、文字集団を素材$A_1 \sim A_m$（mは用いた文字集団の種類の数）として、Cで定義された情報エントロピーを計算した結果が表2である。以後、素材としての文字集団を単語とみなして考察する。

表2の第1列の全単語数は文字集団の種類の数mを表わし、第2列はすべての単語の出現確率を等確率としたときのエントロピーS、則わち前述の最大値Smaxである。第3列はZipfの法則に基く出現確率をもつときの情報エントロピーSの値であり、第4列、第5列は各々効率と冗長度である。一例として、この表の2行目を挙ると、これは単語（文字集団）の個数m＝1,000のときのデータであり、このときS＝7.491、Smax＝9.968で、効率η＝0.7515、冗長度$1-\eta$＝0.2485である。この表2全体から、Zipfの法則の反映により、情報エントロピーS（1単語当りの平均情報量）の効率は70%～80%に低減していることが分かる。このことは、逆に言えば、

表2　入力された文字情報のエントロピー

（1単語当たりの情報量）

全単語数	エントロピー (Non Zipf)	エントロピー (Zipf)	効率η	冗長度$1-\eta$
200	7,646	5,987	0.7830	0.2170
1000	9,968	7,491	0.7515	0.2485
2000	10,969	8,117	0.7400	0.2600
3000	11,554	8,479	0.7339	0.2661
4000	11,969	8,733	0.7296	0.2704
5000	12,291	8,930	0.7265	0.2735
6000	12,554	9,090	0.7241	0.2759
7000	12,776	9,224	0.7220	0.2780
8000	12,969	9,341	0.7203	0.2797
10000	13,290	9,534	0.7174	0.2826

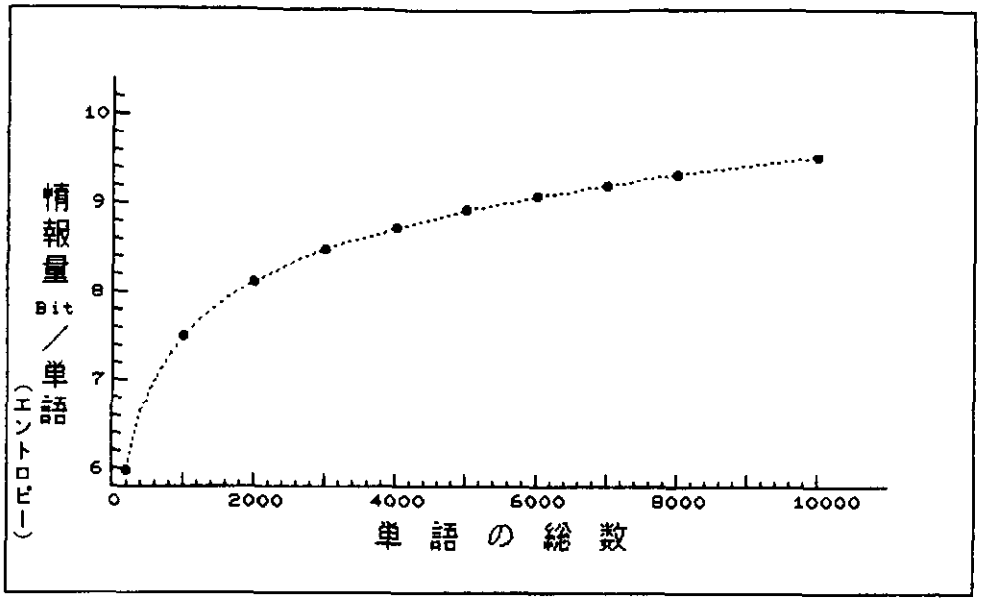


図1　情報エントロピーの推移

入力された文字情報を人工言語系とみたとき、20%から30%の柔軟性（冗長度）を言語系が備えていることを意味する。また全単語数mの増加に伴い、効率は低下し冗長度は増加している。これは一定レベル以上の柔軟性を言語が獲得するためには、一定数以上の単語数が必要であることを示唆している。

図1は、全単語数（異なる文字集団数m）が大きくなるに従って情報エントロピーS（1単語当りの平均情報量）がどのように推移するかをグラフにしたものである。この図1をみると、全単語数mが増加するに従って情報エントロピーSは増加するが、その増加率は徐々に緩かなカーブを描き、全単語数が1万語あたりでは、情報エントロピーはほぼ飽和に達していることが分かる。

〈参考文献〉

(1)　寺岡　宏、矢吹哲夫　1992　知能作用過程のシュミレーション　その1文字情報の読み取り　北星短大紀要：28　113－129

(2)　G. K. Zipf. 1949 Human behavior and the principle of the least effort. Addison, Wesley Press Inc.

(3)　C. E. Shannon. 1948 Mathematical theory of communication. Bell system Tech J. 27　379 and 623

(4)　小野厚夫、川口正昭　1982　情報科学概論　p.12～p.13　培風館

Simulation of the Operating-processes of the Human Intellect

Part 2. Theoritical Study on Literal Information

Hiroshi Teraoka and Tetuo Yabuki

In this paper , we discuss about " entropy" which is useful quantity for calculating an amount of information. The " entropy" is the technical term in physics which was originally devised in thermodynamics and was strictly defined in classical statistical mechanics. Nowadays, it is one of the most universal quantity, which is used not only in physics, but also in many other fields such as economics or biology and so on. In the theory of information, the " entropy" has been first introduced by Shannon, in order to calculate an amount of information.

In this paper , we first explain how to define it in classical statistical mechanics, and secondly how to use it as a measure for calculating an amount of information.
Finnaly, we calculate the average amount, that is to say, entropy in our literal information system. We show the result of it's calculation as a list.