情報理論レポート１(情報量)解答例

提示：2008/10/8(水)　提出：2008/10/15(水)

１．英文の情報量

(1)単語wが情報源Aから生成されたときの自己情報量を$I_A(w)$、情報源Bから生成されたときの自己情報量を$I_B(w)$と表す。このとき、次の自己情報量をそれぞれ求めよ。

（解）

記号αが情報源Aより生成されたときの自己情報量を$i_A(\alpha)$と書き、情報源Bより生成されたときの自己情報量を$i_B(\alpha)$とかく。

$$
\begin{aligned}
&I_A(electronics)\\
&= i_A(e)+i_A(l)+i_A(e)+i_A(c)+i_A(t)+i_A(r)+i_A(o)\\
&\quad +i_A(n)+i_A(i)+i_A(c)+i_A(s)\\
&= -2\times\log\left(0.1183\right)-\log\left(0.0402\right)-2\times\log\left(0.0331\right)-\log\left(0.0849\right)\\
&-\log\left(0.0645\right)-\log\left(0.0731\right)-\log\left(0.0702\right)-\log\left(0.0740\right)-\log\left(0.0723\right)\\
&\simeq 43.3 \quad \left[bit\right]
\end{aligned}
$$

$$
\begin{aligned}
&I_A(\inf ormation)\\
&= i_A(i)+i_A(n)+i_A(f)+i_A(o)+i_A(r)+i_A(m)+i_A(a)\\
&\quad +i_A(t)+i_A(i)+i_A(o)+i_A(n)\\
&= -2\times\log\left(0.0740\right)-2\times\log\left(0.0702\right)-\log\left(0.0206\right)\\
&-2\times\log\left(0.0731\right)-\log\left(0.0645\right)-\log\left(0.0277\right)\\
&-\log\left(0.0876\right)-\log\left(0.0849\right)\\
&\simeq 44.5 \quad [bit]
\end{aligned}
$$

$$
\begin{aligned}
&I_A(system)\\
&= i_A(s)+i_A(y)+i_A(s)+i_A(t)+i_A(e)+i_A(m)\\
&= -2\times\log\left(0.0723\right)-\log\left(0.0186\right)-\log\left(0.0849\right)\\
&\quad -\log\left(0.1183\right)-\log\left(0.0277\right)\\
&\simeq 25.1\,[bit]
\end{aligned}
$$

情報源Bでは、各記号が均等に現れるので、次式が成り立つ。

$$
\begin{aligned}
&i_{B}(a) = i_{B}(b) = \cdots = i_{B}(z) \\
&= -\log P(\alpha) \\
&= -\log \frac{1}{26} \\
&\simeq 4.70\ [bit]
\end{aligned}
$$

よって、情報源 B から生成されたの単語の情報量は文字数だけで定まる。

$$I_{B}(electronics) = 11 \times i_{B}(\alpha) \simeq 51.7\ [bit]$$

$$I_{B}(\inf ormation) = 11 \times i_{B}(\alpha) \simeq 51.7\ [bit]$$

$$I_{B}(system) = 6 \times i_{B}(\alpha) \simeq 28.2\ [bit]$$

これらの結果の注意点は、以下のとおり。
○記号の出現確率が不均等な場合、記号数が同じでも自己情報量は異なる。
○同じ記号列だとしても、どの情報源から生成されたかによって、自己情報量は異なる。

(2)情報源 A のエントロピー（平均情報量）$H(A)$ および情報源 B のエントロピー $H(B)$ をそれぞれ求めよ。

$$
\begin{aligned}
&H(A) = -\sum_{\alpha \in A} P(\alpha) i_{A}(\alpha) \\
&= -\sum_{\beta \in B} P(\alpha) \log P(\alpha) \\
&= -\Big(0.1183\Big)\log\Big(0.1183\Big) - \Big(0.0876\Big)\log\Big(0.0876\Big) - \Big(0.0849\Big)\log\Big(0.0849\Big) - \Big(0.0740\Big)\log\Big(0.0740\Big) \\
&-\Big(0.0731\Big)\log\Big(0.0731\Big) - \Big(0.0723\Big)\log\Big(0.0723\Big) - \Big(0.0702\Big)\log\Big(0.0702\Big) - \Big(0.0645\Big)\log\Big(0.0645\Big) \\
&-\Big(0.0441\Big)\log\Big(0.0441\Big) - \Big(0.0402\Big)\log\Big(0.0402\Big) - \Big(0.0402\Big)\log\Big(0.0402\Big) - \Big(0.0331\Big)\log\Big(0.0331\Big) \\
&-\Big(0.0306\Big)\log\Big(0.0306\Big) - \Big(0.0277\Big)\log\Big(0.0277\Big) - \Big(0.0210\Big)\log\Big(0.0210\Big) - \Big(0.0206\Big)\log\Big(0.0206\Big) \\
&-\Big(0.0189\Big)\log\Big(0.0189\Big) - \Big(0.0186\Big)\log\Big(0.0186\Big) - \Big(0.0176\Big)\log\Big(0.0176\Big) - \Big(0.0171\Big)\log\Big(0.0171\Big) \\
&-\Big(0.0104\Big)\log\Big(0.0104\Big) - \Big(0.0093\Big)\log\Big(0.0093\Big) - \Big(0.0021\Big)\log\Big(0.0021\Big) - \Big(0.0018\Big)\log\Big(0.0018\Big) \\
&-\Big(0.0009\Big)\log\Big(0.0009\Big) - \Big(0.0009\Big)\log\Big(0.0009\Big) \\
&\simeq 4.19 \quad [bit\ /\ 記号]
\end{aligned}
$$

$$
\begin{aligned}
&H(B) = -\sum_{\beta \in B} P(\beta) i_B(\beta) \\
&= -\sum_{\beta \in B} P(\beta) \log P(\beta) \\
&= -26 \times \frac{1}{26} \log \frac{1}{26} \\
&= \log 26 \\
&\simeq 4.70 \ [bit \ / \ \text{記号} \ ]
\end{aligned}
$$

よって、$H(A) \leq H(B)$ が成り立つ。通常の英文の（平均の）情報量は、情報源 B から得られる同じ長さの記号列の（平均の）情報量より、小さいことを意味する。

一般に、同じ記号を生成するような事象系同士 X,Y において、事象の実現確率が不均等な事象系 X のエントロピーより、確率が均等な事象系 Y のエントロピーのほうが大きい。すなわち、次式が成り立つ。

$H(X) \leq H(Y)$

なお、記号の出現確率が均等な場合、ある記号 1 個の自己情報量は、エントロピーと等しい。（平均の意味から考えて、当然である。）