The 25th Symposium on Information Theory
and Its Applications (SITA2002)
Ikaho, Gunma, Japan, Dec. 10–13, 2002

# 単語単位で系列を出力する情報源の性質について

# A Property of Sources which emit Data Sequences by Word Unit

石田 崇* 後藤 正幸† 松嶋 敏泰* 平澤 茂一*

Takashi ISHIDA Masayuki GOTO Toshiyasu MATSUSHIMA Shigeichi HIRASAWA

**Abstract**— Recently, *word-valued source* is proposed as a new class of source models. A word-valued source is defined as a source which has a probability distribution over *word set*.

In this paper, we consider a word-valued source whose word set is not prefix-free and show its properties.

**Keywords**— source coding, word-valued source, entoropy rate, asymptotic equipartition property (AEP)

## 1 はじめに

近年，情報源符号化における情報源モデルとして“言語アルファベット情報源[1]”や“単語単位の情報源[2]”が提案されている．これらの情報源は，情報源アルファベットを有限個連接した“単語”を定義し，この単語集合上に確率分布を与える情報源モデルである．これは，「一般に圧縮の対象となるファイルなどに対しては，単語単位で確率構造を持つ情報源を仮定するのが自然であろう」という考えに基づいている．

西新ら[1]は，可算無限の単語集合について単語単位の定常無記憶（i.i.d.）情報源を“言語アルファベット情報源”と定義し，単語集合が語頭条件を満たすときにはこの情報源がユニフィラー情報源と等価となることや，そのエントロピーレートを導いている．また，後藤ら[2]，石田ら[3],[4]は有限な単語集合について単語単位で定常や定常エルゴードを仮定して，“単語単位の情報源”と呼び，そのエントロピーレートを導いた．さらにこの情報源に対するLZ符号やベイズ符号の漸近的性質についても明らかにしている．

以上の議論では，単語集合には基本的に語頭条件を仮定していた．ここで，語頭条件とは単語集合に属する任意の単語が他の単語の語頭と一致しないことをいう．これによって，単語系列とシンボル系列が1対1に対応し，情報源の解析が容易になるためである．一方，単語集合が語頭条件を満たさない場合の情報源の性質については，西新ら[1]によって，エントロピーレートの上界式が与えられているのみである．

そこで本稿では，まず，西新ら[1]の結果を拡張して語頭条件を満たさない定常エルゴード言語アルファベット情報源に対してエントロピーレートの上界式を与える．次に具体的に語頭条件を満たさないi.i.d.言語アルファベット情報源モデルを定式化してそのエントロピーレートの下界を導出し，その性質について考察を行う．

## 2 言語アルファベット情報源[1],[2]

西新ら[1]が定義した“言語アルファベット情報源”は単語単位で定常無記憶(i.i.d.)を仮定している．一方，後藤ら[2]，石田ら[3],[4]は，単語単位で定常や定常エルゴードとなる情報源を議論の対象としている．そこで本稿では単語単位で定常な情報源まで拡張して，これを“言語アルファベット情報源”と呼ぶことにする．

**定義 1. (言語アルファベット情報源)**

確率変数 $Y$ の無限系列 $\boldsymbol{Y} = Y_1Y_2Y_3\cdots$ を可算アルファベット $\mathcal{Y}$ 上に値をとる定常情報源とする．$\mathcal{X}$ を有限アルファベットとし，$\mathcal{X}^*$ は $\mathcal{X}$ 上の有限系列すべての集合を表すものとする．写像 $\phi$ を $\phi : \mathcal{Y} \to \mathcal{X}^*$ で定める．さらに確率変数 $W = \phi(Y)$ を定義し，$W$ は $\mathcal{W}(\subseteq \cup_{i=0}^{\infty}\mathcal{X}^i)$ 上に値をとるものとする．このとき $W$ の実現値 $w$ を単語，$\mathcal{W}$ を単語集合と呼ぶ．写像 $\phi$ を $Y$ の無限系列 $\boldsymbol{Y}$ に対して拡張して[1]，言語アルファベット情報源 $\boldsymbol{X} = X_1X_2X_3\cdots$ を以下で定義する．

$$\boldsymbol{X} \stackrel{\text{def}}{=} \phi(\boldsymbol{Y}) \tag{1}$$

ここで，

$$\phi(\boldsymbol{Y}) = \phi(Y_1)\phi(Y_2)\cdots = W_1W_2\cdots = \boldsymbol{W} \tag{2}$$

である．

また，$\boldsymbol{Y}$，$\boldsymbol{W}$，$\boldsymbol{X}$ の確率分布を

$$P_{Y^m}(y^m) \stackrel{\text{def}}{=} \Pr\{Y^m = y^m\} \tag{3}$$

$$P_{W^m}(w^m) \stackrel{\text{def}}{=} \Pr\{W^m = w^m\} \tag{4}$$

$$P_{X^n}(x^n) \stackrel{\text{def}}{=} \Pr\{X^n = x^n\} \tag{5}$$

で表す[2]．このとき，

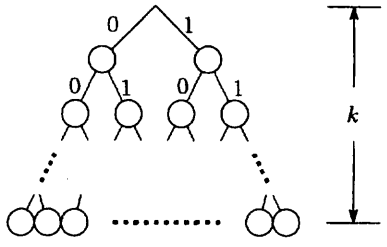
$$P_{W^m}(w^m) = \sum_{\{y^m : w^m = \phi(y^m)\}} P_{Y^m}(y^m) \tag{6}$$

* 早稲田大学 理工学部 経営システム工学科，〒169-8555 東京都新宿区大久保 3-4-1. School of Science and Engineering, Waseda University, 3-4-1 Okubo Shinjyuku-ku, Tokyo 169-8555, Japan.

† 武蔵工業大学 環境情報学部 情報メディア学科，〒224-0015 神奈川県横浜市都筑区牛久保西 3-3-1. Musashi Institute of Technology, Faculty of Environmental and Information Studies, 3-3-1 Ushikubo-nishi, Tuzuki Ward, Yokohama City, 224-0015, Japan.

[1] すなわち，$\phi(\boldsymbol{Y})$ を $\phi(Y_i)$ の連節 $\phi(Y_1)\phi(Y_2)\cdots$ とする．

[2] 本稿では確率変数 $X, Y, W$ やその実現値 $x, y, w$ について長さ $n$ の有限系列を $X^n = X_1X_2\cdots X_n$ のように書く．




$$\mathcal{W} = \{0, 1, 00, 01, 10, 11, 000, \cdots, \underbrace{11\cdots1}_{k}\}$$

$$\|\mathcal{W}\| = \sum_{i=1}^{k} 2^i = 2^k - 1$$

図 1: 単語集合 $\mathcal{W}$ の 2 分木表現

る単語集合クラスに対するベイズ符号の構成法 [3] などがなされている.

しかし, 言語アルファベット情報源はもともとテキストデータなど, 実際に圧縮の対象となるデータの確率構造をより反映させた情報源モデルとして提案されており [2], 一般には語頭条件に制約されないモデルについても考察を行う必要がある.

語頭条件を満たさない言語アルファベット情報源については, 西新ら [1] による i.i.d. 言語アルファベット情報源のエントロピーレート上界式 (式 (10),(11)), その拡張として定常エルゴード言語アルファベット情報源のエントロピーレート上界式 (定理 1) が得られているのみであり, その性質についてはいまだ明らかではないところが多い.

そこで, 本稿では語頭条件を満たさない言語アルファベット情報源をモデル化し, 最も単純な i.i.d. 言語アルファベット情報源に対してその性質の考察を行う.

### 3.1 語頭条件を満たさない i.i.d. 言語アルファベット情報源モデル

本稿で考察する語頭条件を満たさない単語集合 $\mathcal{W}$ 上の i.i.d. 言語アルファベット情報源モデルを以下のように設定する.

**[語頭条件を満たさない i.i.d. 言語アルファベット情報源]**

有限アルファベット $\mathcal{Y}$ 上に値をとる確率変数系列 $\boldsymbol{Y}$ を i.i.d. 情報源, $\mathcal{X} = \{0, 1\}$ とする. ここで, $\mathcal{Y}$ から $\mathcal{W}$ への 1 対 1 写像 $\phi$ を $\phi : \mathcal{Y} \to \mathcal{W} = \cup_{i=1}^{k} \mathcal{X}^i$ で与える.

単語集合 $\mathcal{W}$ を深さ $k$ の完全 2 分木で表現し, 各ノードに単語 $w$ を割り振る (図 1 参照). このとき各単語 $w$ は語頭条件を満足しない. このとき, $\boldsymbol{W}$ のエントロピー (単語単位のエントロピー) は

$$H(\boldsymbol{W}) = -\sum_{w \in \mathcal{W}} P_W(w) \log P_W(w) \ (= H(\boldsymbol{Y})) \quad (21)$$

である. □

この言語アルファベット情報源は語頭条件を満たしていないため, 西新らによるエントロピーレートの上界式 (式 (11)) は成り立っているものの, $H(\boldsymbol{X})$ の収束性などは議論されておらず, この情報源の性質についてはいまだ明らかにされていないところが多い.

そこで, 語頭条件を満たさない言語アルファベット情報源 $\boldsymbol{X}$ についてエントロピーレートの性質を明らかにすることを目的として, $H(\boldsymbol{X})$ の下界についての考察を行う.

### 3.2 エントロピーレート $H(\boldsymbol{X})$ の下界

完全 2 分木の各ノードに対応付けされたすべての単語 $w$ について, ノードの深さ $s$ $(= 1, 2, \cdots, k)$ とその深さにあるノードに対して左から右へ順番につけられた番号 $t \in \mathcal{T}_s = \{1, 2, \cdots, 2^s\}$ $(s = 1, 2, \cdots, k)$ を用いて $w_{s,t}$ と表し, その出現確率を $P_W(w_{s,t})$ と書く.

言語アルファベット情報源 $\boldsymbol{X}$ の $H(\boldsymbol{X})$ の下界を与えるために, 十分長い長さ $n$ のシンボル系列 $x^n$ の出現確率 $P_{X^n}(x^n)$ を評価する. 概略を以下に示す.

i.i.d. にしたがって出力された単語系列 $w^m$ からシンボル系列 $x^n$ への写像は単語集合 $\mathcal{W}$ が語頭条件を満たしていないことから 1 対 1 とはなっていない. しかし, 十分長い長さ $m$ の単語系列 $w^m$ について漸近等分割性 (AEP) が成り立つことから, $x^n$ へ写像される単語系列 $w^m$ の数の上界を与えることによって, シンボル系列 $x^n$ の出現確率 $P_{X^n}(x^n)$ を評価することが可能である.

いま, 単語系列 $w^m$ は i.i.d. 情報源からの出力系列であり, 十分長い長さ $m$ の単語系列 $w^m$ に含まれる各単語 $w$ の個数 $N(w|w^m)$ は,

$$N(w|w^m) \simeq m P_W(w) \quad (22)$$

となる[4]. また, この単語系列 $w^m$ に対応するシンボル系列 $x^n$ の長さ $n$ は式 (22) より

$$n \simeq \sum_{w \in W} |w| \cdot N(w|w^m) \simeq m E[|W|] \quad (23)$$

となり, さらに長さ $s$ $(= 1, 2, \cdots, k)$ の単語の数 $l_s = \sum_{t \in \mathcal{T}_s} N(w_{s,t}|w^m)$ はそれぞれ

$$l_s \simeq m \sum_{t \in \mathcal{T}_s} P_W(w_{s,t}) \quad (24)$$

となる.

十分長い長さ $n$ のシンボル系列 $x^n$ 中には, 式 (24) より長さ $s$ の単語が $m \sum_{t \in \mathcal{T}_s} P_W(w_{s,t})$ 個含まれることがわかるが, $x^n$ だけを観測した場合, 複数の単語系列がこのシンボル系列に写像される可能性があるので, 単語の切れ目については一意に決められない. そこで, 長さ $s$ の単語の数が $l_s \simeq m \sum_{t \in \mathcal{T}_s} P_W(w_{s,t})$ となるような単語の切れ目の入れ方の数を数え上げると, その数 $V$ は

$$V = \frac{m!}{l_1! l_2! \cdots l_k!} \simeq \frac{m!}{\prod_{s=1}^{k} \{m \sum_{t \in \mathcal{T}_s} P_W(w_{s,t})\}!} \quad (25)$$

[4] $\simeq$ は, 漸近的にほぼ等しいことを表す