The 27th Symposium on Information Theory
and Its Applications (SITA2004)
Gero, Gifu, Japan, Dec. 14-17, 2004

Word-valued sourceから出現する系列の単語分割について
Word segmentation of the sequences emitted from a word-valued source

石田 崇† Takashi ISHIDA
松嶋 敏泰† Toshiyasu MATSUSHIMA
平澤 茂一† Shigeichi HIRASAWA

Abstract— Word segmentation is the most fundamental and important process for Japanese or Chinese language processing. Because there is no separation between words in these languages, we firstly have to separate the sequence into words. On this problem, it is known that the approach by probabilistic language model is highly efficient, and this is shown practically. On the other hand, recently, a *word-valued source* (WVS) has been proposed as a new class of source model for the source coding problem. This model is supposed to reflect more of the probability structure of natural languages. We may regard Japanese sentence or Chinese sentence as the sequence emitting from a *non-prefix-free* WVS. In this paper, as the first phase of applying WVS to natural language processing, we present a word segmentation method for the sequence from non-prefix-free WVS. Then, we clarify the performance of word segmentation by numerical computations.

Keywords— word segmentation, word-valued source, non-prefix-free, morphological analysis

1 はじめに

自然言語処理における形態素解析はもっとも基本的で重要な処理である．形態素解析は，文書を構成する文を単語に分割しその語形変化などの情報を付与することであり，音声認識や文字認識，機械翻訳，情報検索など自然言語処理の応用分野において必須の技術である．中でも単語の同定問題は，ヨーロッパ系言語では単語間にスペースを挟む習慣があるため容易であるが，日本語や中国語のように単語間にスペースを入れない（分かち書きをしない）言語においては文を単語に分割することが困難となる．最近では，膨大な量のテキストデータを元にした確率的言語モデルによる自然言語処理が盛んに研究されており [1]，辞書によらない日本語形態素解析に対してもかなりの良い精度を達成している．

その一方，従来から確率モデルが重要な役割を果たしてきた情報源符号化問題においても近年，新たな情報源モデルとして言語アルファベット情報源（Word-valued source: WVS）が提案されている [2, 3]．この情報源モデルは従来のモデルに対して自然言語の確率構造をより反映させたモデルとして考えることができ，情報源の性質や情報源符号化性能の解析がなされている [2, 3, 4, 5]．特に，non-prefix-free WVS[4, 5] は日本語のように単語が語頭条件を満たさず，単語が分かち書きされない言語のモデルとみなすことができる．情報源符号化問題における確率モデルが実際の日本語単語分割に適用された事例もすでに報告されており [6]，情報源モデルや効率的な符号化アルゴリズムの自然言語処理分野への応用が期待される．

本研究では確率モデルとしてWVSモデルを考え，non-prefix-free WVSから出力される人工的なデータ系列に対して単語分割手法を示し，その性能を評価する．その結果，言語モデルの確率構造と単語分割の精度についての検証を行うことを目的とする．

†〒 169–8555 東京都新宿区大久保 3-4-1 早稲田大学理工学部経営システム工学科 School of Science and Engineering,Waseda University Okubo 3-4-1, Shinjuku-ku, Tokyo, 169–8555 Japan. E-mail: ishida@hirasa.mgmt.waseda.ac.jp

2 実データに対する単語分割アプローチ

日本語の実データに対する単語分割に関する研究は数多く存在するが，中でも確率的言語モデルに基づいた単語分割手法は，かなり良い精度を達成していることが報告されている [6, 7, 8]．これらの手法では主に隠れマルコフモデルや N-gram モデル（$N-1$ 重マルコフモデル）が用いられている．

しかし，日本語の実データに対しては文字の数だけでも約3,000 ~ 6,000程度存在し，確率のパラメータ推定におけるゼロ頻度問題やスパースネスなどの問題が生じる．そこで実際には単語単位の確率モデルではなく，文字単位の N-gram モデルが主に用いられており，モデルの次数も低次のものに限定される．また，確率値のスムージング処理や N-gram モデルの次数決定などはその言語に依存して経験的に与えられているのが実情である．

そのため，用いた確率モデルが実際にどの程度自然言語の近似モデルとして有効に機能しているのか，また，このモデルに基づく言語処理システムが，異なる言語に対しても有効に機能するシステムなのかについてはその性能を保証することが困難であるといえる．

そこで本稿では，言語モデルの構造によって単語分割手法の精度がどのような挙動を示すのかを検証するために，実データに対してではなく確率モデルから出力される人工データ系列に対して単語分割性能の評価を行う．本研究では以下で述べる言語アルファベット情報源 (WVS) を用いる．

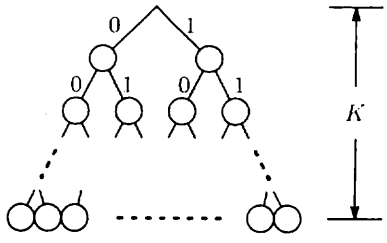
3 言語アルファベット情報源（WVS）

言語アルファベット情報源（WVS）[2, 3] は情報源符号化問題において，自然言語などの実際に圧縮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た．この情報源は，情報源アルファベットの有限系列を単語と定義し単語単位で確率構造を有するモデルと解釈する事ができる．

単語集合 $\mathcal{W}$ において任意の単語が他の単語の語頭に一致していないことを，単語集合 $\mathcal{W}$ は語頭条件を満たしている (*prefix-free*) という．単語集合が語頭条件を満たすとき，単語集合が既知であれば与えられた文に対して単語分割は一意に可能となる．分かち書きを行う言語は単語間の空白を単語の末尾の文字とみなせば語頭条件を満足していると考えることができる．一方，語頭条件を満たさない (*non-prefix-free*) 場合にはたとえ単語集合が既知であったとしても，一意に単語分割を行うことは不可能であり，日本語などはこちらの言語モデルに該当しているといえる．

以下では [4] をもとにして，本研究で用いる non-prefix-free i.i.d. WVS の定義を与える．

定義 1 (non-prefix-free i.i.d. WVS) 情報源アルファベットを $\mathcal{X}$ とし，$\mathcal{X}$ 上に値をとる確率変数を X，その要素を x で表しこれをシンボルとよぶ．x の有限系列を単語とよび w で表し，単語集合を $\mathcal{W}$ とする．いま，$\mathcal{W}$ が $\mathcal{W} = \cup_{k=1}^{K} \mathcal{X}^k$ で与えられているものとする．$\mathcal{W}$ 上に値をとる確率変数の系列 $\boldsymbol{W} = W_1, W_2, \cdots$ を定常無記憶 (i.i.d.) 情報源とするとき，W_i の連接によって生成される確率変数列 $\boldsymbol{X}$ を non-prefix-free i.i.d. 言語アルファベット情報源 (WVS) とよぶ． □



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

$\|\mathcal{W}\| = \sum_{k=1}^{K} 2^k = 2(2^K - 1)$ ($\|\mathcal{W}\|$：単語数)

図 1: 単語集合 $\mathcal{W}$ の 2 分木表現

すなわち，情報源からは単語単位で系列 $w_1^m = w_1 \cdots w_m$ が出力されるが，それぞれの単語が連接された長さ n のシンボルの系列 $x_1^n = x_1 \cdots x_n$ として観測されることを意味しており，このシンボル系列の確率構造 $P_{X^n}(x_1^n)$ に着目したモデルであるといえる[1]．このとき任意の m に対して $n = |w_1| + |w_2| + \cdots + |w_m|$ を満たしている．ここで，$|w|$ は単語 w の長さである．

$\mathcal{X} = \{0, 1\}$ とした場合の単語集合 $\mathcal{W}$ は図 1 のような木構造で表現されるが，この単語集合は明らかに語頭条件を満たしていない．

観測されるシンボルの系列 x_1^n の出現確率と情報源のエントロピーレートは

$$P_{X^n}(x_1^n) = \sum_{w_1^m : x_1^n = w_1^m} P_{W^m}(w_1^m), \tag{1}$$

$$H(\boldsymbol{X}) = \lim_{n\to\infty} E_{P_{X^n}}\left[-\frac{1}{n}\log P_{X^n}(x_1^n)\right], \tag{2}$$

で与えられる [4]．

[4, 5] では，様々な種類の non-prefix-free WVS に対するエントロピーレートの性質が解析されている．WVS の確率構造はシンボル系列と単語系列の写像関係に依存することがわかっている．すなわち，式 (2) からも分かるように，同一のシンボル系列として観測される単語系列の本数によって系列の出現確率が決定する．したがって，単語集合 $\mathcal{W}$ とその確率分布 $P_W(w)$ をどのように与えるかによってシンボル系列と単語系列間の対応関係 (図 2) が大きく変化し，様々な性質の確率構造を持つことになる[2]．情報源のエントロピー $H(\boldsymbol{X})$ は情報源の確率構造の複雑さを図る指標となる．

4 WVS から出力された系列に対する単語分割手法

本章では non-prefix-free WVS から出力された系列に対する単語分割手法の定式化と分割アルゴリズムを提案する．実データに対する単語分割問題においては学習データからのパラメータ推定フェーズと未知データに対する単語分割フェーズとに分けられる．本研究は人工データに対して情報源モデルの特性と単語分割性能を評価するため，ここでは学習が完全に達成された状態（単語集合 $\mathcal{W}$ とその確率分布 $P_W(w)(w \in \mathcal{W})$ が既知）を仮定し，そのもとでの単語分割問題を定式化する．ここでは，WVS から出力されたシンボルの系列 x_1^n に対して，

[1] 本稿では，単語単位での定常無記憶性を仮定しているが，単語単位でのマルコフモデルへも容易に拡張できる．

[2] $P_W(w)$ の与え方によっては語頭条件を満たすモデルと等価になる場合を含んでいる．

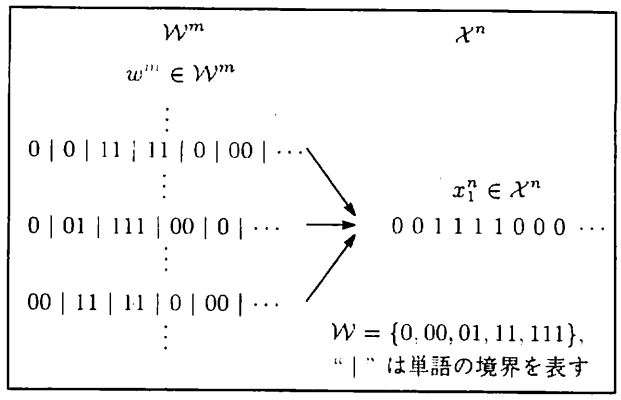


図 2: 単語系列とシンボル系列の関係

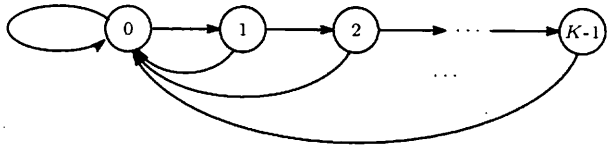


図 3: 状態遷移図

単語の境界を決定して分割することにより情報源から出力された真の単語単位の系列 w_1^m ($n = |w_1| + \cdots + |w_m|$) を推定することが目的となる．

いま，情報源の単語集合 $\mathcal{W}$ とその確率分布 $P_W(w)$ が既知としているので，単語集合 $\mathcal{W}$ が語頭条件を満たすならば，観測された系列 x_1^n から一意に w_1^m が決定される．一方，$\mathcal{W}$ が語頭条件を満たさない場合には，一般に情報源が既知の場合でも x_1^n からは w_1^m を一意に決定することはできない [4]．そこで WVS からの出力系列を隠れマルコフモデルとして捉え，単語分割問題を解決する．

4.1 単語分割モデル

観測されたシンボル x が単語の何番目のシンボルであるかを状態 $s \in S$ として定義する

定義 2 (シンボル x の状態 s) シンボル x が単語 w の先頭から i 番目のシンボルであるとき，状態を $s = i - 1$ と定義する．このとき状態の集合は $S = \{0, 1, \cdots, K-1\}$ となる．ここで，K は単語集合 $\mathcal{W}$ に属する単語の最大長（木の最大深さ）である．また，S 上に値をとる確率変数を S とする． □

シンボルの状態 s は，x が単語 w の先頭の文字となっているとき状態 $s = 0$，2 番目，⋯ の文字となっているとき $s = 1$, ⋯ となる．また，状態 s は，単語集合 $\mathcal{W}$ が図 1 のように表されるとき，x がどの深さの枝に対応しているかを表す．このように定義された状態 $s \in S$ に対して，WVS の性質から状態遷移図は図 3 のようになる．

観測されたシンボル系列 $x_1^n = x_1 x_2 \cdots x_n$ に対して状態遷移系列 $s_1^n = s_1 s_2 \cdots s_n$ が決定されるとそれに対応する単語系列 w_1^m が唯一決まる．すなわち，$s_t = 0$ となるシンボル x_t の直前を単語の境界とする単語系列 w^n に分割する．なお，本研究ではシンボル系列 x_1^n の先頭と末尾は必ず単語の境界となっていることを仮定する．したがって $s_1 = 0$ であり，さらに x_1^n の仮想的な次のシンボル x_{n+1} に対する状態 $s_{n+1} = 0$ として，シンボル系列 x_1^n と状態遷移系列 s_1^{n+1} を考える．

観測系列 x_1^n と状態遷移系列 s^{n+1} の同時確率 $p(x_1^n, s_1^{n+1})$ は以下のように与えられる．

$$p(x_1^n, s_1^{n+1}) = p(s_1)\prod_{t=1}^{n} p(x_t, s_{t+1}|x_{t-s_t}^{t-1}, s_t), \qquad (3)$$

ただし，$u < v$ または $u = v = 0$ のとき $x_v^u = \phi$ (空系列)，$u = v\ (\neq 0)$ のとき $x_v^u = x_u$ とする．ここで，

$$p(s_1) = \begin{cases} 1, & s_1 = 0 \\ 0, & \text{otherwise} \end{cases} \qquad (4)$$

である．単語 w が x の系列により $w = x_{[1]}x_{[2]}\cdots x_{[|w|]}$ と表わされるとき，確率 P_W の周辺和を

$$P_W^{sum}(x_{t-i}^t *) = \sum_{\{w:x_{[1]}=x_{t-i},\cdots,x_{[i+1]}=x_t\}} P_W(w), \qquad (5)$$

と記述することにする (ただし，$i = 0, 1, \cdots, K-1$)．このとき，$p(x_t, s_{t+1}|s_t)$ $(t = 1, 2, \cdots n)$ は，

$$p(x_t, 0|0) = P_W(x_t), \qquad (6)$$

$$p(x_t, 1|0) = P_W^{sum}(x_t *) - P_W(x_t), \qquad (7)$$

$i = 1, 2, \cdots, K-2$ に対して，

$$p(x_t, i+1|i) = \frac{P_W^{sum}(x_{t-i}^t *) - P_W(x_{t-i}^t)}{P_W^{sum}(x_{t-i}^{t-1} *) - P_W(x_{t-i}^{t-1})}, \qquad (8)$$

$i = 2, 3, \cdots, K-1$ に対して，

$$p(x_t, 0|i) = \frac{P_W(x_{t-i}^t)}{P_W^{sum}(x_{t-i}^{t-1} *) - P_W(x_{t-i}^{t-1})}, \qquad (9)$$

によって求められる．

さらに，s_t $(t = 1, 2, \cdots, n)$ の事後確率は次のように計算される．

$$p(s_t|x_1^n) = \sum_{s_1, s_2 \cdots, s_{n-1} \backslash s_t} \frac{p(x_1^n, s_1^{n+1})}{p(x_1^n)}, \qquad (10)$$

$p(x_1^n, s_1^{n+1})$ や $p(s_t|x_1^n)$ は，図 4 のようなトレリスを用い，それぞれの状態遷移に式 (6) ～式 (9) の確率を与えることによって，前向き・後向きアルゴリズムから効率的に計算することができる [1][9].

4.2 状態遷移系列の推定と単語分割

ここでは，以下に示す M1 ～ M3 のそれぞれの方法によって状態遷移系列 s_1^{n+1} の推定，または状態 0 の判定を行い，$s_t = 0$ となるシンボルの直前を単語の境界として分割する．

$$\text{M1: } \hat{s}_t \begin{cases} = 0, & \text{if } p(S_t = 0|x_1^n) > p(S_t \neq 0|x_1^n) \\ \neq 0, & \text{otherwise} \end{cases} \qquad (11)$$

$$\text{M2: } \hat{s}_t = \underset{s_t}{\text{argmax}}\ p(s_t|x_1^n) \qquad (12)$$

$$\text{M3: } \hat{s}_1^{n+1} = \underset{s_1^{n+1}}{\text{argmax}}\ p(s_1^{n+1}|x_1^n) \qquad (13)$$

M1 は各時点 t $(t = 0, 1.2, \cdots n+1)$ において状態の事後確率により状態が 0 か非 0 か (シンボル x_t の直前が境界かそうでないか) の判定による単語分割法である．M2 は各時点 t $(t = 0, 1, 2, \cdots n+1)$ において事後確率が最大の状態 $\hat{s}_t$ からなる状態遷移系列 $\hat{s}_1^{n+1} = \hat{s}_1\hat{s}_2\cdots\hat{s}_{n+1}$ に基づいた単語分割手法である．M3 は事後確率が最大となる状態遷移系列 $\hat{s}_1^{n+1}$ による単語分割法である．これは Viterbi アルゴリズムから求めることができる [1].

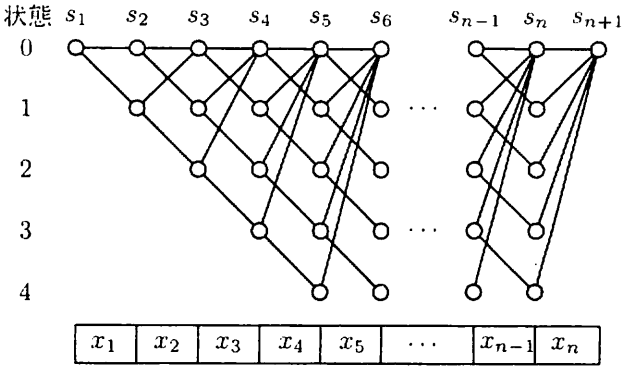


図 4: WVS におけるトレリス図 ($K = 5$ の場合)

5 評価実験

本稿で提案した WVS に対する単語分割手法の性能を評価するため，数値実験による検証を行う．なお，参考として実データに対して用いられている N-gram モデルによる手法 [6] による単語分割も行い結果を比較する．

5.1 実験の条件

WVS の単語集合を $K = 5$ ($||\mathcal{W}|| = 62$) とする．また，$\mathcal{W}^+ = \{w : P_W(w) > 0\}$ と定義する．これは，単語集合 $\mathcal{W}$ のうち正の確率が与えられ，実際に出現する可能性のある単語の集合を表す．$||\mathcal{W}^+|| = 5, 10, 20, 30, 40, 50, 62$ のそれぞれの場合について，乱数によって P_W を与えて 100 個の WVS を生成し，各情報源から出力された系列に対して単語分割を行う．単語分割の評価指標として再現率 ($recall$) と適合率 ($precision$) を計算する [6, 8]．ここで，$True$ を真の単語数 (=単語系列長 m)，Sys を分割アルゴリズムによって分割された単語数，M を単語境界位置の正解数とすると $recall = M/True$, $precision = M/Sys$ である．また，それぞれの WVS の情報源のエントロピー密度レート $H(\boldsymbol{X})$ を計算する [5].

5.2 実験結果

各 $||\mathcal{W}^+||$ での 100 回の分割結果における再現率と適合率の平均値を図 5 ($m = 50$)，図 6 ($m = 500$) に示す．さらに，それぞれの $||\mathcal{W}^+||$ に対して計算した $H(\boldsymbol{X})$ の平均値を表 1 に示す．

表 1: 情報源のエントロピー密度レート $H(\boldsymbol{X})$ の平均値

$\|\mathcal{W}^+\|$	5	10	20	30	40	50	62
$H(\boldsymbol{X})$	0.49	0.71	0.89	0.95	0.98	0.99	0.99

6 考察

評価実験から以下の結果が明らかとなった．

1. 図 5，図 6 から，いずれの手法においても再現率，適合率は $||\mathcal{W}^+||$ の増加とともに値が小さくなる傾向がみられる．これは，表 1 から分かるように，$||\mathcal{W}^+||$ の増加にしたがって WVS の $H(\boldsymbol{X})$ が増加するため，系列の複雑性が増している (1 本の観測系列 x^n に写像される単語系列 w^m がより多く存在する) ことからも自然な結果であるといえる．
2. 図 5，図 6 ともにほぼ同様の結果となっていることから，単語系列長 m は単語分割性能に影響しないのではないかと考えられる．
3. M2 は再現率，適合率のいずれもよい性能を示していることから，一番適した手法であると考えられる．これは，WVS に対して各時点での状態の事後確率による単語分割が優れていることを示している．

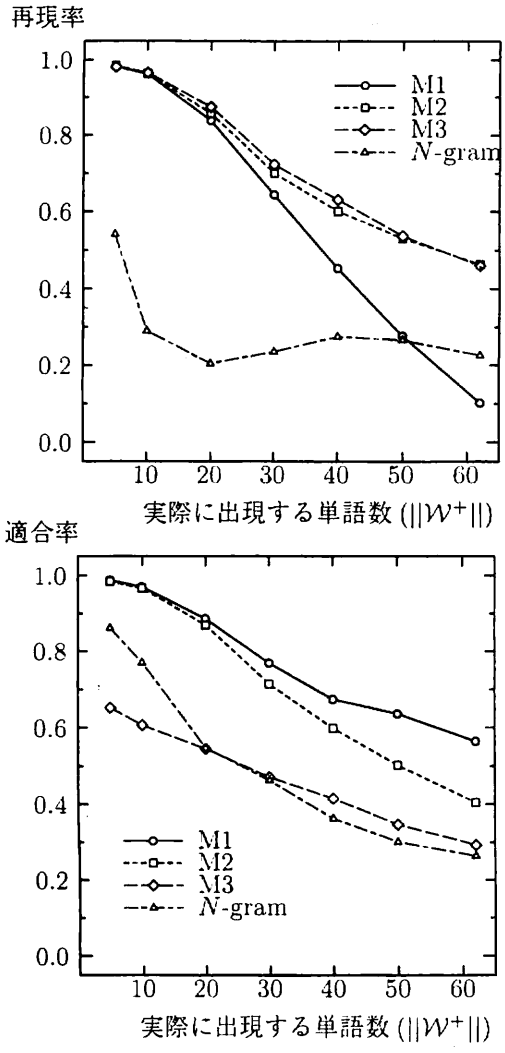


図 5: 各手法における適合率と再現率 ($m = 50$)

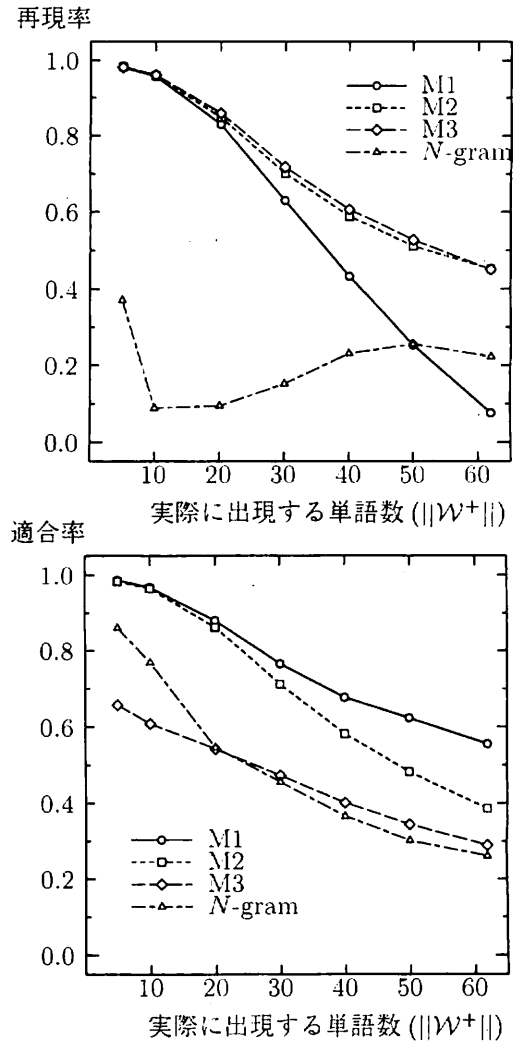


図 6: 各手法における適合率と再現率 ($m = 500$)

4. M1 は適合率は優れているが再現率は $||\mathcal{W}^+||$ の増加に伴って著しく悪化する傾向が見られる．これは，この手法が $s_t = 0$ となる事後確率が 0.5 以上のときに単語にその直前で分割されるため，$H(\boldsymbol{X})$ が増加して状態の事後確率に偏りがなくなると単語分割されにくくなることを示している．
5. M3 は再現率は優れているが適合率は低い．これは M1 とは逆に単語を分割し過ぎる傾向があることを示している．
6. N-gram 法は再現率，適合率ともに低い値となっており，WVS に対して有効には機能していないことが分かる．しかし，この手法は日本語の実データに対して良い性能を示している [6] ことから，依然として実データと WVS モデルとの間には大きな隔たりがあることを示唆している．

7 まとめと今後の課題

本稿では，WVS モデルに対する単語分割問題を定式化して適した分割手法を提案し，数値実験による検証を行った．今後の課題は，より自然言語の構造を反映したモデルに拡張して単語分割問題を扱い，最終的には実データに対して有効な単語分割手法を提案することである．また，今回は有効性の指標が単語境界位置の成否を判定するものであったが，切り出された単語の種類を判定する指標についても検討したい．

8 謝辞

著者の 1 人石田は，日ごろより熱心に議論をしていただく武蔵工業大学 後藤正幸先生，湘南工科大学 小林学先生，早稲田大学 八木秀樹氏に感謝します．

参考文献

[1] 北 研二: 確率的言語モデル, 東京大学出版会, (1999).
[2] M.Nishiara and H. Morita, "On the AEP of word-valued sources," *IEEE Trans. Inform. Theory*. vol.IT-46, no.3, pp.1116-1120, 2000.
[3] M.Goto, T.Matsushima, and S.Hirasawa, "A source model with probability distribution over word set and recurrence time theorem," *IEICE Trans. Fundamentals*, vol.E86-A, no.2, pp.2517-2525, Oct. 2003.
[4] 石田 崇, 後藤 正幸, 松嶋敏泰, 平澤 茂一, "単語単位で系列を出力する情報源の性質について," 第 25 回情報理論とその応用シンポジウム予稿集 (SITA2002), pp.695-698, 2002.
[5] 石田崇, 後藤正幸, 松嶋敏泰, 平澤茂一, "語頭条件を満たさない単語集合をもつ Word-Valued Source の性質について," 電子情報通信学会技術研究報告, IT2003-5, pp.23-28, (2003,5).
[6] 小田裕樹, 北研二, "PPM* 言語モデルを用いた日本語単語分割," 情報処理学会論文誌, vol.41, no.3, pp689-700, 2000.
[7] 山本幹雄, 増山正和, "品詞・区切り情報を含む拡張文字の連鎖確率を用いた日本語形態素解析," 言語処理学会第 3 回年次大会発表論文集, pp.421-424, 1997.
[8] M.Nagata, "A stochastic Japanese morphological analyzer using a forward-DP backward-A* algorithm," *Proc. 15th International Conference on Computational Linguistics*, pp.201-207, 1994.
[9] Y.Ephraim and N.Merhav, "Hidden Markov Processes", *IEEE Trans. Information Theory*, Vol.48, No.6 pp.1518-1569, 2002.