Word segmentation for the sequences emitted from a word-valued source

Takashi Ishida, Toshiyasu Matsushima, and Shigeichi Hirasawa
School of Science and Engineering,
Waseda University,
Okubo 3-4-1, Shinjuku-ku,
Tokyo, 169–8555 Japan.
Email: ishida@hirasa.mgmt.waseda.ac.jp

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 has been proposed as a new class of source model for the source coding problem. This model can be 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 formulate a word segmentation problem for the sequence from non-prefix-free WVS. Then, we examine the performance of word segmentation for the models by numerical computations.

1. Introduction

In the field of natural language processing (NLP), morphological analysis is the most fundamental and important processing. Morphological analysis is the process that divides sentences that constitute a document into words and that gives information including the declensions. It is an indispensable technology for applications of NLP, such as voice recognition, character recognition, mechanical translation, or information retrieval. Word identification in morphological analysis is easy for European languages since they have the habit of inserting a space between words. However, it is very difficult for the languages which do not have a space between words like Japanese or Chinese and so on.

Recently, there are many studies of NLP based on the probabilistic language model (PLM) [5] by use of a huge quantity of text data, and it was shown to be effective for Japanese morphological analysis [8, 6].

On the other hand, in the field of source coding, a word-valued source (WVS) has been proposed recently as a new class of information source model where PLM plays an extremely important role [7, 2]. WVS can be considered as a model in which the probabilistic structure of the natural language is reflected more than the conventional source models. A non-prefix-free WVS, which is defined as a WVS whose word set is not prefix-free, can be regarded as a language model where a spaces are not left between words like Japanese [3, 4]. The case where the probability model of the source coding was applied to the Japanese word segmentation problem has been already reported [8]. The applications of the more effective information source models or coding algorithms to the NLP fields are expected.

Firstly, in this paper, setting the WVS model to a probabilistic model, we formulate a word segmentation problem and algorithms for a non-prefix-free WVS are shown. Then the performance of the algorithms is evaluated by a numerical experiment for the artificial data sequences emitted from WVS model. From the results, the relevance between the probability structure of the models and the performances of the word segmentation is verified.

2. Word-valued source (WVS)

Word-valued source (WVS) [7, 2] was proposed in the source coding problem as a new class of source model in which the probability structure of actual data sequences to be compressed, such as a natural language, was more reflected. WVS can be interpreted as the source which has a probability distribution over a *word set*, where a word is defined as a finite sequence over a source alphabet.

0-7695-2983-6/07 $25.00 © 2007 IEEE
DOI 10.1109/CIT.2007.170

If any word in the word set is not corresponding with a prefixes of any other words, then it is said that the word set is *prefix-free*. When a word set is prefix-free, word segmentation is uniquely possible to the given sentence if the word set is known. It can be considered that the language which leaves spaces between words has a prefix-free word set, where the space between words is regarded as a letter of the tail of a word.

On the other hand, a word set is not prefix-free (*non-prefix-free*), even if a word set is known, it can be said that it is impossible to implement a word segmentation uniquely, and languages which are not written with a space between words, such as Japanese, correspond in this case.

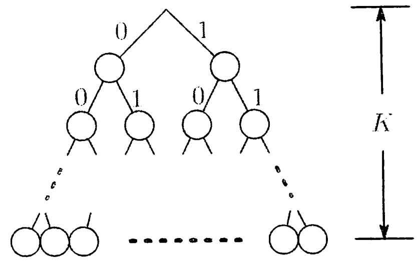
Definition of a non-prefix-free i.i.d. WVS treated in this paper is given as follows based on [4].

Definition 1 (Non-prefix-free i.i.d. WVS)
Let $\mathcal{X}$ be a finite alphabet and X be a random variable which takes a value on $\mathcal{X}$. An element of $\mathcal{X}$ is expressed with x, and we call it *symbol*. Let w be the finite sequence of x, and it is called *word*. We denotes by $\mathcal{W}$ the word set. Here, it is assumed that $\mathcal{W}$ is given by $\mathcal{W} = \cup_{k=1}^{K} \mathcal{X}^k$. If $\boldsymbol{W} = W_1, W_2, \cdots$, which is a sequence of random variable W that takes a value on $\mathcal{W}$, is an independent and identically distributed (i.i.d.) source[1], a random variable sequence $\boldsymbol{X}$ generated by concatenation of W_i is called a *non-prefix-free i.i.d. WVS*. □

For each finite number $n = 1, 2, \cdots$, we denote the sequence of X with length n by $X^n = X_1X_2X_3 \cdots X_n$, and its realization value by $x^n = x_1x_2x_3 \cdots x_n$ respectively. Similarly, for each finite number $m = 1, 2, \cdots$, we use the notation such that $W^m = W_1W_2W_3 \cdots W_m$, and $w^m = w_1w_2w_3 \cdots w_m$.

WVS can be interpreted as follows. A sequence is emitted in a word unit from an i.i.d. source ($w_1^m = w_1 \cdots w_m$), however, we can observe only a symbol sequence with length n ($x_1^n = x_1 \cdots x_n$) which is obtained by concatenation of each words w_i. Here, for any m, $n = |w_1| + |w_2| + \cdots + |w_m|$ holds where $|w|$ is the length of word w. Our interest is in its probability structure $P_{X^n}(x_1^n)$.

Example 1 (Word and word set)
Let $\mathcal{X}$ be $\mathcal{X} = \{0, 1\}$ and a word set $\mathcal{W}$ be $\mathcal{W} = \{0, 01, 101, 111\}$. $\mathcal{W}$ corresponds to a non-prefix-free case. WVS emits the sequence word by word, that is for example $w^4 = w_1w_2w_3w_4 = 01\ 111\ 101\ 0$. However we can only observe the sequence x^9 which is obtained by concatenating the words, that is, $x^9 = 011111010$. We cannot separate the sequence w^4 into words uniquely. Here $n = |w_1| + |w_2| + |w_3| + |w_4| = 2 + 3 + 3 + 1 = 9$. □



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

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

($\|\mathcal{W}\|$ is a cardinality of $\mathcal{W}$)

Figure 1. Binary tree of a word set $\mathcal{W}$

When $\mathcal{X} = \{0, 1\}$, a word set $\mathcal{W}$ is expressed by a binary tree as shown in Fig. 1, and this is non-prefix-free. Hereafter, we call w^m a *word sequence*, and x^n a *symbol sequence* respectively.

The occurrence probability $P_{X^n}(x_1^n)$ and the entropy rate $H(\boldsymbol{X})$ of sequence x_1^n are given by

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

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

respectively [4]. Here, $E_P[\cdot]$ means an expectation over the probability distribution P. We assume the base of logarithm is 2 throughout this paper.

Prefix-free property plays an important role in analyses of the word-valued sources. When $\mathcal{W}$ is prefix-free, the mapping from the word sequences $\mathcal{W}^m$ to the symbol sequences $\mathcal{X}^n$ becomes one-to-one. From an observed symbol sequence x^n, if we know the word set $\mathcal{W}$, we can uniquely determine a certain word sequence w^m which is actually emitted from the source when $n = \sum_{i=1}^{m} |w_i|$. That is, we can see that where the unobserved gaps between each word in the symbol sequence x^n are. By determining where the gaps are in given symbol sequence x^n, we can specify one certain word sequence w^m.

When $\mathcal{W}$ is not prefix-free, on the other hand, the mapping from the word sequences $\mathcal{W}^m$ to the symbol sequences $\mathcal{X}^n$ is generally a many-to-one mapping.

In [3, 4], properties of the entropy rate of various kinds of non-prefix-free WVS were analyzed. The probability structure of WVS depends on the mapping from a word sequence to a symbol sequence. That is, as shown in Eq. (2), the occurrence probability of a symbol sequence is determined by

[1] In this paper, although it is assumed that $\boldsymbol{W}$ is i.i.d. source, the extension to a Markov source is easy.

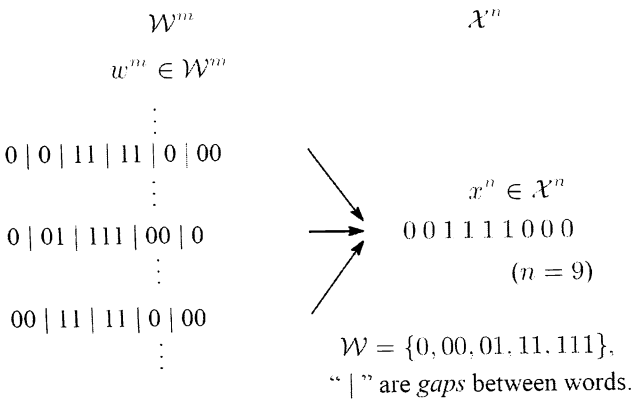


Figure 2. Relationship between word sequence and symbol sequence

the number of the word sequences w^m which is observed as the same symbol sequence x^n.

Therefore, by how the word set $\mathcal{W}$ and its probability distribution $P_{\mathcal{W}}(w)$ are given, the relationship between a symbol sequence and word sequences will change a lot, and will have the various patterns of probability structure[2]. Generally, the appearance probability $P_{X^n}(x^n)$ has a complicated structure depending on the mapping from $\mathcal{W}^m$ to $\mathcal{X}^n$. Figure 2 shows an example of the relation $\mathcal{W}^m$ and $\mathcal{X}^n$. It is found that some word sequences w^m are mapped to one symbol sequence $x^9 = 001111000$. $P_{X^n}(001111000)$ is obtained by the summation of $P_{W^m}(w^m)$ for all w^m mapped to $x^9 = 001111000$.

An entropy rate of the source $H(\boldsymbol{X})$ is regarded as the measure of complexity of the probability structure of its. However, no explicit single letter expression of the entropy rate of the non-prefix-free WVS is known. Only the upper bound and lower bound on the entropy rate has been shown by the present [7, 4].

3. Formulation of the word segmentation problem for the non-prefix-free WVS

In this chapter, we formulate a word segmentation problem for the sequences emitted from a non-prefix-free WVS and the segmentation algorithms are shown. Word segmentation problem for actual text data is divided into the parameter estimation phase from training data and the word segmentation phase to test data.

However, in this paper, our purpose is to investigate the relationship between the properties of WVS and the performance of word segmentation for an artificial data, we assume that the training is surely accomplished, that is the word set $\mathcal{W}$ and its probability distribution $P_{\mathcal{W}}(w)(w \in \mathcal{W})$ are completely known. Under such a situation, a word segmentation problem is formulated.

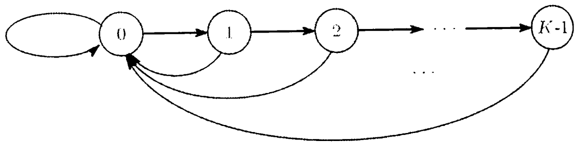


Figure 3. Example of state transition diagram

Here, it is the aim to estimate the sequence of the word sequence emitted from a non-prefix-free WVS by determining the gaps between words and dividing a symbol sequence x_1^n into words w_1^m $(n = |w_1| + \cdots + |w_m|)$. Since we consider the case where the word set $\mathcal{W}$ and the probability distribution $P_{\mathcal{W}}(w)$ of the source are known, if the word set $\mathcal{W}$ is prefix-free, then w_1^m will be determined from the observed sequence x_1^n immediately. On the other hand, generally, it is impossible to determine a word sequence w^m uniquely, even if $\mathcal{W}$ is known when it is not prefix-free [4].

Firstly, a word segmentation model for a non-prefix-free WVS is described below.

3.1. Word segmentation model

For each observed symbol x, a state $s \in \mathcal{S}$ is defined by a symbol of what position in a word the symbol x is.

Definition 2 (State s of the symbol x)
When symbol x is the i-th symbol from the first symbol of the word w, a state s is defined as $s = i - 1$. A set of the states $\mathcal{S}$ is $\mathcal{S} = \{0, 1, \cdots, K - 1\}$. Here, K is a maximum length of the word in the word set $\mathcal{W}$ (the maximum depth of the tree of $\mathcal{W}$). Moreover, the random variable which takes a value on $\mathcal{S}$ is denoted by S. □

For example, the state $s = 0$ when x is the first symbol of the word w, and $s = 1$ when x is the 2nd symbol, $\cdots$ and so on. Moreover, when the word set $\mathcal{W}$ is expressed as shown in Figure 1, state s means the depth of the node corresponding to symbol x. According to the characteristic of WVS, for the state $s \in \mathcal{S}$ defined in this way, it is found that a state transition diagram is obtained as Figure 3.

If a state transition sequence $s_1^n = s_1 s_2 \cdots s_n$ is determined for an observed symbol sequence $x_1^n = x_1 x_2 \cdots x_n$, then the word sequence w^m corresponding to it will be decided uniquely. That is, a symbol sequence x^n is divided into a words uniquely by considering that a gap between words is just before the symbol x with $s = 0$. It assumes that the beginning and the tail of the symbol sequence x_1^n

[2] The case where a WVS has a prefix-free word set depending on how to give $P_{\mathcal{W}}(w)$ is included.

are always a border of a word. Therefore, it is certainly $s_1 = 0$, and letting $s_{n+1} = 0$ be the state for a virtual symbol x_{n+1} following the symbol sequence x_1^n, we consider a symbol sequence x_1^n and state transition sequence s_1^{n+1}.

The joint probability of an observed sequence x_1^n and a state transition sequence s^{n+1}, denoted by $p(x_1^n, s_1^{n+1})$, is given as follows:

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

where $x_v^u = \phi$ (null sequence) when $u < v$ or $u = v = 0$, and $x_v^u = x_u$ when $u = v$ $(\neq 0)$.

Here,

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

Suppose that a word sequence w is expressed as $w = x_{[1]}x_{[2]}\cdots x_{[|w|]}$ with the sequence of x. The marginal probability of P_W is denoted as follows:

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

where $i = 0, 1, \cdots, K-1$.

Then, $p(x_t, s_{t+1}|s_t)$ $(t = 1, 2, \cdots n)$ are calculated by

$$p(x_t, 0|0) = P_W(x_t), \tag{6}$$

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

For $i = 1, 2, \cdots, K-2$,

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

And for $i = 2, 3, \cdots, K-1$,

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

Furthermore, the posterior probability of s_t $(t = 1, 2, \cdots, n)$ is calculated as follows:

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

$p(x_1^n, s_1^{n+1})$ and $p(s_t|x_1^n)$ are efficiently calculable by Forward-Backward algorithm on a trellis as shown in Figure 4 by giving the probability of Eq. (6)-(9) to each state transition [5, 1].

3.2. Estimation of state transition sequence and word segmentation

A state transition sequence s_1^{n+1} is estimated or a judgment of a state 0 is made by three methods (M1-M3) which are shown below. And a symbol sequence x^n is divided for words by considering that the gaps of words are just before the symbol x with $s = 0$.

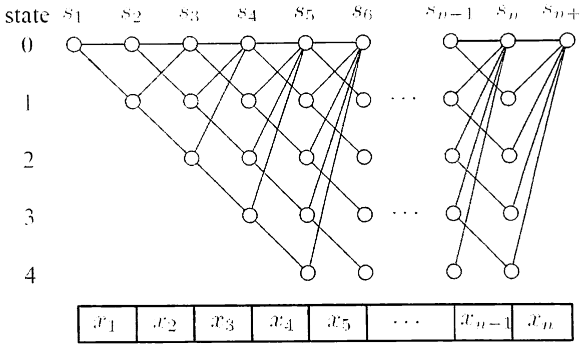


Figure 4. Trellis in word segmentation model in WVS (in case of $K = 5$)

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

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

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

M1 is a word segmentation by judging whether a state is 0 or not 0 (whether there is the gap just before the symbol or not) from the posterior probability of state at each time point t $(t = 0, 1, 2, \cdots n+1)$.

M2 is based on state transition sequence $\hat{s}_1^{n+1} = \hat{s}_1\hat{s}_2\cdots\hat{s}_{n+1}$ which is constituted from a states $\hat{s}_t$ with maximum posterior probability at each time point t $(t = 0, 1, 2, \cdots n+1)$.

M3 is method by state transition sequence $\hat{s}_1^{n+1}$ whose posterior probability is maxima. It can be calculated by the Viterbi algorithm [5].

4. Evaluation experiment

In order to evaluate the performance of the word segmentation algorithms for the non-prefix-free WVS proposed in this paper, the verification by numerical experiment is shown. The word segmentation algorithm by the N-gram model used for the actual text data [8] is also evaluated. These results are compared.

4.1. Conditions of the experiment

A word set of WVS is set to $K = 5$ $(||\mathcal{W}|| = 62)$. We define $\mathcal{W}^+$ as $\mathcal{W}^+ = \{w : P_W(w) > 0\}$, that is, it means a set of words w which may actually appear with a positive occurrence probability $P_W(w) > 0$. About each case of $||\mathcal{W}^+|| = 5, 10, 20, 30, 40, 50$, and 62, 100 WVSs are generated by giving P_W with a random number. Word segmentation algorithms are performed to the sequences emitted from each sources.

Recall and *precision* are calculated as a evaluation indicators concerning a word segmentation [8, 6]. They can be calculated by $recall = M/True$ and $precision = M/Sys$ where $True$ means the number of words in the sequence, that is the length of the word sequence $|w^m| = m$, Sys is the number of words which are divided by the algorithm, and M is the number of the correct cut between words. Moreover, entropy rate $H(\boldsymbol{X})$ of each WVS are also calculated [3].

4.2. Results of the numerical experiment

The average of the recall or the precision in 100 times of word segmentation results for each $||\mathcal{W}^+||$ are shown in Figure 5 ($m = 50$) and Figure 6 ($m = 500$). Furthermore, the mean value of $H(\boldsymbol{X})$ calculated to each $||\mathcal{W}^+||$ is shown in Table 1.

Table 1. Average of the entropy rate $H(X)$

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

5. Consideration

The following results are obtained from the evaluation experiments.

1. It is found that the increase of $||\mathcal{W}^+||$ tends to make the value of recall or precision smaller from Figure 5 and Figure 6. This result is reasonable because $H(\boldsymbol{X})$ of WVS increases according to the increment in $||\mathcal{W}^+||$, as shown in Table 1. That is, the complexity of the sequence is increasing since more word sequences w^m are mapped to one observation sequence x^n).

2. Since Figure 5 and Figure 6 show almost the same results, it is thought that the length of word sequence m has no effect on a word segmentation performance.

3. M2 is considered to have the best performance to WVS since both the recall and the precision show the excellent value. This result suggests that the word segmentation by the posterior probability of the states in each time point is excellent method to WVS.

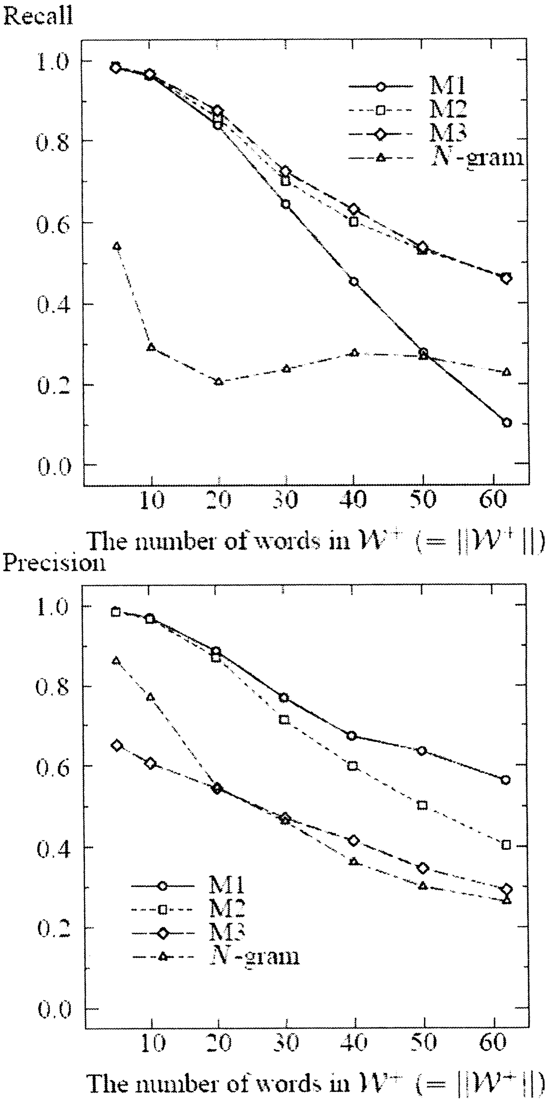


Figure 5. Precision and recall ($m = 50$)

4. Although the precision is excellent for M1, the trend for a recall to get worse remarkably with the increment in $||\mathcal{W}^+||$ is shown. It is explained that since M1 divides a symbol sequence into words just before the symbol whose posterior probability of $s = 0$ is 0.5 or more, in the case where the polarization of the posterior probability is small by the increment of $H(\boldsymbol{X})$, a sequence is hard to be divided.

5. M3 is excellent in the recall, however, it has the lower value of precision. This indicates M3 has a tendency to divide the sequence into words too much contrary to M1.

6. It is found that the recall and the precision of the N-

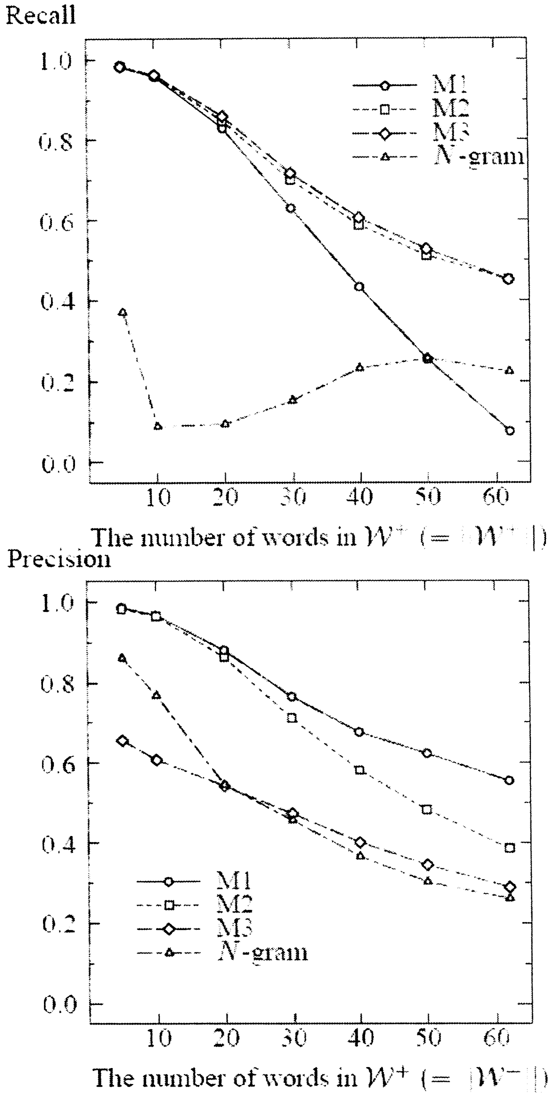


Figure 6. Precision and recall ($m = 500$)

gram method take a lower value, and it is not effect to WVS. However, this method indicates the good performance to the Japanese real text data [8]. It is suggested that there is still a wide gap between a real data and a WVS model.

6. Concluding remarks

In this paper, the word segmentation problem was formulized and the segmentation algorithms for WVS model were proposed. Then the verification by numerical experiments was performed.

A future work is extending to the model which reflected more the structure of the natural language and ultimately proposing the effective word segmentation method which is effective for a Japanese real text data.

Acknowledgement

One of the authors, T. Ishida, wishes to thank Dr. M. Goto, Dr. M. Kobayashi, and Dr. H. Yagi for very helpful discussion. This work was partially supported by Japan Society for the Promotion of Science under Grants-in-Aid for Scientific Research No.18700160.

References

[1] Y. Ephraim and N. Merhav, "Hidden Markov Processes", *IEEE Trans. Information Theory*, Vol.48, No.6 pp.1518-1569, 2002.

[2] 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.10, pp.2517-2525, Oct. 2003.

[3] T. Ishida, M. Goto, T. Matsushima and S. Hirasawa, "Properties of a Word-valued Source with a Non-prefix-free Word Set," in Japanese, *Technical Report of IEICE*, IT2003-5, pp.23-28, 2003.

[4] T. Ishida, M. Goto, T. Matsushima, and S. Hirasawa, "Properties of a Word-Valued Source with a Non-Prefix-Free Word Set," *IEICE Trans. Fundamentals*. vol.E-89A, pp.3710-3723, Dec. 2006.

[5] K. Kita: Probabilistic Language Models, Tokyo Daigaku Syuppan-kai, 1999.

[6] 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.

[7] 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.

[8] H. Oda and K. Kita, "A Japanese Word Segmenter Using a PPM*-based Language Model," Transactions of Information Processing Society of Japan, vol.41, no.3, pp689-700, 2000.