情報処理学会論文誌　数理モデル化と応用　Vol.6　No.1　38–46 (Mar. 2013)

言語モデルの違いによるHMMを用いたテキストセグメンテーションの性能比較

但馬 康宏[1,a)]

受付日 2012年4月19日，再受付日 2012年6月7日，
採録日 2012年7月13日

概要：HMM によるテキストセグメンテーションの問題について，HMM の状態が表す言語モデルを変化させることによる性能の変化を示す．一般に HMM でテキストをモデリングする場合，各状態は単語ユニグラムを言語モデルとして段落を表現する．これに対して本論文では，複数の単語をとりまとめて 1 つの出力記号とする手法を複数提案し，その性能の変化を考察する．評価実験の結果，1 文を出力記号単位とし，単語がその文章に含まれるか否かを確率として持つナイーブな言語モデルが高い性能であることが明らかとなった．また提案手法は，本論文における設定よりも利用できる情報が多くなる教師あり学習の枠組みによるアルゴリズムの性能にはおよばないが，従来法である単語ユニグラムモデルを利用する HMM の性能を上回ることが確認された．

キーワード：テキストセグメンテーション，HMM，生成モデル，確率的言語モデル

Performance Comparison between Different Language Models on a Text Segmentation Problem via HMM

Yasuhiro Tajima[1,a)]

Received: April 19, 2012, Revised: June 7, 2012,
Accepted: July 13, 2012

Abstract: We investigate performances of some text segmentation algorihtms which use HMM. In general, HMM applied to a text segmentation problem uses the word unigram language model to express the text segments. In this paper, we propose new multi-gram language models for a state of HMM, and evaluate them by experiments. From our evaluations, the highest performace model among our proposals is a naive probabilistic vector model for a sentence. In addition, the performances of all our proposed models are higher than that of well known HMM which uses the word unigram language model.

Keywords: text segmentation, HMM, generative model, probabilistic language model

1. はじめに

テキストデータを段落や章，話題など意味のある分割位置で区切ることをテキストセグメンテーションもしくは，段落分割と呼ぶ．この問題に関して，以下のようないくつかの手法が知られている．まずはじめに，TextTiling [1] として知られている変化点を抽出する手法である．テキストデータに対し一定の範囲のテキスト窓を切り取り，その窓内のテキストを特徴付ける特徴量を算出する．テキスト窓をテキストの先頭から末尾まで動かしてゆき，特徴量の変化が大きい位置が分割位置であるとする手法である．たとえば特徴量として，あるテキスト窓内に現れる単語の種類とその出現数をベクトルにしたものを考える．1 つの窓と隣接する窓との間では，2 つのベクトル間のなす角を類似度と見なすことができる．窓を動かしてゆき，類似度が大きく変動する位置が大きく話題の転換する位置だと見な

[1] 岡山県立大学情報工学部情報システム工学科
Department of Systems Engineering, Faculty of Computer Science and System Engineering, Okayama Prefectural University, Soja, Okayama 719–1197, Japan
[a)] tajima@cse.oka-pu.ac.jp

せ，分割位置の候補となる．この手法では，どの程度の変動を分割位置とするかという閾値問題など設定すべきパラメータが性能に大きな影響を与えるが，事前の学習にあたる部分がない点が特徴である．またこの手法では，分割位置の推定はできるが，分割位置に挟まれた段落がどのような話題を扱う段落であるかの推定はできず，別途クラスタリングなどが必要である．

次に HMM を用いた分割手法である [5]．一般的には，単語を 1 つの出力記号とし，HMM の各状態が 1 つの段落や話題を表すものとする．音声認識の分野では音素の抽出などに広く使われており，時系列データの処理での性能の高さがよく知られている．事前に学習データを用いてパラメータを設定することが多く，Baum-Welch などのアルゴリズムが知られている．この手法は，いくつかの発展形があり，状態に到着した時点で出力する記号を確率変数の長さを持った記号列とし，テキストセグメンテーションに適した改良を行う研究 [4] や，出力記号と前状態から現在状態を決定する HMM（MEMM）への改良 [3] などがある．いずれの研究においても，HMM の各状態は段落を表し，出力記号が 1 単語であるので，段落に対する単語ユニグラムによる言語モデルを構築している．

本論文では，HMM の 1 つの状態が表現する言語モデルについて複数の単語の列を表現するモデルを新たに提案する．一般に，複数の単語を扱う言語モデルは n-gram が知られているが，HMM において n-gram モデルを扱おうとすると出力記号数の指数的増大を招き実用的でない．本論文では，この点を考慮した手法を提案する．さらに，単語ユニグラムによるモデルおよび以前の提案による手法 [6] との性能比較を行う．

新たに提案する手法では，HMM における各状態は，1 つの文章を確率的に識別するものとする．この手法は，各状態が段落や話題を表す点は従来手法と同じだが，分割対象のテキストについて 1 文ごとに各状態での受け入れ確率が求められるものとする．その後，テキスト全体において最も受け入れ確率が高い状態遷移系列を求め，互いに違う状態への遷移が段落の切れ目であるとする手法である．状態遷移確率は一般の HMM と同じ扱いができ，空の文も含めた受け入れ確率の和がそれぞれの状態で 1 となるならば，本手法においても Baum-Welch アルゴリズムを利用することができる．

評価実験として，複数のウェブニュースが連なったテキストファイルに対してニュースの記事ごとへの分割を行った．その結果，本手法により従来手法よりも高い性能を得ることができ，特にランダムに話題が移り変わるようなテキストデータに対しては大きな性能向上を得られることが確認できた．

2. HMM による段落分割と状態における言語モデル

実数の集合を R とする．離散型隠れマルコフモデル（HMM）を状態の有限集合 Q，出力記号の集合 B，状態間の遷移確率 $a : Q \times Q \to R$，各状態における出力確率 $b : Q \times B \to R$ にて定義する．任意の $i \in Q$ について，$a(i, \cdot)$ および $b(i, \cdot)$ は確率分布である．初期状態確率分布を $i \in Q$ について $a(0, i)$ と表す．

テキスト t は単語の列 $w_1 w_2 \cdots w_n$ であり，扱うすべてのテキストに出現するすべての単語の集合を W と表す．一般に HMM を用いたテキスト分割は，学習データであるテキスト集合 T を用いて単語の出力モデルである HMM を構成し，分割対象のテキストに対し最適な状態遷移系列を求め，その状態の移り変わりが段落の移り変わりであると見なして分割位置を決定する．

すなわち，以下のように対応付けている．

- テキストにおける段落：$q \in Q$
- テキストを構成する単語：$w \in B$
- 段落 q_1 から段落 q_2 に移り変わる確率：$a(q_1, q_2)$

これはテキストに現れる段落を 1 つの状態とし，その段落を述べる場合に出現しやすい単語の分布を出力記号の分布としてモデル化する手法である．この場合，HMM の各状態は，対応する段落に対する単語ユニグラムの言語モデルを表現しているといえる．

HMM の各パラメータの推定には，EM アルゴリズムである Baum-Welch アルゴリズムがよく知られている．この学習アルゴリズムは，教師なし学習アルゴリズムでありサンプルデータの集合から直接 HMM の各パラメータを推定することができる．

図 1 は，5 つの文章からなる評価対象テキストに対して，学習の完了した HMM を用いて分割を行う例を示している．正解は，最初の 2 文が段落 A であり，続く 2 文が B，最後の 1 文が段落 C である．出力記号が 1 単語である HMM なので，評価対象テキストデータを出力する最適な状態遷移系列の長さは，評価対象テキストデータの単語数 n と等しい．最適な状態遷移系列が図中右下であったと仮定する．本例では，1 つの文の中で割り当てられた状態が最も多いものをその文の分類される段落とする．この場合，最初の文の直後と，最後の文の直前が分割推定位置となる．正解と比較することにより，最初の分割は誤分割，後ろの分割は正しい分割であることが分かる．

2.1 ナイーブな文章生成モデル

本研究では，以下の視点に基づき HMM を用いた段落分割手法を改善する．

- 段落の切れ目は必ず文の終わりであり，文の途中で区切られることはない．

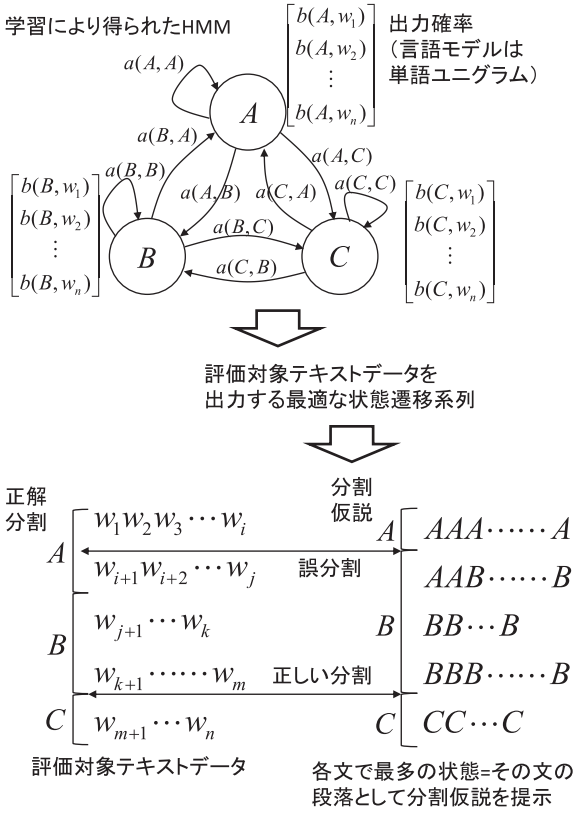

図 1　HMM による段落分割

Fig. 1　Text segmentation via HMM.

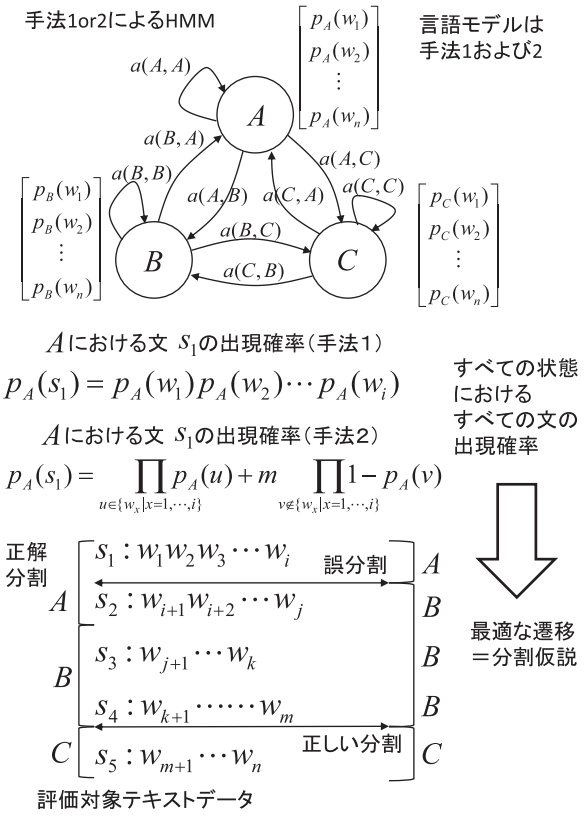

図 2　ナイーブな文章生成モデルと HMM

Fig. 2　Naive models and HMM.

- 複数の単語の組合せで特徴的な用語となる場合がある．

以上の点から，1 文を分割できない範囲と見ることにより，分割性能の向上が期待できる．

一般に，複数の単語を含む範囲を取り扱うには n-gram を出力記号とする HMM とすることが考えられる [7]．しかし，n-gram の出現確率は，単語 1 つの出現率 p の場合に比べ p^n となるため，より多くの学習データが必要であり，学習時間も増加する．本論文では 1 つの文章に対してその出現率を求める方法を定めることにより，1 文を出力記号とする手法を提案する．

まず，1 つの文章 s は単語列 $w_1w_2\cdots w_i$ からなると仮定し，確率変数 x は単語 w について $x=w$ もしくは $x=\neg w$ の 2 値をとるものとする．ある状態 q が 1 文を出力する際にその中に w が含まれている確率を $p_q(x=w)$ とする．以後，確率変数を省略し $p_q(x=w)$ を $p_q(w)$ と表す．この確率は後に述べる学習アルゴリズムの中で再推定される．

以上を用いて，ある文 $s=w_1w_2\cdots w_i$ に対するある状態 $q\in Q$ における出力確率 $p_q(s)$ を以下のように 2 通り提案する．

- 手法 1：文章に含まれる単語の出現確率の積を文章の出現確率とする方法．すなわち，

$$p_q(s)=\prod_{h=1,\cdots,i} p_q(w_h)$$

である．

- 手法 2：文章に含まれない単語も考慮した方法．すなわち，

$$p_q(s)=\prod_{h=1,\cdots,i} p_q(w_h) + m\left(\prod_{u\neq w_h(h=1,\cdots,i)}(1-p_q(u))\right)$$

である．ここで，m は，第 1 項と第 2 項の重みを調整する係数であり，1 文の平均単語数 l と学習データすべての単語の異なり数 v より l/v を基準として，予備実験より定める．

すなわち，HMM における状態遷移確率は従来と同じく，状態 q,r について $a(q,r)$ と表され，各状態は話題を表すが，出力記号は 1 つの文章となる．

図 2 に手法 1 および 2 による評価テキストの分割の流れを示す．本手法による HMM は後述の学習アルゴリズムで獲得される．分割はまず，評価対象テキストの各文章（図では s_1，s_2，s_3，s_4，s_5 の 5 文）に対して，すべての状態（A，B，C）におけるその文章の出力確率の計算から行われ，$i=1,\cdots,5$ について $P_A(s_i)$，$P_B(s_i)$，$P_C(s_i)$ が求められる．この結果を用いて最適な状態遷移系列を求める（図では最適系列が $ABBBC$ であったと仮定した）．これに従い段落分割を行うと，A と B の間の分割位置は誤分割であり，B，C 間の分割は正しい分割となる．状態遷移系列の長さは，単語ユニグラムによる従来法では，評価対

象テキストの単語数であったが，本手法では文の数と一致する．

本手法における HMM の獲得は，以下のように Baum-Welch アルゴリズムが適用できる．状態 $q \in Q$ における 1 文に対する出力確率が定まると，t 番めの文 s_t を出力するまでの前向き確率 $\alpha_t(q)$，後向き確率 $\beta_t(q)$，および γ（$q \in Q$，$r \in Q$）を以下のように定められる．

$$\alpha_1(q) = a(0,q)p_q(s_1)$$
$$\alpha_t(q) = p_q(s_t)\sum_{q' \in Q}\alpha_{t-1}(q')a(q',q)$$
$$\beta_T(q) = 1$$
$$\beta_{t-1}(q) = \sum_{q' \in Q} a(q,q')p_{q'}(s_t)\beta_t(q')$$
$$\gamma_t(q,r) = \frac{\alpha_t(q)a(q,r)p_r(s_{t+1})\beta_{t+1}(r)}{\sum_{q' \in Q}\alpha_T(q')}$$
$$\gamma_t(q) = \sum_{r \in Q}\gamma_t(q,r)$$

ここで T は学習テキストにおける文の数である．文 s_t に出現するすべての単語の集合を W_t とすると，各パラメータの再推定も Baum-Welch のアルゴリズムが適用できる．

$$a(0,q) = \gamma_1(q)$$
$$a(q,r) = \frac{\sum_{t=1}^{T}\gamma_t(q,r)}{\sum_{t=1}^{T}\gamma_t(q)}$$
$$p_q(w) = \frac{\sum_{t:w \in W_t}\gamma_t(q)}{\sum_{t=1}^{T}\gamma_t(q)}$$

特に $p_q(w)$ の再推定については，一般的な記号出力確率ではなく，文 s_t が単語 w を含む確率として，本論文における定義と矛盾なく定められる．さらに，ある文においてある状態に到達する確率を分母とし，そのときの文に単語 w が含まれている割合を再推定値としているため，すべての文を高い確率で出力するような極値に収束する可能性も低く，EM アルゴリズムとしての動作も引き継がれている．

図 3 に手法 1，2 における Baum-Welch アルゴリズムの動きを示す．再推定すべきパラメータは，すべての状態対 $(A,B) \in Q \times Q$ に対する $a(A,B)$ および，すべての状態 $q \in Q$ とすべての単語 w に対する $P_q(w)$ である．学習データのすべての文 s_t について，$P_q(s_t)$ を現時点での HMM から算出し，$\alpha_t(q)$，$\beta_t(q)$，$\gamma_t(q)$ を求める．これらの結果から，上述の再推定式によりパラメータを更新する．

2.2　ポアソン分布による文章生成モデル

1 つの文章においてある単語 w が含まれるか否かの確率 p は微小な確率であり，文章の長さ n との積 np は一定であると仮定できる．すると，ポアソン分布で表すことができる．すなわち，$u = p_q(w)$ を期待値とするポアソン分布

$$PO_{(q,w)}(k) = \frac{u^k}{k!}\exp(-u)$$

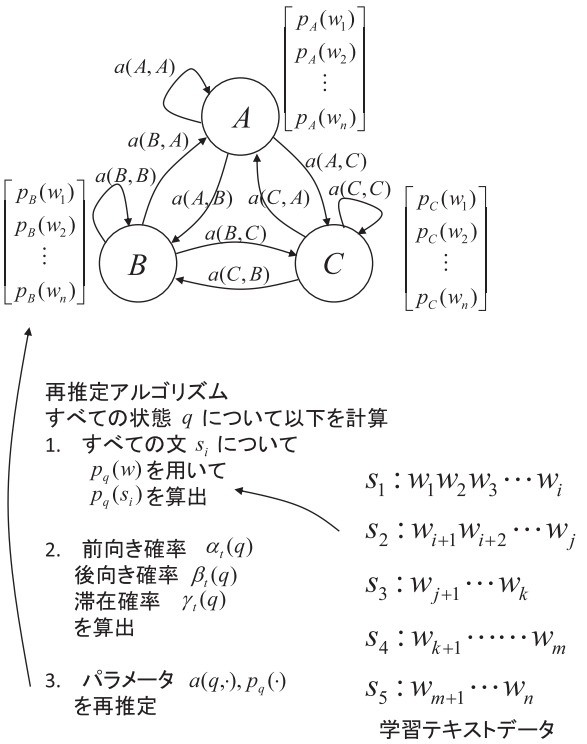

図 3　手法 1,2 における Baum-Welch アルゴリズム
Fig. 3　Baum-Welch algorithm of our method.

を用いて，ある状態から出力される文に単語 w が k 個含まれる確率を求めることができる．以上より，ある状態 $q \in Q$ における文章出力確率の定め方を以下のように定める．

- 手法 3：文章に含まれる単語の出現数をポアソン分布で推定する方法．すなわち，文 s に出現するすべての単語の集合を W_s とし，s に出現する単語 w の個数を k_w すると，

$$p_q(s) = \prod_{w \in W_s} PO_{(q,w)}(k_w)$$

である．

この場合も，文に対する出力確率 $p_q(s)$ が定義できるため，前節と同じ再推定アルゴリズムが利用できる．

図 4 にポアソン分布による評価テキストの分割の流れを示す．手法 1 および 2 の場合と同様に分割はまず，評価対象テキストの各文章（図では s_1，s_2，s_3，s_4，s_5 の 5 文）に対して，すべての状態（A，B，C）におけるその文章の出力確率の計算から行われ，$i = 1,\cdots,5$ について $P_A(s_i)$，$P_B(s_i)$，$P_C(s_i)$ が求められる．この結果を用いて最適な状態遷移系列を求める（図では最適系列が $ABBBC$ であったと仮定した）．これに従い段落分割を行うと，手法 1 および 2 の場合と同様に A と B の間の分割位置は誤分割であり，B，C 間の分割は正しい分割となる．

2.3　ナイーブベイズ識別器を用いた手法

以前我々は，テキストに段落のラベルを付けた学習デー

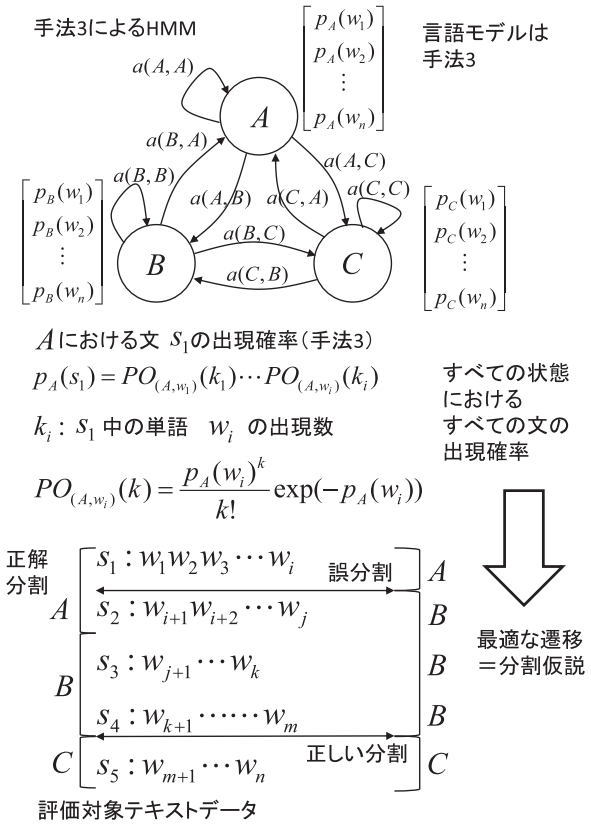

図 4　ポアソン分布による文章生成モデルと HMM
Fig. 4　A poisson distribution model and HMM.

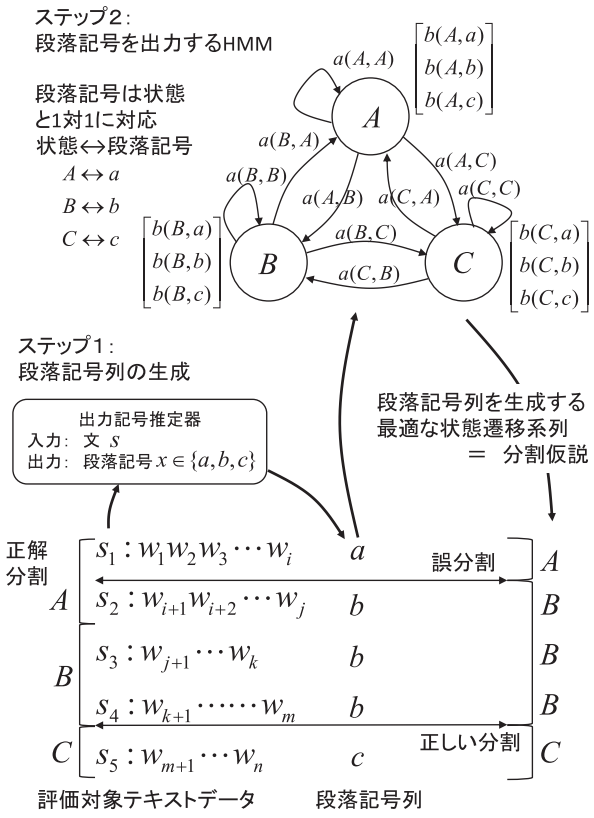

図 5　識別器を用いたモデルと HMM
Fig. 5　A supervised learning model and HMM.

タから，その段落ラベルを出力記号とする HMM を構成し，段落分割を行う手法を提案した [6]．この手法では，学習データを 1 文ごとに分割し，文とラベルとの関係から 1 文に対してどのラベルを割り当てるべきかを決定する分類器を構成する．さらに，学習データであるテキストを 1 文ごとに 1 つのラベルが付いたラベルの記号列に変換し，そのラベルの記号列を出力する HMM を構成する．分割対象のテキストに対しては，分類器を用いて 1 文ごとにラベルを推定し，ラベルの列を作成する．次に作成したラベルの列を生成する最適な状態遷移系列を学習により構成した HMM を用いて推定し，各文がどの話題であるかを決定し，段落分割を行う．この手法では，学習に段落のラベルが付いた正解の学習データが必要であり，本論文における手法とは別の枠組みとなるが，参考のため評価実験を行った．

図 5 に識別器を用いた手法の概要を示す．評価対象テキストの各文に対して，ステップ 1 のナイーブベイズ推定器で文が属する段落を推定する．この推定結果を記号列として並べて，図では *abbbc* という段落記号列が得られたとする．ステップ 2 では，段落記号を出力記号とする HMM において，*abbbc* という出力記号列に対して最適な状態遷移系列を求める．この状態遷移系列が *ABBBC* であったとすると，分割位置は，最初の文の直後（誤分割）と最後の文の直前（正しい分割）となる．

3.　評価実験

3.1　実験データ

評価実験として，ウェブのニュース記事をつなげたものを 1 つのテキストとし，このテキストに対して段落分割を行った．学習データは，ランダムに話題が転換するデータセットとした．ニュース記事の素材の仕様は以下のとおりである．

(1) ウェブニュースの記事のジャンル：5 ジャンル（社会，国内，国際，娯楽，スポーツ）
(2) 各ジャンル平均 1,493 記事ずつ，計 7,467 記事
(3) 1 つの記事の平均長（単語数）：301 単語
(4) 記事の最小長および最大長：最小 14 単語，最大 2,501 単語
(5) 記事集合全体で使われている単語の異なり数：21,943

この記事データから，ランダムに 10 記事を選び結合したものを 1 つの評価テキストとする．学習データとして，評価テキストを 100 テキスト準備し，さらに別の 100 テキストを評価データとしたもを 1 セットとする．4 セットの学習データをそれぞれ data1，data2，data3，data4 と呼び，4 セットの評価データをそれぞれ test1，test2，test3，test4 と呼ぶ．それぞれのデータの詳細は**表 1** のとおりである．

表 1　学習データと評価データ
Table 1　Learning and evaluation data.

	学習データ					評価データ					
	data1	data2	data3	data4	平均	test1	test2	test3	test4	平均	全体平均
1 テキストの最大行数	199	210	224	232	216.25	222	281	262	272	259.25	237.75
1 テキストの最小行数	70	65	60	51	61.50	59	63	58	57	59.25	60.38
1 テキストの平均行数	114.86	116.03	117.60	116.41	116.23	112.33	115.69	114.65	116.48	114.79	115.51
1 文の最大単語数	240	133	151	240	191.0	240	282	240	240	250.5	220.75
1 文の最小単語数	1	1	1	1	1	1	1	1	1	1	1
1 文の平均単語数	26.46	25.88	26.23	26.08	26.16	25.98	26.00	26.24	26.66	26.22	26.19

3.2　評価方法

評価テキスト data1，data2，data3，data4 に対してそれぞれの学習データを用いて HMM を構成する．その後，得られた HMM を用いて評価データ test1，test2，test3，test4 を段落分割し，分割位置の正しさを評価する．評価は，以下の値を比較した．

- テキスト中の 2 つの文に対する誤分類（2 文評価）．これは，文献 [2] における評価尺度である．
- 分割位置の一致に関する精度と再現率および F 値．
- 正しいジャンルに分類されている文章の割合（分類率）．

3.3　2 文評価による結果

正しく段落分割がなされているテキスト（正解データ）を t_r と表し，同じテキストを分割アルゴリズムで分割したもの（仮説データ）を t_h と表す．ともに長さは，n 文であるとする．2 文評価は，t_r における i 番めと j 番めの文章 r_i，r_j と t_h における i 番めと j 番めの文章 h_i，h_j について，段落への分割が一致しているか否かを測る尺度である．すなわち，以下の値 $P_D(t_r, t_h)$ を求める．

$$P_D(t_r, t_h) = \sum_{1 \le i \le j \le n} D(i,j)(\delta_r(i,j) \bar{\oplus} \delta_h(i,j))$$

ここで，$\delta_r(i,j)$ は，t_r において，r_i と r_j が同一の段落に含まれていれば 1 そうでなければ 0 をとる関数であり，$\delta_h(i,j)$ は，t_h において，h_i と h_j が同一の段落に含まれていれば 1 そうでなければ 0 をとる関数である．また，$\bar{\oplus}$ は排他的論理和の否定である．すなわち，両辺が同一の値の場合のとき，かつそのときに限り 1 となる．関数 $D(i,j)$ は，i 番めの文と j 番めの文の位置に対する価値を与える関数である．一般には i と j が遠く離れている場合は低い値をとり，近い場合は高い値を返す．本論文では，以下の 2 種類の関数を用いた．

- 評価対象テキストの全文数を l とする．定数 k（$< l$）に対して，

  $$D_k(i,j) = \begin{cases} 1/(lk - \frac{k(k-1)}{2}) & |i-j| < k \\ 0 & otherwise \end{cases}$$

  とし，$k = 2, 4, 6, 8, 16$ の 5 種類について実験を行う．この関数は，k 個の文までは一定の価値とし，それ以上離れている文どうしの評価はしないという意味を持つ．全文数が l であるので，$i = j$ の場合も含めて $|i-j| < k$ となる i，j の組合せは，$lk - \frac{k(k-1)}{2}$ であるため，上記の値となる．この k の値が大きいほど，離れた文どうしが正しい段落に分類されているかどうかを考慮することとなる．また，k の値が小さいほど，分割位置の近い間違いにのみ注目することとなる．定数 k を段落の平均文数の半分とした場合，ベースラインのアルゴリズム（段落の切れ目を，1. ランダムにする場合，2. すべての文の境界とする場合，3. 一定間隔とする場合，4. テキスト全体を 1 つの段落とする場合）のいずれにおいても評価値 $P_D(t_r, t_h)$ が 0.5 に近くなることが知られている [2]．本実験では，段落の平均文数はおよそ 12 であるため，$k = 6$ を中心とし，分割の傾向をみるため上記の 5 種類とする．

- 定数 $u = 0.2$ として，

  $$D_e(i,j) = u \exp(-u|i-j|)$$

  とする．これは，i と j の差が 3 で 0.1，差が 5 で 0.07 の値となる，緩やかな定数を選んだ．

表 2 に 2 文評価の結果を示す．評価対象としたアルゴリズムは，以下のとおりである．

- 単語ユニグラムを用いた HMM（従来手法）．従来法は，最適状態遷移系列を求めたときに 1 文の中で最も多くとどまった状態を文全体の状態と判定し段落の切れ目を求めた．
- 手法 1（ナイーブな生成モデル）を用いた HMM．
- 手法 2（文章に出現しない単語も考慮したモデル）を用いた HMM．（重みパラメータ $m = 10^{-3}$，これは (1 文の平均単語数 26) / (データ全体の単語の異なり数 22000) の値に基づく）．
- 手法 3（ポアソン分布モデル）を用いた HMM．
- ナイーブベイズ識別器を用いた手法（正解データが必要なモデル）．

関数 D_2，D_4，D_6，D_8 において提案手法が従来法である単語ユニグラムによる HMM の性能を上回った．特に，テストデータの 1 つの段落の平均文数 12 の半分に対応する D_6 では，ベースラインの性能が 0.5 であるのに対し，従

表 2　2 文評価の結果
Table 2　Evaluation by co-occurrence agreement probability.

	D_2					D_4				
	test1	test2	test3	test4	平均	test1	test2	test3	test4	平均
単語ユニグラム（従来法）	0.861	0.869	0.861	0.867	0.869	0.773	0.791	0.770	0.784	0.780
手法 1（ナイーブ）	0.919	0.948	0.943	0.953	0.941	0.835	0.879	0.871	0.889	0.869
手法 2（$m = 10^{-3}$）	0.956	0.958	0.956	0.958	0.957	0.882	0.886	0.883	0.888	0.885
手法 3（ポアソン分布）	0.914	0.916	0.905	0.914	0.912	0.817	0.826	0.802	0.818	0.816
識別器を用いた手法（教師あり学習）	0.967	0.969	0.966	0.969	0.968	0.922	0.927	0.923	0.929	0.925
	D_6					D_8				
	test1	test2	test3	test4	平均	test1	test2	test3	test4	平均
単語ユニグラム（従来法）	0.750	0.771	0.745	0.761	0.757	0.742	0.766	0.737	0.753	0.750
手法 1（ナイーブ）	0.796	0.842	0.831	0.853	0.831	0.775	0.821	0.808	0.831	0.809
手法 2（$m = 10^{-3}$）	0.824	0.827	0.826	0.833	0.828	0.782	0.782	0.783	0.793	0.785
手法 3（ポアソン分布）	0.773	0.787	0.756	0.774	0.773	0.753	0.771	0.737	0.754	0.754
識別器を用いた手法（教師あり学習）	0.897	0.905	0.898	0.906	0.902	0.882	0.893	0.884	0.892	0.888
	D_{16}					D_e				
	test1	test2	test3	test4	平均	test1	test2	test3	test4	平均
単語ユニグラム（従来法）	0.745	0.774	0.740	0.755	0.754	1.87e-2	1.91e-2	1.87e-2	1.87e-2	1.88e-2
手法 1（ナイーブ）	0.761	0.804	0.786	0.807	0.790	1.96e-2	2.05e-2	2.04e-2	2.05e-2	2.03e-2
手法 2（$m = 10^{-3}$）	0.700	0.700	0.698	0.716	0.704	1.93e-2	1.92e-2	1.94e-2	1.93e-2	1.93e-2
手法 3（ポアソン分布）	0.737	0.763	0.727	0.745	0.743	1.91e-2	1.94e-2	1.89e-2	1.89e-2	1.91e-2
識別器を用いた手法（教師あり学習）	0.862	0.881	0.867	0.874	0.871	2.18e-2	2.19e-2	2.19e-2	2.17e-2	2.18e-2

来法が 0.757，手法 1 と 2 がそれぞれ 0.831 と 0.828 であり，性能向上がみられた．手法 3 は 0.773 であり，従来法に比べ，わずかな性能向上となった．しかし，教師あり学習の枠組みである識別器を用いた手法に対しては，いずれも下回る性能となった．

関数 D_{16} においては，手法 2 および 3 が従来法を下回る性能となった．D_k の k 値は大きくなるほど離れた位置の文章を比較対象とするため，性能は低下する傾向にある．特に，k が平均段落行数の 12 を超えている D_{16} では，平均的に 2 つ以上の分割位置をまたいだ評価となる．単語ユニグラム（従来法）は，D_{16} における性能が D_8 での性能に比べ向上していることから，頑強さというよりも，ランダムな分割を行っていると解釈することができる．D_e に関しては，すべての提案手法において従来法を上回る性能となった．

特に手法 1 は，教師あり学習である識別器を用いた手法に近い性能が得られており，本論文における提案手法の有効性が示せたといえる．

3.4　分割位置と分類率による結果

正しく段落分割されているテキストデータ t_r において，段落の切れ目の直前の文番号の集合を B_r とする．同様に分割アルゴリズムを用いて分割したテキストデータ t_h の段落の切れ目の直前の文番号の集合を B_h とする．B_r, B_h の一致について，精度，再現率，F 値を調べる．これを完全一致の結果と呼ぶ．すなわち，精度 p は $p = |(B_r \cap B_h)|/|B_h|$，再現率 r は $r = |(B_r \cap B_h)|/|B_r|$ で計算され，F 値はこれらの調和平均 $F = 2/(1/p + 1/r)$ である．さらに $B'_r = \bigcup_{i \in B_r} \{i, i-1, i+1\}$ および $B'_h = \bigcup_{i \in B_h} \{i, i-1, i+1\}$ とし，精度 $p = |(B'_r \cap B_h)|/|B_h|$，再現率 $r = |(B_r \cap B'_h)|/|B_r|$ および F 値 $F = 2/(1/p + 1/r)$ で求めたものを前後許容による結果と呼ぶ．

さらに，t_r と t_h で同じ段落に分類されている文章の割合を分類率とする．

表 3 に分割位置一致に関する性能を，**表 4** に分類率の性能を示す．比較対象とするアルゴリズムは，2 文評価の場合と同じである．

完全一致の尺度では，手法 1 および 3 の F 値は従来法を上回った．手法 2 の F 値は下回ったが，精度を見ると提案手法の中では最も高い値であり，再現率の低さが F 値の低下に影響している．これは，分割位置の提示が少ないことを意味しているが，前後許容の評価では提案手法の中では低いものの従来法よりも高い F 値となっている．本実験では，各テキストは 10 個の段落を含むため，分割位置はすべて 9 個である．したがって，分割位置の一致では，性能評価を計算する際の分母が 9 と少ないため，結果にばらつきが大きくなる．また，教師あり学習の枠組みを用いたアルゴリズムの性能には，提案手法のいずれも到達していないことも 2 文評価における結果と同様である．

本結果で特徴的な点として，単語ユニグラム（従来法）による結果において，精度が低く，再現率が高いことがあげられる．これは，より多くの位置を分割位置として提示

表 3　分割位置の結果
Table 3　Evaluation by absolute agreement.

	完全一致の精度					前後許容の精度				
	test1	test2	test3	test4	平均	test1	test2	test3	test4	平均
単語ユニグラム（従来法）	0.200	0.202	0.194	0.195	0.198	0.246	0.255	0.244	0.240	0.246
手法 1（ナイーブ）	0.258	0.372	0.351	0.390	0.264	0.371	0.496	0.492	0.535	0.473
手法 2（$m = 10^{-3}$）	0.393	0.362	0.381	0.392	0.382	0.541	0.490	0.554	0.528	0.528
手法 3（ポアソン分布）	0.253	0.245	0.229	0.230	0.239	0.342	0.330	0.314	0.313	0.325
識別器を用いた手法（教師あり学習）	0.541	0.568	0.537	0.558	0.551	0.620	0.647	0.619	0.651	0.634
	完全一致の再現率					前後許容の再現率				
	test1	test2	test3	test4	平均	test1	test2	test3	test4	平均
単語ユニグラム（従来法）	0.763	0.745	0.767	0.768	0.761	0.939	0.953	0.967	0.956	0.954
手法 1（ナイーブ）	0.506	0.502	0.477	0.482	0.492	0.732	0.670	0.691	0.660	0.688
手法 2（$m = 10^{-3}$）	0.242	0.233	0.217	0.247	0.235	0.345	0.312	0.323	0.345	0.331
手法 3（ポアソン分布）	0.564	0.573	0.535	0.559	0.558	0.761	0.761	0.752	0.748	0.756
識別器を用いた手法（教師あり学習）	0.739	0.746	0.742	0.726	0.738	0.848	0.855	0.863	0.857	0.856
	完全一致の F 値					前後許容の F 値				
	test1	test2	test3	test4	平均	test1	test2	test3	test4	平均
単語ユニグラム（従来法）	0.317	0.318	0.310	0.384	0.332	0.390	0.402	0.390	0.384	0.392
手法 1（ナイーブ）	0.342	0.428	0.404	0.431	0.401	0.493	0.570	0.570	0.591	0.556
手法 2（$m = 10^{-3}$）	0.300	0.284	0.277	0.303	0.291	0.421	0.381	0.408	0.418	0.407
手法 3（ポアソン分布）	0.349	0.343	0.321	0.326	0.335	0.472	0.460	0.443	0.441	0.454
識別器を用いた手法（教師あり学習）	0.625	0.645	0.623	0.631	0.631	0.716	0.737	0.721	0.740	0.729

表 4　分類率の結果
Table 4　Evaluation by precision.

	分類率				
	test1	test2	test3	test4	平均
単語ユニグラム（従来法）	0.713	0.750	0.723	0.733	0.730
手法 1（ナイーブ）	0.725	0.772	0.749	0.775	0.755
手法 2（$m = 10^{-3}$）	0.523	0.548	0.516	0.569	0.539
手法 3（ポアソン分布）	0.691	0.730	0.691	0.718	0.708
識別器を用いた手法（教師あり学習）	0.804	0.847	0.828	0.835	0.829

していることを示している．すなわち，細かく分割された段落が多数提示されている．この結果は，2 文評価において，k の値が 6，8，16 と変化しても性能が低下していないことの裏付けとなっている．これに対し提案手法では，精度と再現率のバランスはある程度とれており，2 文評価による結果とも矛盾しない．

分類率による評価では，手法 1 のみが従来法を上回っており，手法 2 および 3 は下回っている．特に手法 2 は提案手法の中でも低い値であるが，分割位置の提示自体が少ない場合，文の分類率は低くなるため完全一致における結果と一貫している．従来法は 2 文評価の結果に比べ分類率が高い値となっているが，これも分割位置が多く提示され，細かく分割されていることの裏付けとなっている．

4. おわりに

テキストセグメンテーションを HMM を用いて行う手法において，各状態が表す言語モデルを複数の単語が扱えるモデルとする方法を提案した．その結果，従来の単語ユニグラムを言語モデルとする手法に比べ，高性能であることが確認できた．提案手法の特徴として，複数の単語を扱い 1 文に対する出力確率を求めることができるが，n-gram による手法に比べて計算の負荷が低いことがあげられる．実際，学習に要する計算時間，評価に要する計算時間ともに単語ユニグラムを用いた従来手法にくらべ大差はなく，いずれも数時間から十数時間程度である．

提案手法の分割性能は，いずれも複数の評価尺度において従来手法を上回り，有効性が示せたといえる．しかし，教師あり学習の枠組みで処理を行うアルゴリズムの性能にはおよばなかった．単語ユニグラムによる従来法では，分割数が多くなり小さな段落が多数作られる傾向があったが，提案手法では解決されていることが，評価実験から明らかとなった．

今後の課題として，教師あり学習の手法に本提案手法を取り込むことが考えられる．教師あり学習の枠組みでは，

時系列データに対する処理は，CRF などの生成モデルが様々な分野で高い性能を示しており，本手法による複数の単語をまとめて取り扱う方法を応用できれば，より高性能な分割器を作ることができると思われる．

参考文献

[1] Hearst, M.A.: Texttiling: segmentaing text into multi-paragraph subtopic passages, *Computational Linguistics*, Vol.23, pp.33–64 (1997).

[2] Beeferman, D., Berger, A. and Lafferty, J.: Statistical models for text segmentation, *Machine Learning*, Vol.34, Nos.1–3, pp.177–210 (1999).

[3] McCallum, A., Freitag, D. and Pereira, F.: Maximum entropy markov models for information extraction and segmentation, *Proc. ICML'00*, pp.591–598 (2000).

[4] Ostendorf, M., Digalakis, V.V. and Kimball, O.A.: From HMM's to segment models: A unified view of stochastic modeling for speech recognition, *IEEE Trans. speech and audio processing*, Vol.4, No.5, pp.360–378 (1996).

[5] Yamron, J.P., Carp, I., Gillick, L., Lowe, S. and van Mulbregt, P.: A hidden markov model approach to text segmentation and event tracking, *Proc. IEEE conf. on Acoustics, Speech and Signal Processing*, Vol.1, pp.333–336 (1998).

[6] 但馬康宏, 北出大蔵, 中林　智, 藤本浩司, 小谷善行：HMM とテキスト分類器による対話の段落分割，情報処理学会論文誌：数理モデル化と応用，Vol.2, No.2, pp.70–79 (2009).

[7] 長野　雄, 鈴木基之, 牧野正三：HMM を用いた複数 n-gram モデルによる言語モデルの構築，情報処理学会研究報告 SLP 40-26, pp.151–156 (2002).

但馬 康宏　（正会員）

1996 年電気通信大学大学院電気通信学研究科博士前期課程修了．同年（株）IHI 入社．2001 年電気通信大学大学院電気通信学研究科博士後期課程修了．博士（工学）．同年東京農工大学助手（助教）．2009 年岡山県立大学准教授，現在に至る．文法推論，機械学習，自然言語処理の研究に従事．電子情報通信学会，人工知能学会各会員．