情報処理学会論文誌　数理モデル化と応用　Vol. 2　No. 2　70–79 (Mar. 2009)

# HMMとテキスト分類器による対話の段落分割

但馬康宏[†1]　北出大蔵[†2]　中野未知子[†2]
中林智[†3]　藤本浩司[†4]　小谷善行[†1]

テキストを段落に分割する問題に対して，本論文ではシナリオに基づいた分割を行う手法を提案する．すなわち分割対象とするテキストは，あるシナリオに基づいたテキストであると仮定し，テキスト内の段落はシナリオに記述された段落のいずれかに分類されるもとのする．本手法では，正確に段落分割された学習データから，1つの文が属する段落を推定するナイーブベイズモデルおよび段落番号の列を出力とするHMMの2つを用いて学習データをモデル化する．分割対象テキストは，1文ごとにナイーブベイズで段落番号を推定され，その段落番号の列に対してHMMの最適な状態遷移系列を求めることにより段落分割を行う．従来，HMMによる段落分割では，単語を出力記号とするHMMを構成することでテキストのモデル化を行うのが一般的であるが，本手法では，段落番号を出力記号とするHMMを利用する点に特徴がある．これにより，特に対話文などの間投詞が多く特徴的な単語の少ないテキストデータに対しても高い分割性能を得ることができる．評価実験として，実際の対話，およびウェブのニュース記事に対して段落分割を行い，本手法の有効性を確かめた．

# A Dialogue Segmentation Method via HMM and a Text Classifier

Yasuhiro Tajima,[†1] Daizo Kitade,[†2]
Michiko Nakano,[†2] Tomo Nakabayashi,[†3]
Koji Fujimoto[†4] and Yoshiyuki Kotani[†1]

We present a new method for text segmentation via HMM and a text classifier. In our method, we suppose that a scenario exists for the target text and we can use learning data about the scenario. Our method has two stages to segment the target text. Every sentence or utterance is classified into a topic in the scenario at the firest stage. Then, in the second stage, the target text is segmented by an HMM which is a model of topic strings generated at the firest stage. Ordinarily, the target text is segmented by an HMM whose output is a string of words. In contrast, our HMM ourputs a string of topic numbers and it absorbs the miss classification in the first stage. For evaluation, we apply our method to a dialogue segmentation task and a news text segmentation task. We can confirm that our method performs better than the ordinal HMM segmentation method.

## 1. はじめに

談話などのテキストからその話題を抽出することは，談話理解における重要な問題の1つである．一般にテキストは，1つのサブトピックで特徴づけられる段落が複数連なる形で構成される場合が多い．テキストをその段落ごとに分割する問題は，テキスト分割（テキストセグメンテーション）と呼ばれる問題である．テキストを段落へ分割することにより，段落が扱うサブトピックに対するキーワード抽出や段落の要約，さらに段落間の関連を調べることによりそのテキストの構成を調べるなどの応用が可能となる．

テキスト分割の代表的な手法として，Hearst[3]による手法がよく知られている（以後Hearst法と呼ぶ）．これはまず，単語の種類を特徴ベクトルの成分とし，テキストのある窓幅内に出現する単語の出現数から特徴ベクトルを算出する．この窓を移動させることにより特徴ベクトルが変化するが，その変化点を抽出することにより分割位置を決定する手法である．この手法では，分割位置を決定する閾値や窓幅など調節すべきパラメータが複数あり，精度の良い結果を得るためにはそれらのパラメータ調整が不可欠である．また，テキスト内の統計的情報を用いて分割精度を上げる研究も多くなされており[4]，統計的な手法による発展も行われている[5]．

また，音声認識の分野ではHMMによる認識手法が研究されており，それをテキスト分割に適用した手法も行われている[1),2)]．すなわち，1つの状態はテキスト内の段落すなわちサブトピックを表すとし，テキストに出現する単語を出力記号とするHMMを機械学習により獲得する．このHMMにおいて，分割対象となるテキストの出現率が最も高くなる状

†1 東京農工大学工学部
Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology
†2 トランスコスモス株式会社
transcosmos inc.
†3 株式会社金融エンジニアリング・グループ
Financial Engineering Group, Inc.
†4 テンソル・コンサルティング株式会社
Tensor Consulting, Inc.

態遷移系列を求めてテキスト分割を行うものである．ここで，HMM の各状態はそれぞれ 1 つのサブトピックを表すため，分割を行う前に分割対象のサブトピックが適切に含まれた学習データを用いて HMM を構成する必要がある．また，HMM をテキスト分割について改良したセグメントモデル（SM）[6] も提案されているが，状態をサブトピックと見立てて利用する点に差はない．

上記いずれの分割手法も，段落分割を正確に行ったテキストデータ（正解データ）を利用せずに実行することができる手法である．しかし実際に分割精度を上げるためには，Hearst 法では正しい分割をもとにしたパラメータ設定が，また HMM による方法では適切な状態数と状態間の接続方法の設計が不可欠である．したがって，従来の手法においても正解データの利用は不可欠であるといえる．さらに，これらの正解データはある一定の共通性を持ったデータの集合であると仮定できる．すなわち，分割対象のテキストがニュース記事がつながったものか，ネット上のあるテーマに関するチャットログなのかなどの情報が既知であると仮定できる．

本論文では，このようなテキストに対する予備知識をシナリオという形で表現し，そのシナリオに沿った段落分割がすでに行われている正解データの利用を前提として，以下のような段落分割手法を提案する．まず，テキストを 1 文や 1 発話ごとに分割し，テキスト分類アルゴリズムによってそれらの文や発話があらかじめ所属する段落を推定する．その後，推定された段落番号を出力記号とする HMM によって分割位置を決定する 2 段階による方法である．すわなち，第 1 段階でその文章や発話が含まれるべき段落を大まかに推定し，第 2 段階の HMM はその段落番号の列から分割位置を推定する．テキスト分類アルゴリズムには，ナイーブベイズを用いた．したがって，第 1 段階の分類を求めるためには，学習データは正確に段落分けがなされており，さらに各段落のサブトピックごとに一意な段落番号が付与されている必要がある．また，第 2 段階の HMM はサブトピックに対応する状態を持つ点は従来手法と同様だが，それぞれの状態は段落番号を出力記号とする HMM である．

本手法の特徴として，相槌や間投詞など特徴的な単語を含まない文が多く出現するテキストに対して，従来手法より分割性能が向上するという点があげられる．対話における発話は，文書として作成されたテキストに比べ一般的な単語のみで構成された文の出現頻度が高く，語の省略が頻繁に行われる．そのようなテキストに対して本手法は有効である．すなわち，第 1 段階における 1 文や 1 発話ごとの推定は間違っていても，それらの段落番号を出力とする HMM でそのゆらぎを吸収でき，本質的にサブトピックが変化した点を抽出することが可能となる．さらに，従来の HMM による分割手法に比べて学習に要する時間が短く済むという点も特徴である．これは特に，Baul-Welch アルゴリズムにおいてリスケーリングの必要な長さのテキストでは，単語を出力記号とした従来手法ではその種類の多さから学習時間が長くなる．その点，本手法における HMM は出力記号の種類がサブトピックの種類，すなわち状態数と同一なためより短い時間で学習を収束させることができる．

本手法の効果確認としてウェブニュース記事に対する段落分割と実際の対話記録に対する段落分割を行い，従来の HMM による手法と比較実験を行った．その結果，対話に対する段落分割において，従来手法よりも良い結果を得ることができた．また，ニュース記事に対する分割でも従来手法と劣らない性能であることを確認した．

## 2. 提 案 手 法

### 2.1 シナリオと分割対象テキスト

本論文で分割対象とするテキストは，段落が扱うサブトピックを**図 1** に示すようなチャートに基づいているものとし，このようなチャートをシナリオと呼ぶ．

シナリオにおける各箱は 1 つの段落を表し，箱の中にはサブトピックを記述するなど，他段落との区別がつくものとする．また，このシナリオは対象テキストとは独立に作成されてもよいものとする．すなわち，テキストが作成された後から分析のためにシナリオを作成し，そのシナリオに沿った段落分割を試みることもできる．したがって，テキストの内容がシナリオに沿っていないことも想定できる．本論文では，シナリオは段落の種類を限定し，

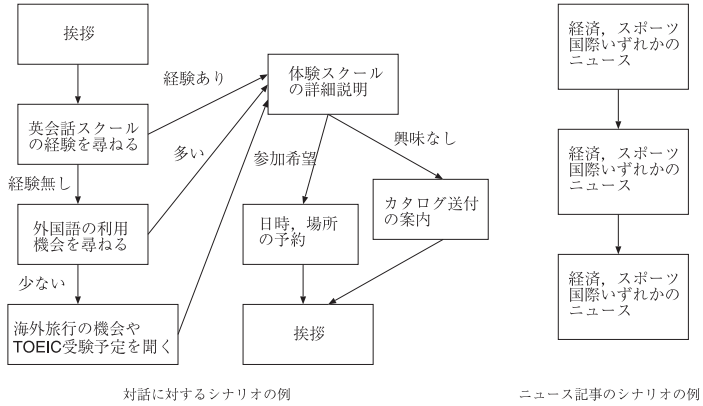


**図 1**　シナリオの例
Fig. 1　An example of scenarios.

正解データを作成するために用いられ，その接続や段落間の関係は利用しない．

従来の HMM による段落分割手法でも，その状態数と状態間遷移の有無を決定する際には，分割対象テキストのサブトピック数に対応した状態数とするなどシナリオに対応する事前知識を利用している．したがって，本論文におけるシナリオは，分割システムに対しては大きな付加情報となるが，シナリオの利用については自然な仮定であるといえ，従来の手法における段落分割の仮定とは相違がある．

以上をふまえ本論文では，以下のような特徴を持つテキストを分割対象として考える．

- 分割対象に関する知識からシナリオが作成されており，利用可能である．
- 分割対象のテキストに対して学習用のデータがあり，学習用データを用いて段落分割のモデルを構成することができる．
- 学習用データでは，シナリオに基づいた正解となる段落分割がすでになされており，シナリオにおける段落に対応する番号が付けられている．すなわち，学習用データから同一の段落番号を持つ段落を抜き出すと，シナリオにおける 1 つの箱に対応する段落を集めることができる．この正解はシナリオに厳密に沿っている必要はない．
- 分割対象のテキストに表れるサブトピックは，すべて学習データにも出現するサブトピックである．

本論文では，シナリオおよび学習データを用いて以下の手順で分割モデルを構成し，そのモデルを用いて段落分割を行う．まず，モデルの構成手順を述べる．

(1)　学習データから各段落，各単語ごとにナイーブベイズで用いるパラメータを抽出する．
(2)　学習データのテキストを 1 文ごとにその文が属している段落番号に置き換え，テキスト 1 つに対して段落番号の列を 1 つ作成する．
(3)　上記で作成した段落番号の列を学習データとして，HMM を機械学習で構成する．

次に，分割対象のテキストに対するモデルの適用方法を示す．

(1)　分割対象のテキストを 1 文もしくは 1 発話ごとに分解し，各文ごとにナイーブベイズで段落番号を推定し，テキストで 1 つの段落番号の列を作成する．
(2)　推定された段落番号の列を最も高い確率で出力する状態遷移系列を求める．
(3)　得られた状態遷移系列から対象テキストの段落分割を行う．

**図 2** に本手法でのデータと処理の流れを示す．

### 2.2　ナイーブベイズによる 1 文ごとの段落推定

テキストの集合 $U$ において，$T \in U$ を 1 つのテキストとする．$T$ は $n$ 文からなるとし，$i$ 番目の文を $u_i$ と表す．文 $u_i$ は $m_i$ 個の単語 $w_1^{(i)}, w_2^{(i)}, \cdots, w_{m_i}^{(i)}$ の連接であるとする．

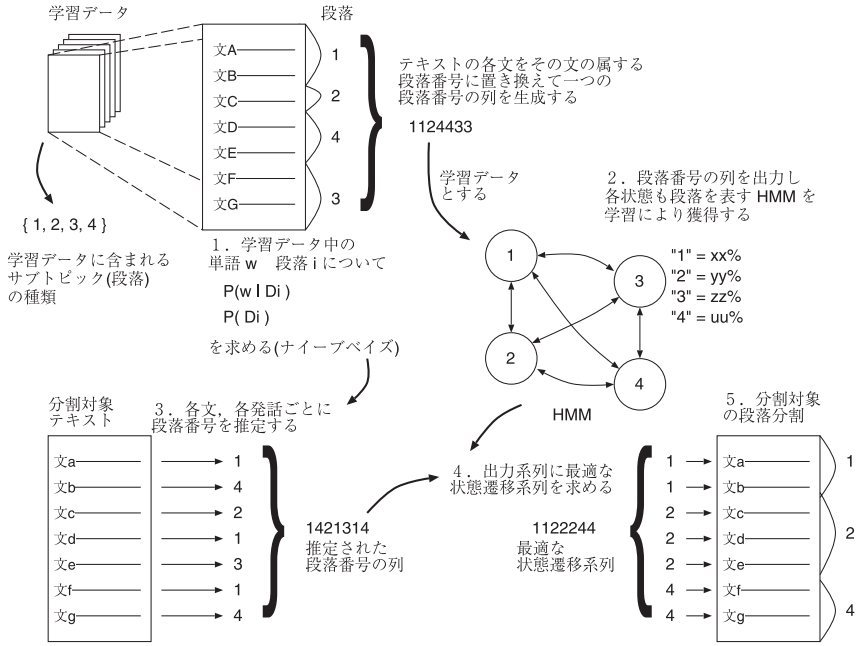


**図 2**　本手法でのデータと処理の流れ
Fig. 2　The overview of our method.

与えられた文が段落番号 $k$ の段落に含まれる事象を $D_k$ と表すと，文 $u_i$ が段落番号 $k$ に含まれる確率は，$P(D_k|u_i)$ であり，

$$P(D_k|u_i) = \frac{P(u_i|D_k)P(D_k)}{P(u_i)}$$

として求められる．ここでさらに，

$$P(u_i|D_k) = P(w_1^{(i)}|D_k)P(w_2^{(i)}|D_k)\cdots P(w_{m_i}^{(i)}|D_k)$$

と近似し，単語 $w$ について，

$$P(w|D_k) = \frac{\sum_{u\in H_k} c(w,u)}{\sum_{u\in H_k}\sum_{v\in W_u} c(v,u)}$$

とする．ただし，$H_k$ は学習データにおいて段落番号 $k$ に属するすべての文の集合であり，$W_u$ は文 $u$ に現れるすべての単語の集合，$c(w,u)$ は文 $u$ における単語 $w$ の出現回数である．

上記の方法で文 $u_i$ に付けられた段落番号を $o(u_i)$ とする．テキスト $T = u_1 u_2 \cdots u_n$ に対する段落番号の列 $o(u_1)o(u_2)\cdots o(u_n)$ を出力する HMM を次節で構成する．

### 2.3　HMM による段落分割

学習データの各文をその文が属する段落番号で置き換え，学習データのテキスト 1 つに対して段落番号の列を 1 つ作成する．この段落番号の列の集合を学習データとして HMM を構成する．ここで HMM の隠れ状態はそれぞれ段落番号に対応している．すなわち，この HMM は，出力記号も各状態もそれぞれが段落番号に対応している．HMM の学習アルゴリズムは Baum-Welch アルゴリズムとした．

学習の終了した HMM は，学習データをシナリオに沿ってモデル化したものと見ることができる．このモデルを使い分割対象のテキストを段落分割する．分割対象のテキスト $T = u_1 u_2 \cdots u_n$ から，前節のナイーブベイズにて 1 文ごとに段落番号の推定を行い，段落番号の列 $O = o(u_1)o(u_2)\cdots o(u_n)$ を得る．その後，$O$ を出力する最適な状態遷移系列 $q_1 q_2 \cdots q_n$ を Viterbi アルゴリズムで求める．ここで，各 $q_i$（$i = 1, 2, \cdots, n$）は HMM の状態である．最終的に文 $u_i$ は段落 $q_i$ に属すると推定され，段落番号の変化点，すなわち，$q_j \neq q_{j+1}$ となる位置で段落分割を行う．

一般的な HMM による段落分割と比べ本論文における HMM は，テキストそのものを出力することはできない．しかし，本論文における HMM の特徴である対話テキストに対する分割では有効である．

本手法における学習後の HMM の各状態は，学習データを効率良く表現できる段落に対応する．ここで，学習開始時には各状態がどの段落に対応するかは定められていないため，各状態がどの段落を表すかは，学習データにおける段落の遷移に依存して変化する．したがって，HMM において $i$ 番目の状態が $i$ 番めの段落に対応するか否かは保証されない．しかし，学習後の各状態において，その状態が最も高い確率で出力する出力記号がその状態に割り当てられた段落となるので，段落内容の把握は可能である．これは，従来の HMM を用いた段落分割手法でも同様である．本研究においては，テキストの段落分けを目的とするため，従来の手法と同じく段落番号と状態番号の一致までは考慮しないものとする．また，各文ごとに段落が交互に入れ替わるような極端な学習データの場合は，2 つ以上の段落が同一の状態に割り当てられたり，同一の段落が 2 つ以上の状態で表現される可能性もある．しかし一般には，学習データを効率良く表現できる段落と状態の対応は，学習データにおける段落番号 1 つに状態 1 つが対応する場合が多いと考えられる．どのようなデータが極端なものとなるかは，今後の研究課題である．

## 3. 評 価 実 験

本手法の効果確認として，実際の対話に対する段落分割とウェブニュース記事のテキストデータに対する段落分割を行った．

### 3.1　対話に対する段落分割

ある架空の英会話教室への勧誘を目的とした対話を設定し，以下の条件で実際の対話を行い，その書き起こし記録に対して本手法を適用し評価を行った．

- 被験者は，勧誘オペレータ 4 名，被勧誘者 62 名である．被勧誘者 1 名に対して 1 対話が実施され，計 62 対話のテキストデータを利用した．対話は電話で行われ，被勧誘者は英会話教室からの電話であることだけが事前に知らされ，対話の主題については教えられていない．
- シナリオは，**図 3** に示す 2 種類を用いた．対話の主題が勧誘であるので，事前にオ

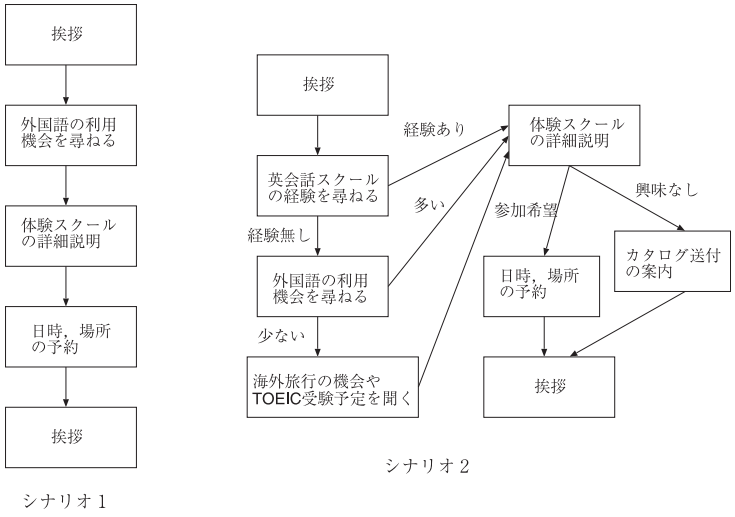


**図 3**　対話実験に対するシナリオ
Fig. 3　Scenarios for the dialogue segmentation.

ペレータはトークスクリプトと呼ばれる話の進め方のメモを作成した．シナリオ 2 はオペレータの作成したトークスクリプトをもとにしたシナリオである．シナリオ 1 は実験後，分析の段階でオペレータ以外の者が対話記録を見て作成したシナリオであり，Left-to-Right モデルの HMM を想定している．

- 対話記録のデータは以下のとおりである．
  - 対話数：62
  - 実験全体での発話数：10,358
  - 1 つの対話あたりの平均発話数：167.06
  - 1 つの発話に含まれる平均単語数：10.58
  - 1 つのテキストに含まれる平均単語数：1,767.53
  - すべての対話における単語数の合計：109,587
- 以上の対話に対して，本手法を適用するが，ナイーブベイズにて推定するテキストの長さを 1 文ではなく，1 発話ごととした．これは，文としての切れ目が曖昧な発話が存在することによる．付録 A.1 に対話のサンプルを示す．
- テキストの形態素解析には，mecab 0.91 を使用した．
- 評価は，62 対話を 5 分割交差検定で行った．それぞれのデータ集合に含まれる対話の数は，**表 1** のとおりである．

比較対象として，以下の 2 種類の分割方法も同時に行った．

**HMM：** 従来から知られている単語を出力記号とした HMM による分割結果．この場合，1 つの発話の中で最も多くの単語が対応づけられた状態をその発話が属する段落として，段落境界が常に発話と発話の間となるようにした．

**bayes：** ナイーブベイズで 1 発話の属する段落を推定した結果をそのまま採用した場合の結果．すなわち，本論文の提案手法における第 1 段階の出力をそのまま結果として利用した場合である．

ここで分割対象のテキストが 1,500 語程度を超えると計算精度不足により Baum-Welch アルゴリズムを実行することが難しくなる．そこで本論文では，リスケーリングを行った．具体的には以下の手順で学習中の前向き確率および後ろ向き確率を置き換える．状態数が $N$ である HMM の各状態を $Q_k$（$k = 1, 2, \cdots, N$）と表し，状態 $Q_k$ から $Q_l$ への遷移確率を $a_{kl}$，状態 $Q_k$ における出力記号 $o$ の出力確率を $b_k(o)$ と表す．また，Baum-Welch アルゴリズムにおいて，時刻 $i$ における状態 $Q_k$ の前向き確率を $\alpha_i(Q_k)$，後向き確率を $\beta_i(Q_k)$ と表す．このとき前向き確率 $\alpha_i(Q_k)$ を

$$
\begin{aligned}
\hat{\alpha_1}(Q_k) &= \alpha_1(Q_k) \\
\bar{\alpha_{i+1}}(Q_k) &= \sum_{j=1}^{N} \hat{\alpha_i}(Q_j) a_{kj} b_k(o(u_{i+1})) \\
c_{i+1} &= \frac{1}{\sum_{j=1}^{N} \bar{\alpha_{i+1}}(Q_j)} \\
\hat{\alpha_{i+1}}(Q_k) &= c_{i+1} \bar{\alpha_{i+1}}(Q_k)
\end{aligned}
$$

で定められる $\hat{\alpha_i}(Q_k)$ で置き換え，同様に後向き確率 $\beta_i(Q_k)$ を

$$
\begin{aligned}
\hat{\beta_n}(Q_k) &= \beta_n(Q_k) \\
\bar{\beta_i}(Q_k) &= \sum_{j=1}^{N} a_{kj} b_j(o(u_{i+1})) \hat{\beta_{i+1}}(Q_k) \\
\hat{\beta_i}(Q_k) &= c_i \bar{\beta_i}(Q_k)
\end{aligned}
$$

なる $\hat{\beta_i}(Q_k)$ で置き換えて学習を進める方法である．ここで，$n$ はテキスト $T$ に含まれる文の数，すなわち HMM への 1 つの学習データの長さであり，$o(u_i)$ は時刻 $i$ における出力記号，すなわち文 $u_i$ に対するナイーブベイズによる推定段落番号である．

まず，シナリオ 1 での分割結果を示す．表中の F 値は，精度 p，再現率 r を用いて

$$F = \frac{2pr}{p + r}$$

として求められる．**表 2** は，段落の分割位置が正確に正しい場合のみに正解，そうでない場合を不正解とした場合の分割の精度および再現率である．最下段の “発話割合” は，テキスト中の全発話のうち正しい段落番号が付けられた発話の割合である．この表より，本手法が最も高い性能となっていることが分かる．**表 3** は，正しい分割位置から前後 1 発話までは正解と見なした場合の性能である．いずれも本手法の F 値が最も高く，次いで HMM となり，1 発話ごとのナイーブベイズによる分類は再現率は高いが，精度が低い値となっている．これは，発話ごとに段落番号が変化するため，段落分割点を過剰に生成しているためである．

次に同じ対話データに対して，シナリオ 2 で段落分割を行った場合の結果を示す（**表 4**,

**表 1**　データ集合とそのサイズ
Table 1　The size of dialogue data.

| データ名 | data1 | data2 | data3 | data4 | data5 |
|---|---|---|---|---|---|
| 対話数 | 13 | 13 | 12 | 12 | 12 |

**表 2**　シナリオ 1 での分割性能
Table 2　Performances on the scenario 1.

| 手法 | data1 | data2 | data3 | data4 | data5 | 平均 |
|---|---|---|---|---|---|---|
| 本手法（精度） | 0.4625 | 0.4423 | 0.4583 | 0.2708 | 0.5417 | 0.4355 |
| 本手法（再現率） | 0.4744 | 0.4423 | 0.4653 | 0.2708 | 0.5417 | 0.4395 |
| 本手法（F 値） | 0.4679 | 0.4423 | 0.4618 | 0.2708 | 0.5417 | 0.4375 |
| HMM（精度） | 0.3231 | 0.3681 | 0.3667 | 0.3708 | 0.3738 | 0.3600 |
| HMM（再現率） | 0.3154 | 0.3692 | 0.3708 | 0.3667 | 0.3833 | 0.3605 |
| HMM（F 値） | 0.3192 | 0.3687 | 0.3687 | 0.3687 | 0.3785 | 0.3603 |
| bayes（精度） | 0.1216 | 0.1406 | 0.0968 | 0.1060 | 0.1099 | 0.1155 |
| bayes（再現率） | 0.7436 | 0.6538 | 0.6528 | 0.6875 | 0.7292 | 0.6935 |
| bayes（F 値） | 0.2091 | 0.2314 | 0.1686 | 0.1836 | 0.1910 | 0.1980 |
| 本手法（発話割合） | 0.8968 | 0.8597 | 0.8729 | 0.8921 | 0.8189 | 0.8684 |
| HMM（発話割合） | 0.7802 | 0.8061 | 0.8403 | 0.7815 | 0.8699 | 0.8149 |
| bayes（発話割合） | 0.7101 | 0.8061 | 0.8403 | 0.7815 | 0.8699 | 0.6986 |

**表 3**　シナリオ 1 での分割性能（前後 1 発話許容）
Table 3　Performances on the scenario 1 (1 sentence permission).

| 手法 | data1 | data2 | data3 | data4 | data5 | 平均 |
|---|---|---|---|---|---|---|
| 本手法（精度） | 0.6346 | 0.7115 | 0.6667 | 0.5417 | 0.7500 | 0.6613 |
| 本手法（再現率） | 0.6474 | 0.7115 | 0.6806 | 0.5417 | 0.7500 | 0.6667 |
| 本手法（F 値） | 0.6410 | 0.7115 | 0.6735 | 0.5417 | 0.7500 | 0.6640 |
| HMM（精度） | 0.5846 | 0.6445 | 0.6667 | 0.5917 | 0.7190 | 0.6404 |
| HMM（再現率） | 0.5654 | 0.6462 | 0.6750 | 0.5833 | 0.7333 | 0.6395 |
| HMM（F 値） | 0.5748 | 0.6453 | 0.6708 | 0.5875 | 0.7261 | 0.6400 |
| bayes（精度） | 0.1465 | 0.1918 | 0.1311 | 0.1378 | 0.1474 | 0.1515 |
| bayes（再現率） | 0.8782 | 0.9231 | 0.8958 | 0.9167 | 0.9791 | 0.9180 |
| bayes（F 値） | 0.2511 | 0.3177 | 0.2287 | 0.2396 | 0.2563 | 0.2601 |

**表 4**　シナリオ 2 での分割性能
Table 4　Performances on the scenario 2.

| 手法 | data1 | data2 | data3 | data4 | data5 | 平均 |
|---|---|---|---|---|---|---|
| 本手法（精度） | 0.1985 | 0.2528 | 0.3123 | 0.3215 | 0.2358 | 0.2646 |
| 本手法（再現率） | 0.3458 | 0.2772 | 0.4349 | 0.4804 | 0.3271 | 0.3741 |
| 本手法（F 値） | 0.2522 | 0.2645 | 0.3635 | 0.3852 | 0.2741 | 0.3100 |
| HMM（精度） | 0.1298 | 0.1303 | 0.1648 | 0.1519 | 0.1080 | 0.1367 |
| HMM（再現率） | 0.7996 | 0.5240 | 0.7913 | 0.7013 | 0.6732 | 0.6975 |
| HMM（F 値） | 0.2233 | 0.2087 | 0.2728 | 0.2497 | 0.1862 | 0.2287 |
| bayes（精度） | 0.0990 | 0.0933 | 0.1361 | 0.1227 | 0.1003 | 0.1103 |
| bayes（再現率） | 0.8083 | 0.7276 | 0.8040 | 0.7755 | 0.8293 | 0.7893 |
| bayes（F 値） | 0.1763 | 0.1654 | 0.2328 | 0.2119 | 0.1790 | 0.1936 |
| 本手法（発話割合） | 0.5660 | 0.6960 | 0.6218 | 0.6365 | 0.7013 | 0.6451 |
| HMM（発話割合） | 0.5496 | 0.6403 | 0.5818 | 0.5433 | 0.5472 | 0.5716 |
| bayes（発話割合） | 0.5113 | 0.5226 | 0.5244 | 0.5175 | 0.5532 | 0.5261 |

**表 5**　シナリオ 2 での分割性能（前後 1 発話許容）
Table 5　Performances on the scenario 2 (1 sentence permission).

| 手法 | data1 | data2 | data3 | data4 | data5 | 平均 |
|---|---|---|---|---|---|---|
| 本手法（精度） | 0.3510 | 0.4597 | 0.4781 | 0.4723 | 0.4325 | 0.4391 |
| 本手法（再現率） | 0.5944 | 0.5016 | 0.6746 | 0.6863 | 0.5799 | 0.6081 |
| 本手法（F 値） | 0.4414 | 0.4797 | 0.5596 | 0.5596 | 0.4954 | 0.5100 |
| HMM（精度） | 0.1504 | 0.2075 | 0.1948 | 0.2045 | 0.1394 | 0.1791 |
| HMM（再現率） | 0.9032 | 0.8666 | 0.9519 | 0.8709 | 0.8420 | 0.8859 |
| HMM（F 値） | 0.2579 | 0.3348 | 0.3234 | 0.3313 | 0.2392 | 0.2980 |
| bayes（精度） | 0.1194 | 0.1228 | 0.1702 | 0.1599 | 0.1183 | 0.1381 |
| bayes（再現率） | 0.9667 | 0.9429 | 0.9603 | 0.9872 | 0.9633 | 0.9645 |
| bayes（F 値） | 0.2125 | 0.2173 | 0.2889 | 0.2753 | 0.2108 | 0.2416 |

**表 5**）.

いずれも本手法が最も F 値が高い結果となった．シナリオ 1 の場合と同じく，精度は本手法が最も高く，再現率はナイーブベイズが最も高い値となる．シナリオ 1 の場合との相違点は，従来型の HMM による結果が悪化し，ナイーブベイズによる結果に近い値となった点である．このことから，状態数が増え状態間の接続が複雑になると従来法では性能の悪化が著しいことが分かる．

### 3.2　ニューステキストに対するトピック分割

本手法と従来型の HMM による段落分割およびナイーブベイズでの分割の性能を従来法で評価対象として多く取り上げられているニュース記事のトピック分割において比較した．

ウェブのニュース記事より，国内，海外，経済，エンターテイメント，スポーツ，テクノロジーの 6 つのトピックの記事を集め，以下の 2 種類のデータを作成した．

(1)　Left-to-Right モデルに沿ったシナリオ．**図 4** におけるシナリオ 3 に基づき，6 つのトピックのうち 4 つのトピックがランダムに選ばれ，その 4 つのトピックがシナリオ 3 の順で出現するテキストデータ．すなわち，国内，海外，経済，エンタテイメント，スポーツ，テクノロジのうち，2 つのトピックがランダムに削られたデータである．以後，データセット 3 と呼ぶ．

(2)　すべてのトピックがランダムに出現するシナリオ．6 つのトピックの記事からランダムに 4 つの記事を選択して，1 つのテキストデータとしたもの．以後，データセット

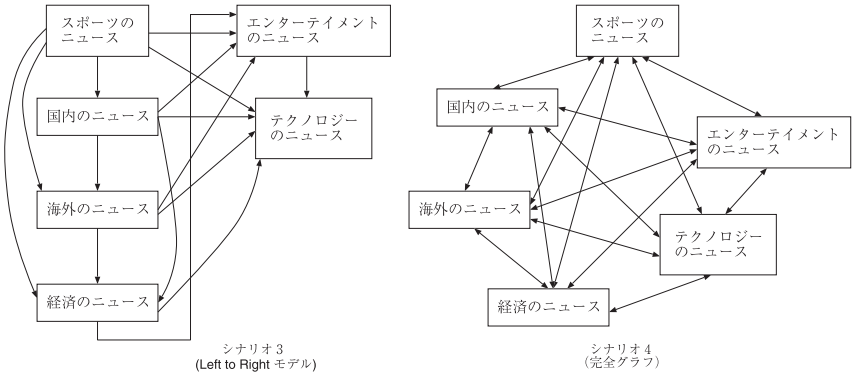


**図 4**　ニュース記事に対するシナリオ
Fig. 4　Scenarios for the news text segmentation.

**表 6**　データセット 3, 4 の仕様
Table 6　Descriptions for data set 3 & 4.

| データセット | テキスト数 | 1 記事の平均文数 | 1 記事の平均単語数 |
|---|---|---|---|
| 3 | 200 | 91.54 | 1,868.32 |
| 4 | 200 | 96.69 | 2,000.92 |

4 と呼ぶ．

データセット 3 およびデータセット 4 の内容を**表 6** に示す．

**表 7**，**表 8** にデータセット 3 に対する結果を示す．分割位置が厳密に正しい場合を正解とした評価では，本手法よりも従来手法による HMM の方が F 値が良い結果となっている．これは，ニュースのテキストは，各トピックごとに特徴的な単語が明確であり，従来手法での分割が効果的であることを示している．しかし，分割位置の前後 1 文までを正解とした評価では，本手法の方が F 値が高く，本手法をニューステキストに適用した場合でも十分な性能が得られることが分かる（表 8）．

**表 9**，**表 10** は，データセット 4 に対する結果である．分割位置が正確である場合，1 文の前後を認めた場合ともに本手法が最も F 値が高くなっている．特に従来手法の HMM による分割では，Left-to-Right 型では本手法を上回る性能であったが，トピックがランダムに出現するモデルになるとナイーブベイズによる方法と大きな差がなくなっている．これは，従来手法がデータセット 4 のようなテキストのモデルにうまく対応できないことを示している．対照的に本手法では，複雑なモデルになるほど F 値は低下するが，従来手法に比べ顕著な低下は見られない．

**表 7**　シナリオ 3 での分割性能
Table 7　Performances on the scenario 3.

| 手法 | data1 | data2 | data3 | data4 | data5 | 平均 |
|---|---|---|---|---|---|---|
| 本手法（精度） | 0.4681 | 0.4694 | 0.4083 | 0.5111 | 0.5078 | 0.4729 |
| 本手法（再現率） | 0.4500 | 0.4611 | 0.4194 | 0.5083 | 0.5278 | 0.4733 |
| 本手法（F 値） | 0.4589 | 0.4652 | 0.4138 | 0.5097 | 0.5176 | 0.4731 |
| HMM（精度） | 0.5517 | 0.3797 | 0.6316 | 0.5086 | 0.5586 | 0.5260 |
| HMM（再現率） | 0.5958 | 0.2861 | 0.6667 | 0.5500 | 0.6000 | 0.5397 |
| HMM（F 値） | 0.5729 | 0.3263 | 0.6487 | 0.5285 | 0.5786 | 0.5328 |
| bayes（精度） | 0.0755 | 0.0747 | 0.0709 | 0.0807 | 0.0815 | 0.0767 |
| bayes（再現率） | 0.8222 | 0.8778 | 0.8556 | 0.8778 | 0.8611 | 0.8589 |
| bayes（F 値） | 0.1383 | 0.1376 | 0.1309 | 0.1479 | 0.1488 | 0.1407 |
| 本手法（発話割合） | 0.7354 | 0.7374 | 0.7110 | 0.7435 | 0.7678 | 0.7390 |
| HMM（発話割合） | 0.8307 | 0.5392 | 0.8290 | 0.7722 | 0.8006 | 0.7543 |
| bayes（発話割合） | 0.5409 | 0.5311 | 0.5182 | 0.5512 | 0.5543 | 0.5391 |

**表 8**　シナリオ 3 での分割性能（前後 1 文許容）
Table 8　Performances on the scenario 3 (1 sentence permission).

| 手法 | data1 | data2 | data3 | data4 | data5 | 平均 |
|---|---|---|---|---|---|---|
| 本手法（精度） | 0.6750 | 0.6139 | 0.5431 | 0.6167 | 0.6333 | 0.6164 |
| 本手法（再現率） | 0.6444 | 0.5722 | 0.5556 | 0.6111 | 0.6389 | 0.6044 |
| 本手法（F 値） | 0.6594 | 0.5923 | 0.5492 | 0.6139 | 0.6361 | 0.6104 |
| HMM（精度） | 0.6294 | 0.4033 | 0.7524 | 0.5978 | 0.6189 | 0.6004 |
| HMM（再現率） | 0.6792 | 0.3069 | 0.7889 | 0.6458 | 0.6667 | 0.6175 |
| HMM（F 値） | 0.6534 | 0.3486 | 0.7702 | 0.6209 | 0.6419 | 0.6088 |
| bayes（精度） | 0.0952 | 0.0824 | 0.0805 | 0.0881 | 0.0907 | 0.0874 |
| bayes（再現率） | 0.9944 | 0.9889 | 0.9444 | 0.9722 | 0.9722 | 0.9744 |
| bayes（F 値） | 0.1737 | 0.1522 | 0.1483 | 0.1616 | 0.1659 | 0.1604 |

## 4. おわりに

シナリオが想定できるテキストの段落分割に対して，ナイーブベイズと HMM を組み合わせた分割手法を提案し，その効果を確かめた．特に，対話などの不完全な文章や間投詞などが多いテキストに対する段落分割において，従来手法よりも性能が高いことを確認した．また，ニューステキストにおけるトピック分割でも，従来手法に劣らない性能であることを

**表 9**　シナリオ 4 での分割性能
Table 9　Performances on the scenario 4.

| 手法 | data1 | data2 | data3 | data4 | data5 | 平均 |
|---|---|---|---|---|---|---|
| 本手法（精度） | 0.3125 | 0.2967 | 0.2432 | 0.3569 | 0.3605 | 0.3140 |
| 本手法（再現率） | 0.6111 | 0.5778 | 0.4389 | 0.6500 | 0.5444 | 0.5644 |
| 本手法（F 値） | 0.4136 | 0.3920 | 0.3130 | 0.4608 | 0.4337 | 0.4035 |
| HMM（精度） | 0.1233 | 0.1152 | 0.0855 | 0.0896 | 0.0985 | 0.1024 |
| HMM（再現率） | 0.7208 | 0.7653 | 0.7431 | 0.7333 | 0.7208 | 0.7367 |
| HMM（F 値） | 0.2106 | 0.2003 | 0.1533 | 0.1597 | 0.1733 | 0.1798 |
| bayes（精度） | 0.0657 | 0.0800 | 0.0548 | 0.0781 | 0.0759 | 0.0709 |
| bayes（再現率） | 0.8806 | 0.8778 | 0.8417 | 0.8778 | 0.8528 | 0.8661 |
| bayes（F 値） | 0.1222 | 0.1466 | 0.1030 | 0.1434 | 0.1394 | 0.1311 |
| 本手法（発話割合） | 0.6939 | 0.6881 | 0.6010 | 0.7378 | 0.6186 | 0.6679 |
| HMM（発話割合） | 0.6663 | 0.6475 | 0.6004 | 0.6179 | 0.6111 | 0.6286 |
| bayes（発話割合） | 0.5724 | 0.5588 | 0.5029 | 0.6053 | 0.5365 | 0.5551 |

**表 10**　シナリオ 4 での分割性能（前後 1 文許容）
Table 10　Performances on the scenario 4 (1 sentence permission).

| 手法 | data1 | data2 | data3 | data4 | data5 | 平均 |
|---|---|---|---|---|---|---|
| 本手法（精度） | 0.3952 | 0.3693 | 0.3168 | 0.4044 | 0.4625 | 0.3896 |
| 本手法（再現率） | 0.7528 | 0.7417 | 0.5778 | 0.7528 | 0.7139 | 0.7078 |
| 本手法（F 値） | 0.5183 | 0.4931 | 0.4092 | 0.5261 | 0.5613 | 0.5026 |
| HMM（精度） | 0.1633 | 0.1412 | 0.1120 | 0.1137 | 0.1211 | 0.1302 |
| HMM（再現率） | 0.9528 | 0.9625 | 0.9569 | 0.9486 | 0.9125 | 0.9467 |
| HMM（F 値） | 0.2789 | 0.2463 | 0.2004 | 0.2030 | 0.2138 | 0.2290 |
| bayes（精度） | 0.0756 | 0.0842 | 0.0614 | 0.0837 | 0.0846 | 0.0779 |
| bayes（再現率） | 0.9889 | 0.9417 | 0.9500 | 0.9694 | 0.9500 | 0.9600 |
| bayes（F 値） | 0.1404 | 0.1546 | 0.1154 | 0.1541 | 0.1553 | 0.1441 |

確認した．

本手法は，正解データの利用が前提となっているため，シナリオが想定できるテキストという条件が必要である．しかし，シナリオの作成はテキストとは独立に行えるため，本手法を利用する際の妨げにはならないと見込まれる．今後の課題として，シナリオを用いずに本手法を適用する方法などがあげられる．

## 参 考 文 献

1) 越仲孝文，奥村明俊，磯谷亮輔：HMM の変分ベイズ学習によるテキストセグメンテーション及びその映像インデキシングへの応用，信学論，Vol.J89-D, No.9, pp.2113–2122 (2006)
2) 今井　亨，R. Schwartz，小林彰夫，安藤彰男：話題混合モデルによる放送ニュースからの話題抽出，信学論，Vol.J81-D-II, No.9, pp.1955–1964 (1998)
3) Hearst, M.A.: Texttiling: Segmentaing text into multi-paragraph subtopic passages, *Computaional Linguistics*, Vol.23, pp.33–64 (1997)
4) 別所克人：クラスタ内変動最小基準に基づくテキストセグメンテーション，情報処理学会論文誌，Vol.47, No.3, pp.957–967 (2006)
5) Beeferman, D., Berger, A. and Lafferty, J.: Statistical models for text segmentation, *Machine Learning*, Vol.34, Nos.1-3, pp.177–210 (1999)
6) 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)

## 付　　録

### A.1　対話のサンプル

各行の先頭の “O:” はオペレータの発話を表し，“C:” は被勧誘者の発話を表す．発話の最後に付与されている “()” 内の数字がシナリオ 2 での正解の段落番号を表す．

**対話** *no.28*

O：あのー、XXX 様のお宅でしょうか？(0)
C：はい。(0)
O：はい。YYY 様でいらっしゃいますでしょうか？(0)
C：はい。(0)
O：あの、お忙しいところ恐れ入ります。私、トランスコスモス英会話スクールのオオタと申します。(0)
C：はい。(0)
(中略)
O：少々お時間をいただいてもよろしいでしょうか？(0)
C：はい、大丈夫です。(0)
O：ありがとうございます。トランスコスモス英会話スクールなんですけれども、来年で開校 30 周年を迎えることになりまして、これを機会に無料の体験スクールの開催を計画しております。(1)
C：はい。(1)
O：はい。あの、YYY 様は、先日ですね、あの私共の英語学習に関するセミナーにご参加

いただいたということなんですけれども、あの、これまでになにか、英会話スクールへのご通学などをお考えになったということはございますでしょうか？(1)
C：あ、はい。えーっと、いろいろ考えてはいるんですけれども、どっかスクールに行くとかいうのは全然ないので。(1)
O：あ、さようでございますか。(1)
C：はい。(1)
O：それでは何か、英会話スクールへのご通学をお考えになったということは、あの、特に実践的な会話のお勉強をされたいというようなご希望が、(1)
C：そうですね。はい。(1)
O：あ、さようでございますか、ありがとうございます。あの、それではですね、具体的になにか英語をお使いになる機会などが、あの、ございますでしょうかね？(2)
C：は、はい。国際会議とかで。(2)
O：はい。(2)
C：使いますんで。(2)
O：あの、ビジネス関係でお使いになるということでしょうか？(2)
C：はい。(2)
O：ありがとうございます。あの、今回の私どもの体験スクールなんですけれども、(4)
C：はい。(4)
O：あの、まず構内をご見学いただきまして、その後に一般的な私どものレッスンを体験していただくというようなですね、全体といたしまして 1 時間程の内容となっております。(4)
C：あ、はい。(4)
O：はい。あの先ほどの、あの、(4)
C：今回のそれはなんか、け (4)
O：はい。(4)
C：見学だけなんですか?(4)
O：あ、はい。(4)
C：なにか、あの、そ、その紹介みたいなかたちで終わってしまう、体験レッスンていうだけなん、話ですか？そしたら、あの、(4)
O：はい。(6)
C：パンフレットか何かいただければそれで、(6)
O：あ、(6)
C：いらないというか。(6)
O：さようでございますか。(6)
C：はい。(6)
O：はい。あの、特にですね、あの、私ども、あの、ネイティブな外国人講師として、優秀な講師をそろえておりますので、あの、そちらの実際のレッスンがどのような仕方で行われるかということをですね、あの、校舎に来ていただきまして、あの、肌で感じていただきたいという点とですね、(6)
C：あー。(6)
（中略）
O：あの、もしよろしければ資料だけでも送らせていただきたいんですけれども (6)
C：はい。(6)
O：いかがでしょうか？ (6)
C：あ、じゃ、それだけお願いします。(6)
O：あ、はい。ありがとうございます。それではセミナーの際におうかがいしたご住所に (6)
C：はい。(6)
O：お送りいたしますので、(6)
C：はい。(6)
O：よろしくお願いいたします。(7)
C：あ、はい。お願いします。(7)
O：あの、それでは本日はお時間いただきまして、ありがとうございました、はい、失礼いたします。(7)

(平成 20 年 8 月 21 日受付)
(平成 20 年 10 月 9 日再受付)
(平成 20 年 10 月 23 日採録)

**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．同年石川島播磨重工業（株）入社．2001 年電気通信大学大学院博士後期課程修了．同年東京農工大学工学部助手．2007 年同大学共生科学技術研究院助教．博士（工学）．計算学習理論とくに文法推論およびその応用の研究に従事．電子情報通信学会，人工知能学会各会員．

**北出　大蔵**

2002 年一橋大学経済学部卒業．数理経済学専攻．トランスコスモス入社後，データ/テキストマイニングを用いた新サービスの企画/開発に従事．現在，VOC コンサルタントとして，販売促進や解約防止等のマーケティング活動に，WEB サイトやコールセンターで集めた顧客の声の分析結果を活用するためのソリューションを展開中．近著に『アウトバウンドの本』（リックテレコム．2005.10）がある．

**中野未知子**

1998 年國學院大學文学部卒業．日本文学専攻．2000 年トランスコスモス入社．コールセンターのマネジメントを経て，テキストマイニング技術を活用した顧客の声分析サービス開発に参画．現在は，顧客の声分析アナリスト育成や大学生，社会人を対象とした能力開発サービスの企画・運用に取り組んでいる．

**中林　　智**

自由学園最高学部数理科学研究室にて，文章表現の解析に取り組む．2004 年卒業．同年（株）金融エンジニアリング・グループ入社．コンサルタントとして主に金融機関のリテール部門における信用リスク分析に従事．

**藤本　浩司**

1985 年上智大学理工学部数学科卒業．1999 年東京農工大学大学院工学研究科博士後期課程修了．博士（工学）．1985 年より日本アップジョン社（現ファイザー（株）），アメリカンエキスプレス（Japan）社，金融エンジニアリンググループ（株）にて，生物統計，人工知能，データベースマーケティング，広告・宣伝，信用リスク分析に従事．2007 年よりテンソル・コンサルティング（株）を創業し，ビジネス数理モデリングの研究開発に取り組んでいる．

**小谷　善行**（正会員）

1949 年生．1977 年東京大学大学院博士課程修了．同年東京農工大学勤務．現在，同大学教授．人工知能，知識処理，自然言語処理，知識獲得，ゲームシステム，教育工学の研究に従事．電子情報通信学会，人工知能学会等各会員．前コンピュータ将棋協会会長．