社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ＳＬＰ－51　（９）
2004／5／28

HMMの変分ベイズ学習によるテキスト文書の話題分割法

越仲　孝文†　　磯　　健一†　　奥村　明俊†

† NEC メディア情報研究所
神奈川県川崎市中原区下沼部 1753
E-mail: †koshinak@ap.jp.nec.com

あらまし　確率モデルに基づくテキスト分割法を提案する．left-to-right 型の離散 HMM をテキスト生成モデルと考え，テキスト分割を HMM のパラメータ推定問題として定式化する．パラメータ推定法として，最尤推定およびベイズ推定 (変分ベイズ法) を用いて，日本語ニュース番組を各ニュース項目へ分割する評価実験を行い，最尤推定に比べてベイズ推定が精度よくテキストを分割できることを示す．さらに，従来法として Hearst 法を取り上げ，従来法と比べた提案法の利点や課題を明らかにする．

キーワード　テキスト分割，隠れマルコフモデル，パラメータ推定，変分ベイズ学習．

An HMM-based text segmentation method using variational Bayes approach

Takafumi KOSHINAKA†, Ken-ichi ISO†, and Akitoshi OKUMURA†

† Media and Information Res. Labs., NEC
1753, Shimonumabe, Nakahara-ku, Kawasaki, Kanagawa, Japan
E-mail: †koshinak@ap.jp.nec.com

Abstract　This paper presents a new text segmentation method based on stochastic modeling. When supposing a generative model of a text document to be a discrete left-to-right hidden Markov model (HMM), a transition between topics in the text document corresponds to a state transition in the HMM, and text segmentation can be formulated as model parameter estimation using the text document. Compared to the traditional maximum likelihood approach, advantage of the Bayes approach (Variational Bayes) is shown by some experiments, which evaluate segmentation accuracy in segmenting Japanese broadcast news programs into each news article. Comparison between the proposed method and a conventional method, well-known Hearst's method, is also presented in this paper. The comparison shows the proposed method to be encouraging.

Key words　Text segmentation, hidden Markov models, parameter estimation, variational Bayes framework.

1. はじめに

テキストを意味的なまとまり，すなわち話題ごとに分割するセグメンテーションは，自然言語処理の要素技術として重要なばかりでなく，音声認識と併用されることによって，大規模ビデオアーカイブの検索支援などに利用可能なメタデータ生成に直接役立つことから，その重要性は近年増してきている [5] [6].

従来，テキスト分割においては，Hearst 法[3] に端を発する，変化点検出に基づく手法が広く用いられてきた．Heast 法では，テキスト上の各位置において，前後にそれぞれ一定単語数を覆う窓を設け，各窓内の単語の出現頻度に関する類似度 (余弦類似度) を計算する．この類似度を，着目する位置周辺における (語彙的) 結束度と考え，類似度の低い位置，すなわち結束の緩い位置を，話題の変化点として検出する．

変化点検出に基づく手法では，性能を確保するためにいくつかのパラメータを経験的に決めておく必要がある．例えば窓幅は，想定される話題長に応じて決める必要がある．また，変化点を頑健に求めるための平滑化パラメータも同様である．

図 1 は，後述の実験でも使用するテキストデータに対して，Hearst 法を適用した結果である．窓幅を適切に設定できれば (窓幅 120)，真の話題境界付近において比較的顕著な谷 (極小点) が現れる．しかし，それができない場合 (窓幅 60)，真の話題境界以外の位置にも目立つ谷が現れ，真の話題境界のみを選び取るのは極めて困難に思われる．

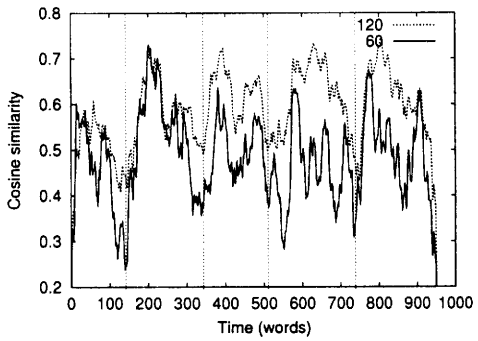

図 1 Hearst 法の適用例: 窓幅 60, 120 単語で, 隣接窓間の余弦類似度をプロット. 真の話題境界を鉛直線で併記. ここから平滑化を行い, 極小点を話題境界として検出する.

また, パラメータの最適値は, 個々の入力テキストごとに異なってしかるべきである. しかしながら実際応用上は, 平均的な入力に対してうまく動作するように設計された 1 組のパラメータセットで, すべての入力を処理することになる. つまり, 調整すべきパラメータの存在が, 平均からはずれたテキストが入力された場合に性能のロスを生じる要因となっている. さらにいえば, そもそも窓幅は話題長が決まって初めて設定が可能となる量であり, テキスト分割の前に窓幅を決めるのは本末転倒ともいえる.

変化点検出に基づく手法はオンラインアルゴリズムであり, 逐次入力されるテキストから即時的に話題境界を出力できるというメリットはある. しかしその一方で, 入力テキスト全体をみて処理を行うバッチ処理タイプのアルゴリズムがもしあれば, オンラインタイプのアルゴリズムよりも性能面で原理的に有利であり, 有用性は高いと筆者らは考える.

本稿では, バッチ処理を基本としたテキスト分割方式を提案する. テキストの生成モデルとして left-to-right 型の離散 HMM を仮定し, モデルを入力テキストに当てはめるパラメータ推定問題として, テキスト分割問題を定式化する. その際のパラメータ推定法としては, 一般的に用いられる最尤推定に加えて, ベイズ推定 (変分ベイズ学習) を検討する. 実際のテキストデータに本手法を適用し, テキストのように本質的にスパースなデータ系列の分割問題に対するベイズ推定の有効性を示す.

2. HMM の学習とテキスト分割

基本概念: テキストの生成モデルとして, 図 2 に示すような N 状態 left-to-right 型離散 HMM を仮定する. モデルは状態 $I=0$ から動作を開始し, 遷移確率 a_i に従って状態遷移をくり返し, 最終的に状態 $F=N+1$ に達した時点で動作を終了する. また, 状態遷移のたびごとに, その時点での状態 $i \in \{1,\cdots,N\}$ に対応する確率分布 (多項分布 b_{ij}; $j=1,\cdots,L$) に従って単語を 1 つ出力する. 動作終了時には, N 個の話題が連なる, 長さ T の単語列 $O=(o_1,\cdots,o_t,\cdots,o_T)$ としてテキストが生成されている. ここに $N \leqq T$.

モデルからテキストを生成するのとは逆に, テキスト単語列 $O=(o_1,\cdots,o_T)$ を所与のデータとして, データへのモデルの当てはめ, すなわち学習を行った場合, 学習結果にはテキストの意味的構造が含まれる. 入力テキストを用いた HMM の学習と, 学習結果からの意味的構造の抽出により, テキスト分割が実現できる.

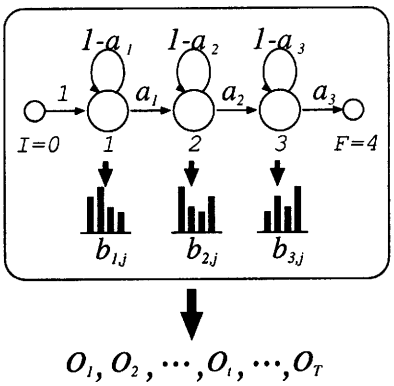

図 2 テキスト生成の確率モデル (話題数 3 の場合)

最尤推定に基づくテキスト分割: 入力テキスト O を学習データとして, HMM $\theta=\{a_i, b_{ij} \mid i=1,\cdots,N,\quad j=1,\cdots,L\}$ の最尤推定を考える. HMM の最尤推定は, よく知られているように, EM アルゴリズムの一種である Baum-Welch アルゴリズムによって実現される. Baum-Welch アルゴリズムでは, パラメータ再推定の中間変数として, 以下の前向き変数 $\alpha_t(i)$ および後向き変数 $\beta_t(i)$ が使用される.

$$\begin{cases} \alpha_t(i) = P(z_{ti}=1, o_1,\cdots,o_t \mid \theta) \\ \beta_t(i) = P(o_{t+1},\cdots,o_T \mid z_{ti}=1, \theta) \end{cases} . \tag{1}$$

ここに z_{ti} は, 時刻 t における状態が i であれば 1, そうでなければ 0 をとる 2 値変量とする.

式 (1) より, 時刻 t における状態が i である確率は,

$$P(z_{ti}=1 \mid O, \theta) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \tag{2}$$

と計算することができる. 式 (2) は, テキスト O 中の各々の単語 o_t を N 個のグループのうちのどれか 1 つに確率的に分類する, 一種のセグメンテーションを表している.

テキスト O を用いた N 状態 left-to-right 型 HMM の学習は, (単語の出現頻度分布の観点で) 均質な N 個の区間をテキスト O から見出すことを意味する. よって, 学習収束後に式 (2) を計算すれば, 意味的なまとまりとしてのテキスト分割結果を得ることができる.

ベイズ推定に基づくテキスト分割: パラメータ推定法として, 最尤推定の代わりにベイズ推定を用いた場合は, ベイズ的テキスト分割が実現される. 最尤推定と比べたベイズ推定の主な違いは, モデルパラメータ θ の値を推定 (点推定) するのではなく, パラメータの分布を推定する点である. この分布が学習データ

の多寡に依存するため，学習データ量を反映した推定が行われる．データ規模に比して語彙数の多いテキスト文書のようにスパースなデータ系列に対しては，ベイズ推定が有効と考えられる．

HMM のように隠れ変数を伴うモデルのベイズ推定については，近年，変分ベイズ学習 (VB) と呼ばれる，EM アルゴリズムのベイズ版と呼ぶべき方法論が確立され[1]，多くのモデルでベイズ推定が可能となっている．本稿でも，left-to-right 型離散 HMM のベイズ推定に VB 学習を用いる．

VB 学習の概念や導出の詳細については，例えば解説記事[7]等を参照していただくこととし，以下では left-to-right 型離散 HMM における定式化の結果を示すに止める．

（1） パラメータ θ の事前分布として以下の形を仮定する．

$$p(\theta) = \prod_{i=1}^{N} \mathcal{B}(a_i \mid \kappa_{0,i}, \kappa_{1,i}) \times \prod_{i=1}^{N} \mathcal{D}(b_{i,1}, \cdots, b_{i,L} \mid \lambda_{i,1}, \cdots, \lambda_{i,L}).$$

ここに $\mathcal{B}(x \mid a, b) \propto x^{a-1}(1-x)^{b-1}$ はベータ分布，$\mathcal{D}(x_1, \cdots, x_n \mid \phi_1, \cdots, \phi_n) \propto x_1^{\phi_1 - 1} \cdots x_n^{\phi_n - 1}$ は (n 項)Dirichlet 分布である．これらの分布仮定は，二項分布および多項分布の共役事前分布がそれぞれベータ分布および Dirichlet 分布であることによる．共役性を仮定することにより，パラメータの事後分布 $p(\theta \mid O)$ も上の形で表せる点に注意．なお，パラメータの分布を規定するパラメータ $\kappa_{0,i}$, $\kappa_{1,i}$, λ_{ij} を超パラメータと呼ぶ．本稿では，事前分布を規定する超パラメータを $\kappa_{0,i}$, $\kappa_{1,i}$, λ_{ij} で表し，事後分布を規定する超パラメータには，右肩添字 (l) を付けて，$\kappa_{0,i}^{(l)}$, $\kappa_{1,i}^{(l)}$, $\lambda_{ij}^{(l)}$ と表すことにする (l は VB 学習の反復回数)．

（2） (ベイズ的) 前向き変数および後向き変数は以下の漸化式で算出される．

$$\alpha_1(i) = \begin{cases} \exp(B_{i,o_1}) & i = 1 \\ 0 & \text{otherwize} \end{cases},$$

$$\alpha_{t+1}(i) = \alpha_t(i-1)\exp\left(A_{i,i-1} + B_{i,o_{t+1}}\right) + \alpha_t(i)\exp\left(A_{0,i} + B_{i,o_{t+1}}\right),$$

$$\beta_T(i) = \begin{cases} \exp(A_{1,N}) & i = N \\ 0 & \text{otherwize} \end{cases},$$

$$\beta_{t-1}(i) = \beta_t(i)\exp(A_{0,i} + B_{i,o_t}) + \beta_t(i+1)\exp(A_{1,i} + B_{i+1,o_t}).$$

ただし，

$$A_{0,i} = \Psi\left(\kappa_{0,i}^{(l)}\right) - \Psi\left(\kappa_{0,i}^{(l)} + \kappa_{1,i}^{(l)}\right),$$

$$A_{1,i} = \Psi\left(\kappa_{1,i}^{(l)}\right) - \Psi\left(\kappa_{0,i}^{(l)} + \kappa_{1,i}^{(l)}\right),$$

$$B_{ik} = \Psi\left(\lambda_{ik}^{(l)}\right) - \Psi\left(\sum_{j=1}^{M} \lambda_{ij}^{(l)}\right).$$

また，$\Psi(x)$ は digamma 関数で，$\Psi(x) = \Gamma'(x)/\Gamma(x) = (\log\Gamma(x))'$．なお，$\alpha_{t,0} = \beta_{t,N+1} = 0$ $(t = 1, \cdots, T)$ とする．

（3） 隠れ変数の期待値を次式によって計算する．

$$\overline{z_{t,i}} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)},$$

$$\overline{z_{t,i}z_{t+1,i}} = \frac{\alpha_t(i)\exp\left(A_{0,i} + B_{i,o_{t+1}}\right)\beta_{t+1}(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)},$$

$$\overline{z_{t,i}z_{t+1,i+1}} = \frac{\alpha_t(i)\exp\left(A_{1,i} + B_{i+1,o_{t+1}}\right)\beta_{t+1}(i+1)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}.$$

（4） 超パラメータを次式によって更新する．

$$\kappa_{0,i}^{(l+1)} = \kappa_{0,i} + \sum_{t=1}^{T-1} \overline{z_{t,i}z_{t+1,i}},$$

$$\kappa_{1,i}^{(l+1)} = \kappa_{1,i} + \sum_{t=1}^{T-1} \overline{z_{t,i}z_{t+1,i+1}} + \delta_{N,i},$$

$$\lambda_{ik}^{(l+1)} = \lambda_{ik} + \sum_{t=1}^{T} \delta_{k,o_t}\overline{z_{t,i}}.$$

ここに δ_{ij} はクロネッカのデルタ．

VB 学習においては，事前分布の超パラメータ $\kappa_{0,i}$, $\kappa_{1,i}$, λ_{ik} に適当な値を設定した上で，(2)〜(4) を反復し，事後分布の超パラメータ $\kappa_{0,i}^{(l)}$, $\kappa_{1,i}^{(l)}$, $\lambda_{ik}^{(l)}$ を (点) 推定することで，パラメータの事後分布を得る．

ベイズ推定におけるテキスト分割は，最尤推定の場合に式 (2) で示したのと同様の手続きで，前向き変数 $\alpha_t(i)$ と後向き変数 $\beta_t(i)$ の組合せで得られる．ただし，ベイズ推定で計算される前向き変数および後向き変数は，パラメータ θ について期待値をとり周辺化した量である点に注意．さらにいえば，ベイズでは式 (2) の左辺にもパラメータ θ に関する期待値操作が入るので，左辺は θ を消去した形になる．

<u>話題数の推定</u>: テキスト分割においては，テキストにいくつの話題が含まれているかは一般に未知である．この場合，何通りかの話題数 (状態数) 仮説にわたってそれぞれ HMM を学習し，学習結果から 1 つの話題数を選択することになる．これは，パタン認識ではモデル選択問題として古くから研究されているテーマである．

最尤推定においては，赤池情報基準 (AIC) や記述長最小 (MDL) 基準といったモデル選択基準がある．一方，ベイズ推定においては，VB 学習の枠組内でモデルの事後分布が計算できる．よって，事後確率最大となるモデルを選ぶことで，話題数を決定することができる．

HMM のように隠れ変数を伴うモデルは一般に非正則モデルであり，AIC や MDL が理論的基盤とする推定量の漸近正規性が成り立たない．したがって，AIC や MDL によるモデル選択では理論的基盤が危うい．これに対して，VB によるベイズ的モデル選択は，推定量の漸近正規性を必要としないという点で，HMM のモデル選択に適しているといえる．ただし，VB 学習はパラメータ事前分布の共役性という別の仮定を必要とする点に注意．

3. 評価実験

前節で提案したテキスト分割法の有効性を実験的に確かめる．実験に使用するテキストデータは，日本語ニュース番組 (15 分 ×5 回分) の書き起こしまたは音声認識結果とする．この場合，番組で伝えられる個々のニュース項目が話題に相当するので，テキストをニュース項目単位に分割することがテキスト分割の目的となる．データに含まれる話題総数は 66，また 1 話題あたりの単語数は最小で 16，最大で 699，平均では約 180(読上げ時間にして 74 秒程度) である．なお，1 回の放送で使われる語彙数は 718〜778．

実験に使用する書き起こしテキストについては，あらかじめ形態素解析器にかけて語切りをし，その表記文字列のみを用いる．簡単化のため，活用語の活用形正規化や品詞情報の利用などは考えない．また，助詞などの不要語 (stop words) をテキストからあらかじめ除外しておくことは一般に有効とされているが，不要語の選び方による恣意性を排除するため，不要語の除去はあえて行っていない．さらに，テキスト分割の標準的な評価では，文境界を既知として，文境界の中から話題境界を選ぶというタスクが一般的であるが，ここでの実験では，すべての単語境界から話題境界を選ぶという，幾分困難なタスクでの評価を行う．これは，テレビ番組のように話し言葉を含むテキストでは，文境界が必ずしも明確ではないからである．

認識結果テキストを評価に用いる場合も，形態素解析が不要な点を除いて条件は書き起こしテキストと同様．なお，認識結果テキストの単語正解精度は 80.5〜88.4%の範囲にあり，平均して 84.5%であった．

テキスト分割の精度を測る評価尺度としては，co-ocurrence agreement probability(CoAP) を用いる [2]．CoAP は自由度の高い尺度であるが，よく行われているように，一定単語数 (= 平均話題長 /2) だけ離れた 2 単語が同一の話題に属するか否かを正しく判定できた割合として算出する．

ここでは，話題数既知の評価実験を行う．つまり，例えば 5 つのニュース項目が続く区間 (話題数 5) を評価データからすべて取り出し，5 状態の HMM を用いて分割し，分割精度を算出する．このような実験を，話題数 5〜10 にわたって行い，それぞれの分割精度を調べる．

テキスト分割精度 (CoAP) を図 3 に示す．最尤推定に基づく分割 (ML) とベイズ推定に基づく分割 (Bayes) とを比べた場合，書き起こしでも認識結果でも，明らかにベイズ推定が優位である．学習で得られた単一のパラメータ値 θ に基づいてテキスト分割を計算する最尤推定に対して，ベイズ推定は θ のあり得る範囲にわたる積分という形で分割を算出する．この点で，ベイズ推定は出現頻度の低い単語に過度に影響されることなく，安定したテキスト分割が行えているものと推測される．

一方，話題数と分割精度の関係をみると，話題数の増加に伴い，分割精度が徐々に低下していることがわかる．個々のデータの分割結果をみると，真の話題境界の配置に比べて，出力された話題境界の配置に偏りがある，局所最適とみられるケースが，話題数の多い長いテキストにおいて散見された．これはある程度予想していた現象であり，今後改善策を検討したい．

書き起こしと認識結果との精度比較に関しては，やはり書き起こしと比べて，誤りを含んだ認識結果で精度の低下がみられる．ただ，後述する従来法 (Hearst 法) の結果と比べると，低下の幅は非常に小さいといえる．

なお，事前分布 ($\kappa_{0,i}$, $\kappa_{1,i}$, λ_{ik}) の設定については，予備実験で何種類かを検討したが，概ね，$1 < \kappa_{0,i}, \kappa_{1,i}, \lambda_{ik} \leq 1.1$ の範囲内で分割精度は安定しており，CoAP にして 0.01 程度の変動がみられるのみであった．

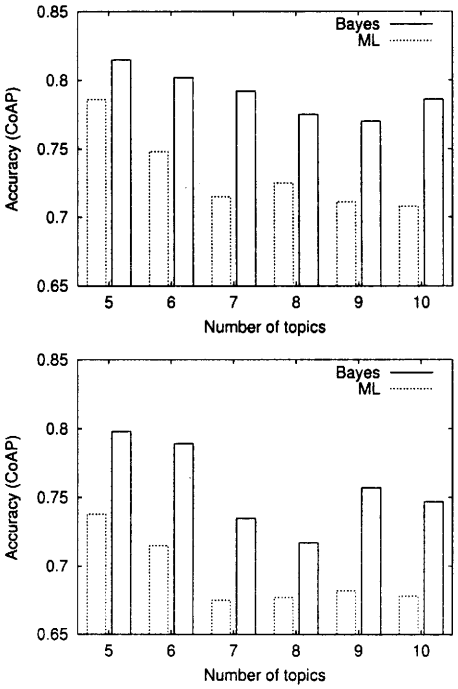

図 3　話題数既知 (5〜10) での提案法の分割精度 (CoAP): 書き起こし (上) および認識結果 (下) に対する分割精度．

話題数 5 の書き起こしテキストの中から，結果の一例を図 4 に示す．この図では，第 t 番目の単語 o_t が第何番目の状態に属するかを期待値

$$\begin{cases} \text{ML:} & \displaystyle\sum_{k=1}^{N} kP\left(z_{tk}=1 \mid O,\theta\right) \\ \text{Bayes:} & \displaystyle\sum_{k=1}^{N} kP\left(z_{tk}=1 \mid O\right) \end{cases}$$

で表し (太線), 正解 (細線) とともに示してある. ベイズ推定による分割結果 (Bayes) の方が, 話題の遷移を正確に抽出できていることがわかる.

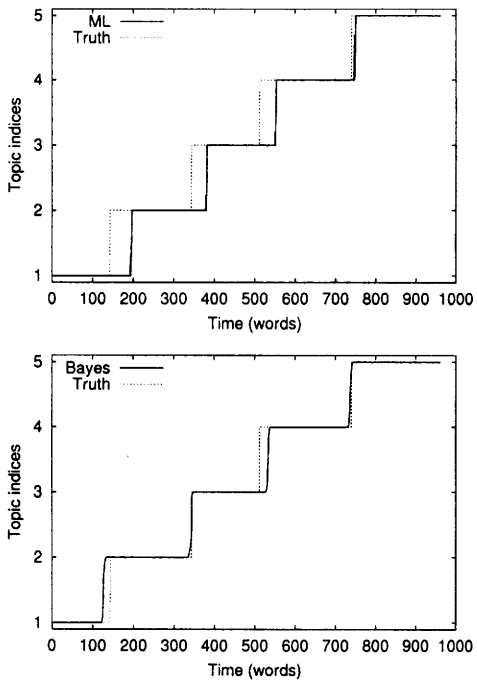

図 4 テキスト分割例: 最尤推定 (ML, 上) とベイズ推定 (Bayes, 下).

ここで参考までに, Hearst 法において同じ評価データを分割した場合の分割精度を示す. Hearst 法は, オンラインアルゴリズムであるという性格上, テキストの話題数を固定した評価が難しい. したがって, 話題数を既知として行った先の実験結果とは単純に比べることはできない. しかしながら, 定性的な違いについて論ずることは可能と考え, 以下に示す.

実験データおよび条件は先の実験と同一とした. ただし, 先の実験は話題数既知であり, 出力される話題数が常に真の話題数に等しくなるのに対して, 本実験では, 出力される話題数を外から指定することが容易でないため, 話題数未知での実験としている点に注意が必要である.

Hearst 法の動作を規定するパラメータとしては, 窓幅と平滑化回数 (局所平均化フィルタの適用回数) のみを考える. 局所平均化フィルタのサイズは 3 単語に固定した (フィルタサイズの変更は平滑化回数の変更と効果が類似するため). 話題境界を認定するための極小点の検出では, 文献[3] にならい, 極小点の左右の極大点との高低差の和を評価基準として, 評価値の大きい極小点から順に話題境界として認定する方法を採った. ただし, 一定値以下の長さの話題が生ずる場合には, その極小点は棄却する. その際の話題長のしきい値は, 窓幅と同一とした. 話題境界として認定可能な極小点はすべて認定する.

パラメータの設定値は, 窓幅が 50, 75, 100, 125, 150 の 5 通り, 平滑化回数が 64, 128, 256, 512, 1024, 2048 の 6 通りとし, 両者の組合せで 30(5×6) 通りの実験を行った. 結果, 書き起こしテキスト (図中の "Manual") においては窓幅 125, 平滑化回数 1024 の組合せが, 認識結果テキスト (図中の "Auto.") においては窓幅 125, 平滑化回数 512 の組合せが, それぞれもっともよい分割精度を示した. これらの最適点から, パラメータの 1 つである窓幅を変更した場合の分割精度 (話題数 5〜10 のすべてのテキストの平均値) の変化を示したのが図 5 である. なお, 分割精度が最大となるように平滑化回数の値を選んでいるので, 一部クローズドな条件での評価結果となっている点に注意.

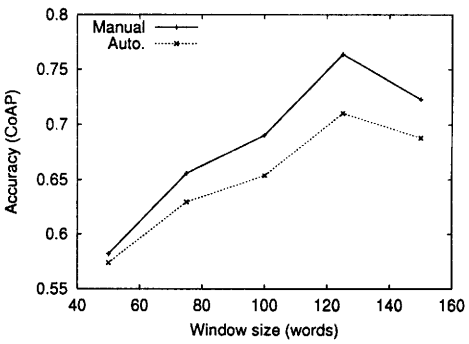

図 5 Hearst 法の分割精度 (CoAP): 窓幅の変動と精度の関係. 書き起こし (Manual) および認識結果 (Auto.).

図 5 から, 分割精度がパラメータ (窓幅) に対して敏感であることがわかる. 今回の実験で使用したデータの話題長は, この図の横軸のレンジよりもかなり広い範囲にわたって分布する. したがって, 窓幅というパラメータを設定することは, 精度劣化の危険を伴うといえる. この点で, 調整が必要なパラメータを持たない提案法には利点がある.

Hearst 法に関するもう 1 つの実験として, 評価データを 2 分割し, 一方でパラメータ (窓幅および平滑化回数) を最適化し, 他方で分割精度の評価を行うオープン評価を行った. より具体的には, 評価データに含まれる 5 回分のニュース番組のうちの 4 回分でパラメータの最適化を行い, 残りの 1 回で分割精度評価を行う実験を 5 回くり返す, leave-one-out 評価を行った.

結果を図 6 に示す. 書き起こしテキスト (図中の "Manual") では, 最適なパラメータの組合せは常に窓幅 125, 平滑化回数 1024 となり, データが変動してもパラメータの最適値が変わらないという好ましい傾向を示した. 結果として分割精度も比較的高い値となっている. しかしながら, 認識結果テキスト (図中の "Manual") における評価では, 最適な窓幅と平滑化回数の組合せは, 2048 と 100, 512 と 125, 256 と 125 という 3 通りに分かれ, 結果として著しい分割精度の劣化を招いている. 認識結果のような誤り含みのデータに対して, 提案法はそれほどの精度劣化を見せていないこと (図 3) から, 提案法がノイズに対して頑健に動作し得ることがわかる.

ただし, Hearst 法は話題数の多い長いテキストにおいても分割精度が低下するような傾向はなく, この点は顕著な利点といえる. 提案法の評価結果には, 長いテキストにおいて分割精度

が低下する傾向が現れているが, 原理的には提案法にテキストの長さに依存するような性質はないはずであり, 局所最適問題が精度低下の主たる要因と考えている. この問題は, 提案法の改善のポイントとして今後検討したい.

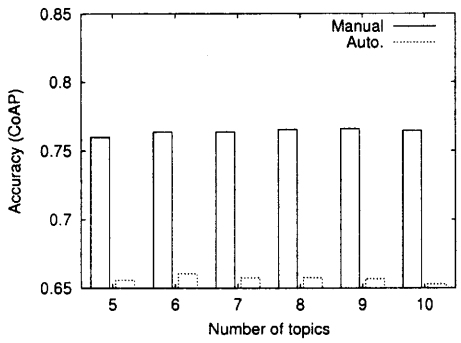

図 6 Hearst 法の分割精度 (CoAP): 書き起こし (Manual) および認識結果 (Auto.) に対する分割精度.

4. 関連研究

Hearst 法は, 同一単語のくり返しという, 語彙的結束性のもっとも素朴な形態を話題表現に用いたものである. Hearst 法以降, より高度な形態で語彙的結束性を考慮する研究が多くなされている.

Stokes ら [5] は, 語彙的結束性の形態として, (a) 同一単語のくり返し, (b) 類義語のくり返し: "警察" と "警官," (c) 上位/下位概念の関係にある単語: "凶器" と "ナイフ," (d) 全体と部分の関係にある単語: "車" と "エンジン," (e) 統計的に共起しやすい単語: "Osama bin Laden" と "the World Trade Centre," を挙げ, これらの単語間の関係を利用することの重要性を述べている.

例えば別所らの方法 [6] は, Hearst 法の変化点検出の考え方に, 上記 (e) を取り入れたものといえる. すなわち, 大規模コーパスから単語の共起頻度を抽出することにより, 単語の概念的な類似性を実ベクトルで表現する. 入力テキストの話題境界は, 実ベクトル列の変化点として検出することができる.

これらの語彙的結束性に関する研究は, パタン認識に例えるなら, 特徴量設計の研究に相当すると考えられる. つまり, 話題の表現にはいかなる特徴量が適しているかという問題に対して解答を与えようというのが, 上述の研究である.

さて, 周知の通りパタン認識においては, 特徴量の研究にも増して, 識別モデルの研究が盛んに行われている. そこで, モデルという観点でテキスト分割の研究を俯瞰すると, Hearst 法をはじめとして, ほとんどが, 局所的領域ごとに特徴量の変動を捕らえる, ボトムアップ的かつ逐次的なパタンマッチング・検出のモデルに留まっているといえる.

本稿は, テキスト分割におけるモデル面からの提案であり, 今後のこの分野でのモデル研究の重要性を主張するものである. 本稿で提案したモデルは, 特徴量に関する制約はなく, 例えば概念ベクトル [6] などにも特に障害なく適用することができる. また, 確率モデルの手法に則しており, 単語の共起関係のような統計的情報を活用するなど, 種々の拡張も容易であると考えている.

5. おわりに

left-to-right 型の離散 HMM をテキスト生成の確率モデルと考え, テキスト分割を HMM のパラメータ推定問題として定式化した, バッチ処理タイプのテキスト分割手法を提案した. 2 種のパラメータ推定法, すなわち最尤推定とベイズ推定 (変分ベイズ学習) を用いて, ニュース番組を各ニュース項目へ分割する評価実験を行い, 最尤推定に比べてベイズ推定が精度よくテキストを分割できることを示した. ベイズ推定は, 未知データに対する汎化性を改善する方策という見方が一般的であるが, 今回の実験結果は, ベイズ推定がテキスト分割のようなデータマイニングの問題に対しても有効であることを示している.

今後の課題としては, まず第一に, 話題数未知の条件下での提案法の評価が挙げられる. 話題数を自動的に決定できる枠組が, テキスト分割の実際応用上不可欠であることは言を待たない. 特にベイズ的モデル選択が, スパース性の高い実データに対してどの程度有効に働くかを, 今後確認したい.

他の課題としては, 学習の局所最適回避策の検討, 事前に集められた大規模コーパスの効果的利用法の検討などを考えている. また, tied-mixture タイプの HMM を用いた場合の定式化と評価実験を行いたい. 状態数よりも少数の多項分布を全状態が共有する tied-mixture 構造の HMM を導入することは, 確率的 LSA [4] の動的モデルへの拡張として興味深いだけでなく, テキストに内在するデータスパースネスの問題をさらに軽減する効果が期待できると考えている.

文 献

[1] Attias, "Inferring parameters and structure of latent variable models by variational Bayes," Proc. 15th Conf. on Uncertainty in Artificial Intelligence, 1999,

[2] Beeferman, Berger & John Lafferty, "Statistical models for text segmentation," Machine Learning, Vol.34, No.1-3, pp.177–210, 1999,

[3] Hearst, "Multi-Paragraph Segmentation of Expository Text," 32nd. Annual Meeting of the Association for Computational Linguistics, 1994,

[4] Hofmann, "Probabilistic Latent Semantic Indexing," Proc. 22nd Int'l Conf. on R&D in Information Retrieval (SIGIR'99), 1999,

[5] Stokes, Carthy & Smeaton, "Segmenting broadcast news streams using lexical chains," STarting AI Researchers Symposium (STAIRS2002), 2002,

[6] 別所, 大附, 松永, 林, "概念ベクトルを用いたテキストセグメンテーションのニュース音声への適用," FIT, 2002,

[7] 上田, "ベイズ学習 [III] —変分ベイズ学習の基礎—," 信学誌, Vol.85, No.7, pp.504–509, 2002.