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

別所克人†

テキストをトピック単位に分割するテキストセグメンテーションは，テキストを構造化するための重要な要素技術の 1 つである．本論文では，テキストを単語の意味表現の 1 つである概念ベクトルの系列に変換し，ベクトルの系列を分割するクラスタ列で，クラスタ内のベクトルの変動量の総和であるクラスタ内変動が最小となるものをトピック区間列とする手法を提案する．提案手法の特徴は，テキストの 1 区間の意味的なまとまりの度合いを該区間内のベクトルの変動量により判断する点と，テキストの局所的な範囲内の情報のみでなく，テキスト全体のベクトルの分布情報に基づき，セグメンテーションを行う点にある．新聞記事を用いた評価実験の結果，局所的な範囲内でトピック境界を判断する従来手法よりも高精度であることを確認した．

Text Segmentation Based on Minimum Within-cluster Variation Criterion

Katsuji Bessho†

A new text segmentation method is proposed. In the proposed method, text is converted into a concept vector sequence, which corresponds to semantic word representations. Then, the concept vector sequence is segmented into topic clusters to minimize within-cluster variation. The characteristics of the proposed method are to assess the degree of semantic cohesiveness using vector variation within each segment, and to segment text based on both local text information and vector distribution within the entire text. The results of experiments done on newspaper articles have shown that the proposed method yields a much higher segmentation accuracy than the conventional method, which locally determines topic boundaries.

1. はじめに

テキストは一般に複数のトピック区間から構成されている場合が多い．テキストをトピック単位に分割することによって，テキスト処理に関する様々な効用が得られる．テキストセグメンテーションにより，照応や省略を解決する手がかりが得られ，また個々のトピック区間からキーワードや要約を抽出することができ，トピック区間どうしの関係を解析しテキスト全体の談話構造を抽出することが可能となる．この結果，人によるテキスト全体の内容の把握が容易となる．また，複数の文書に対する検索においても，文書単位でなくトピック単位で行うことにより精度向上が期待できる[1]．

テキストセグメンテーションの手法として，語彙的結束性と呼ばれるテキスト上での同一語彙や関連語彙の出現情報を利用するものがある．関連語彙の例としては，類義性のある語彙や共起性のある語彙があげられる．

同一語彙のみの出現情報を利用するものとして，Hearst による手法（本論文では Hearst 法と呼ぶことにする）[2] がある．この手法では，テキスト中の単語列における一定間隔の単語境界（基準点と呼ぶ）の前後に一定の単語数の窓を設定し，各窓ごとに，単語の窓内の出現頻度ベクトルをとり，そのベクトル間の余弦測度を当該基準点の結束度とする．次に，結束度の微弱な振動を除去するため，基準点の結束度を，当該基準点とその前後一定数の基準点の結束度の平均に変換する（結束度の平滑化）．トピック境界では，この結束度が極小となっていると期待されるので，結束度が極小となる基準点（極小点と呼ぶ）で，結束度の谷の深さ（depth score と呼ぶ）の大きいものからトピック境界として出力する．仲尾は，上記の Hearst 法をベースにして，テキストの話題の階層的な構成を自動認定する手法を提案している[3]．

類義性のある語彙の出現情報を利用するものとして，Morris らは，シソーラス上の同一クラスに属する語の

† 日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

連鎖からトピック境界を求める手法を提案している[4]．望月らは，Morris らがあげたような語彙的連鎖の情報に加え，複数の表層的手がかりを組み合わせて，トピック境界を検出することを行っている[5]．

共起性のある語彙の出現情報を利用するものとして，Kozima らは，テキストの各位置の近傍の単語列の結束度を，英語辞書から規則的に構成された意味ネットワーク上の活性伝播によって計算し，この結束度が極小となる位置をトピック境界とする手法を提案している[6]．Brants らは，訓練テキストから推定された PLSA モデルのパラメータによって，対象テキスト中の各ブロックごとの単語確率分布を求め，隣接ブロック間の類似度が極小となる位置をトピック境界とする手法を提案している[7]．

共起性のある語彙の出現情報を利用する手法として，筆者はすでに，テキストを単語の意味表現の 1 つである概念ベクトルの系列に変換することによって，テキストセグメンテーションを行うことを提案した[8]．概念ベクトルの系列は単語の意味の変遷を表していると考えられるので，この系列の変化を利用してテキストの分割が行えることが期待できる．文献 8) においては，Hearst 法における窓に対応するベクトルとして，窓に含まれる単語の概念ベクトルの重心をとる手法が提案されているが，これを本論文では概念ベクトル結束度法と呼ぶ．概念ベクトルを用いることにより，異なる単語でも意味的に近ければ，類似度を高くすることができるため，Hearst 法よりも一般に高精度となる．

しかしながら，Hearst 法や概念ベクトル結束度法では，一定サイズの窓をスライドさせていくので，窓幅よりも短いトピック区間が窓に含まれているとき，対応するベクトルは当該トピック区間の意味を適切に表していないという問題がある．このため，窓幅よりも短いようなトピック区間の検出が困難である．また，テキストの局所的な範囲においてトピック境界を判断しているため，きわめて長いトピック区間が細断され，このような長いトピック区間の検出が困難である傾向がある．

これらの課題を解決するため，本論文では，テキストを概念ベクトルの系列に変換したうえで，ベクトルの系列を分割するクラスタ列で，クラスタ内のベクトルの変動量の総和であるクラスタ内変動が最小となるものをトピック区間列とする手法を提案する．テキストの局所的な範囲内の情報のみでなく，テキスト全体のベクトルの分布情報に基づき，あらゆる分割のクラスタ集合としての妥当性を検証し，クラスタ集合として最適なものを選択するので，より高精度なトピック区間の検出が行えると考えられる．

なお，クラスタリングをベースとするテキストセグメンテーションの手法としては，文献 9)，10) で提案されている手法等がある．これらの手法では，テキスト中の各パラグラフを，単語を座標軸とするベクトルで表現した後，ベクトルの集合をクラスタリングする．クラスタリングの結果，連続する同一クラスタのパラグラフの列がトピック区間となる．また，テキストセグメンテーションではないが，要約の研究として，文献 11)，12) では，文やパラグラフを，単語を座標軸とするベクトルで表現し，そのベクトルの集合を，k-means 法の一種によってクラスタリングし，テキスト内のトピック数の推定を行っている．

これら既存手法と比較した場合，本論文で提案する手法の特徴は，(1) 単語そのものをベクトルで表現する，(2) テキストの 1 区間の意味的なまとまりの度合いを該区間内の単語のベクトルの変動量で定義する，(3) ベクトルの変動量の総和であるクラスタ内変動が最適となる分割を求める，という点にある．(3) に関しては，文単位でのセグメンテーションを行う際，文の一次元配列という制約下での文集合のクラスタリングとなる．上記既存手法のような一次元配列の制約なしのクラスタリングでは，異なるトピック区間に属する似たような意味を持つ文が同一クラスタに属したり，逆に同一トピック区間に属する意味的に遠い文が別クラスタに属したりするという問題があると考えられる．本論文では，(3) の最適解を求めるための効率的なアルゴリズムも提案する．

以下，2 章で概念ベクトルについて説明する．次に，3 章で提案するアルゴリズムについて説明する．4 章で評価結果について述べ，5 章でまとめを述べる．

2. 概念ベクトル生成

本章では，文献 13) で提案されている潜在的意味解析に基づいた概念ベクトルの生成法について述べる．概念ベクトルの生成においては，まずコーパスを形態素解析した後，内容語以外を除去する．残った異なり単語の集合から，高頻度語の部分集合を 2 つとる．一方の部分集合中の単語を概念語と呼び，もう一方の部分集合中の単語を共起語と呼ぶ．任意の概念語と任意の共起語との間の 1 文中に共起する頻度をカウントし，各行が概念語に対応し，各列が共起語に対応しているような共起行列 X を作成する（**表 1** 参照）．共起行列の各行ベクトルは，対応する概念語の共起パターンを表しており，この行ベクトルを共起ベクトルと呼

表 1 共起行列の例
Table 1 An example of a co-occurrence matrix.

		共起語				
		...	貿易	...	歌手	...
	...	...	...	...	...	...
概念語	関税	...	301	...	2	...
	...	...	...	...	...	...
	オペラ	...	4	...	73	...
	...	...	...	...	...	...

ぶ．ある 2 単語に対応する共起ベクトルが近ければ，共起パターンが似ているので，この 2 単語は意味的に近いということが推測される．ただし，このままではデータのスパースネス性があることをはじめとして，テキストデータから抽出される単語の情報にはつねに欠落があると予想されるため，ベクトル間の類似度の精度は低いと考えられる．また，一般に共起ベクトルの次元数は非常に大きなものとなるため，計算量も無視できないものとなる．このため共起行列を特異値分解により，次元数を縮退させた行列に変換する．ここで特異値分解にかける前に，精度向上のため，共起行列 X の各要素値として，共起頻度の平方根をとる．X を $p \times q$ の行列としたとき，特異値分解により X は，以下のように分解できる．

$$\underset{p\times q}{X} = \underset{p\times r}{U}\ \underset{r\times r}{\Sigma}\ \underset{r\times q}{V^T}$$

ここで，添字 T は行列の転置を表す．$r = \mathrm{rank}\, X \leq \min(p, q)$，$U^T U = V^T V = I$（$I$：単位行列）であり，$\Sigma = (\delta_{ij})$ としたとき，$\delta_{ii} \geq \delta_{jj} > 0 \left(1 \leq {}^{\forall}i \leq {}^{\forall}j \leq r\right)$，$\delta_{ij} = 0\,(i \neq j)$ である．$\delta_{ii}\,(1 \leq i \leq r)$ を X の特異値と呼ぶ．V^T の r 個の行ベクトルは，q 次元空間中の正規直交基底であり，X の第 i 番目の q 次元行ベクトルは，この正規直交基底の張る r 次元部分空間において，$U\Sigma$ の第 i 番目の r 次元行ベクトルで表される．ここで，$1 \leq {}^{\forall}r' \leq r$ に対し，U の最初の r' 列，V^T の最初の r' 行，Σ の最初の r' 行，r' 列をとり，

$$\underset{p\times q}{X'} = \underset{p\times r'}{U'}\ \underset{r'\times r'}{\Sigma'}\ \underset{r'\times q}{V'^T}$$

とする．$U'\Sigma'$ の第 i 番目の行ベクトルは，$U\Sigma$ の第 i 番目の行ベクトルの 1 番目から r' 番目までの座標をとったものであり，$U\Sigma$ の第 i 番目の行ベクトルを，V'^T の行ベクトルの張る r' 次元部分空間に射影して得られるものである．X の第 i 番目の q 次元行ベクトルは，$U'\Sigma'$ の第 i 番目の r' 次元行ベクトルに射影される．V'^T の行ベクトルが張る r' 次元部分空間は，X の各行ベクトルとその射影した点との距離の自乗和が最小となる r' 次元部分空間であり，その意味で X の行ベクトルの分布に最もあてはまりのよい r' 次元部分空間である．この V'^T の行ベクトルが張る r' 次元部分空間を概念空間と呼ぶ．$U'\Sigma'$ の各行ベクトルは，U' の対応する行ベクトルを，各座標ごとに対応する特異値の割合で伸縮したものである．U' の行ベクトルをその長さで割って単位ベクトルに正規化したものを，対応する単語の概念ベクトルと呼ぶ．

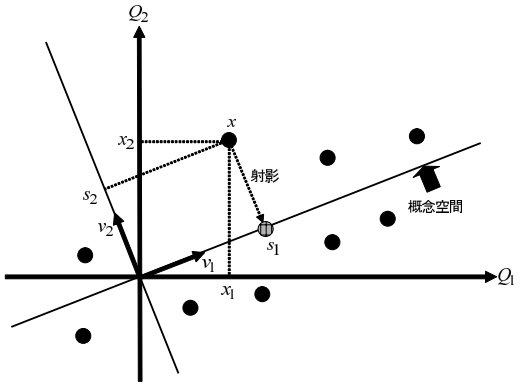

図 1 概念空間への射影
Fig. 1 Projection onto the concept space.

図 1 では，$q = 2$ であり，X の各行ベクトルは $\mathcal{Q}_1$，$\mathcal{Q}_2$ を座標軸とする空間上の点として表されている．r の値は 2 であり，V^T の 2 個の行ベクトル v_1，v_2 が正規直交基底となっている．X のある行ベクトル $x = (x_1, x_2)$ は，この正規直交基底の張る空間において，(s_1, s_2) と表され，これが $U\Sigma$ の対応する行ベクトルである．$r' = 1$ としたとき，v_1 の張る空間が概念空間であり，x は $U'\Sigma'$ の対応する 1 次元行ベクトル (s_1) に射影される．

3. 提案手法

提案手法においては，セグメント対象テキストを形態素解析して単語に分割し，得られた単語のうち，内容語のみに対応する概念ベクトルを付与する．以後，ベクトルを付与された内容語の系列を処理対象とする．

本論文では，トピック間の境界は文間の境界と仮定する．ただし，単語間の境界としてもアルゴリズム上，本質的な違いはない．

テキストを分割するトピック区間列（各区間は文集合）において，同一区間に含まれる概念ベクトルは，比較的類似性が高い（幾何学的距離が近い）と予想される．異なるトピックを表す区間のそれぞれで，内部に含まれる概念ベクトル集合をとったとき，2 つの集合はクラスタ集合としてよく分離されていると推測される．このことから，テキストを分割する区間列で，

クラスタ集合として妥当である（よく分離されている）ものがトピック区間列である可能性が高いと考えられる．そこで，クラスタ集合の妥当性の指標を定義し，それに基づき提案手法のアルゴリズムを述べる．

3.1 用語の定義

クラスタ集合の妥当性の指標と，提案手法のアルゴリズムを説明するためにいくつかの用語の定義が必要となるため，本節でそれらの定義を述べる．**図 2** は，定義するオブジェクトの間の典型的な関係を示す．

(1) 単語と単語番号

ベクトルを割り当てられた単語の系列 $w_1, w_2, \ldots, w_x$ に対し，単語 w_i の添数 i を単語番号と呼ぶ．$1 \leq i < j \leq x$ に対し，$w_i = w_j$ ということはありうる．

(2) 単語の重み

各単語番号 i に対し，単語 w_i の重み a_i を定めておく．セグメント対象テキストが音声認識結果等の場合は，a_i として認識信頼度等をとることが考えられるが，本論文では，通常のテキストを対象とするため，任意の i に対し $a_i = 1$ とする．

(3) 単語のベクトル

単語 w_i のベクトルを v_i と表記する．

(4) 単語クラスタ

単語番号の集合 $\{1, 2, \ldots, x\}$ の任意の部分集合を単語クラスタと呼ぶ．

(5) 単語クラスタ W に関する単語の重み和 $vol_w(W)$

$$vol_w(W) \underset{def}{=} \begin{cases} \sum_{k \in W} a_k & W \neq \phi \text{ のとき} \\ 0 & W = \phi \text{ のとき} \end{cases}$$

(6) 単語クラスタ W に関する単語の重み付き重心ベクトル $M_w(W)$

$$M_w(W) \underset{def}{=} \begin{cases} \left(\sum_{k \in W} a_k v_k\right) \Big/ vol_w(W) & vol_w(W) \neq 0 \text{ のとき} \\ \mathbf{0} & vol_w(W) = 0 \text{ のとき} \end{cases}$$

(7) 単語クラスタ W のコスト $cost_w(W)$

$$cost_w(W) \underset{def}{=} \begin{cases} \sum_{k \in W} a_k \|v_k - M_w(W)\|^2 & vol_w(W) \neq 0 \text{ のとき} \\ 0 & vol_w(W) = 0 \text{ のとき} \end{cases}$$

(8) 単語クラスタ集合のコスト

互いに排反な単語クラスタの集合 $\{W_1, W_2, \ldots, W_p\}$ があったとき，この単語クラスタ集合のコストを，

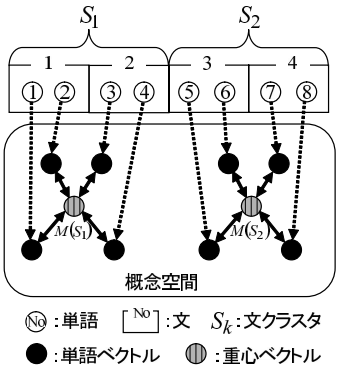

図 2 オブジェクト間の関係
Fig. 2 Relations among objects.

$$cost_w\left(\{W_1, W_2, \ldots, W_p\}\right) \underset{def}{=} \sum_{k=1}^{p} cost_w(W_k)$$

と定義する．

(9) 文と文番号

テキスト中の文の系列を $s_1, s_2, \ldots, s_y$ とする．文 s_i の添数 i を文番号と呼ぶ．

(10) 文クラスタ

文番号の集合 $\{1, 2, \ldots, y\}$ の任意の部分集合を文クラスタと呼ぶ．

(11) 文クラスタと単語クラスタ

文クラスタ S に対し，S 内の文番号の文に含まれる単語番号の全体を $W(S)$ と表記する．

(12) 文クラスタ S に関する単語の重み和 $vol(S)$

$$vol(S) \underset{def}{=} vol_w(W(S))$$

(13) 文クラスタ S に関する単語の重み付き重心ベクトル $M(S)$

$$M(S) \underset{def}{=} M_w(W(S))$$

(14) 文クラスタ S のコスト $cost(S)$

$$cost(S) \underset{def}{=} cost_w(W(S))$$

(15) 文クラスタ集合のコスト

互いに排反な文クラスタの集合 $\{S_1, S_2, \ldots, S_p\}$ があったとき，この文クラスタ集合のコストを，

$$cost\left(\{S_1, S_2, \ldots, S_p\}\right) \underset{def}{=} cost_w\left(\{W(S_1), W(S_2), \ldots, W(S_p)\}\right) = \sum_{k=1}^{p} cost_w(W(S_k)) = \sum_{k=1}^{p} cost(S_k)$$

と定義する．

(16) 連続する文番号列に関する各種記号
任意の連続する文番号の列 $S=\{i,i+1,\ldots,j\}$ に対し，以下のように定義する．
(a) 連続する文番号列に関する単語の重み和

$$vol\,(i,j) \underset{def}{=} vol\,(S)\,, vol\,(i) \underset{def}{=} vol\,(i,i)$$

(b) 連続する文番号列に関する単語の重み付き重心ベクトル

$$M\,(i,j) \underset{def}{=} M\,(S)\,, M\,(i) \underset{def}{=} M\,(i,i)$$

(c) 連続する文番号列のコスト

$$cost\,(i,j) \underset{def}{=} cost\,(S)\,, cost\,(i) \underset{def}{=} cost\,(i,i)$$

3.2 クラスタ集合の妥当性の指標

互いに排反な文クラスタの集合 $\{S_1,S_2,\ldots,S_p\}$ があったとき，$T=S_1\cup S_2\cup\ldots\cup S_p$ とおく．以下の関係式が成り立つことが計算により証明される[14]．

$$cost\,(T)=\sum_{k=1}^{p} cost\,(S_k)+ \sum_{k=1}^{p} vol\,(S_k)\,\|M\,(S_k)-M\,(T)\|^2 \qquad (1)$$

左辺を，このクラスタ集合の全変動と呼ぶ．また，右辺の第1項をクラスタ内変動と呼び，第2項をクラスタ間変動と呼ぶ．クラスタ内変動は3.1節で定義したクラスタ集合のコストである．

$\{S_1,S_2,\ldots,S_p\}$ が，テキストの分割によって得られる文クラスタの列とした場合，全変動は一定である．したがって，クラスタ内変動が小さいほど，クラスタ間変動は大きくなり，各クラスタ間はよく分離されているといえる．このように，クラスタ内変動の小ささが，クラスタ集合としての妥当性を示す指標となる．

関係式 (1) から分かることは，あるクラスタ列をさらに細分割して得られるクラスタ列のクラスタ内変動は，必ず分割前のクラスタ列のクラスタ内変動以下となるということである．クラスタ内変動は，クラスタ数が1つの場合，最も大きく，各クラスタが1文の場合，最も小さくなる．したがって，クラスタ集合として妥当かどうかは，クラスタ数を固定した場合のクラスタ列の集合の中で意味がある．トピック区間列は，その分割数をとるクラスタ列の中で，クラスタ内変動が最小になっていることが推測される．したがって，クラスタ数を固定したうえでクラスタ内変動最小基準を満たすクラスタ列が，トピック区間列の候補といえる．

3.3 分割アルゴリズム

3.2節での考察に基づき，提案手法は以下のような方式をとる．まず，任意の分割数 p に対し，テキストを p 個に分割するクラスタ列のコストの最小値を求める．次に，この最小値の分布からトピック数の推定を行う．最後に，分割数がこのトピック数であり，かつ，コスト最小のクラスタ列を求め，トピック区間列とする．提案手法のアルゴリズムは，計算量の爆発を抑えるような動的計画法である．以下，アルゴリズムを順次述べる．

1文コストの算出 各文番号 i に対し，文番号 i に関する単語の重み和 $vol\,(i)$，単語の重み付き重心ベクトル $M\,(i)$，コスト $cost\,(i)$ を計算し，保持する．

任意区間コストの算出 テキスト中の文番号の最大値を y としたとき，任意の連続する文番号列 $\{i,i+1,\ldots,j\}\,(1\le i<j\le y)$ のコストを**図3**のアルゴリズムにより算出していく．

このアルゴリズムでは，文番号列 $\{i,\ldots,j\}\,(i<j)$ に関する重み和と重み付き重心ベクトルおよびコストを，それまでに求まっている，文番号列 $\{i,\ldots,j-1\}$ と $\{j\}$ に関するそれらの値から求める．図3の4行目の重み付き重心ベクトルに関する式と，5行目のコストに関する式が成り立つことは計算により証明される[15]．

文数 y に対する時間計算量は $O\left(y^2\right)$ である．任意の区間のコストを3.1節の定義式に基づいて計算するやり方に比べ，本アルゴリズムは効率的なものとなっている．

なお空間計算量については，重み和については，

1. $for\,(i=1;i\le y-1;i++)\,\{$
2. $\quad for\,(j=i+1;j\le y;j++)\,\{$
3. $\quad\quad vol\,(i,j)=vol\,(i,j-1)+vol\,(j)\,;$
4. $\quad\quad M(i,j)=\dfrac{vol\,(i,j-1)}{vol\,(i,j-1)+vol\,(j)}M(i,j-1)+ \dfrac{vol\,(j)}{vol\,(i,j-1)+vol\,(j)}M\,(j)\,;$
5. $\quad\quad cost\,(i,j)=cost\,(i,j-1)+cost\,(j)+ \dfrac{vol\,(i,j-1)\bullet vol\,(j)}{vol\,(i,j-1)+vol\,(j)}\|M(i,j-1)-M(j)\|^2\,;$
6. $\quad\}$
7. $\}$

ただし，$vol\,(i,j)=0$ のとき，
$M\,(i,j)=\mathbf{0}$，$cost\,(i,j)=0$

図3 任意区間コスト算出アルゴリズム
Fig. 3 Algorithm that can compute the cost of any segment.

$vol(i,j-1)$，$vol(i,j)$ の 2 つの領域のみ使用し，重み付き重心ベクトルについては，$M(i,j-1)$，$M(i,j)$ の 2 つの領域のみ使用して，以降の処理で不要になる値をアルゴリズムの過程で捨て，使用領域を節約することができる．ただし，任意の $(i,j)\,(1 \le i \le j \le y)$ に対する算出した $cost(i,j)$ は，次の最小コストを求める処理に使用するため，保持する．

最小コストの算出　テキスト中の文番号の最大値を y としたとき，$1 \le q \le y$ であるような q に対し，q 番から連続する文番号の列 $\{q, q+1, \ldots, y\}$ の p 個のクラスタへの分割

$$\begin{aligned}\Big\{\ & T_1 = \{q, q+1, \ldots, n_2-1\},\\ & T_2 = \{n_2, n_2+1, \ldots, n_3-1\},\\ & \quad\ldots\\ & T_i = \{n_i, n_i+1, \ldots, n_{i+1}-1\},\\ & \quad\ldots\\ & T_p = \{n_p, n_p+1, \ldots, y\}\Big\}\end{aligned}$$

のコスト $cost\big(\{T_1, T_2, \ldots, T_p\}\big)$ の最小値を $e(p,q)$ と表す．任意のクラスタ数 $p\,(1 \le p \le y)$ に対する $e(p,1)$ を，**図 4** のアルゴリズムにより求める．

このアルゴリズムでは，$e(p,q)$ を，それまでに求まっている，文番号列 $\{q, q+1, \ldots, r-1\}$ のコストと，文番号列 $\{r, r+1, \ldots, y\}$ を $p-1$ 個に分割するクラスタ列のコストの最小値との和の中で，最小となる値を求めることにより取得する．**図 5** は，本アルゴリズムが動作する 2 次元配列 (p,q) であり，各点に $e(p,q)$ が対応付けられている．$e(p,q)$ を求めるにあたり，$e(p-1,q+1), \ldots, e(p-1,y-p+2)$ の値を使用することを示している．

図 4 の 5 行目では，$e(p,q)$ をとる分割開始位置 r の集合を，後述する最小コストをとるクラスタ列を求める処理に使用するため保持している．$e(p,q)$ をとる分割開始位置 r をすべて採用せず，最小値のみ，あるいは最大値のみを採用するというバリエーションも考えられる．最小値のみを採用する場合，同じ最小コストをとる同一分割数の分割の中でも，テキストを前方から後方に見ていって，より早く見つかったトピック境界を採用することを意味する．

本アルゴリズムの文数 y に対する時間計算量は $O\left(y^3\right)$ である．本アルゴリズムでは，最小コストをとりえない分割のコスト計算を避け，計算量の爆発を防いでいる．

なお空間計算量については，p を固定したとき，$e(p-1,q)$ と $e(p,q)$ の 2 つの領域のみ使用して，以降の処理で不要になる値をアルゴリズムの過程で捨て，使用領域を節約することができる．ただし，任意の p に対する $e(p,1)$ は，後述するトピック数の推定に使用するため保持する．

1. $1 \le q \le y$ であるような q に対し
   $e(1,q) = cost(q,y)$ とおく;
2. $for\,(p=2; p \le y; p++)\,\{$
3. $\quad for\,(q=1; q \le y-p+1; q++)\,\{$
4. $\quad\quad e(p,q) =$
   $\min\limits_{q+1 \le r \le y-p+2} [cost(q, r-1) + e(p-1, r)]$
   として求める;
5. $\quad\quad e(p,q)$ をとる r の集合を
   $D_{p,q} =$
   $arg \min\limits_{q+1 \le r \le y-p+2} [cost(q, r-1) + e(p-1, r)]$
   として記憶しておく;
6. $\quad\}$
7. $\}$

図 4　最小コスト算出アルゴリズム
Fig. 4　Minimum cost calculation algorithm.

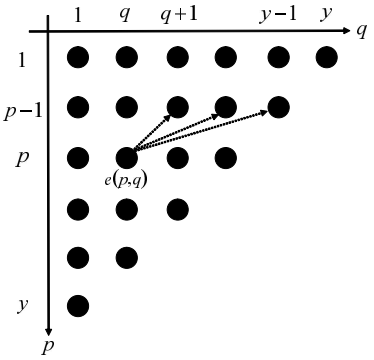

図 5　2 次元配列 (p,q)
Fig. 5　Two-dimensional array (p,q).

トピック数の推定　トピック数 $p'\,(2 \le p' \le y)$ があらかじめ与えられていたならば，図 4 の 2 行目で $p = p'$ で，かつ，3 行目で $q = 1$ のときに，5 行目の処理が終了した時点で，2 つのループから抜け終了する．そして，テキストを p' 個に分割するクラスタ列で，コストが $e(p',1)$ であるものを後述する処理により求める．

トピック数があらかじめ与えられていない場合に，トピック数を推定する方法について述べる．3.2 節の関係式 (1) から，分割数 p の増加にともない，$e(p,1)$ は単調減少していく．各トピック区間においては，概念ベクトルが凝集していると推測されるため，分割数がトピック数に達したときに，$e(p,1)$ の減少速度は低下すると考えられる．この減少速度は，分割数 $p\,(2 \le p \le y)$

に対し，$diff(p) = e(p-1,1) - e(p,1)$ と表される．**図 6** は，あるテキストにおいて，分割数 p に対する $diff(p)$ の値と，$diff(p)$ の平均 μ と標準偏差 σ により標準化した値 $(diff(p) - \mu)/\sigma$ を，プロットしたものである．p が増加するにつれ，コストは最初は急激に減少するが，ある時点から減り方が緩慢になることが分かる．そこで標準化した値が，ある値 α に最も近くなる p で，最小のものを推定トピック数とする．この推定トピック数を p' としたとき，テキストを p' 個に分割するクラスタ列で，コストが $e(p',1)$ であるものを後述する処理により求める．

トピック区間列の取得 所与の，または推定されたトピック数 p に対し，テキストを p 個に分割するクラスタ列で，コストが $e(p,1)$ であるものを，以下のようにして求める．求めるべき分割の i 番目のクラスタの最初の文番号を $sid(i)$ とする．$sid(1)$ は 1 であ

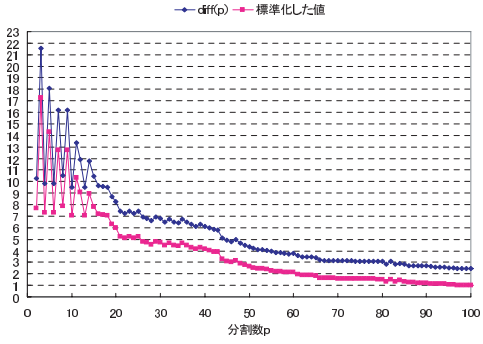

図 6 分割数 p に対する $diff(p)$ と標準化した値
Fig. 6 $diff(p)$ and the standardized value, where p is the number of segments.

1. $Search(p,s,t)$ {
2. $if\,(s == 1)$ {
3. $sid(1), \ldots, sid(p)$ を各トピック区間の開始文番号として出力;
4. } $else$ {
5. $D = D_{s,t}$;
6. $while\,(D \neq \phi)$ {
7. $k = \min_{r \in D}(r)$;
8. D から k を削除;
9. $sid(p-s+2) = k$;
10. $Search(p, s-1, k)$;
11. }
12. }
13. }

図 7 トピック区間列取得アルゴリズム
Fig. 7 Algorithm for searching topic sequences.

る．**図 7** のアルゴリズム $Search(p,s,t)$ を用いて，$Search(p,p,1)$ を実行することで，すべての分割を求めることができる．$Search(p,s,t)$ は，t 番目以降の文番号列を s 個に分割する分割数 p のクラスタ列における $sid(p-s+2), \ldots, sid(p)$ を取得するアルゴリズムである．提案手法では，このようにして得られたクラスタ列をトピック区間列と認定する．

以上のようにして，提案手法のアルゴリズムにより，クラスタ内変動最小基準を満たすクラスタ列を導出することができる．

4. 評価実験

4.1 実験データ

概念ベクトルを生成するためのコーパスとして，CD-毎日新聞 2000 年版 1 年分の 106,614 記事の見出しと本文の部分を用いた．形態素解析後に名詞，動詞，形容詞等の内容語のみを残し，異なり単語の集合を得た．異なり単語数は 109,949 であり，これらすべてを概念語とし，また，頻度順位が上位 51 番目以降の 1,500 語を共起語とした．この概念語集合と共起語集合から共起行列を生成し，特異値分解により 750 次元の概念ベクトルを生成した．

セグメント対象テキストとして，CD －毎日新聞 2001 年版の科学，文化，経済，社会，国際の各分野から，100 記事ずつ抽出し，各分野ごとに，記事の本文を連接することにより，5 つのテキストを作成した．精度評価においては，各記事を 1 トピックと仮定した．このように評価実験では，1 記事分の意味的まとまりを持った異なるトピックが隣接しているようなフラットな構造のテキストを対象とした．各テキストに関する情報は**表 2** のとおりである．表 2 においてベースラインとは，正解範囲の文境界数を，文境界の総数で割った値であり，値が小さいほど，セグメンテーションの難度が高い．±0 は，トピック境界のみを正解とし，±1 は，トピック境界から前後 1 文の範囲まで正解とすることを表す．また表 2 で示されているように，1 テキスト内においても，記事内の文数は，記事によってかなりばらつきがある．

表 2 セグメント対象テキストの情報
Table 2 Overview of test texts for segmentation.

分野	記事数	総文数	ベースライン		記事内の文数	
			±0	±1	最小	最大
科学	100	3,252	3.0%	9.1%	6	72
文化	100	2,449	4.0%	12.1%	6	131
経済	100	1,442	6.9%	20.6%	4	63
社会	100	1,085	9.1%	27.4%	3	74
国際	100	1,079	9.2%	27.6%	4	39

4.2 実験内容

対象テキストに対するセグメンテーション実験を，提案手法と概念ベクトル結束度法および Hearst 法のそれぞれで行った．対象テキストを形態素解析し，名詞，動詞，形容詞等の内容語のみを残し，処理対象の系列とした．提案手法と概念ベクトル結束度法では，残った内容語に対応する概念ベクトルを付与した．そして，このようにしてベクトルを付与された内容語の系列を処理対象とした．

提案手法においては，図 4 の 5 行目の処理において，$e(p,q)$ をとる分割開始位置 r の集合として，最小値のみを採用することとした．また，正解トピック数を与えた場合と，トピック数を推定した場合の両方の精度を測定した．セグメント対象テキストとは別の新聞記事を連接したテキストに対する予備実験により，3.3 節のトピック数の推定で述べた値 α として 1.3 をとることとした．

概念ベクトル結束度法と Hearst 法では，正解トピック数を与えた場合のみ精度を測定した．片側の窓幅を，1 記事あたりの平均処理対象単語数の小数点以下を四捨五入した値とし，基準点は 1 単語間隔でとった．結束度の平滑化では，各基準点とその直前，直後の基準点の結束度の平均をとった．極小点の depth score を求め，極小点を直近の文境界で一番前方のものに変換した．得られた文境界に，変換前の極小点の depth score の最大値を割り当てたうえで，depth score の大きい文境界から正解トピック境界数分だけ出力して，精度を測定した．

精度の指標として，以下の式で表される再現率，適合率と F 値を採用した．

再現率 = 正解出力境界数/ 正解トピック境界数

適合率 = 正解出力境界数/ 出力境界数

F 値 = (2 × 再現率 × 適合率)/ (再現率 + 適合率)

正解とする出力境界を，正解トピック境界と完全一致しているもののみとする場合（±0 と表す）と，正解トピック境界から前後 1 文の範囲までのものとする場合（±1 と表す）について，精度値を算出した．

また，精度の別の指標として，一定の距離（1 以上の整数値 k とする）にある 2 つの文の組（s_i と s_{i+k}）で，同一正解トピック区間内にあるものが別の出力トピック区間にあったり，別の正解トピック区間にあるものが同一出力トピック区間にあったりするようなものの割合である誤り率を採用した．誤り率が低いほど，精度が高い．距離 k として，対象テキストにおける 1 トピックあたりの平均文数の半分の値の小数点以下を四捨五入した値をとった☆．

実験では，各手法の処理時間についても計測した．

以下，4.3 節で精度結果を，4.4 節で処理時間の結果を述べる．

4.3 セグメンテーション精度

正解トピック数を与えた場合の提案手法の精度を**表 3** に，トピック数を推定した場合の提案手法の精度を**表 4** に，概念ベクトル結束度法の精度を**表 5** に，Hearst 法の精度を**表 6** に示した．なお表 4 において，分野名の下の括弧で囲まれた数字は，推定したトピック数を表す．表 3〜表 6 より，概念ベクトル結束度法や Hearst 法に対する提案手法の優位性が見てとれる．概念ベクトル結束度法と Hearst 法は，アルゴリズム的にはほとんど同じだが，概念ベクトル結束度法は Hearst 法より精度が高く，概念ベクトルを用いることにより，精度が向上することが分かる．提案手法は，その概念ベクトル結束度法よりも精度がさらに高い．

各手法ごとに，開始文番号と終了文番号が 1 つの正

表 3 提案手法の精度（正解トピック数を与えた場合）
Table 3 Accuracy of the proposed method (when correct topic number is assigned).

分野	正解範囲	再現率	適合率	F 値	誤り率
科学	±0	72.7%	72.7%	72.7%	11.6%
	±1	81.8%	81.8%	81.8%	
文化	±0	72.7%	72.7%	72.7%	11.9%
	±1	80.8%	80.8%	80.8%	
経済	±0	83.8%	83.8%	83.8%	8.4%
	±1	86.9%	87.9%	87.4%	
社会	±0	82.8%	82.8%	82.8%	7.6%
	±1	86.9%	90.9%	88.8%	
国際	±0	93.9%	93.9%	93.9%	1.8%
	±1	98.0%	98.0%	98.0%	

表 4 提案手法の精度（トピック数を推定した場合）
Table 4 Accuracy of the proposed method (when the topic number is estimated).

分野	正解範囲	再現率	適合率	F 値	誤り率
科学	±0	72.7%	68.6%	70.6%	12.5%
(106)	±1	82.8%	78.1%	80.4%	
文化	±0	81.8%	73.6%	77.5%	9.7%
(111)	±1	90.9%	82.7%	86.6%	
経済	±0	83.8%	83.0%	83.4%	8.7%
(101)	±1	86.9%	87.0%	86.9%	
社会	±0	72.7%	84.7%	78.3%	10.0%
(86)	±1	80.8%	95.3%	87.5%	
国際	±0	90.9%	94.7%	92.8%	2.8%
(96)	±1	94.9%	98.9%	96.9%	

☆ 誤り率の考えと距離 k の好ましいとり方については，文献 16) で提示されているが，文献 16) では一定の距離にある 2 つの単語の組を用いている．

表 5 概念ベクトル結束度法の精度
Table 5 Accuracy of the method based on the cohesion scores of the concept vectors.

分野	正解範囲	再現率	適合率	F 値	誤り率
科学	±0	33.3%	33.3%	33.3%	32.8%
	±1	41.4%	42.4%	41.9%	
文化	±0	29.3%	29.3%	29.3%	34.7%
	±1	42.4%	46.5%	44.4%	
経済	±0	54.5%	54.5%	54.5%	24.0%
	±1	63.6%	65.7%	64.6%	
社会	±0	57.6%	57.6%	57.6%	21.7%
	±1	69.7%	75.8%	72.6%	
国際	±0	57.6%	57.6%	57.6%	23.6%
	±1	67.7%	74.7%	71.0%	

表 6 Hearst 法の精度
Table 6 Accuracy of Hearst method.

分野	正解範囲	再現率	適合率	F 値	誤り率
科学	±0	28.3%	28.3%	28.3%	34.4%
	±1	33.3%	35.4%	34.3%	
文化	±0	23.2%	23.2%	23.2%	40.1%
	±1	34.3%	38.4%	36.3%	
経済	±0	35.4%	35.4%	35.4%	34.4%
	±1	50.5%	52.5%	51.5%	
社会	±0	38.4%	38.4%	38.4%	34.2%
	±1	53.5%	62.6%	57.7%	
国際	±0	53.5%	53.5%	53.5%	28.4%
	±1	60.6%	70.7%	65.3%	

解トピック区間と完全一致するような出力トピック区間の文数の分布を見たところ，提案手法，概念ベクトル結束度法，Hearst 法の順に，分布の範囲が狭まっていく傾向が見られた．概念ベクトル結束度法や Hearst 法では，長すぎる正解トピック区間や，逆に短すぎる正解トピック区間は，再現できていない傾向がある．これに対して，提案手法では，正解トピック数を与えた場合でも推定した場合でも，文化と経済のテキストにおける最大文数の正解トピック区間以外の，最大文数および最小文数の正解トピック区間を再現できていた．

提案手法では，任意の区間におけるベクトルの変動量を考慮し，あらゆる分割の中から，クラスタ集合として最適な分割を選択するため，様々な長さのトピック区間を含むテキストに対しても，窓内の局所的な範囲の情報しか用いない手法と比べ，より的確にトピック区間を検出できるといえる．

提案手法において，ベースラインが下降するにつれ精度値は低下していく傾向が見られるが，ベースラインが最も低い科学のテキスト（±0 のときに 3.0%）に対しても約 70% の F 値であり，提案手法が，トピック区間の平均長が比較的長いテキストに対しても，高精度にトピック区間を検出できることが分かる．

表 7 セグメンテーションの処理時間（秒）
Table 7 Segmentation processing time (second).

分野	総文数	提案手法		概念結束	Hearst
		推定なし	推定あり		
科学	3,252	239.4	2,019.1	101.6	348.4
文化	2,449	139.7	883.9	56.4	171.8
経済	1,442	47.8	200.4	24.8	52.3
社会	1,085	27.1	92.6	14.6	16.8
国際	1,079	27.8	91.8	16.7	37.0

提案手法においてトピック数を推定した場合の推定トピック数は，いずれの分野のテキストに対しても，正解トピック数と著しい隔たりはない．トピック区間の平均長にかかわらず，トピック数推定に用いる α の指定した値（$= 1.3$）のあたりで，クラスタが 1 トピック相当の意味的なまとまりにまで達すると考えられる．

α の値を，1.3 から ±0.1 だけ変更し，$\alpha = 1.2$ または 1.4 でトピック数を推定した場合，推定トピック数の変更前と変更後の差の絶対値の平均は 3.9 であり，精度の変更前と変更後の差の絶対値の平均は，F 値：1.5%（正解範囲：±0），1.4%（正解範囲：±1），誤り率：0.9% であった．α の値を微妙に変化させた場合に，推定トピック数や精度は，大きく変化することはなく，比較的安定しているといえる．

評価実験では，異なるトピックの記事が隣接しているようなテキストを対象とし，その記事境界を検出したが，実際には 1 つのトピックを構成する，より細かい小トピックへの分割も必要になる．小トピック内でも，その内部の概念ベクトル集合は，隣接する小トピックの概念ベクトル集合とは，クラスタ集合として分離されていると推測されるため，トピック数推定のための α の値をより小さく調節することにより，小トピック区間の検出が行えると期待できる．ただし，異なるトピックの記事間の場合ほど，クラスタ集合として鮮明に分離されていないことも考えられ，精度向上のために，手がかり語等の言語的特徴の使用といった補完処理も必要になる可能性もある．

4.4 セグメンテーションの処理時間

表 7 に，実験における各手法のワークステーション上での処理時間を示す．この処理時間は，形態素解析結果中の内容語の系列を得てからセグメンテーション結果を得るまでの時間である．表 7 において，提案手法の推定なしとは，正解トピック数（$= 100$）を与えた場合を意味し，推定ありとは，トピック数を推定した場合を意味する．概念結束とは，概念ベクトル結束度法を意味し，Hearst とは Heasrt 法を意味する．

Heasrt 法と概念ベクトル結束度法の処理時間は，テキストの総文数ではなく，処理対象単語数に依存する．

また，結束度算出のための窓幅も，処理対象単語数に比例するため，両手法の処理時間は，処理対象単語数の自乗のオーダとなる．

提案手法は，トピック数 100 を与えた場合の処理時間は，総文数にかかわらず，テキストの局所的な範囲の情報を用いる他の 2 手法と比べ，大きな違いはないといえる．しかし，トピック数を推定した場合の処理時間は，他の 2 手法と比べ長く，特に科学のテキストのようなサイズのテキストに対しては，かなりの長時間を要する．これは，トピック数を推定する方法が，あらゆる分割数に対する最小コストを求めることに起因する．

文献 11)，12) では，準最適解を求めることにより計算量削減を行っている．提案手法で計算量を削減する方法として，以下の方法が考えられる．テキストを 2 分割するコスト最小のクラスタ列を 1 つ求め，以降，これまでに求めた分割位置を固定したまま，1 回分割してコスト最小となるクラスタ列を 1 つ求める操作を繰り返す．任意の分割数に対する最小コストを求めた後に，トピック数の推定を行い，分割数が推定したトピック数であるクラスタ列をトピック区間列と認定する．任意の分割数に対して最小コストを求める処理の時間計算量は，総文数 y に対し $O\left(y^2\right)$ であり，比較的短い時間で準最適解を求めることができる．

ただし，任意の分割数に対して最小コストを求めるトピック数推定の方法では効率上問題が残っており，分割数がトピック数に達した時点で最小コスト算出処理が停止するような仕組みが，さらなる効率化のために必要である．

5. ま と め

本論文では，テキストを単語の意味表現の 1 つである概念ベクトルの系列に変換し，クラスタ内変動最小基準に基づいてテキストセグメンテーションを行う手法を提案し，その有効性を確認した．

今後の課題として，4.3 節や 4.4 節で述べた，より細かいレベルのトピックへの分割や，計算効率の良いトピック数推定処理に関する研究があげられる．このほか，概念ベクトルを用いる手法では，セグメント対象テキスト中の単語の中で，対応する概念ベクトルのない単語は，処理において考慮されないという問題がある．これに対し筆者は，初期の概念語集合に含まれない単語の概念ベクトルを推定する手法を検討しており，推定した概念ベクトルをテキストセグメンテーションに適用することにより，さらなる精度向上を期待できると考えている[17]．

参 考 文 献

1) Hearst, M.A. and Plaunt, C.: Subtopic Structuring for Full-Length Document Access, *16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.59–68 (1993).
2) Hearst, M.A.: TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, *Computational Linguistics*, Vol.23, No.1, pp.33–64 (1997).
3) 仲尾由雄：語彙的結束性に基づく話題の階層構成の認定，自然言語処理，Vol.6, No.6, pp.83–112 (1999).
4) Morris, J. and Hirst, G.: Lexical Cohension Comuputed by Thesaural Relations as an Indicator of the Structure of Text, *Computational Linguistics*, Vol.17, No.1, pp.21–48 (1991).
5) 望月 源，本田岳夫，奥村 学：複数の表層的手がかりを統合したテキストセグメンテーション，自然言語処理，Vol.6, No.3, pp.43–58 (1999).
6) Kozima, H. and Furugori, T.: Segmenting Narrative Text into Coherent Scenes, *Literary and Linguistic Computing*, Vol.9, pp.13–19 (1994).
7) Brants, T., Chen, F. and Tsochantaridis, L.: Topic-Based Document Segmentation with Probabilistic Latent Semantic Analysis, *Proc. CIKM 2002*, pp.211–218 (2002).
8) 別所克人：単語の概念ベクトルを用いたテキストセグメンテーション，情報処理学会，Vol.42, No.11, pp.2650–2662 (2001).
9) Salton, G., Singhal, A., Buckley, C. and Mitra, M.: Automatic Text Decomposition Using Text Segments and Text Themes, *Proc. Hypertext'96, the Association for Computing Machinery*, pp.53–65 (1996).
10) Caillet, M., Pessiot, J., Amini, M. and Gallinari, P.: Unsupervised Learning with Term Clustering For Thematic Segmentation of Texts, *Proc. RIAO 2004*, pp.26–28 (2004).
11) Nomoto, T. and Matsumoto, Y.: A New Approach to Unsupervised Text Summarization, *Proc. ACM SIGIR 2001*, pp.26–34 (2001).
12) Hu, P., He, T. and Ji, D.: Chinese Text Summarization Based on Thematic Area Detection, *Proc. ACL-04 Workshop Text Summarization Brances Out*, pp.112–119 (2004).
13) Schütze, H.: Automatic Word Sense Discrimination, *Computational Linguistics*, Vol.24, No.1, pp.97–123 (1998).
14) 奥野忠一，芳賀敏郎，久米 均，吉澤 正：多変量解析法，日科技連 (1971).
15) 宮本定明：クラスター分析入門—ファジィクラ

スタリングの理論と応用，森北出版 (1999).
16) Beeferman, D., Berger, A. and Lafferty, J.: Statistical Models for Text Segmentation, *Machine Learning*, Vol.34(1-3), pp.177–210 (1999).
17) 別所克人：未知語の概念ベクトル推定手法，信学技報，pp.59–64 (2004).

(平成 17 年 3 月 10 日受付)
(平成 17 年 12 月 2 日採録)

別所 克人（正会員）

1992 年大阪大学理学部数学科卒業．1994 年同大学大学院理学研究科数学専攻修士課程修了．同年日本電信電話（株）入社．現在，NTT サイバーソリューション研究所メディアコンピューティングプロジェクト勤務．自然言語処理の研究に従事．電子情報通信学会，言語処理学会各会員．