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

別所 克人†

本論文では，単語の意味表現の 1 つである概念ベクトルを用いて，テキストを意味的なまとまりの単位に分割するテキストセグメンテーションの方法を提案する．単語の概念ベクトルは，セグメント対象のテキストと同じ分野のテキストコーパスにおける単語の共起関係から導出する．この概念ベクトルを用いたテキストセグメンテーションの方法として，時系列分析の一手法である区分的線形回帰分析による方法と，単語列の結束度による方法について述べる．両提案手法によってテキストの意味段落境界を認定する実験を行い，境界認定精度である再現率と適合率を統合した尺度である F 尺度を算出した．その結果，区分的線形回帰分析による方法は 71.5%の精度を持ち，単語列の結束度による方法は 77.9%の精度を持つことを確認した．

Text Segmentation Using Word Conceptual Vectors

Katsuji Bessho†

We describe a novel method of segmenting a text into thematic units. The method uses word conceptual vectors, which are based on the co-occurrence of words in a text corpus of the same field of the target text. We explain two segmentation methods; one is based on piecewise linear regression, a kind of time series analysis, and the other on cohension scores of word lines. An experiment to decide semantic paragraphs' boundaries in the text was done by using these two methods. The F-measure obtained by the piecewise linear regression method was 71.5%, and the F-measure obtained by the method using cohension scores of word lines was 77.9%.

1. はじめに

テキストを意味的なまとまりの単位である意味段落に分割するテキストセグメンテーションは，テキスト情報の処理に関する様々な分野の 1 つの基礎となるものである．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，また個々の意味段落にキーワードや要約が付与でき，意味段落間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，人間によるテキスト全体の内容の把握が容易となるばかりでなく，複数のテキストに対する自動分類や検索の精度向上も図ることができる．米国の NIST と DARPA が共同主催している TDT 評価プロジェクト（http://www.nist.gov/speech/tests/tdt/index.htm）では，テキストセグメンテーション技術の応用である，ニュース等のデータを時系列に並べたストリームから新しいトピックを検出してユーザに通知するようなトピック検出技術に関するコンテストを行っている．

従来からテキストセグメンテーションの研究者は，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多く利用してきた．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 は，各単語の境界に対し，前後一定の単語数の窓をとり，各窓に含まれる単語の出現頻度ベクトルの余弦測度をこの境界近傍の結束度として計算し，結束度が極小となる境界位置を段落境界候補とする方法を提案している[2]．また，この結束度を平滑化することにより，精度が向上することを示している[3]．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方法を提案している[9]．

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連鎖から段落境界候補を求める方法を提案している[8]．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，段落境界を検出することを行っている[7]．

† 日本電信電話株式会社 NTT サイバースペース研究所
NTT Cyber Space Laboratories, NTT Corporation

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段落境界候補とする方法を提案している[5]．また Ferret は，コーパスから作成した意味ネットワーク上の活性伝播によって結束度を計算する方法をとっている[1]．

本論文では，共起性のある語彙の出現情報を利用した新しいテキストセグメンテーションの方法を提案する．語彙の共起情報としては，NTT が開発した発想誘導型情報検索システム[4),12)] によって生成される概念ベースを用いる．発想誘導型情報検索システムは，Schütze らによる論文[10),11)] の情報のみに基づいて作成したものであり，テキストコーパスを入力として，各単語にその共起パターンをベクトル化した意味表現（概念ベクトルと呼ぶ）を対応付け，単語とその概念ベクトルの対の集合である概念ベースを生成する．ある 2 単語に対応するベクトル値が近ければ，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．

セグメント対象のテキスト中の各単語に，この概念ベース中のベクトルを対応付けて得られるベクトル列は，単語の意味の変遷を表していると考えられるので，このベクトル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．ただし，概念ベースを生成するには多量のコーパスを必要とするので，学習用コーパスで概念ベースを生成しておき，その概念ベースを別のテキストのセグメンテーションに適用する形をとるのが一般的である．セグメント対象のテキストに出現する単語の共起パターンを得るために，学習用コーパスは，セグメント対象のテキストと同じ分野であるものをとる．

本論文では，セグメント対象のテキストから得られるベクトル列を利用したテキストセグメンテーションの方法として，時系列分析の一手法である区分的線形回帰分析による方法と，単語列の結束度による方法について述べる．

以下，まず 2 章で概念ベースについて説明する．次に，3 章で区分的線形回帰分析による方法について説明し，4 章で単語列の結束度による方法について説明する．5 章で両手法の実験結果について述べ，6 章で他手法との比較を，7 章でまとめを述べる．

2. 概念ベース

発想誘導型情報検索システムは，与えられたテキストコーパスに対し，まずコーパス中の単語間の共起頻度を記録した共起行列を生成する（**表 1**）．表 1 の共起行列において，行と列の単語群はいずれも異なる単語の集合であり，行列の (i, j) 成分は i 行目の単語と j 列目の単語との共起頻度である．ここで共起頻度は，一方の単語の出現位置から一定の窓幅をとったときの，他方の単語の出現回数である．

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

	...	貿易	...	歌手	...
...	...	...	...	...	...
関税	...	301	...	2	...
...	...	...	...	...	...
オペラ	...	4	...	73	...
...	...	...	...	...	...

各行ベクトルは，対応する単語の共起パターンを表す．ただしこのままでは，ベクトルの次元数が多すぎるため計算量が多くなり，またデータのスパースネス性が生じる．そこで発想誘導型情報検索システムは，共起行列を特異値分解（SVD: Singular Value Decomposition）により，次元数を縮退させた行列に変換する．

$n \times p$ の行列 X を共起行列としたとき，特異値分解により共起行列 X は，以下のように分解できる．

$$\underset{n\times p}{X} = \underset{n\times r}{U}\ \underset{r\times r}{S}\ \underset{r\times p}{V^T}$$

ここで，$r = \operatorname{rank} X \leq \min(n, p)$ であり，添字 T は行列の転置を表す．行列 U，V は，$U^TU = V^TV = I$（I：単位行列）を満たす．また S は r 次の対角行列であり，対角成分 (i, i) の値（X の特異値という）は正の数で，i が増えるにつれ等しいままか減少する．r 以下の任意の自然数 k をとり，U，V の先頭の k 列をとって得られる行列をそれぞれ U'，V'，S の先頭の k 行 k 列をとって得られる正方行列を S' とし，積として，

$$\underset{n\times p}{X'} = \underset{n\times k}{U'}\ \underset{k\times k}{S'}\ \underset{k\times p}{V'^T}$$

をとる．$\operatorname{rank} X' = k$ であり，X' の第 i 番目の行ベクトルは，X の第 i 番目の行ベクトルを p 次元空間内のある k 次元空間 W に射影したものである．W は，X の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 k 次元空間であり，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 k 次元空間である．V' の k 個の列ベクトルは W の正規直交底であり，X' の第 i 番目の行ベクトルを，この正規直交底のなす座標の成分で表したものが，$U'S'$ の

第 i 番目の行ベクトルである．X の第 i 行目の単語の p 次元の行ベクトルは，射影により $U'S'$ の第 i 行目の k 次元の行ベクトルに変換される．

$U'S'$ 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ているものなので，本研究では，変換後のベクトルを $U'S'$ ではなく，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とする．変換後のベクトルを概念ベクトルと呼び，単語とその概念ベクトルの対の集合を概念ベースと呼ぶ．

本研究では，概念ベクトルの次元数 k として 100 次元をとっている．

3. 区分的線形回帰分析による方法

時系列分析（time series analysis）は，離散的な時点に対しあるデータ値が対応しているとき，データ値の変化を直線や 2 次曲線等に近似して，データ値の傾向的な変動を分析する手法である．時系列分析においてデータ値の変動を 1 つの線形モデル（直線）だけで説明するのが困難な場合，時点列をいくつかの区分に分けて，それぞれの区分ごとに線形モデルを与える手法が区分的線形回帰分析（piecewise linear regression）である．区分的線形回帰分析は，直観的にいってデータ値に大きな段差が生じる箇所を時点列の境界として認定し，時点列をセグメントする手法である（**図 1**）．以下，3.1 節でデータ値が実数の時点列に対する区分的線形回帰分析のアルゴリズムを述べ，3.2 節でデータ値がベクトルの時点列に区分的線形回帰分析を適用することによってテキストセグメンテーションを行う方法を述べる．

3.1 区分的線形回帰分析のアルゴリズム

時点列を $s(1), s(2), \ldots, s(N)$ $(s(1) < s(2) < \ldots < s(N))$ とし，各時点 $s(i)$ $(1 \leq i \leq N)$ に対し，実数のデータ値 $Y(s(i))$ が対応付けられているとする．時点列 $s(i)$ $(1 \leq i \leq N)$ を区分的線形回帰分析によりセグメントするアルゴリズムは以下のとおりである．このアルゴリズムは文献 6) に基づく．なお，アルゴリズム中，F 検定を行う箇所があるため，検定の有意水準値を α と定めておく．自由度 (k_1, k_2) の F 統計量 $F(k_1, k_2)$ の上側確率が α となる値を $F_\alpha(k_1, k_2)$ とする．$P\big(F \geq F_\alpha(k_1, k_2)\big) = \alpha$ である．

【アルゴリズムの開始】

step1 各点 $s(i)$ $(1 \leq i \leq N)$ をただ 1 つの点からなるクラスタ（孤立クラスタと呼ぶ）として固定する．固定したクラスタをフィクストクラスタと呼び，フィクストクラスタリストに登録する．

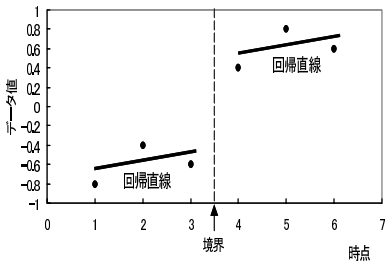


図 1 区分的線形回帰分析の概念
Fig. 1 Concept of piecewise linear regression.

step2 最初と最後の 2 つの点を除く各点 $s(i_0)$ に対して，点列 $s(i_0 - 1), s(i_0), s(i_0 + 1)$ からなる仮のクラスタをつくる．固定していない仮のクラスタをポテンシャルクラスタと呼び，ポテンシャルクラスタリストに登録する．

step3 ポテンシャルクラスタリスト内の新規に生成された任意のポテンシャルクラスタを，$t(1), t(2), \ldots, t(M)$ $(t(1) < t(2) < \ldots < t(M))$ としたとき，$t(j)$ $(1 \leq j \leq M)$ を説明変量，$Y(t(j))$ $(1 \leq j \leq M)$ を目的変量として，線形回帰分析を行う．線形回帰分析は，直線 $\hat{Y}(t(j)) = \beta_0 + \beta_1 t(j)$ をとったとき，目的変量との残差 $\epsilon(t(j)) = Y(t(j)) - \hat{Y}(t(j))$ の自乗和 $\sum_{j=1}^{M} \epsilon(t(j))^2$ を最小にするような回帰係数

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

を求めることである．回帰係数 β は，

$$P = \begin{pmatrix} 1 & t(1) \\ 1 & t(2) \\ \vdots & \vdots \\ 1 & t(M) \end{pmatrix}, \quad Q = \begin{pmatrix} Y(t(1)) \\ Y(t(2)) \\ \vdots \\ Y(t(M)) \end{pmatrix}$$

と表したとき，$\beta = \left(P^T P\right)^{-1} P^T Q$ として計算する．$\hat{Y}(t(j))$ $(1 \leq j \leq M)$ を $t(j)$ $(1 \leq j \leq M)$ の回帰推定値という．

次に，当該ポテンシャルクラスタの適合度 ϕ を，

$$\phi = \sum_{j=1}^{M} \frac{\epsilon(t(j))^2}{M - 2}$$

として計算する．

ポテンシャルクラスタ内の全ポテンシャルクラスタの中で，最小の適合度を持つポテンシャルクラスタを特定する．

step4 step3 で特定した適合度最小のポテンシャルクラスタを固定すべきか否かの検定を行う．

ポテンシャルクラスタが，3 つの孤立クラスタから成り立っていれば，無条件に検定合格とする．

ポテンシャルクラスタが，1 つの孤立クラスタ $t(0)$ と，もう 1 つの非孤立フィクストクラスタ $t(j)$ $(1 \leq j \leq M)$ から成り立っている場合，以下のように行う．$Y(t(0))$ を $t(0)$ に対するデータ値，$\hat{Y}(t(j))$ $(1 \leq j \leq M)$ を非孤立フィクストクラスタ $t(j)$ $(1 \leq j \leq M)$ における回帰推定値，$\hat{Y}^*(t(j))$ $(0 \leq j \leq M)$ をポテンシャルクラスタ $t(j)$ $(0 \leq j \leq M)$ における回帰推定値，ϕ を非孤立フィクストクラスタ $t(j)$ $(1 \leq j \leq M)$ の適合度としたとき，以下の F 統計量を求める．

$$F(1, M-2) = \left[\left(Y(t(0)) - \hat{Y}^*(t(0))\right)^2 + \sum_{j=1}^{M}\left(\hat{Y}(t(j)) - \hat{Y}^*(t(j))\right)^2\right] \Big/ \phi$$

$F(1, M-2) < F_\alpha(1, M-2)$ ならば検定合格とし，$F(1, M-2) \geq F_\alpha(1, M-2)$ ならば検定不合格とする．

ポテンシャルクラスタが 2 つの非孤立フィクストクラスタから成り立っている場合，以下のように行う．一方の非孤立フィクストクラスタを $t(j)$ $(1 \leq j \leq G)$，もう一方の非孤立フィクストクラスタを $t(j)$ $(G+1 \leq j \leq G+H)$ とする．$\epsilon(t(j))$ を非孤立フィクストクラスタ $t(j)$ $(1 \leq j \leq G)$ における $t(j)$ の回帰推定値と $Y(t(j))$ の残差とし，$\epsilon^*(t(j))$ を非孤立フィクストクラスタ $t(j)$ $(G+1 \leq j \leq G+H)$ における $t(j)$ の回帰推定値と $Y(t(j))$ の残差とし，$\epsilon^{**}(t(j))$ をポテンシャルクラスタ $t(j)$ $(1 \leq j \leq G+H)$ における $t(j)$ の回帰推定値と $Y(t(j))$ の残差としたとき，以下の F 統計量を求める．

$$F(2, G+H-4) = \left\{\left[\sum_{j=1}^{G+H}\epsilon^{**}(t(j))^2 - \sum_{j=1}^{G}\epsilon(t(j))^2 - \sum_{j=G+1}^{G+H}\epsilon^*(t(j))^2\right] \Big/ 2\right\} \Big/ \left\{\left[\sum_{j=1}^{G}\epsilon(t(j))^2 + \sum_{j=G+1}^{G+H}\epsilon^*(t(j))^2\right] \Big/ (G+H-4)\right\}$$

$F(2, G+H-4) < F_\alpha(2, G+H-4)$ ならば検定合格とし，$F(2, G+H-4) \geq F_\alpha(2, G+H-4)$ ならば検定不合格とする．

検定が合格のとき step5 に進み，不合格のとき終了する．

step5 適合度が最小のポテンシャルクラスタを構成するフィクストクラスタをフィクストクラスタリストから外し，代わりに当該ポテンシャルクラスタをフィクストクラスタとして登録する．また，この新しいフィクストクラスタと交わりを持つポテンシャルクラスタをポテンシャルクラスタリストから外し，代わりにこのフィクストクラスタとその直前，直後のフィクストクラスタとを別々に結合して新たなポテンシャルクラスタを作り，ポテンシャルクラスタリストに登録する．この段階でフィクストクラスタリストの要素数が 2 個以上なら step3 に進み，1 個なら終了する．

図 2 クラスタリストの変遷の例
Fig. 2 Transition of cluster lists.

【アルゴリズムの終了】

図 2 は，本アルゴリズムによるフィクストクラスタリストとポテンシャルクラスタリストの変遷の例を表したものである．時点列として $1, 2, \ldots, 7$ があり，それぞれに対してデータ値が対応付けられているとすると，まず各時点を孤立クラスタとする（フィクストクラスタリスト 1）．次に 3 つの連続する孤立クラスタをポテンシャルクラスタとする（ポテンシャルクラスタリスト 1）．図ではクラスタを，それに含まれる時点を直線でつないで示している．ポテンシャルクラスタ 2，3，4 が適合度最小であり，固定すべきか否かの検定が合格だったので，時点列 2，3，4 をフィクストクラスタとする（フィクストクラスタリスト 2）．次にフィクストクラスタ 2，3，4 とその前後の孤立クラスタ 1 および 5 とを別々に結合して新たなポテンシャルクラスタ 1，2，3，4 および 2，3，4，5 を作る（ポテンシャルクラスタリスト 2）．以下，同様にクラスタリストの更新を続けていき，フィクストクラスタがただ 1 つになった段階（フィクストクラスタリスト 5）で処理を終了する．

3.2 テキストセグメンテーションの方法

十分な量の学習用コーパスから概念ベースを生成し

ておけば，学習用コーパスと同じ分野のセグメント対象のテキストに対し，区分的線形回帰分析を用いて分割を行うことができる．まず，セグメント対象のテキストを形態素解析して単語に分割し，得られた各単語のうち，自立語のみに概念ベース中の対応するベクトルを付与する．これによってベクトルの系列が得られる．3.1 節で述べた区分的線形回帰分析は，データ値が実数の系列を対象としているので，ベクトルの各座標ごとに得られる実数値の系列のそれぞれに対し区分的線形回帰分析を行う．

ここで，セグメント対象のベクトルの系列として，単語単位のベクトルの列をとったならば，以下の理由で意味段落への適切なセグメンテーションが行えないと考えられる．

- 一般に単語単位のベクトル列の変動は大きすぎ，1 つの意味段落内で安定した傾向を示さない．
- 1 文内の単語の出現順序は文法的制約によっており，必ずしも意味の変遷を反映しているとは限らない．

1 つの意味段落内のすべての単語のベクトルの重心は，その意味段落の意味を表現するベクトルと見なせる．これを意味段落のベクトルと呼ぶことにすると，一般に意味段落内の 1 つの単語のベクトルよりも，複数の単語のベクトルの重心の方が，意味段落のベクトルにより近い値をとる確率が高くなる．そこでセグメント対象のベクトル列として，連続する複数の単語のベクトルの重心の列をとる．こうすると，1 つの意味段落内でベクトルの系列は，単語単位のベクトルの系列に比べ安定した傾向を示す．連続する複数の単語の集合としては，1 文内の単語集合や，一定個数の連続する単語の集合等をあげることができる．

以下，具体的な 1 テキストのセグメンテーションの精度に基づいて，区分的線形回帰分析によるテキストセグメンテーションの適切な方法を構成する．

大量の紙媒体の記事等の文字を OCR でテキスト化したデータや，ニュース音声等を音声認識ソフトでテキスト化したデータをはじめ，大量の未整理の電子化された記事群等のテキストデータを活用することは今後ますます増大していくと予想される．このようなテキストデータからユーザが所望するトピックを検出するトピック検出技術においては，テキストを個々の記事へセグメントし，個々の記事を意味的なまとまりの単位として扱えるようにしておくことが非常に有効な手がかりとなる．

このような応用をふまえ，本研究では，セグメント対象テキストとして新聞記事を連接したものをとり，個々の記事を意味段落ととらえて，テキストの記事境界を認定することを考える．セグメント対象テキストは，タグがついていず，各行が改行で区切られたプレーンテキストとする．2 つ以上の文を含む行が存在するとき，各行を形式段落と呼び，そのテキストは形式段落の書式情報があると呼ぶことにする．実験に用いるテキスト中の記事数は 300 であり，各記事には形式段落の書式情報がある．各記事は平均 12 文からなる少量の完結したテキストであり，各記事の内容が 1 つの意味段落に相当すると考えられたので，精度認定に問題ないと判断した．

ここで形式段落の書式情報がある場合，意味段落の境界の候補は形式段落間の境界であると仮定する．すなわち少なくとも形式段落の中途に意味段落の境界はないとする．このとき意味段落境界すなわち記事境界の候補数は 1,620 となる．また正解境界数は 299 である．1 形式段落内の単語のベクトルの重心をとり，形式段落のベクトル列をつくる．先頭の形式段落から 1 番から順に番号をふっていき時点番号とする．ベクトルの各座標ごとに得られる実数値の系列のそれぞれに対し，ある有意水準値の下で区分的線形回帰分析を行う．これにより，各座標ごとに境界候補の集合が得られる．

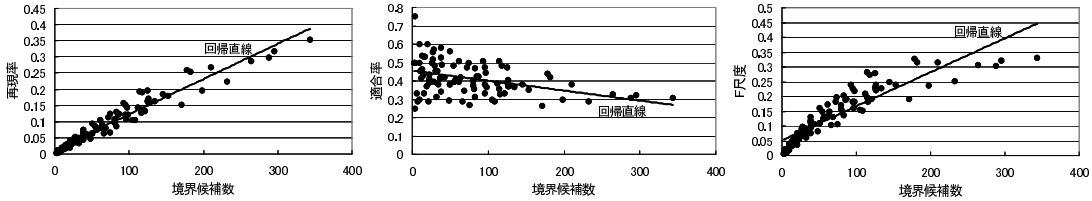
各座標ごとのセグメンテーションの精度は，再現率と適合率および両者を統合した尺度である F 尺度によって測定する．ここで，各尺度は以下の式により定義する．

$$\text{再現率} = \frac{\text{正解の境界候補数}}{\text{正解境界数}}$$

$$\text{適合率} = \frac{\text{正解の境界候補数}}{\text{境界候補数}}$$

$$\text{F 尺度} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}}$$

図 3 (a) は，有意水準値が 10^{-6} の下でセグメントを行った場合の，100 個の各座標の境界候補数と再現率の組をプロットしたものであり，図 3 (b) は同じく境界候補数と適合率の組を，図 3 (c) は境界候補数と F 尺度の組をプロットしたものである．1 つの座標は某かの観点を表していると考えられるので，観点によってテキストのセグメント結果に多様性があることが分かる．図 3 (a) では，境界候補数が多くなるほど再現率が高くなる正の相関（相関係数：0.97）が見られ，図 3 (b) では，境界候補数が少ないほど適合率が高いものが増えるゆるやかな負の相関（相関係数：−0.41）が見られる．図 3 (c) を見ると，境界候補数が最も多い座標が最も高い F 尺度の値 33.0%をとる．



(a) 境界候補数と再現率 (b) 境界候補数と適合率 (c) 境界候補数と F 尺度

図 3 境界候補数と精度の関係
Fig. 3 Relationship between the number of outputted boundaries and accuracy.

境界候補数が少ないほど適合率が高いものが増える傾向があることを鑑み，境界候補数の少ないものから複数の座標を選び，それら複数の座標の値の系列に対するセグメントを考えることによって，単独の座標によるよりも高い精度を出すことを試みる．複数の座標の値の系列に対するセグメントの境界候補の集合は，考慮する各座標の境界候補の集合の論理和となる．なぜなら，どれか 1 つの座標の境界候補においては，その座標の値の大きな段差ができているため，複数の座標の値の系列においても境界候補となるからである．境界候補数の少ないものから複数の座標を選んだ場合，各座標の適合率は比較的高い場合が多いため，1 つの座標のときの精度よりも，適合率は高く維持されたまま，再現率が向上することが期待できる．

境界候補数の少ないものから座標を選ぶにあたり，複数の座標の下での境界候補数が，正解境界数と直近になるまで座標をとることにする．その結果，複数の座標の下での精度は，再現率 41.5%，適合率 40.4%，F 尺度 40.9%であり，単独の座標による F 尺度よりも高い精度となる．

上記の再現率，適合率のほかに有意水準値を 10^{-7}，10^{-8}，10^{-9}，10^{-10}，10^{-11} のそれぞれにした場合の，複数の座標下での再現率と適合率をプロットして，順に直線でつないだものが，**図 4** における A である．一般に，有意水準値を小さくして，個々の座標の境界候補数を少なくした方が，個々の座標の適合率は高くなり，複数の座標をとったときの精度も高くなる傾向がある．

複数の座標の下での境界候補数が，正解境界数の 2/3 に直近になるまで座標をとった場合と，正解境界数の 1/2 に直近になるまで座標をとった場合のそれぞれについて，同様に各有意水準値の下での精度をプロットして直線でつないだものが，図 4 における B と C である．A と比べ，適合率はほぼ保たれたまま，再現率だけ低くなっている．正解境界数は実際にセグメントを行う際は不明であるが，境界候補数をある限度まで増やすにつれ，適合率はほぼ維持したまま，再現率を増加させられることが分かる．

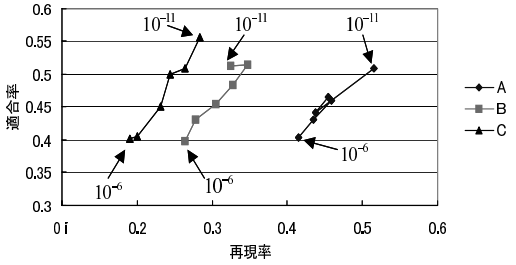


図 4 各有意水準値の下での精度
Fig. 4 Accuracy corresponding to each significance level.

A，B，C それぞれのセグメント結果において，有意水準値が小さいほど精度は高くなる傾向があることを鑑み，異なる有意水準値下でのセグメント結果を組み合わせることにより，適合率を維持したまま再現率をさらに向上させることを考える．ある有意水準値の下でのセグメント結果において，隣接する境界候補間に 2 つ以上の記事が存在することがある．このような記事群は，その有意水準値の下では同じ意味段落にあることを意味している．より厳しい（大きい）有意水準値でセグメントを行えば，このような記事群をより詳細レベルの意味段落に分割できる．そこで，ある有意水準値の下でのセグメント結果において，ある数以上の形式段落が存在する隣接する境界候補間に，より大きい有意水準値の下での境界候補がある場合，その境界候補を追加することにする（**図 5**）．これは，ある観点の下でのセグメント結果において長い意味段落があった場合，そこの意味段落だけ別の観点でより細かくセグメントすることに該当すると解釈できる．一番最初のセグメントは精度が比較的高い，小さな有意水準値で行うので，このセグメント結果の合成により，適合率はほぼ維持したまま，再現率を向上させることができる．

セグメント対象のテキストの 1 記事あたりの平均形式段落数は 5.4 段落である．A における有意水準値：10^{-11} の下でのセグメント結果において，2 記事分の 12 段落以上の間隔がある境界候補間に，有意水

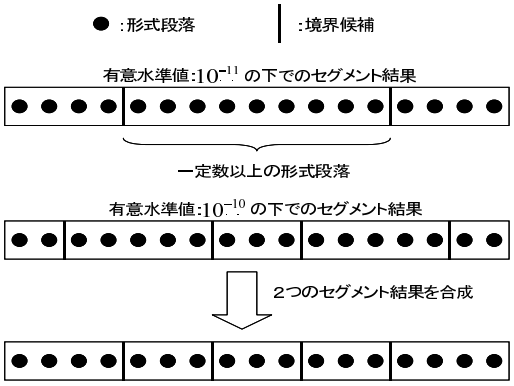


図 5 セグメント結果の合成
Fig. 5 Composition of segmentation results.

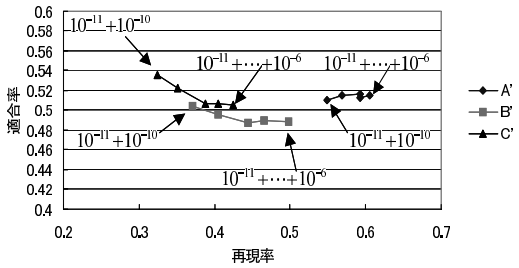


図 6 合成したセグメント結果の精度
Fig. 6 Accuracy of composite segmentation results.

準値：10^{-10} の下での境界候補があるとき，この境界候補を追加してセグメント結果を更新する．次に，この新しいセグメント結果において，12 段落以上の間隔がある境界候補間に，有意水準値：10^{-9} の下での境界候補があるとき，この境界候補を追加してセグメント結果を更新する．以下，順次有意水準値：10^{-8}，10^{-7}，10^{-6} の境界候補を同様に追加し，セグメント結果を更新していく．更新された各セグメント結果に対する再現率と適合率をプロットして直線で結んだのが**図 6** の A' である．B，C において同様の操作を行った結果が，図 6 の B'，C' である．いずれも，適合率はほぼ維持されたまま，再現率が上昇していることが分かる．

以上述べた精度の傾向は，形式段落の書式情報のない別のテキストでも同様であった．このテキストの各行は 1 文のみからなっており，意味段落境界すなわち記事境界の候補を文間の境界と仮定している．実験では，1 文内の単語のベクトルの重心の列に対し，区分的線形回帰分析を行った．

区分的線形回帰分析によるテキストセグメンテーションの方法をまとめると，以下のようになる．以下の説明では，テキストに形式段落の書式情報がない場合を記述している．テキストに形式段落の書式情報がある場合は，以下の説明中の「文」は「形式段落」に置き換える．

(1) セグメント対象のテキストを形態素解析して単語に分割し，得られた各単語のうち，自立語のみに概念ベース中の対応するベクトルを付与する．

(2) 単語列を連続する複数の単語の集合ごとに分割し，それぞれの単語群に対し，その中に含まれる単語のベクトルの重心をとる．

(3) 先頭の単語群から 1 番から順に番号をふっていき，ベクトルの時点列とする．ベクトルの各座標ごとに得られる実数値の系列のそれぞれに対し，ある有意水準値の下で区分的線形回帰分析を行う．得られた境界候補が文間の境界と一致していない場合は，境界候補を直近の文間の境界に代える．

(4) 境界候補数の少ないものから複数の座標を，境界候補集合の論理和の要素数が一定数に直近になるまでとる．

(3)，(4) の処理を，いくつかの異なる有意水準値の下で行う．

(5) 最も小さい有意水準値の下でのセグメント結果において，隣接する境界候補間に一定数以上の文があり，かつ次に小さい有意水準値の下での境界候補がある場合，その境界候補を境界候補集合に追加する．更新されたセグメント結果に対し，その次に小さい有意水準値の下での同様な条件を満足する境界候補を追加する操作を続ける．

4. 単語列の結束度による方法

セグメント対象のテキスト中の各単語に，概念ベース中のベクトルを付与して得られるベクトル列から，単語列の結束度を計算して境界候補を出す方法について述べる．まず，この方法のベースとなる Hearst 法のアルゴリズムについて述べる．

Hearst 法では，単語間の境界位置の前後に，一定の単語数の窓を設定し，各窓に含まれる単語の出現頻度ベクトルの余弦測度（cosine measure）をこの境界近傍の結束度（cohension score）として計算する．境界位置を i，左の窓を b_l，右の窓を b_r とし，単語 t の b_l，b_r における出現頻度をそれぞれ ω_{t,b_l}，ω_{t,b_r} としたとき i における結束度 C_i は，

$$C_i = \frac{\sum_t \omega_{t,b_l}\omega_{t,b_r}}{\sqrt{\sum_t {\omega_{t,b_l}}^2 \sum_t {\omega_{t,b_r}}^2}}$$

である．

結束度を計算する境界位置は，一定の単語数の刻み幅でとる．

表 2 セグメント対象のテキストのデータ
Table 2 Data of target texts of segmentation.

	見出しなしのテキスト	見出しありのテキスト
記事数：(a)	365	365
正解境界数：(a)-1	364	364
文数：(b)	5,062	4,530
1 記事あたりの平均文数：(b)/(a)	13.9	12.4
形式段落数：(c)	1,955	1,970
1 記事あたりの平均形式段落数：(c)/(a)	5.4	5.4
1 形式段落あたりの平均文数：(b)/(c)	2.6	2.3
テキスト中の自立語ののべ数：(d)	58,360 (100%)	50,688 (100%)
テキスト中の自立語でベクトルを付与された語ののべ数：(e)	51,017 (87.4%)	43,982 (86.8%)
テキスト中の自立語の異なり数	8,907 (100%)	8,637 (100%)
テキスト中の自立語でベクトルを付与された語の異なり数	6,433 (72.2%)	6,219 (72.0%)
1 記事あたりの平均自立語数：(d)/(a)	159.9	138.9
1 記事あたりの平均の，ベクトルを付与された自立語数：(e)/(a)	139.8	120.5
1 形式段落あたりの平均の，ベクトルを付与された自立語数：(e)/(c)	26.1	22.3

次に，各境界位置の結束度を，当該境界位置とその前後一定数の境界位置の結束度の平均に変換する．この操作を結束度の平滑化といい，結束度の微弱な振動を除去するために行う．

意味段落の境界では，結束度が極小となっていると期待される．結束度が極小となる境界位置（極小点と呼ぶ）を i，極小点の左側で単調減少している部分の開始位置を l，右側で単調増加している部分の終了位置を r とし，それぞれの結束度を C_i，C_l，C_r としたとき，極小点 i に対し，谷の深さを示す以下の depth score と呼ばれる値 d_i を計算する．

$$d_i = (C_l - C_i) + (C_r - C_i)$$

depth score の大きい極小点から，境界候補として出力する．

本論文で提案する単語列の結束度を用いる方法では，まずセグメント対象のテキストを形態素解析して単語に分割し，得られた各単語のうち，自立語のみに概念ベース中の対応するベクトルを付与する．以後，ベクトルを付与された自立語のみを処理の対象とする．単語間の境界位置の前後に，一定の単語数の窓を設定し，各窓ごとに，その窓に含まれる単語のベクトルの和（または重心）を計算する．各窓に対応する重心ベクトルは，窓の意味を表現するベクトルと見なせる．よって，左右の和ベクトル（または重心ベクトル）の余弦測度を，この境界位置の結束度として計算する☆．後の処理は，Hearst 法と同様である．

☆ 余弦測度はベクトル間の角度で決まるので，ベクトルとして和ベクトルをとっても重心ベクトルをとっても余弦測度は同じである．

5. 評価実験

5.1 実験データ

評価実験においては，3.2 節で述べたのと同じ理由により，セグメント対象テキストとして新聞記事を連接したものをとり，各記事を 1 つの意味段落ととらえて，テキストの記事境界の認定精度を測定する．テキストは，タグがついていないプレーンテキストとする．

概念ベースを生成するための学習用コーパスとしては，日経全文記事データベース日本経済新聞 CD-ROM 版 95 年版 2 月分の約 14,000 記事の見出しと本文の部分を用いた．概念ベースの次元は 100 次元であり，学習用コーパスにおける出現頻度の高いものから 20,000 語の単語をとり，それらの単語のベクトルを概念ベース中の単語ベクトルとした．

セグメント対象のテキストは，3.2 節で扱ったテキストとは異なるものとして，日経全文記事データベース日本経済新聞 CD-ROM 版の 96 年版から記事をピックアップし連接して作成したものをとる．実際のテキストでは，意味段落の最初に見出しや，見出しではないがその意味段落を要約するような一行があることが多い．見出し相当の行の存在によりセグメンテーションの精度にどのような影響があるかをみるため，作成テキストとして，本文のみの記事を連接したものと，見出しと本文からなる記事を連接したものの 2 つをとる．それぞれ見出しなしのテキスト，見出しありのテキストと呼ぶことにする．

各テキストのデータは**表 2** のとおりである．表 2 においては，見出しも 1 つの形式段落と見なしている．テキストを形態素解析するにあたっては，日本語形態素解析システム：茶筅（Chasen）version2.0 を使用し

た．テキスト中の自立語ののべ数の約 87%，異なり数の約 72%にあたる自立語に概念ベースのベクトルが割り当てられている．

テキストに形式段落の書式情報がないものとしたとき，意味段落境界すなわち記事境界の候補は文間の境界とし，あるものとしたとき，形式段落間の境界と仮定する．実験では，両方の場合について精度を測定することとする．また実験では，記事境界との完全一致を正解とする場合と，1 文または 1 形式段落のずれまでを正解と見なす場合の双方について精度を測定することとする．

5.2 両提案手法の評価実験

見出しなし，見出しありの各テキストに対し，境界候補が文境界の場合と形式段落境界の場合のそれぞれについて区分的線形回帰分析による方法の精度を測定する．

境界候補が文境界の場合は，単語ベクトルの重心をとる単語群の単位として，先頭から 1 文単位にとる場合と，形式段落長程度の長さにとる場合について区分的線形回帰分析を行う．1 形式段落は 2〜3 文であり，1 形式段落あたり平均の，単語ベクトルの数が二十数個であるので（表 2 参照），形式段落長程度の長さとして，2 文単位にとる場合と，20 個の単語ベクトルの単位にとる場合の 2 つの場合について行う．形式段落長程度の長さにとる場合は，1 つの単語群が記事境界をまたぐことがあり，そのような単語群のベクトルは，意味表現として曖昧なものになる．20 個の単語ベクトル単位にとる場合，各座標のセグメント結果において，出力された境界候補を直近の文境界に変換する．

境界候補が形式段落境界の場合は，単語ベクトルの重心をとる単語群の単位として，先頭から 1 形式段落単位にとる場合について区分的線形回帰分析を行う．

境界候補数の少ないものから複数の座標をとる際は，複数の座標の下での境界候補数が正解境界数の 364 に直近になるまで座標をとる．

境界候補が文境界のときの各場合について，有意水準値：10^{-11} の下で，3.2 節の (3)，(4) の処理を行ったときの，完全一致の精度は**表 3** のとおりである．いずれのテキストも 1 文単位よりも 2 文単位あるいは 20 単語単位の方が精度が高い．後者の 2 ケースの方が原理的に適合率の期待値がやや高くなることと，単語ベクトルを 1 文単位よりも形式段落長程度分集めた方が，1 つの意味段落内でベクトル値が安定する傾向があることが原因だと考えられる．見出しなしのテキストでは，20 単語単位の方が 2 文単位の方よりも精度が高く，見出しありのテキストでは，逆に 2 文単位の方が 20 単語単位の方よりも精度が高い．これは，見出しなしのテキストの場合では，20 単語単位の区切り位置の方がたまたま記事境界により近くなることが多く，見出しありのテキストの場合では，2 文単位の区切り位置の方が記事境界により近くなることが多いからだといえる．そこで，見出しなしのテキストに対しては，より精度の高い 20 単語単位に重心ベクトルをとる方法で，また見出しありのテキストに対しては，より精度の高い 2 文単位に重心ベクトルをとる方法で，有意水準値：10^{-12}，10^{-11}，10^{-10}，10^{-9}，10^{-8}，10^{-7}，10^{-6} のそれぞれの下で，3.2 節の (3)，(4) の処理を行った後，(5) の処理を行うこととする．(5) の処理を行う際の条件である隣接境界候補間の長さとしては，見出しなしのテキストの方は，1 記事あたりの平均文数が 13.9 文なので 2 記事分の 28 文とし，見出しありのテキストの方は，1 記事あたりの平均文数が 12.4 文なので 2 記事分の 26 文とする．

表 3 境界候補が文境界のときの有意水準値：10^{-11} の下での完全一致の精度

Table 3 Accuracy of complete accordance corresponding to significance level: 10^{-11}.

テキストの種類	重心をとる単位	再現率	適合率	F 尺度
見出しなし	1 文単位	12.6%	12.1%	12.4%
	2 文単位	15.9%	15.4%	15.7%
	20 単語単位	20.3%	20.6%	20.5%
見出しあり	1 文単位	14.0%	14.3%	14.2%
	2 文単位	23.4%	22.4%	22.8%
	20 単語単位	16.5%	16.2%	16.3%

境界候補が形式段落境界の場合は，いずれのテキストに対しても，有意水準値：10^{-11}，10^{-10}，10^{-9}，10^{-8}，10^{-7}，10^{-6} のそれぞれの下で，3.2 節の (3)，(4) の処理を行った後，(5) の処理を行うこととする．(5) の処理を行う際の条件である隣接境界候補間の長さとしては，いずれのテキストも 1 記事あたりの平均形式段落数が 5.4 段落なので 2 記事分の 12 段落とする．

上記の条件のもとで，区分的線形回帰分析による方法を行ったときの精度は，**表 4** のようになる．

同一テキストに対して，単語列の結束度による方法の精度も測定する．単語列の結束度による方法を以下の条件で行う．

(1) テキストの形態素解析結果における単語のうち，ベクトルを付与された自立語のみを処理の対象とする．1 記事あたり平均の，ベクトルを付与された自立語数は，見出しなしのテキストは 139.8 語であり，見出しありのテキストは 120.5 語である．そこで，

表 4 区分的線形回帰分析による方法の精度
Table 4 Accuracy of the method based on piecewise linear regression.

テキストの種類	境界候補	正解の判定基準	再現率	適合率	F 尺度
見出しなし	文境界	完全一致	28.3%	20.7%	23.9%
		1 文ずれ許容	47.5%	41.0%	44.0%
	形式段落境界	完全一致	47.5%	38.1%	42.3%
		1 段落ずれ許容	62.6%	68.1%	65.2%
見出しあり	文境界	完全一致	27.5%	20.5%	23.5%
		1 文ずれ許容	52.2%	40.9%	45.8%
	形式段落境界	完全一致	59.6%	46.7%	52.4%
		1 段落ずれ許容	68.4%	74.8%	71.5%

表 5 単語列の結束度による方法の精度
Table 5 Accuracy of the method based on cohension scores of word lines.

テキストの種類	境界候補	正解の判定基準	再現率	適合率	F 尺度
見出しなし	文境界	完全一致	51.4%	51.4%	51.4%
		1 文ずれ許容	59.6%	72.0%	65.2%
	形式段落境界	完全一致	61.3%	61.3%	61.3%
		1 段落ずれ許容	69.8%	81.3%	75.1%
見出しあり	文境界	完全一致	53.8%	53.8%	53.8%
		1 文ずれ許容	65.7%	74.2%	69.7%
	形式段落境界	完全一致	61.0%	61.0%	61.0%
		1 段落ずれ許容	70.6%	86.8%	77.9%

各単語の境界位置の前後それぞれに，窓幅が 1 記事の 3 分の 1 程度の長さ（いずれのテキストも 40 語）の窓をとり，左右の各窓ごとに，その窓に含まれる単語のベクトルの和（または重心）を計算し，左右の和ベクトル（または重心ベクトル）の余弦測度を結束度として計算する．

(2) 各単語境界位置に対する結束度を，当該境界位置とその直前，直後の境界位置の結束度の加重平均に変換する．

(3) 極小点となる境界位置の depth score を求め，depth score の大きな境界位置から，境界位置を直近の文境界（境界候補が形式段落境界の場合は形式段落境界）に変換したうえで出力する．ただし，同じ文境界（形式段落境界）は重複して出力しない．

単語列の結束度による方法では，正解境界数である 364 と同数の境界候補を出力することとする．上記の条件のもとで，単語列の結束度による方法を行ったときの精度は，**表 5** のようになる．

5.3 考　　察

区分的線形回帰分析による方法の F 尺度は，形式段落の書式情報の有無に応じて，それぞれ最高 71.5%，45.8%である．また，単語列の結束度による方法の F 尺度は，形式段落の書式情報の有無に応じて，それぞれ最高 77.9%，69.7%である．すべての場合において，区分的線形回帰分析による方法よりも単語列の結束度による方法の方が精度が高い．両手法とも，テキストの局所的な範囲の単語ベクトルの重心である，その範囲の意味を表現するベクトルに着目している点では共通している．区分的線形回帰分析による方法では，重心ベクトルそのものの変遷から重心ベクトル間の段差を検出するのに対し，単語列の結束度による方法では，重心ベクトル間の余弦測度の変遷をみてその極小点を検出する．余弦測度の方が，重心ベクトル間の差異をより精密に測定するので，意味段落の境界をより精密に検出するのだと考えられる．

区分的線形回帰分析による方法の精度は，単語列の結束度による方法との比較において，境界候補が文境界の場合よりも形式段落境界の場合の方が良好である．境界候補が文境界の場合，20 単語単位や 2 文単位で単語ベクトルの重心をとっているので，2 つの記事にまたがる単語群のベクトルが意味を的確に表現できない．これに対し，形式段落境界の場合，形式段落単位に単語ベクトルの重心をとっているので，2 つの記事にまたがるようなベクトルがないことと，重心をとる単語ベクトルの数が多く，重心ベクトルが記事内で比較的安定することが原因だといえる．

また，区分的線形回帰分析による方法の精度は，境界候補が形式段落境界のとき，見出しなしのテキストよりも見出しありのテキストの方が高い．見出し部分には記事の特徴を表す単語が集中して出現するため，見出し部分の重心ベクトルと次の形式段落の重心ベクトルとの間でクラスタが形成されやすいことがその原

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

テキストの種類	境界候補	正解の判定基準	再現率	適合率	F 尺度
見出しなし	文境界	完全一致	28.0%	28.0%	28.0%
		1 文ずれ許容	40.7%	48.4%	44.2%
	形式段落境界	完全一致	40.9%	40.9%	40.9%
		1 段落ずれ許容	54.1%	66.8%	59.8%
見出しあり	文境界	完全一致	38.7%	38.7%	38.7%
		1 文ずれ許容	47.0%	57.7%	51.8%
	形式段落境界	完全一致	45.6%	45.6%	45.6%
		1 段落ずれ許容	57.7%	69.8%	63.2%

因だと考えられる．

同様に単語列の結束度による方法の精度も，ほとんどの場合，見出しなしのテキストよりも見出しありのテキストの方が高い．見出し部分には記事の特徴を表す単語が集中して出現するので，見出し以降の部分と意味的類似性が高く，記事境界周辺の depth score が見出しなしのテキストに比べ高くなる場合が多いのがその原因だと考えられる．

両手法とも記事の見出しのみならず，一般に意味段落の最初に，その意味段落を要約するような行がある場合，直前の境界を検出する可能性は高いといえる．

評価実験において，区分的線形回帰分析による方法では，境界候補数の少ない複数の座標をとるときや，異なる有意水準値の下でのセグメント結果を合成する際に，正解境界数の情報を用いている．また，単語列の結束度による方法でも，窓幅や出力境界候補数の設定に正解境界数の情報を用いている．しかし，実際には正解境界数は一般に不明である．区分的線形回帰分析による方法は，3.2 節の予備実験の結果が示すように，出力境界候補数を増やしていった場合，適合率をほぼ維持させることができ，精度の大幅な低下を招く可能性は低い．実際にセグメンテーションを行う際は，区分的線形回帰分析による方法も単語列の結束度による方法も，1 つの意味段落の平均サイズを想定したうえで各パラメータ値を適宜定めることになる．セグメント対象テキストから最適な精度を出すパラメータ値を推定することは今後の課題である．

6. 他手法との比較

5 章の評価実験で用いたテキストに対して，Hearst 法によるセグメントも行い，精度の比較を行う．Hearst 法によるセグメントでは，同一の形態素解析結果における全自立語を処理の対象とする．1 記事あたり平均の自立語数は，見出しなしのテキストは 159.9 語であり，見出しありのテキストは 138.9 語である．そこで，各単語の境界位置の前後それぞれに，窓幅が 1 記事の 3 分の 1 程度の長さ（見出しなしのテキスト：50 語，見出しありのテキスト：40 語）の窓をとり，左右の各窓に含まれる単語の出現頻度ベクトルの余弦測度を結束度として計算する．その他の条件は，単語列の結束度による方法と同一とする．

Hearst 法では depth score の分布から，出力すべき境界候補の depth score の閾値の基準例を示しているが，最適な精度を出す閾値の設定は難しく，閾値によっては精度が大きく変動してしまう可能性がある．このため評価実験においては，正解境界数である 364 と同数の境界候補を出力することとする．上記の条件の下で，Hearst 法によるセグメントを行ったときの精度は，**表 6** のようになる．

区分的線形回帰分析による方法の F 尺度は，Hearst 法に比べ，境界候補が文境界の場合は低いが，境界候補が形式段落境界の場合は 1.4%〜8.3%高い．形式段落が平均 2，3 文以上からなりたっているようなテキストに対し，区分的線形回帰分析による方法は有効に働くといえる．

単語列の結束度による方法の F 尺度は，Hearst 法に比べ，境界候補が文境界の場合は 15.1%〜23.4%高く，境界候補が形式段落境界の場合は 14.7%〜20.3%高い．**図 7** は見出しなしのテキスト中の単語（ベクトルを付与されていないものを含む）間の境界に対応する，本手法と Hearst 法の平滑化した結束度をプロットしたものである．境界候補が文境界の場合の本手法と Hearst 法によるセグメント結果の境界候補と正解境界もプロットしている．

Hearst 法では，記事境界では左右の窓に共通して含まれる単語が少ないので結束度は低くなる．しかし，1 記事内の単語境界位置においても左右の窓に共通して含まれる単語が少ないことがあり，左右の窓の意味的類似性が高いにもかかわらず，depth score が記事境界周辺と同じくらい高くなる場合が多い．このため，depth score が相対的に高い単語境界の中に 1 記事の中途にあるものが頻出し，出力した境界候補の中に正

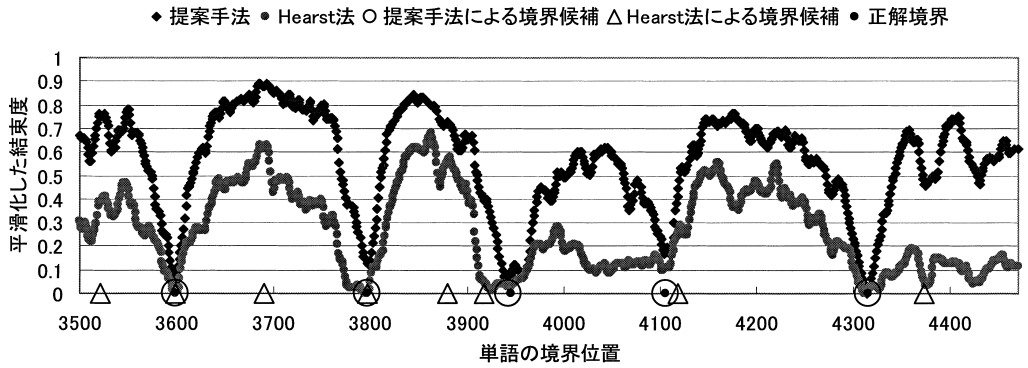


図 7 単語列の結束度による方法と Hearst 法との比較
Fig. 7 Comparison between the proposed method and Hearst method.

解でない境界候補が多く含まれる．

これに対し単語列の結束度による方法では，仮に記事境界の左右の窓に共通して含まれる単語がない場合であっても，左右の窓の意味的類似性はまったくないわけではないので，結束度は 0 でない場合が多い．このように記事境界において，本手法の結束度は Hearst 法に比べやや高くなる．しかし，1 記事内においても，各単語境界位置の左右の窓に共通して含まれる単語がたとえ少なくても意味的類似性は高いので，結束度は記事境界に比べ恒常的に高くなり，記事境界周辺の depth score はそれ以外の depth score に比べ高くなる．これが，本手法と Hearst 法との大きな精度の差の原因となっている．

両提案手法と，共起性のある語彙の出現情報を利用する関連研究との比較を述べる．文献 5) の手法は，辞書における単語の共起情報を利用している．辞書から導出される共起情報は，セグメント対象テキストの文脈と直接関係がないため，必ずしも文脈をよく反映したセグメンテーションを行えるとは限らない．これに対し，本研究の手法では，セグメント対象テキストと同分野のコーパスにおける単語の近接性に基づく情報を利用しており，文脈をよく反映したセグメンテーションを行えると考えられる．次に，文献 1) の手法はコーパスにおける単語間の相互情報量を用いており，本研究の手法で用いている特異値分解で生成される共起情報とは異なるものである．文献 1) では，実験の条件が違うという事実はあるものの，Hearst 法より精度が劣っていたと報告している．上述の 2 つの従来手法のアルゴリズムは本研究の手法と異なっており，単語の意味を表すベクトル値の変化そのものに着目してセグメンテーションを行うところが，本研究の手法の従来手法にない新しい点である．

7. おわりに

本論文では共起性のある語彙の出現情報を利用したテキストセグメンテーションの方法として，単語の共起パターンをベクトル化した概念ベクトルを利用する方法を述べた．具体的には，テキスト中の単語に付与された概念ベクトルの列を，区分的線形回帰分析によってセグメントする方法と，単語列の結束度によってセグメントする方法を述べた．

本研究の手法は，大量の未整理の記事群等のテキストを対象としたトピック検出技術において有力な手段となる可能性を持っている．両提案手法とも書き言葉の文書に多く見られる，形式段落の書式情報があるようなテキストのセグメントに有効であり，特に単語列の結束度による方法は，話し言葉をテキスト化した文書のような形式段落の書式情報がないテキストのセグメントにおいても有効であると考える．

単語列の結束度による方法は，単語境界の前後の窓の間の類似度をより精密にすることによって，さらなる精度の向上を期待できる．今後の課題は，より性能の高い単語列の結束度による方法をベースとして，窓の間のより性能の高い類似度の尺度を考案することである．また，表層的な手がかり語の情報も併用し精度への相乗効果を検証していく．

謝辞 本研究を進めるにあたりご指導いただいた小原永 NTT サイバースペース研究所メディア処理プロジェクトマネージャ，加藤恒昭東京大学助教授，NTT コミュニケーション科学基礎研究所の笠原要氏，NTT アドバンステクノロジ（株）の曽根正氏，島田茂夫氏，ならびに NTT サイバースペース研究所メディア処理プロジェクト言語メディアグループの皆様に深く感謝いたします．

参 考 文 献

1) Ferret, O.: How to thematically segment texts by using lexical cohesion?, *Proc. COLING-ACL '98*, pp.1481–1483 (1998).
2) Hearst, M.A.: Multi-Paragraph Segmentation of Expository Text, *32nd Annual Meeting of the Association for Computational Linguistics*, pp.9–16 (1994).
3) Hearst, M.A.: TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, *Computational Linguistics*, Vol.23, No.1, pp.33–64 (1997).
4) Kato, T., Shimada, S., Kumamoto, M. and Matsuzawa, K.: Idea-Deriving Information Retrieval System, *Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp.187–193 (1999).
5) Kozima, H. and Furugori, T.: Segmenting Narrative Text into Coherent Scenes, *Literary and Linguistic Computing*, Vol.9, pp.13–19 (1994).
6) Mcgee, V.E. and Carleton, W.T.: Piecewise Regression, *Journal of the American Statistical Association*, Vol.65, No.331, pp.1109–1124 (1970).
7) 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).
8) 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).
9) 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).
10) Schütze, H.: Dimensions of Meaning, *Proc. Supercomputing '92*, pp.787–796 (1992).
11) Schütze, H. and Pedersen, J.O.: A Cooccurrence-Based Thesaurus and Two Applications to Information Retrieval, *Proc. RIAO '94*, pp.266–274 (1994).
12) 熊本 睦，島田茂夫，加藤恒昭：概念ベースの情報検索への適用―概念ベースを用いた検索の特性評価，情報処理学会研究報告，Vol.SIG-ICS 115, pp.9–16 (1999).

(平成 13 年 3 月 8 日受付)
(平成 13 年 9 月 12 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 サイバースペース研究所メディア処理プロジェクト勤務．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．