

# 語彙的結束性と単語重要度に基づく テキストセグメンテーション

平尾 努<sup>†</sup> 北内 啓<sup>†</sup> 木谷 強<sup>††</sup>

情報検索・文書要約・文書分類などに代表される文書処理技術の精度向上や文書の構造解析に必要な基盤技術としてテキストセグメンテーション技術が注目されている。テキストセグメンテーションの代表的な手法として、文書内の同一語の繰り返しによる語の結束性を用いる TextTiling 法が提案されている。しかし、TextTiling 法では比較的短い文書を処理対象とした場合には十分な精度を得られるとは限らない。本稿では、文書内の単語の共起を考慮した語の結束性に基づくテキストセグメンテーション法と文書内における文の重要度の変化に基づくテキストセグメンテーション法をそれぞれ提案し、さらに、両手法を相補的に統合することで短い文書でも適用可能である高精度なテキストセグメンテーション法を新たに提案する。1件が20文程度の新聞記事326件を対象として、その形式段落を話題境界の正解データに用いて提案手法を評価した。その結果、形式段落との完全一致では適合率31.8%、再現率21.3%、形式段落の近傍での出力も正解に含めた場合には適合率64.7%、再現率43.5%であり、既存手法より高精度なテキストセグメンテーション法を実現できた。

## Text Segmentation Based on Lexical Cohesion and Word Importance

TSUTOMU HIRAO,<sup>†</sup> AKIRA KITAUCHI<sup>†</sup> and TSUYOSHI KITANI<sup>††</sup>

Text segmentation is one of the fundamental technologies indispensable to analyzing the document structure and improving the accuracy of text processing technology, such as information retrieval, text summarization, and text categorization. TextTiling method have been proposed as a major text segmentation method, which is based on lexical cohesion regarding the repetition of the same words in a document. However, using TextTiling method for short documents, we do not always get enough accuracy. In this paper, we first propose two text segmentation methods; one is based on lexical cohesion considering co-occurrences of words, and the other is based on the changes of the importance of the each sentence in a document. Next, we propose a new text segmentation method integrating these two methods to get high accuracy even for short documents. We evaluate our method in the view of precision and recall rates. In our experiment, we use paragraph borders in newspaper articles as correct segment borders on 326 newspaper articles each of which contains about 20 sentences. As the result, it is shown that our methods are more accurate compared with the conventional methods; 31.8% on the precision rate and 21.3% on the recall rate in case that paragraph borders in newspaper articles regarded as correct segment borders, and 64.7% on the precision rate and 43.5% on the recall rate in case proximity of paragraph borders regarded as correct segment borders.

### 1. はじめに

近年、インターネットに代表されるネットワークの発達や、大容量メディアの低価格化により、大量の電

子化文書が広く流通している。いわゆる情報の洪水の中においては、如何に必要な情報を効率良く選択するかが重要となる。このような状況を背景として情報検索・文書要約・文書分類などに代表される文書処理の研究が盛んに行われている。しかし、複数の話題を含む文書に対してこれらの処理を文書全体に施す場合には必ずしも十分な精度を得られるとは限らない。この原因の一つとして、文書中の重要な話題を特定したうえで処理を実行していないことがあげられる<sup>1)~5)</sup>。

このような問題を解決する技術として文書を意味的

<sup>†</sup> 株式会社 NTT データ 技術開発本部 オープンシステムセンタ  
Open Systems Center, Research and Development  
Headquarters, NTT DATA CORPORATION

<sup>††</sup> 株式会社 NTT データ 技術開発本部 北米技術センタ  
Technical Center of California, Research and Development  
Headquarters, NTT DATA CORPORATION

なまとまりで分割するテキストセグメンテーション技術が注目されている。文書を話題ごとのまとまり(セグメント)の集合として考えることで情報検索・文書要約・文書分類などの文書処理技術の精度向上が期待できる<sup>6)~9)</sup>。また、文書構造の解析や文書の可読性の向上など幅広い応用も期待できる。このように、文書処理の基盤技術としてテキストセグメンテーションは重要な技術である。

本稿では、比較的短い文書に対応できるテキストセグメンテーション法として、文書内の語の共起を考慮した語彙的結束性に基づくテキストセグメンテーション法と単語重要度に基づくテキストセグメンテーション法をそれぞれ提案し、さらに両手法を相補的に統合したより高精度なテキストセグメンテーション法を提案する。326件の新聞記事を用いて、その形式段落を正解とした場合の適合率・再現率、F-measureにより精度を評価し、提案手法が有効であることを示す。

## 2. 研究の前提

一般に、論文、マニュアルなどの長い文書であれば、章や節などにより文書は構造化されている場合が多い。ここで、話題の粒度を粗く考えた場合には章や節が話題を表す単位であると考えることが可能である。しかし、話題の粒度を細かく考えた場合には、文書の章や節内にも複数の話題が存在する。同様に、短い文書では、話題の粒度を粗く考えた場合には1文書に対して1話題であるとも考えられるが、話題の粒度を細かく考えた場合には短い文書にも複数の話題が存在すると考えられる。このように文書における話題を議論する際には、その粒度を考慮する必要がある。文書における話題構造を理解することは様々な文書処理技術に有効なことであるが、そのためには、粒度の粗い話題だけでなく粒度の細かい話題も考慮する必要がある。そこで本稿では、先に述べた大きな文書内における章や節の内部や短い文書に存在する粒度の細かい話題を認定することを目的としたテキストセグメンテーションについて議論する。

比較的短い文書(領域)をテキストセグメンテーションの対象とした場合には、テキストセグメンテーションによって文書に設ける話題数を文書に含まれる実際の話題数に対し、それ以下にするということが必要である。すなわち、本稿ではテキストセグメンテーションによって文書に設ける話題数を実際の話題数と同等かそれ以下として、適合率に対して再現率がそれを上回らないことを前提とする。その理由を以下にあげる。

文書に含まれている話題数より多くの話題を設ける

例として、21文から成る文書が5個の話題(段落)から構成されていた場合に1文を1段落とするテキストセグメンテーションを考える。すると、話題境界候補の出力は20箇所となる。ここで、出力した話題境界の候補に実際の話題境界の正解が含まれる率は4/20となるので、適合率は20%、出力した話題境界候補に含まれる正解が実際の話題境界の正解に占める率は4/4となるので、再現率は100%となる。このことから、文書の実際の話題数より多くの話題を設けた場合には、高い再現率を得ることができるが、テキストセグメンテーションそもその目的からは大きくはずれていることがわかる。このように短い文書では、実際の話題数を越えて多くの話題を設けた場合には有意な語彙的連鎖が切断されるため、文書に設けた各話題が意味的なまとまりを表さない。すなわち、テキストセグメンテーションの出力数を考慮せずに再現率を重視することはふさわしくないからである。

## 3. 従来の研究とその問題点

テキストセグメンテーションに関する手法はいくつか提案されているが、長い文書を対象として粒度の粗い話題を認定するために提案されている場合が多い。本章では、テキストセグメンテーション技術の従来研究について説明し、その代表的な手法であるTextTiling法<sup>9)</sup>とその問題点について述べる。

### 3.1 従来の研究

文書に複数の話題が存在する場合、文書中の各話題に対応して意味的なつながりを持った語の連鎖である語彙的連鎖が存在する。語彙的連鎖を認定し、各語彙的連鎖ごとに文書を分割することがテキストセグメンテーションである。

テキストセグメンテーションの手法としては、以下に示す併合型・分割型の2種の手法がある。

**併合型** 文書内の文や単語を最小単位とし、隣接する単位を結合する手法

**分割型** 分割されていない文書から話題分割のための境界を探索する手法

併合型の手法としては、重要語の続く隣接単位を連結する手法が提案されている<sup>8)</sup>。分割型の手法としては、以下に示す手法が提案されている。

- シソーラス上での類義語が連続して出現する部分を語彙的連鎖として捉え、その開始位置、終了位置、連鎖の出現しないギャップの位置にスコアを与え、その総和より話題の境界を推定する手法<sup>7)</sup>
- シソーラス上の類義語が連続している部分を語彙的連鎖として捉え、さらに接続詞、助詞などの

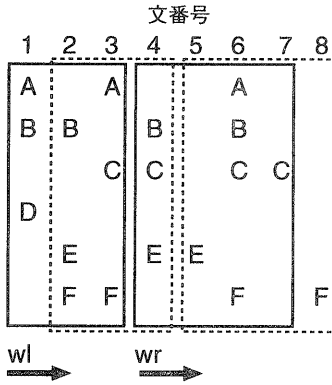


図1 TextTiling 法  
Fig. 1 TextTiling method

表層情報を組み合わせて話題の境界を推定する手法<sup>6)</sup>

- テキスト中のある基準点に対して左右同数の単語を含む窓を設け、左右の窓の類似度が高い部分を語彙的連鎖とみなし、類似度が極小となる基準点を話題の境界とする TextTiling 法

併合型の 8), 分割型の 6), 7) は再現率が適合率を上回っており, 2 章で述べた前提条件を満たさない. TextTiling 法は適合率が再現率を上回っており, 2 章で述べた前提条件を満たしている. しかし, この手法は長い文書を対象とした手法であり, 短い文書を対象とした場合には十分な精度が得られるとは限らない. 以下, TextTiling 法とその問題点を説明する.

### 3.2 TextTiling 法

TextTiling 法では, 文書の意味的に関連の深い部分には, 同一の語が繰り返し出現するという性質を利用している. 文書中のある基準点に対してその左右に同数の単語を包含した窓を設け, 左右の窓の類似度(結束度)を求め, 基準点を一定間隔でずらしながら類似度の変化に着目し, グラフにおける類似度の極小点を話題の境界と推定する手法をとっている. 窓間の類似度は, つぎに示す cosine measure で表される.

$$sim(wl, wr) = \frac{\sum_t f(t_{wl})f(t_{wr})}{\sqrt{\sum_t f(t_{wl})^2 \sum_t f(t_{wr})^2}} \quad (1)$$

ここで,  $wl$  と  $wr$  は, それぞれ左窓と右窓であり,  $f(t_{wl})$  と  $f(t_{wr})$  は, それぞれ, 単語  $t$  の左窓, 右窓における出現頻度である.

図 1 の例において, 文書の 3 文目と 4 文目の境界を基準点として窓が包含する単語数を 9 とすると, 基準点に対する左右の窓における各単語 ( $A \sim F$ ) の出現頻度は, 以下に示すとおりである.

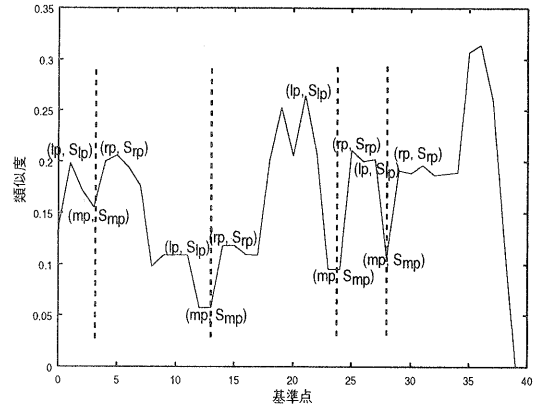


図2 TextTiling 法によるセグメンテーションの例  
Fig. 2 An example of text segmentation using TextTiling method

$$\begin{aligned} f(A_{wl}) &= 2, f(A_{wr}) = 1 \\ f(B_{wl}) &= 2, f(B_{wr}) = 2 \\ f(C_{wl}) &= 1, f(C_{wr}) = 3 \\ f(D_{wl}) &= 1, f(D_{wr}) = 0 \\ f(E_{wl}) &= 1, f(E_{wr}) = 2 \\ f(F_{wl}) &= 2, f(F_{wr}) = 1 \end{aligned}$$

よって, 式 (1) を用いて類似度  $sim(wl, wr)$  は, 以下のとおり計算される.

$$sim(wl, wr) = \frac{2 + 4 + 3 + 0 + 2 + 2}{\sqrt{15 \cdot 19}} = 0.77$$

上述の手法を用いて, 基準点を文書の先頭から末尾に向かって一定間隔で移動しながら各基準点における左右の窓の類似度をプロットすると図 2 に示すようなグラフになる. ここで, 類似度が極小値をとる基準点, すなわち, 左右の窓の結束性が極小となる位置を話題の境界とする. ただし, 類似度の微妙な揺れを無視するため, 極小点  $mp$  の類似度  $S_{mp}$  と, 左側の極大点  $lp$  における類似度  $S_{lp}$ , 右側の極大点  $rp$  における類似度  $S_{rp}$  の差を考慮し, 以下の式で depth score (以下,  $d$ ) を求め,  $d$  がしきい値  $d_{th}$  を越えた場合に話題の境界とする.  $\bar{S}$  は類似度の平均,  $\sigma$  は類似度の分散である.

$$d = (S_{lp} - S_{mp}) + (S_{rp} - S_{mp}) \quad (2)$$

$$d_{th} = \bar{S} - \sigma/2 \quad (3)$$

### 3.3 TextTiling 法の問題点

TextTiling 法には以下の 2 つの問題がある.

- (1) 窓が小さくなると左右の窓における類似度が顕著に低くなるため, 短い文書では左右の窓の正確な類似度が計算できない.
- (2) 文書の先頭, 末尾付近では左右同一の単語数を

包含する窓を設定できない。

(1) について、左右の窓に含まれる単語数が数百以上であれば、両窓に同一語が出現することが期待できる。特にマニュアルや論文などの複数の章から構成される大きな文書を対象とした場合に有効である<sup>5)</sup>。しかし、短い文書では、小さな範囲内で同一の語が繰り返し出現することは稀である。よって、短い文書を対象とする場合には、設定できる窓幅が小さくなることもあり、左右の窓の類似性はみられなくなる場合が多くなる。左右の窓の類似度が0となる基準点が多くなると語彙的連鎖があっても認定できないため、テキストセグメンテーションの精度に悪影響を及ぼす。

(2) について、基準点に対して単純に左右に窓を設ける手法では、文書の先頭付近では左側の窓、文書の末尾付近では右側の窓が包含する単語が少なくなることは避けることができない。

上記の問題点は、特に短い文書では顕著に現れ、語彙的連鎖認定における大きな問題点となる。

#### 4. 語彙的結束性と単語重要度に基づくテキストセグメンテーション

本章では、比較的短い文書にも適用可能なテキストセグメンテーション法を提案する。そこで、3.3節で述べた Texttiling 法の2つの問題点を解決するために、文書内における単語間の共起を考慮した語彙的結束性に基づく手法と単語重要度に基づく手法をそれぞれ提案し、さらに両手法を相補的に統合することによって、より高精度なテキストセグメンテーション法を提案する。

##### 4.1 単語間の共起を考慮した語彙的結束性に基づくテキストセグメンテーション

文書内の小さな範囲の中で特定の語が繰り返し使用されることは稀であるが、言い換えた語やその関連語が使用されることは多い。そこで、語彙的連鎖の認定のために同一の語のみを対象とするのではなく、シソーラスを用いる手法が提案されている<sup>6),7)</sup>。しかし、既存のシソーラスは語彙不足のため語の関係が十分に反映されているとは言えない。また、「不況 → 倒産」のような連想関係も登録されていない。語彙的連鎖を認定するためには、類似関係にある語だけではなく、このような連想関係にある関連語もあわせて考慮する必要がある。

そこで、本稿では3.3節で述べた問題点の(1)を解決するために語の共起情報を利用する。ここで述べる語の共起とは、ある文書内で着目する語に対して他の語がその近傍で出現することである。そして、ある語

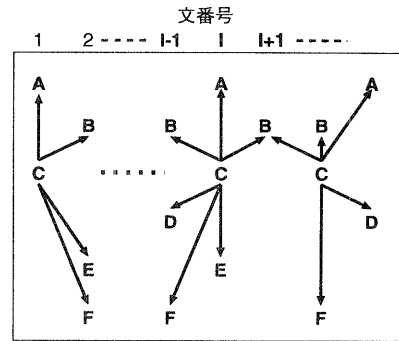


図3 文書内での単語共起の例

Fig. 3 An example of term co-occurrences in a document

に対してその近傍で頻繁に出現する語を共起語と呼ぶことにする。共起語と着目した語の間には何らかの関連がある場合が多く、この考え方は情報検索における検索語拡張にも適用されている<sup>10)</sup>。共起語は以下のように定義できる。

ある文書中において、着目する単語  $t$  が出現する文の番号を、 $l_1, l_2, \dots, l_i, \dots, l_n$  とするとき、すべての  $l_i$  に対して前後の文の番号  $l_i - 1, l_i + 1$  を考え、それらすべての集合を  $L$  と表す。

$$L = \{l_1 - 1, l_1, l_1 + 1, \dots,$$

$$l_i - 1, l_i, l_i + 1, \dots, l_n - 1, l_n, l_n + 1\}$$

$L$  に出現する単語  $t'$  のうち  $L$  における総出現頻度  $f_L(t')$  がしきい値  $N$  を超える単語を、 $u_1, u_2, \dots, u_j, \dots, u_k$  と表し、それらを単語  $t$  の共起語と呼ぶ。この手法を用いて文書に出現するすべての単語  $t$  に対して共起語を決定する。次に、3.2節で説明した窓に単語  $t$  が1回出現するごとに共起語である  $u_1, u_2, \dots, u_j, \dots, u_k$  も  $\alpha$  回出現したと考えて左右の窓における  $u_1, u_2, \dots, u_k$  の出現回数に  $\alpha$  を加算する。ただし、 $\alpha$  は共起語の重み係数 ( $0 \leq \alpha \leq 1$ ) である。

たとえば、図3で単語  $C$  に着目すると、 $N = 3$  とした場合には、 $A, B, F$  が共起語となり、 $C$  が1回出現すると  $A, B, F$  も同時に  $\alpha$  回出現したと考え、左右の窓における  $A, B, F$  の出現数に  $\alpha$  を加算し、(1)式の  $f(t_{wl})$  と  $f(t_{wr})$  を修正する。仮に、左窓に  $C$  が1回、右窓に  $C$  が2回出現した場合には、 $f(A_{wl}) + \alpha$  を左窓における  $A$  の出現回数、 $f(A_{wr}) + 2\alpha$  を右窓における  $A$  の出現回数と考え、類似度を計算する。

図4に示す同一の語のみの繰り返しによる計算では、類似度が0である基準点が連続しているため極

\* 他の共起語についても同様に  $C$  の出現回数に応じて  $\alpha$  を加算する。

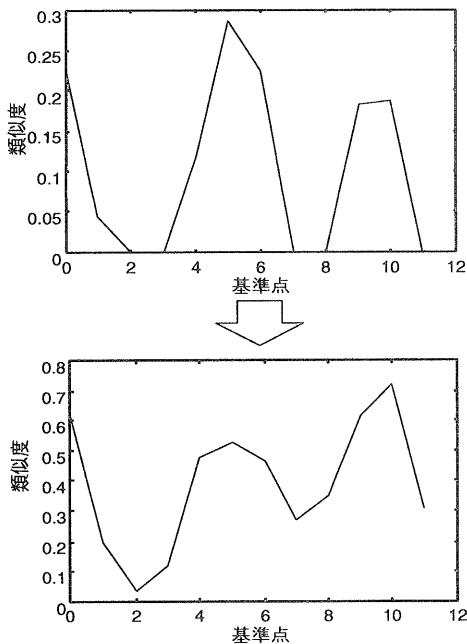


図4 共起を考慮したことによる類似度の変化

Fig. 4 Change of similarity using term co-occurrences

小点を認定できない場合にも、共起を考慮することによってそれを認定できるようになる。このように語の共起語を考慮することによって、短い文書にも適用できるようになる。

ただし、共起語を考慮することにより類似度が全体的に大きくなるため3.3節での(2)式、(3)式は不適切となる。本稿では、極小点  $mp$  において、グラフの窪みが大きい点を優先して話題境界とする手法をとる。以下に処理の流れを説明する。

ある区間における極小値  $mp$  に対して、左側の最大値  $S_{l_{mx}}$  と右側の最大値  $S_{r_{mx}}$  を用いて  $d$  を以下の式で定義する。

$$d = \min(S_{l_{mx}}, S_{r_{mx}}) - S_{mp} \quad (4)$$

$d$  を用いて以下の処理で話題境界を認定する(図5)。

**Step1** テキストセグメンテーションの対象となる区間  $D$  の初期値として文書全体を与える。

**Step2** 区間  $D$  内の全ての極小点  $mp$  に対して(4)式で  $d$  を決定する。

**Step3** Step2 で求めた  $d$  の最大値が、 $d \geq d_{th}$  を満たした場合に  $d$  を与える極小点  $mp$  を境界として、区間  $D$  を分割する。

**Step4** Step3 で分割したそれぞれの区間  $D_n$  に対して、Step2 ~ Step3 を繰り返す。

たとえば図5では、区間  $D_0$  における、極小点  $mp_1$ 、 $mp_2$ 、 $mp_3$  に対する左側の最大値  $S_{l_{mx}}$ 、右側の最大値

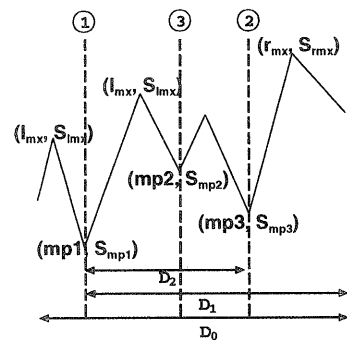


図5 話題境界の認定法

Fig. 5 A method for determining segment borders

$S_{r_{mx}}$  は図に示すとおりとなる。ここで、極小点  $mp_1$  における  $d$  が最大であり、かつ  $d_{th}$  を超えるため区間  $D_0$  を極小点  $mp_1$  で分割する。つぎに、区間  $D_1$ 、についての処理を行い、 $mp_3$  で区間  $D_1$  を分割する。区間  $D_2$  についても同様に  $mp_2$  で区間  $D_2$  を分割する。このように、ある区間内において最大でかつ  $d_{th}$  を超える  $d$  を与える極小点  $mp$  にて文書を分割するという手順でテキストセグメンテーションを行う。ただし、 $d_{th}$  はパラメータであり、実験により最適値を決定する。

## 4.2 単語重要度に基づくテキストセグメンテーション

本節では、3.3節で説明した問題点(2)を解決するために単語重要度に基づき計算した文の重要度の文書内での変化に着目したテキストセグメンテーション法を提案する。

### 4.2.1 単語重要度と文の重要度

まず、本稿で用いる単語重要度の算出法とそれを用いた文の重要度の定義について説明する。一般的に単語の重要度を求める方法は、ある文書に閉じて求める手法<sup>11)</sup>や文書集合全体から求める手法<sup>12)</sup>がある。本稿では、情報検索分野で広く用いられている手法である後者を用いる。文書集合DB中のある文書に出現する単語  $t$  の重要度  $W(t)$  は、単語  $t$  の出現文書頻度  $df(t)$  と文書集合DBの大きさ  $|DB|$  を用いて以下に定義される<sup>\*</sup>。

$$W(t) = \log \left( \frac{|DB|}{df(t)} \right) \quad (5)$$

文書に含まれる任意の1文  $s$  の重要度  $I(s)$  を以下のように定義する。 $K$  を  $s$  に出現するそれぞれの単語の出現頻度の総和とする。このとき、 $I(s)$  は  $s$  に出現するそれぞれの単語の重要度  $W(t)$  を用いて、以下の式

<sup>\*</sup> 式(5)は一般的に  $IDF$  といわれる。

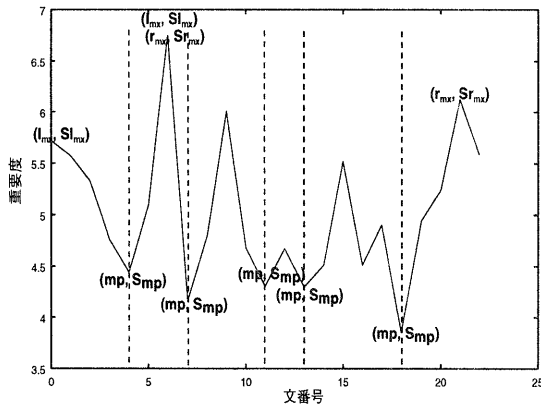


図6 単語重要度を利用したセグメンテーションの例

Fig. 6 An example of text segmentation based on word importance

で定義する。

$$I(s) = \sum W(t)/K \quad (6)$$

すなわち、 $s$  に出現する単語  $t$  の重要度の総和を  $s$  に出現する総単語数  $K$  で正規化したものである。

4.2.2 文の重要度変化に基づく語彙的連鎖の認定  
単語重要度に着目したテキストセグメンテーションの手法としては、重要語の連続部分を語彙的連鎖として捉え、隣接する単位(重要語)を結合する手法が提案されている<sup>8)</sup>。このように重要語は文書における話題認定のためののがかりとなる情報である。4.2.1 節では単語の重要度に基づいて算出する文の重要度について説明した。本節では、この文の重要度の変化に着目することで語彙的連鎖を認定する。

ここでは、以下の2つの仮定を導入する。

仮定1 文書中の各話題に対する重要文は隣接して出現する。

仮定2 話題境界となる文は重要度が低い。

これは、文書に複数の話題が存在する場合、各話題に対する重要文は隣接していると考え、重要文が続く部分を連鎖とし、重要度が低くなった部分で連鎖が切れるとすることである。4.2.1 節で定義した式(6)を用いて文の重要度を計算し、文書内における文の重要度をグラフに表した例として図6を示す。グラフにおいて、文の重要度の変化に着目し、重要度が極小値となる文で話題が切れると考える。

このように、文書中で文の重要度が変化する原因は、ある話題に対して重要個所が続いた後、話題が収束するに従い、一旦、重要度が低くなり、話題が変化することで再び重要度が増加することによると考える。

4.1 節で説明した、基準点に対する左右窓の類似性

表1 出力境界文番号の異なり例

Table 1 An example of outputs by text segmentation based on lexical cohesion and word importance

文書	語彙的結束性	単語重要度
1	5, 14	5, 11
2	10	2, 4, 10
3	7	4, 16
4	—	2, 4, 14
5	13	5, 13
6	8	5, 8
7	11, 14	7, 15
8	11	4
9	8, 13	15
10	—	4, 10

に着目するのではなく、文の重要度変化にのみ着目することによって、文書中の位置にかかわらず同一の信頼性でテキストセグメンテーションを行うことができる。よって、3.3 節で説明した問題点(2)を解決できる。

なお、 $d$  については、4.1 節と同様に求め、 $d$  のしきい値  $d_{th}$  をパラメータとして実験により最適値を決定する。

### 4.3 語彙的結束性と単語重要度を相補的に用いたテキストセグメンテーション

4.1 節で提案した手法では、文書内の単語の共起を考慮することによって3.3 節で説明した問題点(1)を解決したが、基準点に対して同数の単語を含む左右の窓の類似度に着目してテキストセグメンテーションを行っているため、問題点(2)を解決できない。一方で、4.2 節で提案した手法では、文書の先頭および末尾付近でも信頼性の高いテキストセグメンテーションを実現できる。

20 文程度からなる新聞記事 10 件\* を対象として、4.1 節で提案した語彙的結束性に基づく手法と 4.2 節で提案した単語重要度に基づく手法によりテキストセグメンテーションを行い、その出力に含まれる正解を分析した結果を表1に示す。表中の数字は文書の話題境界となる文の番号である。表1より、語彙的結束性に基づく手法では文書の先頭および末尾付近における正解を認定できていないことがわかる。一方、単語重要度に基づく手法は文書中の位置における制約を受けないため文書中の広い領域の正解を認定できている。さらに両手法に共通の正解が少ないこともわかる。これは、テキストセグメンテーションののがかりとする情報が異なるからであると考えられる。そこで、両手法を相補的に組み合わせることでより高精度なテキストセ

\* 5.1 節で説明する実験データよりランダムに抽出した新聞記事。

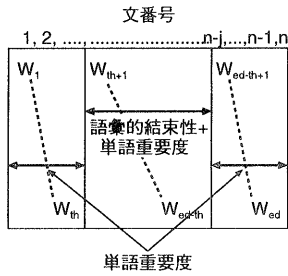


図7 語彙的結束性と単語重要度の統合

Fig. 7 The integration of method based on lexical cohesion and word importance

表2 実験データ

Table 2 Data used for the experiments

文書数	平均文数	平均話題境界数	平均単語数
326	19.2	5.06	142.4

グメンテーション法が実現できると考える。

つまり、文書の先頭および末尾から  $w_{lh}$  番目の単語が出現するまでの領域では、単語重要度に基づくテキストセグメンテーション法による出力を用い、それ以外の文書半ばの領域では、互いに出力傾向の異なる語彙的結束性に基づくテキストセグメンテーション法と単語重要度に基づくテキストセグメンテーション法の出力の和集合を用いる (図7)。

## 5. 実験

4章で説明した各提案手法を実装し、テキストセグメンテーションの従来手法である TextTiling 法との精度比較による評価実験を実施した。実験データと実験方法について以下に説明する。

### 5.1 実験データ

評価実験の対象としては、検索精度評価用テストコレクション *BMIR-J2* <sup>☆13)</sup> に収録されている新聞記事から、文数が20文程度<sup>☆☆</sup>のランダムに選んだ326記事とし、その形式段落を正解データとして用いた。表2に実験データの詳細を示す。

新聞記事は、CD-ROMで公開されているために電子化データの入手が用意であり、書き手が熟練者であるために文書の質が高いとも言える。以上の理由より、評価対象として新聞記事を利用した。

また、新聞記事の形式段落を正解データとして用い

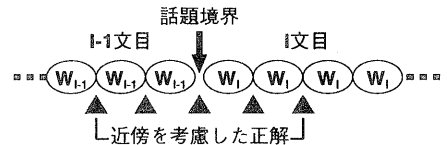


図8 話題境界の近傍を正解に含めた評価

Fig. 8 The Evaluation in which proximity of segment borders are regarded as correct

た理由としては、記事における意味的なまとまりが明確であり形式段落と意味段落が一致すると考えることと人手により正解データを作成することが困難であることによる。人手により正解データを作成する手法としては、複数の被験者によるテキストセグメンテーションの実験を行い、その結果から被験者間で一致の割合が高い結果を正解データとする手法がある。しかし、このような手法では人と時間のコストが多にかかるため、大規模な正解データを作成することは困難である。さらに、テキストセグメンテーションの個人ごとの揺れが予想されるため、妥当性のある正解データが作成されるとも限らない。よって、本稿では書き手に対して信頼のおける新聞記事を用い、その形式段落は記事中の話題と対応しているものと考えて形式段落を正解データとして用いることにした。なお、新聞記事の形式段落を正解データとすることの妥当性については、6.2節でさらに考察する。

処理対象単語としては、日本語形態素解析システム「茶釜」<sup>14)</sup>を使用した結果から名詞および未定義語を用いた。

### 5.2 評価指標

評価指標としては、情報検索分野で一般的に用いられる適合率 (Precision)、再現率 (Recall) を用いる。適合率、再現率は以下の式で求められる。

$$\text{適合率} = \frac{\text{出力結果に含まれる正解境界数}}{\text{出力境界数}} \quad (7)$$

$$\text{再現率} = \frac{\text{出力結果に含まれる正解境界数}}{\text{正解境界数}} \quad (8)$$

さらに、上記の適合率 (P)・再現率 (R) を用いて、F-measure でもあわせて評価する。

$$F\text{-measure} = \frac{2PR}{P+R} \quad (9)$$

また、TextTiling 法では基準点の移動を文単位ではなく単語単位で行っている。そのため、その出力は必ずしも文末にはならず、文の途中に現れることが多くなる。そこで、評価の際には話題境界の候補とする出力に対して前後数単語以内に実際の話題境界が含まれていれば正解としている。図8の例で仮に  $l-1$  文目

☆ (社)情報処理学会・データベースシステム研究会が、新情報処理開発機構との共同作業により、毎日新聞 CD-ROM'94 データ版を基に構築した情報検索システム評価用テストコレクションを使用。

☆☆ 複数の話題を含んでいることを想定して設定した。

表 3 実験パラメータ

Table 3 Parameter used for the experiments

	語彙的結束性	単語重要度	ハイブリッド
$d_{lth}$	0.01 ~ 0.20	-	-
$d_{wth}$	-	0.1 ~ 2.0	-
words	20 ~ 60	-	-
$w_{th}$	-	-	20 ~ 60

の終りが実際の話題境界であった場合に、テキストセグメンテーションの出力に対して 2 単語以内に話題境界が含まれている場合を正解とする。すると、 $l-1$  文目の末尾から 2 単語目から  $l$  文目の先頭から 2 単語目までのいずれかが出力された場合に正解となる。このように実際の話題境界に対してその近傍での出力も正解として評価している。本稿で評価する手法は基準点の移動を 1 文単位で行い、話題境界の候補となる文を出力しているが、基準点の移動を単語単位で行う場合と同様にその出力の近傍に正解がある場合も評価するため、実際の話題境界の正解を表す文が  $x$  番目の文であった場合に、 $x-1$  番目の文を出力した場合も正解とする。

### 5.3 評価実験

評価では cross validation により、326 文書を 10 分割し、9/10 の文書を学習用データセット、残りの 1/10 文書を評価用データセットとする。評価は 10 分割したそれぞれのデータセットに対して行う (図 9)。

学習の際に推定するパラメータは、語彙的結束性に基づく手法では 4.1 節で説明したしきい値 ( $d_{lth}$ )、窓が包含する単語数 (words)、単語重要度に基づく手法では 4.2 節で説明したしきい値 ( $d_{wth}$ )、さらに語彙的結束性と単語重要度を相補的に組み合わせた手法 (以下、ハイブリッド手法) では、4.3 節で説明した語彙的結束性を用いた手法に制限を設ける場合のしきいとなる語数  $w_{th}$  である。TextTiling 法における窓が包含する単語数については、語彙的結束性に基づく手法と同じ単語数 (words) とし、 $d_{th}$  については式 (3) により計算した結果を利用した。なお、各手法におけるパラメータは表 3 に示す通りに変化させた。

#### 5.3.1 語彙的結束性を利用した手法の評価

10 通りの学習用データセットを用いて決定したそれぞれのパラメータ (words,  $d_{lth}$ ) での評価結果とその平均を表 4 に示す。カッコ内の数値は形式段落の近傍での出力も正解に含めた場合の値である。各評価用データセットに対して、再現率より適合率が高くなっているため、2 章で述べた前提条件を満たしていることがわかる。しかし、再現率は適合率に比べて大幅に低くなっている。4.2 節で説明したしきい値  $N$ 、共起

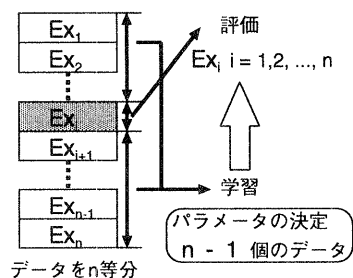


図 9 Cross validation

Fig. 9 Cross validation

表 4 Cross validation による語彙的結束性を利用した手法の評価結果

Table 4 The results of text segmentation based on lexical cohesion by cross validation

データセット	$d_{lth}$	words	適合率 (%)	再現率 (%)
1	0.1	30	40.7 (72.4)	16.9 (31.3)
2	0.1	53	17.4 (64.4)	7.27 (22.4)
3	0.1	29	28.2 (68.5)	7.53 (20.8)
4	0.1	31	31.0 (56.0)	11.4 (22.6)
5	0.1	50	56.5 (84.8)	10.3 (22.9)
6	0.1	42	38.1 (69.6)	9.50 (19.8)
7	0.1	30	38.9 (78.9)	13.7 (28.6)
8	0.1	53	42.4 (74.6)	10.1 (24.0)
9	0.1	46	22.2 (65.3)	8.16 (23.0)
10	0.1	41	48.1 (77.6)	10.5 (22.5)
平均			36.4 (71.2)	10.5 (23.8)

語の重み係数  $\alpha$  は経験的に  $N = 6$ ,  $\alpha = 0.3$  に設定した。なお、左右の窓は words に最も近い単語数を含む文単位の領域で設け、基準点の移動は 1 文単位で行った。

#### 5.3.2 単語重要度を利用した手法の評価

10 通りの学習用データセットを用いて決定したパラメータ ( $d_{wth}$ ) での評価結果とその平均を表 5 に示す。カッコ内の数値は形式段落の近傍での出力も正解に含めた値である。5.3.1 節の結果と同様、各評価用データセットに対して、再現率より適合率が高くなっている。適合率では語彙的結束性に基づく手法より若干低い値であるが再現率では上回っている。

#### 5.3.3 ハイブリッド手法の評価

10 通りの学習用データセットを用いて決定したパラメータ ( $w_{th}$ ) での評価結果とその平均を表 6 に示す。カッコ内の数値は形式段落の近傍での出力を正解に含めた場合の値である。また、語彙的結束性に基づく手法の words,  $d_{lth}$ 、単語重要度に基づく手法の  $d_{wth}$  については、それぞれの評価実験で決定した値 (表 4, 表 5) を用いた。5.3.1 節、5.3.2 節の結果と同様、適合率が再現率より高くなっている。適合率は、語彙的



表5 Cross validation による単語重要度を利用した手法の評価結果

Table 5 The results of text segmentation based on word importance by cross validation

データセット	$d_{wth}$	適合率 (%)	再現率 (%)
1	1.9	51.1 (67.0)	19.0 (29.8)
2	1.5	26.7 (58.5)	13.7 (28.2)
3	1.5	25.6 (57.9)	15.4 (34.0)
4	1.5	29.5 (60.9)	19.6 (39.7)
5	1.6	22.2 (51.7)	12.2 (25.1)
6	1.5	37.0 (58.4)	23.0 (37.7)
7	1.5	40.5 (68.3)	19.0 (35.0)
8	1.5	31.3 (72.2)	17.2 (42.2)
9	1.5	28.8 (56.2)	19.5 (39.3)
10	1.6	28.4 (67.2)	19.6 (40.7)
平均		32.1 (61.2)	17.8 (35.2)

表6 Cross validation によるハイブリッド手法の評価結果  
Table 6 The results of hybrid text segmentation by cross validation

データセット	$w_{th}$	適合率 (%)	再現率 (%)
1	47	41.2 (67.4)	25.3 (44.1)
2	32	31.1 (61.5)	20.6 (41.7)
3	41	24.4 (59.8)	16.5 (39.0)
4	45	27.7 (58.1)	22.7 (42.5)
5	38	22.5 (52.5)	13.3 (31.8)
6	41	36.8 (64.2)	26.0 (46.9)
7	52	38.1 (72.2)	20.8 (42.9)
8	37	32.4 (71.6)	22.1 (48.0)
9	35	29.1 (64.2)	24.6 (53.6)
10	37	34.9 (75.8)	21.5 (44.2)
平均		31.8 (64.7)	21.3 (43.5)

結束性に基づく手法、単語重要度に基づく手法と比較すると共に低いが、再現率では提案手法の中で最も高い値である。なお、語彙的結束性を利用した手法と単語重要度を利用した手法の使いわけは、 $d_{wth}$ に最も近い単語数を含む文単位の領域で行った。

### 5.3.4 TextTiling 法の評価

5.3.1 節で決定したパラメータ  $words$  での評価結果とその平均を表7に示す。カッコ内の数値は形式段落の近傍での出力を正解に含めた場合の値である。形式段落との完全一致ではどの提案手法と比較しても最も低い値であるが、形式段落の近傍も正解に含めることにより大幅に精度が向上している。なお、左右の窓は  $words$  に最も近い単語数を含む文単位の領域で設け、基準点の移動は1文単位で行った。

## 6. 考 察

5章での評価実験の結果から提案手法の有効性について考察する。

表7 Cross validation による TextTiling 法の評価結果  
Table 7 The results of TextTiling method by cross validation

データセット	$words$	適合率	再現率
1	30	25.7 (59.1)	14.0 (36.7)
2	53	29.6 (59.8)	16.8 (36.2)
3	29	30.4 (62.8)	17.3 (39.9)
4	31	33.5 (64.4)	19.4 (42.4)
5	50	32.3 (57.7)	14.8 (35.0)
6	42	26.5 (56.4)	17.3 (34.9)
7	30	30.5 (61.1)	18.1 (41.2)
8	53	31.5 (64.4)	12.8 (30.6)
9	46	40.2 (60.0)	21.0 (40.1)
10	41	31.9 (60.4)	14.9 (38.9)
平均		31.0 (61.6)	14.4 (37.5)

### 6.1 精度について

一般に、 $p$ 文からなる文書において、正解となる境界が  $q$  個である場合に、 $r$  個の境界を正解の候補として出力した場合、その出力中における正解数の期待値  $E$  は以下の式で表される。

$$E = \sum_{m=1}^{\min(q,r)} \frac{m \times {}_q C_m \times {}_{p-1-q} C_{r-m}}{{}_{p-1} C_r} \quad (10)$$

このとき、適合率・再現率の理論値はそれぞれ以下の式で計算できる。

$$\text{適合率} = \frac{E}{r}, \quad \text{再現率} = \frac{E}{q} \quad (11)$$

ただし、正解となる境界に対してその近傍を正解に含める場合には、正解となる話題境界の配置によって同数の出力でも期待値が異なるため、(11)式のように定式化できない。よって、正解となる境界に対してその近傍を正解に含めた場合の適合率・再現率の理論値は出力の全ての組合せを数えあげることでも求める。

今回の評価実験において各手法が出力した話題境界候補数の平均  $q$ 、正解の話題境界数の平均  $r$  を用いて (11) 式で求めた適合率・再現率の理論値の平均と正解となる境界に対してその近傍を正解に含めた場合の適合率・再現率の理論値の平均、および各手法の話題境界候補の平均出力数、適合率・再現率の実験値を表8、表9に示す。さらに、表8と表9より求めた各手法の F-measure を表10に示す。

#### 6.1.1 形式段落と完全一致の場合

表8より、各手法ともに適合率が再現率を上回っており、2章で述べた前提条件は満たしている。適合率で比較した場合では、語彙的結束性に基づく手法、単語重要度に基づく手法の順で優れていることがわかる。語彙的結束性に基づく手法の適合率は、TextTiling法と比較して約5%の精度向上がみられ、単語間の共起を

表 8 各手法における適合率・再現率の実験値と理論値

Table 8 The theoretical and practical value of precision and recall in each method

手法	平均出力数	実験値		理論値	
		適合率 (%)	再現率 (%)	適合率 (%)	再現率 (%)
TextTiling	2.75	31.0	14.4	27.2	16.6
語彙的結束性	1.69	36.4	10.5	23.0	8.92
単語重要度	2.92	32.1	17.8	28.2	15.3
ハイブリッド	3.48	31.8	21.3	27.3	18.2

表 9 各手法における近傍を正解に含めた場合の適合率・再現率の実験値と理論値

Table 9 The theoretical and practical value of precision and recall of in each method in the case proximity of segment borders regarded as correct

手法	実験値		理論値	
	適合率 (%)	再現率 (%)	適合率 (%)	再現率 (%)
TextTiling	61.6	37.5	49.4	28.7
語彙的結束性	71.2	23.8	49.2	26.2
単語重要度	61.2	35.2	49.4	28.8
ハイブリッド	64.7	43.5	49.4	28.8

表 10 各手法の F-measure

Table 10 The value of F-measure in each method

手法	近傍を正解に含めない	近傍を正解に含める
TextTiling	19.7(%)	46.6(%)
語彙的結束性	16.3(%)	35.7(%)
単語重要度	22.9(%)	44.7(%)
ハイブリッド	25.5(%)	52.0(%)

利用したことによる効果が確認できる。さらに適合率の理論値との比較では、TextTiling 法が約 3% 上回っていることに対して約 13% 上回っており既存手法より優れていることがわかる。実験データからランダムに抽出した 65 件の記事について、得られた共起関係を目視で調査したところ、連想、類似、上位/下位関係をあらわす有効な共起関係は約 30% であり、無意味な共起関係が多く含まれていることがわかった。よって、共起を考慮して左右の窓の類似度を計算する際には、共起語の重み係数  $\alpha$  を低く設定しなければテキストセグメンテーションの精度に悪影響をおよぼすことが予想される。本論文では  $\alpha = 0.3$  と設定したことで無意味な共起関係の影響をおさえることができたと考えられる。

単語重要度に基づく手法の適合率は、TextTiling 法と比較して約 1% の精度向上がみられ、適合率の理論値に対しては約 4% 上回っている。TextTiling 法に対してはわずかに精度が向上している。ハイブリッド手法の適合率は、TextTiling 法と比較して 1% 弱の精度向上しかみられないが、適合率の理論値に対しては約 5% 上回っている。

再現率で比較した場合は、ハイブリッド手法、単語重要度に基づく手法の順で優れていることがわか

る。ハイブリッド手法の再現率は TextTiling 法と比較して約 7% の精度向上がみられる。再現率の理論値との比較では、TextTiling 法が約 2% 下回っていることに対して約 3% 上回っており既存手法より優れていることがわかる。単語重要度に基づく手法も TextTiling 法と比較して約 7% の精度向上がみられ、再現率の理論値との比較でも約 2.5% 上回っている。TextTiling 法と語彙的結束性に基づく手法の実験値が理論値を下回っていることから、単語重要度に着目することの有効性がわかり、4.2.2 節で導入した仮定に妥当性があったこともわかる。

さらに、適合率・再現率をもとに計算した F-measure (表 10) でも、ハイブリッド手法が最も優れていることがわかる。TextTiling 法と比較して、約 6% の精度向上がみられる。ハイブリッド手法と TextTiling 法と比較した場合には、TextTiling 法よりも出力が多いにもかかわらず、適合率・再現率ともに精度が向上することからその有効性がわかり、4.3 節で説明した語彙的結束性に基づく手法と単語重要度に基づく手法の統合に妥当性があったことがわかる。

以上、適合率による比較からは単語間の共起を利用した語彙的結束性に着目することの有効性を確認できた。再現率による比較からは単語重要度に着目する

表 11 人手によるテキストセグメンテーションの評価結果  
Table 11 The results of text segmentation by human's judge

文書数	平均出力数	適合率 (%)	再現率 (%)	F-measure (%)
30	4.21	12.0 (75.6)	9.50 (55.7)	10.6 (64.1)

ことの有効性を確認できた。さらにハイブリッド手法はその出力数が TextTiling 法よりも多いにもかかわらず、適合率・再現率ともに TextTiling 法を上回り、F-measure においても最も高精度であることより語彙的結束性と単語重要度を相補的に組み合わせることで高精度なテキストセグメンテーション法を実現できることがわかった。

### 6.1.2 形式段落の近傍での出力を正解に含めた場合

表 8 に対して、話題境界の正解に対しその近傍での出力も正解に含めた結果が表 9 である。表 9 も表 8 と同様の傾向を示しているが、正解近傍の出力を正解に含めることによって適合率では 29%～35%、再現率では 13%～23%精度が向上している。

表 8 と同様に、適合率で比較した場合には語彙的結束性に基づく手法、再現率で比較した場合にはハイブリッド手法が最も優れていることがわかる。F-measure による比較でもハイブリッド手法が最も優れている。表 8 では各手法と比較して最も低い値であった TextTiling 法も形式段落の近傍での出力を正解に含めることによって大幅に精度が向上しているが、適合率・再現率、F-measure のすべてにおいてハイブリッド手法が上回っており、ハイブリッド手法の有効性がわかる。

以上、適合率・再現率による比較、適合率・再現率の理論値に対する比較、F-measure の比較から、2 章で述べた前提条件のもとで、表 8 に対する考察と同様、ハイブリッド手法が TextTiling 法より優れていることが確認できた。

話題境界の正解として形式段落の近傍での出力も含めて評価したことにより、適合率・再現率は大幅に向上した。これは、形式段落との完全一致では話題境界の認定に失敗する場合でもその近傍を話題境界として認定していることが多いことを示している。本稿では、2 章でも述べたとおり比較的短い文書を対象としているが、このような文書では、文書中の各話題に重なりが生じている可能性がある。よって、明確な話題の境界が認定できなくとも、実際の話題境界の近傍を認定できることは有効である。

20 文程度の新聞記事に対して提案手法の有効性を示せたことで、短い文書や文書領域に対して有効なテキストセグメンテーション法を提案できた。提案手法

を章や節などの文書構造を表す情報と組み合わせることによって細かい話題の粒度に対応したテキストセグメンテーションが行える。よって、様々な文書処理技術などに対応でき、それらの精度向上にも役立つことが期待できる。

### 6.2 正解データと評価法の妥当性について

評価実験における話題境界の正解データとして、新聞記事の形式段落を用いたが、形式段落と意味段落は必ずしも一致するわけではない<sup>7)</sup>。しかし、意味段落の正解データを人手で作成する場合には、5.1 節で述べたとおり、話題分割に対する人間の判断の揺れや時間と手間が多くかかることにより困難である。本節では、人手によるテキストセグメンテーションの結果から、形式段落の近傍を正解として評価することに対する妥当性と話題境界の正解として新聞記事の形式段落を用いたことの妥当性を考察する。

表 2 に示した実験データよりランダムに 30 件の文書を抽出した。これらの文書を対象として被験者 24 名にテキストセグメンテーションを行なってもらい、その結果を評価した。評価指標としては、5.1 節で説明した指標を用いた。評価結果の平均を表 11 に示す。なお、形式段落の情報を取り除いた文書を対象として、設ける話題数の制限などは行わず、各個人の独断でテキストセグメンテーションを行ってもらった。カッコ内の数値は形式段落の近傍での出力も正解に含めた場合の値である。

表 11 より、形式段落との完全一致の場合では、適合率 12%、再現率 9.5%と非常に低い値であるが、形式段落の近傍での出力も正解に含めた場合には、適合率 75.6%、再現率 52.2%と大幅な精度の向上がみられた。このことから、新聞記事の形式段落の近傍で人間は話題が変わったと判定していることがわかる。記事の形式段落に対して完全に一致せず、近傍で一致する理由については、被験者は文書を書くために特別な訓練などを受けていないため、話題境界の認定に曖昧性があること、短い文書においては、文書中の話題に重なりが生じていることなどによって話題境界が曖昧な場合があることによると考える。よって、新聞記事の形式段落に対して、その近傍も正解に含める評価することには妥当であったと考える。

記事に設ける話題の数に関しては、新聞記事の話題

境界数が平均約5個であることに対して人間が判断した場合には平均約4.2個であった。被験者と記事の書き手の間に話題の粒度に対する認識が異なることが原因となり、実際の記事の話題数と被験者が設けた話題数に違いが生じたと考える。つまり、被験者の考える話題の粒度は記事の書き手の考える話題の粒度よりも、若干、粗いと考える。被験者と記事の書き手の間に話題の粒度に関して若干の違いはあるものの被験者が判断する話題の境界がほぼ形式段落の近傍であることから新聞記事の形式段落を話題境界の正解として用いたことは妥当であったと考える。

## 7. おわりに

本稿では、比較的短い文書を対象として、適合率を重視したテキストセグメンテーション法を提案した。既存手法の問題点を解決するため、単語の共起を考慮した語彙的結束性に基づくテキストセグメンテーション法と単語重要度を利用した文の重要度の変化に着目したテキストセグメンテーション法とを提案した。さらに、両手法を相補的に統合することで、高精度なテキストセグメンテーション法を実現できた。新聞記事を対象としてその形式段落を正解データに用いて評価実験を行い、適合率・再現率で評価した結果、形式段落と完全一致の場合には、適合率31.8%、再現率21.3%、形式段落の近傍を正解に含めた場合には、適合率64.7%、再現率43.5%であった。既存手法であるTextTiling法との比較では、形式段落との完全一致の場合には、適合率で0.8%、再現率で6.9%、形式段落の近傍を正解に含めた場合には、適合率で3.1%、再現率で6%の精度向上がみられ、その有効性を確認した。

今後の予定として、新聞記事以外の文書での有効性の確認、さらに情報検索等の文書処理へ組み込んだ場合の精度向上への寄与の確認を考えている。

## 参考文献

- 1) 望月源, 本多岳夫, 奥村学: 語彙的連鎖を利用した文書検索, 情報処理学会研究報告 NL-117-19, pp. 137-144 (1997).
- 2) 望月源, 岩山真, 奥村学: 抄録を利用した検索, 言語処理学会 第4回年次大会ワークショップ論文集, pp. 22-29 (1998).
- 3) 望月源, 岩山真, 奥村学: 語彙的連鎖に基づくパッセージ検索, 情報処理学会研究報告 NL-127-6, pp. 39-46 (1998).
- 4) 平尾努, 木谷強: 単語の重要度に基づくテキストの要約, 情報処理学会研究報告 FI-49-6, pp. 41-47 (1998).
- 5) 仲尾由雄: 文書の意味的階層構造の自動認定に基

づく要約作成, 言語処理学会 第4回年次大会ワークショップ論文集, pp. 72-79 (1998).

- 6) 望月源, 本田岳夫, 奥村学: 複数の知識の組合わせを用いたテキストセグメンテーション, 情報処理学会研究報告 NL-109-7, pp. 47-54 (1995).
- 7) 本田岳夫, 奥村学: 語彙的結束性に基づいたテキストセグメンテーション, 情報処理学会研究報告 NL-102-4, pp. 25-32 (1994).
- 8) 西沢信一郎, 中川裕志: 名詞の文書内頻度を利用したテキストセグメンテーション, 情報処理学会研究報告 NL-117-20, pp. 145-152 (1997).
- 9) Hearst, M. A.: TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, *Association for Computational Linguistics*, Vol. 23, No. 1, pp. 111-112 (1997).
- 10) Xu, J. and Croft, W. B.: Query Expansion Using Local and Global Document Analysis, *Proceedings of the 19th ACM SIGIR*, pp. 4-11 (1996).
- 11) 原正巳, 中島浩之, 木谷強: テキストのフォーマットと単語の範囲内重要度を利用したキーワード抽出, 情報処理, Vol. 38, No. 2, pp. 299-309 (1997).
- 12) Salton, G. and McGill, M.: *Introduction to Modern Information Retrieval*, McGrawHill, New York (1983).
- 13) 木谷強ほか: 日本語情報検索システム評価用テストコレクション (BMIR-J2), 情報処理学会研究報告 DBS-114-3, pp. 15-22 (1998).
- 14) 松本裕治, 北内啓, 山下達雄, 今一修, 今村友明: 日本語形態素解析システム『茶釜』version 1.0 使用説明書, Information Science Technical Report NAIIST-IS-TR97007, Nara Institute of Science and Technology (1997).

(平成11年12月20日受付)

(平成12年3月30日採録)

(担当編集委員 安達 淳)

平尾 努 (正会員)



1995年関西大学工学部電気工学科卒業。1997年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了。同年、NTTデータ通信(株)(現、(株)NTTデータ)入社、

現在に至る。自然言語処理、情報検索に関する研究開発に従事。

**北内 啓**

1996年京都大学理学部数学科卒業。1998年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了。同年、NTTデータ通信(株)(当時)入社、現在に至る。自然言語処理の研究に従事。

**木谷 強 (正会員)**

1960年生。1983年慶応義塾大学工学部電気工学科卒業。同年、日本電信電話公社(現NTT)入社。1988年NTTデータ通信(株)に転籍。形態素解析、情報抽出、情報検索に関する研究開発に従事。現在、(株)NTTデータ技術開発本部北米技術センタ部長。博士(工学)。平成10年度坂井記念特別賞受賞。