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

2005－NL－166（4）
2005／3／10

語の近接性に基づいた意味段落境界の判定手法

中野滋徳[†]　足立顕[††]　牧野武則[†]

本稿は語と語の近接性に着目して語彙結束度を求め，意味的にまとまった段落（意味段落）に分割する手法について提案する．ことばの組み合わせが無数にあるのに文章表現がばらばらに遊離せずにまとまりをもつ働きには，文に生起する語と語の近接する位置的関係にあるという仮説のもとに，文間結束度，話題結束度を求め，これらが意味結束度を表すとして意味段落境界の判定を行った．タイトル部を除去した社説記事を連結し，記事の区切りが本手法による境界判定と一致したときを正解とする評価実験を行った．その結果，再現率で 63.9%，適合率で 27.1%の結果を得た．本手法は小さなテキストに対しても有効であり，小説に対しても適用し考察を行った．

キーワード: 意味段落，結束性，近接性

Text segmentation based on nearness of words

Shigenori Nakano [†] Akira Adachi [††] and Takenori Makino[†]

This paper presents a new method of text segmentation using lexcical cohesion based on nearness of words in texts. Although sentences are written by innumerable combination of words, they are able to make up meaningful contents without confusion. We attempted to segment a text into paragraphs that contains topics on the hypothesis that nearness of words between sentences generates a settled content of text. Each splitted point between articles which are connected continuously indicates to be a boundary of contents. The method is applied to testify how many boundaries are identified with splitted points. The result of test by the mothod showed 63.9% in recall and 27.1 % in precision. This method is also effective for small texts.

Keywords: *text segmentation, cohesion, nearness of words*

1. はじめに

本稿は，テキスト中で生起する語の位置的な近接関係を尺度化した語彙的結束度に基づき，意味段落境界の判定手法について述べる．

テキストを意味的にまとまった段落に分割することで，文構造が把握しやすくなり文書理解を支援することができる．また情報検索技術やテキスト要約技術の応用にも期待ができる．

意味的にまとまった単位に分割するためには，そのまとまりを作っている要素を明らかにする必要がある．実際にことばの組み合わせが無数にあるのに，文章表現がばらばらに遊離せずにまとまりをもった文脈を作るのは，文を構成することばに依存していることは明らかである．

林[1)]は「この文脈とほとんど同義語で CONTEXT という言葉が使われているが，text は texture（織物）と同根で，con の『結び着け』により『糸を組んで織りなす』といった原義から,『言葉を結びつけて意味を織りなす』のような意味になってきたのがコンテクストであり文脈であろう」と述べている．この「糸」の役割を担うものが語の近接性にあるとして，同一語句が文間をまたがる位置的関係に焦点

† 東邦大学 大学院理学研究科 情報科学専攻
Department of Information Sciences Toho University
†† 富士通株式会社 報道メディアシステム統括部
Division of Media Solution, Fujitsu Co. Limitted

を当てた．

テキスト中に生起する語の位置的近接関係によって文と文とを引き合う牽引力が発生する．その結果，語彙的結束度を生むという仮説を設けた．意味段落境界の判定実験を通して本手法の有効性を示す．評価実験には新聞社説を対象とし，小説にも適用して考察を行った．

本稿は，2節で意味段落の位置づけを確認し，3節で段落分けに関する関連研究について述べる．4節で語の位置的近接関係によって尺度化した文の結束モデルの説明を行い，このモデルに基づいた実験の詳細を5節6節で紹介する．7節で考察を行い8節でまとめる．

2. 意味段落

時枝が「文章論」[2]を提唱して以来，文章と文との間に存在するひとかたまりの話題を表す中間的単位として,「文段」[3]「論理的段落」[4]「文塊」[1]等の定義がなされてきた．一方，国語教育では改行一字下げで目に確認できる段落を「形式段落」と呼び，この形式段落の単位で意味を重視した段落の整理に「意味段落」が使われてきた．

しかしこれらの単位に明確なコンセンサスが得られていないのが現状である．したがって本稿では，意味段落をテキストにおける「ひとかたまりの話題を表す中間的単位」の境界に判別することを指し，これを「意味段落」に分割すると呼ぶ．以後，意味段落を単に段落と呼ぶこととし，形式段落と区別する必要があるときに意味段落と表記する．

段落分割の研究には，語彙的結束性によるものと手がかり語によるものがある．日本語の場合，特に文脈上で理解できることは省略されるのが普通で，文そのものが必ずしも文法的に適格であるとは限らないという現実がある．一方，文脈を維持する上で必要な言葉は繰り返し提示するという特徴がある．そこで境界判定手法としては語彙的結束性に限定した．

実験では新聞記事社説を対象とする．新聞記事は1行13字組みを前提とする関係から，なるべく10行以内で改行するという原則[5]がある．形式段落が修辞的側面をもっている[6]ことを認識した上で，語彙的結束性の適用は文単位に行い，形式段落は少なくとも意味段落の候補であるという立場をとった．

3. 関連研究

段落分割の手法としてシソーラスによる類義語に着目した研究や共起情報を組み込んだ研究，さらに語の類似性に基づいた研究が報告されている．

本田[7]らは，テキスト中の意味的に関連がある語の集まりに対して，シソーラス上での類義関係が連続して出現する部分を連鎖としてとらえ，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，その総和より段落境界を推定する試みを行っている．

豊浦[8]らは，文を越えて係り受け関係にある語群を文脈的に同じ話題に関する結束性がそこに存在するとして段落の推定を行っている．

平尾[9]らは，左右に一定の単語幅をもった窓を設け，左右の窓に出現する語の類似度のスコアと，前後の文に共起して出現する共起語を考慮したスコアから，語彙的結束性と共起語を統合した段落分割手法を提案している．

仲尾[10]の語彙的結束性に基づく話題の階層構成を求める研究でも，ある一定の結束度計算用窓においてどのくらい同一語彙が出現するかを測定し，窓部分の類似性から話題単位の境界認定を行っている．話題の階層性を推定する点で特徴をもっている．しかし窓単位による測定のため，文書の先頭や末尾が正確に処理できないという問題が残る．

本手法はこれまでとは異なった新しい方法で，極めて単純に語彙的結束性が求まる点に特徴があり，比較的小さな単位のテキストに対しても有効である．

4. 文の結束モデル

意味的にまとまりをもった意味段落を解析するにあたり文をまたがった文脈を扱うことになる．この文と文とによって表される文脈には，大きく整合性(coherence)と結束性(cohesion)の概念がある[11]．整合性とは文と文との論理関係を表し，結束性とは文と文とのつながりを明示する表層的な結びつきを表し，照応関係，接続表現，文の情報構造などの要素から成り立つ．

本稿では情報構造の側面に焦点を当て，2つの言語的背景をもとに当モデルを考案した．

> (1) 文章は情報伝達メカニズムをもつ．
> (2) 主題の存在するところに反覆語句が多く出現する．

前者から文間結束度を導き出し，後者から話題結束度を導き出した．そしてこの2つの総和を意味段落を同定する結束度とし，本稿ではこれを意味結束度と呼ぶ．

意味結束度 ＝ 文間結束度 ＋ 話題結束度

それぞれの内容を個々に述べる．

4.1 文間結束度

北原[13] は,「言葉による表現は多くの場合，情報を伝達するために用いられるもので，情報伝達においては，ある前提があって，その前提のもとに未知の情報が新しく伝達される」と述べている．

つまり冒頭部分を除いては，既出の文の情報（語句）を前提としてそれに新しい情報（語句）を組み込みながら伝達情報がなされる．ある話題から別の話題に転換するときには必ず新しい情報が提供され，話題転換後に元の話題に戻るときには，その話題に関係する既出の語句を提示するというメカニズムがある．

この情報伝達メカニズムに着目すると，同一の反覆語句をもった文間には，前文の既知情報（語句）を引用し後文に引用されるという，語句の近接する関係が文と文との結びつきの強さを示すことになる．

図 1 に示す例では，文 i の「公益」（文 i に属する語のサフィックスを j とする）と同一語句は，文 i を起点にしてテキスト先頭方向に最も近い文 f_j と，末尾方向に最も近い文 r_j とに存在している．この「公益」に関する文間の結びつきの強さは，文 f_j と文 r_j との文間距離で定まる．

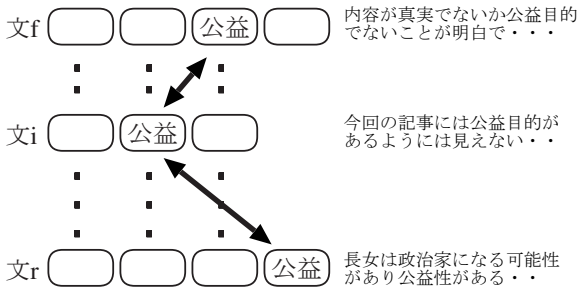

図 1 文間の結束を示す事例

そこでこの文間の結びつきの強さを，同一語句が表れる文間距離の二乗に反比例するとして尺度化した．この尺度化したものを文間結束度と呼ぶこととする．

文間結束度の求め方を示す．

文 i に含まれる語句を W_{ij} とするとき，語句 W_{ij} による文間結束度を P_{ij} とし，文 i 全体の文間結束度を P_i とする．

W_{ij} と同一の語句をもった，テキスト先頭方向で最も近い距離に位置する文 f_j とテキスト末尾方向で最も近い距離に位置する文 r_j とすると，W_{ij} によってもたらされる文間結束度 P_{ij} は式 (1) を用いて算出する．

但し，文間距離 $i - f_j$，$r_j - i$ には上限 (これを α とする) を設ける．その上限については実験により最適値を定める．

$$P_{ij} = \frac{1}{(i - f_j)^2} + \frac{1}{(r_j - i)^2} \qquad (1)$$
$$\text{但し，} (i - f_j \leq \alpha\ ,\ r_j - i \leq \alpha)$$

さらに文 i の文間結束度 P_i は式 (2) を用いて算出する．n は文 i に含まれる語の数である．

$$P_i = \sum_{j=1}^{n} P_{ij} \qquad (2)$$

同一文内での同一の語の反復を無視すれば，異なり語としての反復は必ず文間をまたがることになる．

4.2 話題結束度

話題を提示するときは「は」などの助詞によって取り立てられた場合に話題の存在を示唆することがあるが，強調・対比の意味に使われることも多い．さらに文を越えていくときに自明のことは繰り返さないという略題[14]（題目語の省略）が発生するため，手がかり語として利用するのは簡単ではない．

一方，主題の存在するところに反覆語句が多く出現し，全体を貫く主題はテキスト全体に表れ，小さい主題は一部分に集中して表れる傾向があることが報告されている[12]．このことは題目語の省略も話題が継続して文間距離が増加するに従い省略された語も再提出され，意味の補強が行われる．

そこで反覆語の出現区間と出現頻度に着目し，話題を構成する範囲には，同一の語句が集中して表れる特徴を尺度化する．

図 2 に示す例は，テキスト全体の話題として「プライバシー」が確認でき,「差し止め」「訴訟」「公益」「表現の自由」は部分的な小話題として確認できる．

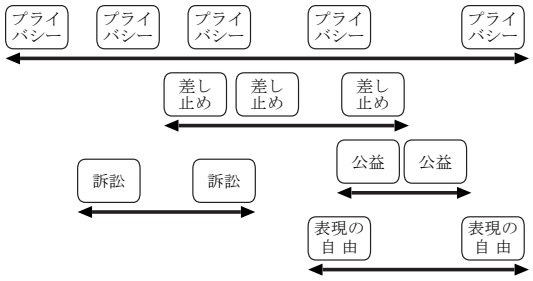

図 2 話題結束度を示す例

この反覆語句の出現区間とその区間内に出現する反覆頻度をもとに，個々の話題のまとまりに対する密度分布を構成するとして尺度化した．この尺度化したものを話題結束度と呼ぶ．

出現区間内に存在する全ての文に対してこの話題結束度を加算することにより，全ての反覆語による

総和を文単位の話題結束度とする．

話題結束度の求め方を示す．

文 i の話題結束度を Q_i，テキスト全体に出現する反覆語（異なり語）を R_j とし，反覆語 R_j による話題結束度を Q_{ij} とする．

テキストの先頭から見て，反覆語 R_j が初めて出現する文を文 f_j，最後に出現する文を文 r_j とする．そして文 f_j から文 r_j までに出現する反覆語 R_j の頻度を h_j としたとき，反覆語 R_j による話題結束度を Q_{ij} を式 (3) を用いて算出する．

但し，文間距離 $r_j - f_j$ には上限 (これを α とする) を設ける．その上限については実験により最適値を定める．

$$Q_{ij} = \frac{h_j - 1}{r_j - f_j + 1} \quad (r_j - f_j \leq \alpha) \tag{3}$$

文単位の話題結束度 Q_i を式 (4) を用いて求める．n は文 i に含まれる語の数である．

$$Q_i = \sum_{j=1}^{n} Q_{ij} \quad (f_j \leq i \leq r_j) \tag{4}$$

4.3 意味結束度

文単位の意味結束度 S_i は式 (5) に示すように，文間結束度 P_i に話題結束度 Q_i を重み付けして加算したしたものを話題結束度とする．この重み付け要素を β とし，実験により最適値を定める．

$$S_i = P_i + \beta \times Q_i \tag{5}$$

5. 実験方法

実験に用いたデータ，境界判定基準，評価方針について述べる．

5.1 実験に用いたデータ

実験に用いた文書は読売新聞社説[17]2004 年 3 月から 2005 年 1 月までの 617 件を対象とした．2004 年 3 月の 1ヶ月分を訓練セットとし，残りの 10ヶ月分を評価セットとした．

1 日分の社説が 2 つ以上の記事で構成されている場合は別々の記事に分割した．実験に一般性をもたせるためにタイトル行は除去し，本文中に小見出しがある場合は対象から除外した．テキストに存在する形式段落は一旦はずして文の単位に分解した．

形態素解析には茶筌[16] を使い，名詞，形容動詞語幹，ナイ形容詞語幹となるものを解析の対象とし，未知語はサ変名詞として取り扱う．本稿ではこれらを名詞的語句と呼ぶ．

名詞的語句に限定した背景には，出現語彙の分析例[15] で，文章の中で繰り返し表れるものに重要語と無性格語とがあり，文脈を支える語として，名詞（形式名詞などの無性格語は除く）以外の語句が主題語や特徴語になることがほとんどないと指摘していることに基づいた．

複合名詞の取り扱いは該当文書内でその組み合わせが限定される場合は，複合名詞として取り扱った．例えば「公益」「目的」の組み合わせが複数存在したとしても,「公益的」（茶筌では「公益」が名詞一般として取り出される）のように違った表記があれば別々に取り扱った．記事の中で 4 文字を 2 文字で表現する略称表記が多く見られる．例えば関西電力（関電），文芸春秋（文春），京都大学（京大）などは異表記ではあるが同一の語として取り扱った．世界保健機関（WHO）などのように，括弧付き表記でふりがな以外は別名としてWHOも単独の語句として取り扱った．

5.2 境界判定基準

意味結束度が大きく落ち込んだ箇所（極小値）が段落境界の可能性が高いと考えられる．極小値の左側の単純減少落差と右側の単純増加との落差から *depth score* による判定方法があるが，本稿では傾斜地点の境界判定精度を考慮して以下の方法を採用した．

まず極大値 S_i から右隣接する 2 つ目の極小値 m_{i+1}（最終は文末）と結んだ線分に対して，極小値 m_i から下ろした垂線の長さを d_i とする．閾値 d_{th} に対して $d_i \geq d_{th}$ ならば，m_i を段落境界候補とする．閾値 d_{th} は実験により最適値を定める．

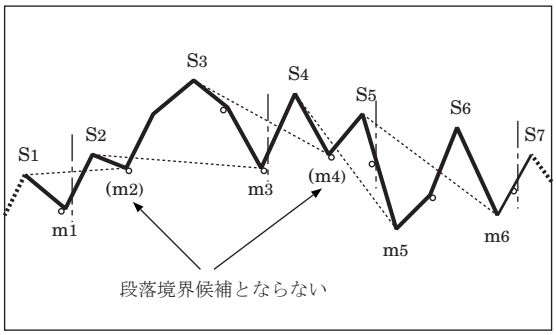

図 3 境界判定例

図 3 の m_2, m_4 はそれぞれ線分 S_2, m_3，線分 S_4, m_5 に下ろした垂線の長さ d_2, d_4 が閾値 d_{th} 未満のため段落境界候補とせず，m_1, m_3, m_5, m_6 を段落境界候補とする例である．

さらに，段落境界の確定は形式段落を条件とするため，境界候補近傍の形式段落に対して，以下に示すアルゴリズムで段落境界を確定する．

極大値 S_i，極小値 m_i に対応する文をそれぞれ

B_{S_i}，B_{m_i} で表し，m_i が境界候補になったとして説明する．

Step1 B_{m_i} が形式段落ならば B_{m_i} を段落境界とする（図 3 の m_1, m_3）．

Step2 $B_{S_{i-1}+1}$ から B_{m_i-1} の範囲で m_i に最も近い形式段落を段落境界とする（図 3 の m_5）．

Step3 B_{m_i+1} から文 $B_{S_{i+1}-1}$ の範囲で m_i に最も近い形式段落を段落境界とする（図 3 の m_6）．

Step4 境界確定ができないときは処理を終了し，次の境界候補に対して **Step1** から **Step3** を繰り返す．

5.3 評価の方針

手法の評価には記事を連結して 1 つのテキストとしたとき，この連結テキストの記事の区切りは少なくともまとまった意味の境界に相当する．そこで本手法による境界判定結果と比較して評価する．

評価指標には情報検索分野で一般的に用いられる再現率（*Recall*），適合率（*Precision*）を採用した．再現率，適合率をそれぞれ以下の式で求める．

$$\text{再現率}= \frac{\text{出力結果に含まれる正解境界数}}{\text{記事境界数}} \times 100$$

$$\text{適合率}= \frac{\text{出力結果に含まれる正解境界数}}{\text{記事境界出力数}} \times 100$$

境界判定基準では意味結束度の極小値を境界候補として，その近傍の形式段落を段落境界とした．

図 4 は極小値を中心に形式段落 a,b が連続している例である．判定基準ではケース 1 で示すように極小値となる形式段落 b が境界として判定され，前の意味段落に所属するとしたが，ケース 2 のように b から意味段落が始まると見ることができる．

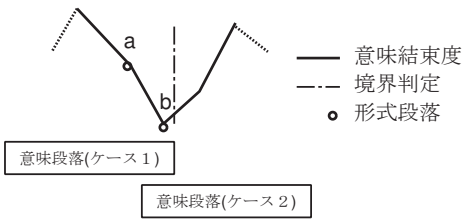

図 4 境界判定例

したがって前の意味段落に所属するケース 1 を前境界，後の意味段落に所属するケース 2 を後境界と呼び，双方を正解として評価する．

6. 評価実験

4 章で述べた手法を実装して予備実験を行った．

表 1 に訓練セットと評価セットについての詳細を示す．両セットとも同質の傾向にあることがわかる．なお文数はタイトル行を除去して連結したときの文数を示す．

種　類	記事数	形　式 段落数	文　数	記事平均 形式段落数	記事平均 文　数
訓練セット	60	809	1,496	13.5	24.9
評価セット	557	7,539	14,168	13.5	25.4

表 1 訓練セットと評価セットの記事傾向

まず，本手法で解析した具体例を示す．予備実験対象として，林の「文章論の基礎問題」（*p203〜226*）[1] の中から段落分割の模範例を評価した．

6.1 解　析　例

「くらしの中のまるい形」という 49 文の模範例では，文塊という定義での段落区切りが示されている．本稿ではこの文塊を意味段落と同義ととらえて実験を行った．

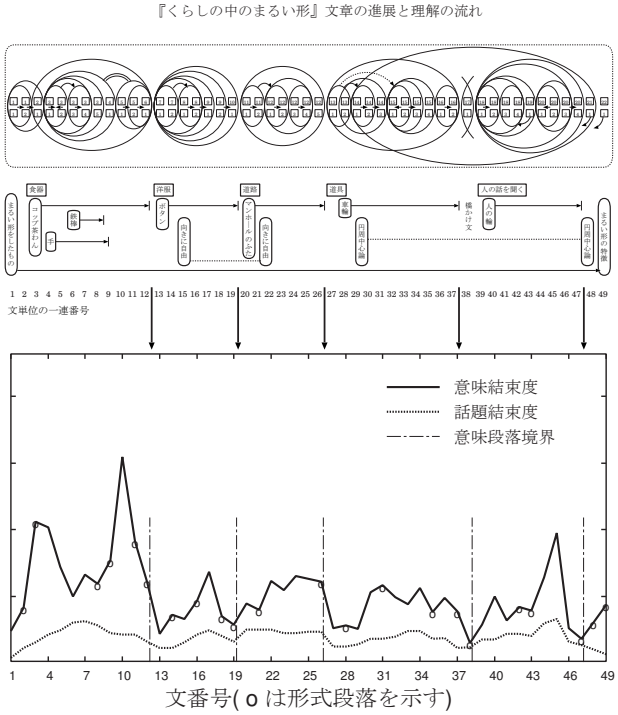

図 5 くらしの中のまるい形

図 5 の上段は「文章論の基礎問題」の掲載内容で，下段は本手法で求めた意味結束度（実線部分）の推移である．点線部分は話題結束度の推移を示し，実線部分と点線部分の差が文間結束度を示す．

模範例の段落区切り点を文番号で示すと，12,19,26,37,(38),47,(48)(49) となる．() は段落の所属が明確にされていない文である．例えば (38) はどちらにも

属さぬ「橋かけ文」または「渡りの文」としており，(48) はさらなる大文塊を作り，(49) は冒頭文と首尾照応して全体を括るとしている．

そこで本手法による段落境界判定結果と照合すると，文番号で 37 と (38) の部分で食い違いがある．(38) はどちらにも属さぬ文という位置づけであることから，本手法による境界判定結果と全てが一致する．

例文には「丸（まる）い」という形容詞が 16 回出現している．本手法では名詞的語句だけを解析対象としたにもかかわらず，良好な結果であった．

同じく図 6 に，訓練セットとした新聞社説の中からの解析事例を示す．

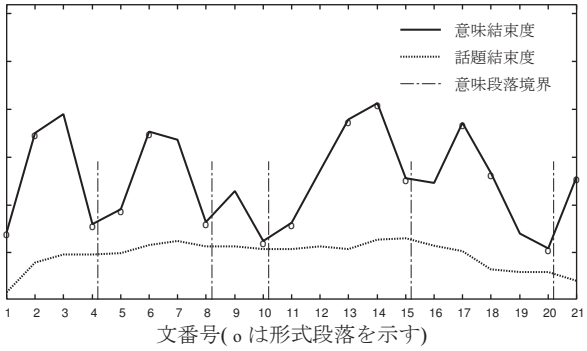

図 6 プライバシーの侵害

21 文，15 の形式段落をもつ記事を 6 段落に分割した（原文は最終ページに掲載）．第 1 段落で「プライバシー侵害事件の概要」，第 2 段落で「関連話題」，第 3 段落は「本題に入る前準備」，第 4 段落で「プライバシーの保護」，第 5 段落で「公益目的と表現の自由」，第 6 段落で「まとめ」という意味上でのまとまりが認められる．段落境界判定結果としては良好と判断した．

6.2 予備実験

模範例や訓練セットでの解析結果を通して，境界判定の閾値 d_{th} を $sd_s/3$ に設定した．sd_s は意味結束度の標準偏差値を指す．

さらに文間距離の上限 α を設定するため，訓練セットにおける反覆語の文間距離と出現頻度を調査し，図 7 に示すグラフの落ち込む点に注目して，文間距離の上限を $\alpha = 19$ に設定した．

訓練セットの記事を連結したテキストを対象に，段落境界の判定結果と記事単位の区切りとがどの程度一致するか実験を行った．

表 2 に文間結束度と話題結束度の加算比率 β を変化させたときの再現率，適合率を示す．

段落境界出力数とは本手法で出力した意味段落数

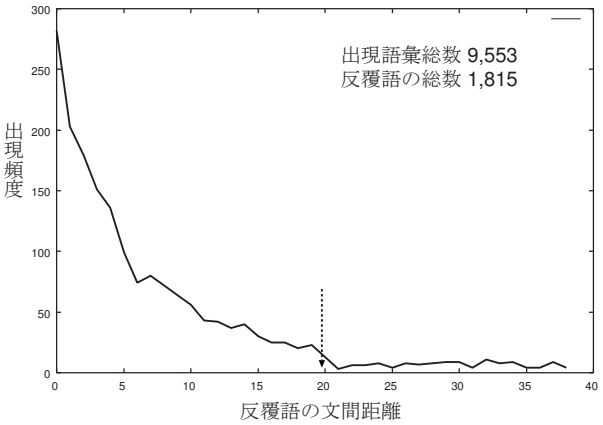

図 7 反覆語の文間距離と出現頻度

である．これには記事区切り以外の箇所の意味段落境界を含むので，1 記事単位で境界判定した結果と一致する段落境界を取り除くことにより，記事区切りとして出力した境界数を推定した．これを記事境界出力数と呼び，適合率の計算に用いた．

段落境界の正解数は記事数と同じ 60 件で，形式段落数は 809 件である．

β	段落境界出力数	記事境界出力数	前境界数	後境界数	正解数	再現率	適合率
0.1	275	146	19	19	38	63.3	26.0
0.2	276	146	20	19	39	65.0	26.7
0.3	277	148	21	20	41	68.3	27.7
0.4	278	148	22	19	41	68.3	27.7
0.5	277	147	22	19	41	68.3	27.9
0.6	273	146	23	19	42	70.0	28.8
0.7	272	146	23	18	41	68.3	28.1
0.8	272	146	23	18	41	68.3	28.3
0.9	272	145	23	18	41	68.3	28.3
1.0	265	144	23	18	41	68.3	28.5

表 2 訓練セットによる予備実験結果

予備実験では $\beta = 0.6$ のとき，再現率 70%，適合率 28.8%の結果を得た．加算比率として $\beta = 0.6$ を採用する．

6.3 評価実験

評価セットの記事を連結したテキストを対象に評価実験を行った．実験結果を表 3 に示す．

記事数	段落境界出力数	記事境界出力数	前境界数	後境界数	正解数	再現率	適合率
557	2,525	1,312	237	119	356	63.9	27.1

表 3 評価セットによる評価実験結果

正解となるべき境界（記事数）は 557 件，段落境

界出力数 2,525 件のうち 1,312 件が記事境界出力数である．正解と一致した前境界数は 237 件，後境界数が 119 件となり，総合計 356 件の 66.6%が前境界であった．実験結果として再現率 63.9%，適合率 27.1%を得た．

7. 考　　察

評価実験の結果をもとに提案手法の有効性について考察する．

7.1 精度について

一般に，k 個の形式段落からなる文書において，正解となる境界が m 個である場合に，システムが n 個の境界を出力した場合，正解出力の期待値は，式 (6) となる．

$$E = \sum_{i=1}^{mim(n,m)} \frac{i \times {}_mC_i \times {}_{k-1-m}C_{n-i}}{{}_{k-1}C_n} \qquad (6)$$

適合率，再現率はそれぞれ

$$\text{適合率} = \frac{E}{n}\ ,\ \text{再現率} = \frac{E}{m}$$

で計算できる．

今回の評価実験において境界候補となる形式段落数 7,539，記事境界出力数 1,312，正解数（記事数）557 を用いて上式の再現率・適合率の理論値の平均を求めると表 4 を得た．

形式段落数	記事境界出力数	正解数	再現率	適合率
7,539	1,312	557	17.4	7.4

表 4 正解出力の理論値

評価結果と比較すると再現率で 46%，適合率で 20%も理論値を上回っている．従来の研究報告では評価対象が異なるため，横並びに比較することはできないが，適合率が 27.1%で少し低いが，再現率で 70%となった点で提案手法が段落境界の判定に有効に働いているといえる．

なお評価実験で正解となる境界から後ろに 1 文（1 段落）ズレて境界と認定できなかったものが評価実験で 127 件（記事全体の 22.8%）存在した．判定できなかった理由として，記事冒頭部の表現に特徴がある．例えば「極めて憂慮すべき事態だ。」「今度は、なんとしても、成功させてほしい。」「きょうは何の日か。」「ワラにもすがる思いだったのだろう。」などのような，タイトル文を前提とした文（同時に形式段落でもある）が冒頭にあるため，境界判定が 1 つズレたものである．

これらも正解相当としたとき，再現率で 86.7%，適合率で 36.8%となる．

7.2 提案手法の特徴

語が生起する位置的関係を手がかりにした本手法は反覆語が比較的多いテキストに対して有効である．結束度計算窓を設けて類似度から段落境界を求める手法は，テキストの先頭と終了位置での境界判定が困難であり，比較的大きな規模のテキストを対象としている．提案手法は，小さな単位のテキストにも対応できる点で優れている．

しかし反覆語句や名詞的語句があまり存在しないテキストや，同じ表現を避けて類義語を用いるようなテキストに対して，提案手法は必ずしも有効とはいえない．

特に論説文以外のテキストにおいて，反覆する名詞的語句が出現しない文が続く場合がある．このような場合の手法の拡張として，移動平均などによる平滑化の適用が有効である．図 8 は芥川龍之介の「鼻」[18] に本手法を適用して平滑化し，段落分けしたものである☆．

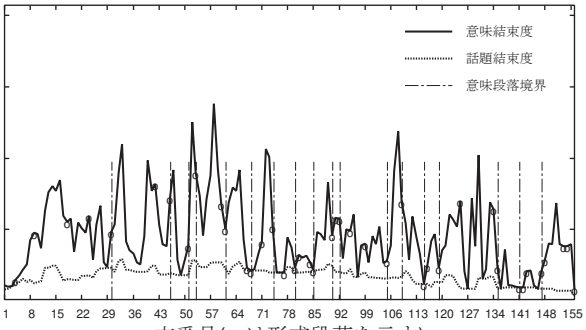

図 8 鼻（芥川龍之介）

登場人物の発言場面の切り替え (69,75,81,86) や話題転換 (31,52,120,148)，さらには内供の内面描写の移り変わり (106,136,142) など，まとまりのある意味段落に分割した結果を得た．() 内の数字は該当する文番号である．

19 分割のうち 11 段落が境界として妥当と判断できた．

8. ま と め

本稿は語の近接する位置関係から語彙的結束性を求め，意味段落境界の判定を行った．

もともと意味段落についての定義にコンセンサスが得られていないため，評価方法に一般的な尺度が存在しない．意味的まとまりを人為的に作り出した

☆ 「と言った。」「と答えた。」の 1 文 1 形式段落になっている箇所を前の段落に統合した．

記事の連結テキストを対象にした段落境界判定実験から，再現率 63.9%，適合率 27.1%の結果を得た．

極めて簡単な方法で語彙的結束性が求まることと，小さい単位のテキストに対しても境界判定ができる点に特徴がある．

提案手法では形式段落を意味段落候補としたが，形式段落が必ずしも境界とはならない可能性もある．今後，形式段落だけでなく文単位を対象とした境界判定方法にも取り組む必要がある．

参 考 文 献

1) 林四郎: “文章論の基礎問題”. 三省堂, (1998).
2) 時枝誠記: “日本語文法口語編”. 岩波書店, (1950).
3) 市川孝: “国語教育のための文章論概説”. 教育出版, (1978).
4) 塚原鉄雄: “「論理的段落」と「修辞的段落」『表現研究』4 号”. 表現学会, (1966).
5) “記者ハンドブック”. 株式会社共同出版社, *pp.*10-11. (2001).
6) 外山滋比古: “日本語の個性”.*pp.17-29*, 中書新書 433, (1976).
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”. 情報処理研究会, *NL102-4* (1994).
8) 豊浦潤，木山次郎，伊藤慶明，岡隆一: “共起関係に基づくテキストの話題境界推定の試み”. 言語処理学会第 2 回年次大会, (1996).
9) 平尾努，北内啓，木谷強: “単語重要度と語彙的結束性を利用したテキストセグメンテーション”. 情報処理研究会, *NL130-6* (1999).
10) 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).
11) 田窪行則，西山佑司，三藤恵，片桐恭弘: “談話と文脈”. 岩波書店, *pp.97-106*, (2004).
12) 馬場俊臣: “「主要語句の連鎖」と「反覆語句」との交渉”. 永野賢 (編). 文章論と国語教育. 朝倉書店, *pp.97-106*, (1986).
13) 北原保雄: “表現文法の方法”. 大修館書店, (1996).
14) 三上章: “象は鼻が長い”. くろしお出版, (1960).
15) 田中章夫: “抄録のための言語処理 朝倉新日本語講座 6”. 朝倉書店, (1983).
16) 松本裕治，北内啓，山下達雄，平野善隆，松田寛, 高岡一馬, 浅原 正幸: “日本語形態素解析システム『茶筌』 version 2.3.3 使用説明書, 奈良先端科学技術大学院大学”. (2000).
17) “YOMIURI ON-LINE”. 読売新聞社, http://www.yomiuri.co.jp/.
18) “青空文庫”. http://www.aozora.gr.jp/ 底本「芥川龍之介全集 1」ちくま文庫，筑摩書房, (1986).

資　　料

「出版禁止命令　プライバシーの侵害は明らかだ」

読売新聞 2004 年 3 月 18 日朝刊社説

文番号	形式段落	記事本文 （罫線は意味段落境界）
1	1	プライバシーの侵害であることは明らかだ。
2	2	元外相、田中真紀子衆院議員の長女のプライバシーに関する記事を掲載した「週刊文春」最新号について、東京地裁は出版禁止の仮処分決定をした。
3	3	決定は長女側の「プライバシー侵害」の主張を「相当」と認め、記事を切除または抹消しなければ、販売してはならない、と発行元の文芸春秋に命じた。
4		文春は同地裁に異議を申し立てた。
5	4	出版の差し止めを命ずる仮処分や判決は、小説や単行本などでは出ているが、販売部数が多く、影響力の大きい週刊誌に対しては極めて異例のことだ。
6	5	一部のメディアによる、露骨なプライバシー侵害の記事などが氾濫するなかで被害者が、メディア側を相手取る名誉棄損訴訟が最近、相次いでいる。
7	6	一連の訴訟では、個人の名誉やプライバシーを重視する裁判所の判断が示されている。
8		認められる損害賠償額も高額化する傾向にある。
9	7	だが、こうした判決は、被害者の「事後の救済」を目指すもので、「事前の救済」とはならない。
10		今回の出版禁止の決定には、やむを得ない面がある。
11	8	プライバシーの権利は、重要な基本的人権の一つとして定着しつつある。
12	9	田中元外相は政治家という公人であっても、長女は私人であり、そのプライバシーは保護されなければならない。
13		それを認めた決定は、田中元外相にかこつけた記事を掲載する出版社側の姿勢を厳しく戒めたものといえる。
14	10	出版の差し止めが許されるケースについて最高裁の判例は、「内容が真実でないか、公益目的でないことが明白で、被害者が重大で回復困難な損害を被る恐れがある時」に限定している。
15	11	差し止めによって、「表現の自由」などが侵害される恐れがあるため、極めて例外的な手段として認めたものだ。
16	12	しかし、今回の記事に「公益目的」があるようには見えない。
17		文春側は仮処分の審尋で、「政治家になる可能性がある人に関する記事であり公益性がある」と主張したが、説得力はない。
18	13	出版の差し止めでは、裁判所は、表現の自由か、個人の名誉やプライバシーの権利か、を選ぶ重い判断を迫られる。
19	14	表現の自由は、民主主義に不可欠である。
20		出版の差し止めには、慎重さが特に必要であり、認める場合の基準は明確でなければならない。
21	15	だが、「表現の自由」を振りかざしてプライバシーを侵害するようなことが横行すれば、かえって民主主義社会の根幹を崩しかねない。