情報処理学会論文誌 Vol. 49 No. 12 4025–4038 (Dec. 2008)

ウェブ検索を利用したしきい値選択型テキストセグメンテーション

阿部直人[†1] 内山俊郎[†1]
内山匡[†1] 奥雅博[†1]

テキストセグメンテーションは与えられたテキストを内容に応じて意味段落に分割する手法である．著者らは事前に学習データを必要としない手法として名詞検索法を提案した．しかし，話し言葉で書かれたテキストに対して算出した連結度や段落境界の決定方法が局所的な内容の変動の影響を受けやすく，テキストセグメンテーションの精度に悪影響を与えるという問題があった．そこで，本論文では助詞と助動詞を除くすべての単語を用いた検索語と関連語の抽出方法を検討する．また，局所的な内容変動を吸収する連結度の算出方法と，それに基づいて境界位置を決定するしきい値の自動選択を行うテキストセグメンテーション手法を提案する．実際のニューステキストやブログテキストを用いた実験を行った．その結果，名詞検索法に対し F 値で 14.9 ポイントの改善が見られ，提案手法の有効性を確認できた．

A Text Segmentation Using WWW Search and Automatic Threshold Selection

Naoto Abe,[†1] Toshio Uchiyama,[†1]
Tadasu Uchiyama[†1] and Masahiro Oku[†1]

Text segmentation is to split a text into sub-paragraphs according to the content. We proposed a text segmentation method based on World Wide Web search from the viewpoint that the training data set was not needed beforehand. However, the performance of our method was degraded for the text written in spoken language. Therefore, in this study, we propose a text segmentation method that uses all words except for a particle and a auxiliary verb in search words and related words. Then, the proposed method finds the contextual relation among several sentences and selects a threshold to determine the boundaries of coherent paragraphs automatically. We examined the performance of proposed method using real-world news texts and blog texts. The experimental result showed the 14.9 points increase of f-value in comparison with the previous method.

1. はじめに

テキストセグメンテーションは，与えられたテキストを内容に応じて意味段落に分割する手法である．たとえば，様々なテキスト（特にブログテキスト）からあるキーワードに対する意見や評判を抽出する際に，テキストセグメンテーションを行うことで抽出精度の向上が期待されている．従来ではキーワードに対して限られた範囲，またはテキスト全体を分析の対象としたため，意見や評判が不足するという問題，逆に余分な評価も含まれるという問題があった．これに対し，テキストセグメンテーションを用いればテキストを内容的なまとまりである意味段落に分割できる．その結果，キーワードに関連する部分テキストを収集することが可能となり，上記の問題を解決できる．上記のほかにも，テキストセグメンテーションは書き起こしテキストの文書整形やスニペット生成，テキスト要約への応用が期待される．

これまでにテキストセグメンテーションに関する多くの手法が提案されており，与えられた学習データから求めた統計的情報や言語的情報に基づく方法[1)–15)]，あるいは事前に学習データを必要としないテキストセグメンテーション手法[16)–18)] がある．前者の場合，たとえば与えられたテキストに関連する記事を学習データから選択し，その情報に基づいてテキストセグメンテーションを行う手法が西脇らにより提案されている[2)]．しかし，十分な精度でテキストセグメンテーションを行うためには大規模な学習データが必要となる．現在新聞コーパスなどの規模の大きい言語資源コーパスを学習データとして利用可能であるものの，コーパスに含まれる内容によりセグメンテーションを行えるテキストの対象が制約されるという問題が残る．これに対し，後者の方法はテキスト内の単語だけを使用するため幅広くテキストセグメンテーションを実行できる．学習データを使用しないテキストセグメンテーション手法として代表的な Hearst 法[16)] は，テキスト内の単語の出現頻度に基づいて文どうしのつながりをコサイン距離により算出し意味段落の境界を求める．一方で，Hearst 法の性能は表記揺れや単語の出現の仕方に影響を受けやすく，過分割が発生しテキストセグメンテーションの精度が低下するという問題がある．上記の 2 つの問題点をふまえ，著者らは事前に学習データを使用しない手法としてウェブ検索を利用した名詞の出現頻度に基づく方法（名詞検索法）を提案した[18)]．名詞検索法は Hearst 法に対して，テキスト内の名詞を検索語としてウェブ検索を行い，検索結果によって得られた情報を用いて単語ベクトルを

†1 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

拡張する方法である．ウェブ上には莫大なテキストが蓄積されている点から，広範囲のジャンルの記事をウェブ検索によって収集できる．そして，得られた記事から単語を抽出することで，テキストの内容のジャンルを表す単語や異なる表記の単語など，検索語に関連する単語が幅広く得られる．実際のニューステキストを用いた比較実験の結果から，Hearst 法よりも良いセグメンテーション結果を得ることが確認されている．

しかし，名詞検索法ではブログテキストのような話し言葉で書かれたテキストに対して，テキストセグメンテーションの精度が低下するという問題があった．主な原因は以下にあげる 3 点である．

(1)　名詞だけでは検索語や関連語を適切に抽出できないことが多いため，内容の関連性を十分に比較できない．

(2)　テキストの内容が局所的に変動するため，隣接文間の比較では過分割が発生しやすい．

(3)　テキストによって内容の変化の強さが異なるため，1 つの固定しきい値では十分に段落境界を抽出できない．

そこで，本論文では上記の問題点を解決するため，単語の抽出方法を工夫し，意味段落の境界の明瞭さを保存しつつ局所的な内容の変動による影響を低減させる連結度の算出方法を提案する．そして，算出した連結度に基づいて，意味段落間の内容の異なり度合いを考慮しつつ段落境界の位置を決定するためのしきい値を自動的に選択する方法を示す．また，実際のニューステキストとブログテキストを用いて従来手法との比較実験を行い，ニューステキストとブログテキストの両方に対して提案手法が有効であることを検証する．

本論文では，2 章において提案手法の説明を行い，3 章で実験の詳細を述べる．そして，4 章で考察を行い，最後に 5 章で本論文のまとめを行う．

2. 提案手法と名詞検索法

2.1 ウェブ検索に基づくテキストセグメンテーションの概要

名詞検索法と提案手法におけるウェブ検索を利用したテキストセグメンテーションの概要を**図 1** に示す．初めに，与えられたテキストに対して，各文から検索語として使用する単語を抽出する．ここで，本論文において文とは句点「。」で区切られた 1 文のことを指す．次に，得られた検索語を用いてウェブ上で AND 検索を行い，検索結果上位にあるウェブ上の記事（関連記事と呼ぶ）から単語（関連語と呼ぶ）を抽出する．検索結果の上位に含まれる記事を使用する理由は，上位であるほど検索語が同時に多く含まれ，検索語に対して強く関連する記事が得られるからである．また，関連記事内の出現頻度が高い単語は検索語に関

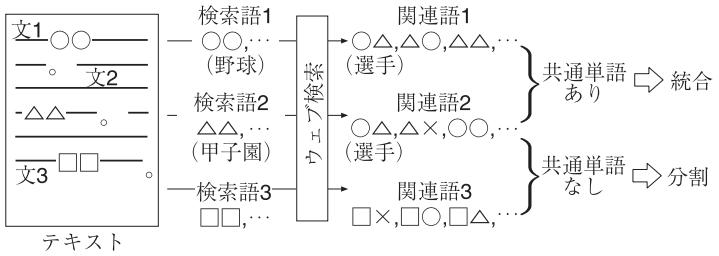

図 1　ウェブ検索を利用したテキストセグメンテーション法の概要
Fig. 1　Overview of www search based text segmentation.

係する単語であることが多く，実際にこの性質を利用して与えられた単語の専門分野を判定する研究がある[19]．そこで，ウェブ検索に基づくテキストセグメンテーションでは検索語と関連語を用いて単語が変化する文と文の間（たとえば，図 1 の文 2 と文 3）を意味段落の境界としてとらえ，分割を行う．

2.2 名詞検索法との差異

名詞検索法では，初めに各文から抽出した名詞のみを検索語として使用しウェブ検索を行う．そして，得られた検索語と関連語を用いて隣接文間で結合度を調べ意味段落の境界を求める．しかし，ブログテキストのような話し言葉で書かれたテキストでは以下のような傾向が見られ，テキストセグメンテーションの精度が低下するという問題があった．

- 名詞だけでは検索語や関連語を適切に抽出できないことが多いため，内容の関連性を十分に比較できない．
- テキストの内容が局所的に変動するため，隣接文間の比較では過分割が発生しやすい．
- 内容の変化の強さはテキストによって異なるため，1 つの固定しきい値では段落境界を十分に抽出できない．

上記の問題を解決するために，(1) 名詞以外の単語も使用した検索語と関連語の抽出方法，(2) 局所的な内容変動の影響を低減させる連結度の算出方法，(3) 意味段落間の内容の異なり度合いを考慮したしきい値の設定方法，の 3 点に関して名詞検索法に改良を加えた手法を提案する．以降の節で提案手法の詳細を述べる．

2.3 提案手法の概要

2.3.1 検索語の抽出

提案手法では与えられたテキストにある N 個の文に対して，各文から助詞や助動詞（た

表 1　図 2 のテキストに対する検索語と関連語
Table 1　An example of search words and related words for the text shown in Fig.2.

文番号	検索語	関連語
1	ドライブ，高速道路	料金，道路，ETC, 渋滞，車，北海道，通行止め，工事
2	ドライブ, 運転，慣れる，札幌, 長距離，日帰り，網走	北海道，風，走る
3	ガソリン代	ガソリン，車，価格，節約，ガソリンスタンド，カード，高騰，給油，家計
4	ドライブ，温泉	露天風呂，楽しむ，ホテル，風呂，楽しめる，満喫，日帰り，食べる
5	ゆっくり，温泉，食べる，絶景，眺める，美味しい	料理，旅館，ホテル，観光
6	ドライブ，温泉，開拓，楽しめる，日帰り	観光，楽しむ，日帰り温泉，料金，満喫
7	ゴルフ，始める，打ちっぱなし，練習	打つ，ボール，ゴルフ練習，当たる，教える，スコア
8	ゴルフ，ショット，仕事，打ちっぱなし，弾道，練習	スコア，打つ，飛距離，スイング
9	うまい，気長，見つける，打てる，練習	無理，欲しい，通る，打つ，立つ

とえば,「を」,「られる」）を除くすべての単語を抽出し検索語として使用する．各文に対して形態素解析を行い，名詞など活用形のない単語はそのまま使用する．また，動詞など活用形のある単語は，すべて終止形に変換した単語を検索語として使用する．形態素解析は JTAG [20] を用いて実行し単語を抽出した．**表 1** に**図 2** で示すテキスト（$N = 9$）に対して検索語を抽出した結果を示す．終止形を使用する理由は以下の 2 点である．

- 「見て」・「見た」・「見る」など活用形のある単語を 1 つにまとめ，検索語が膨大になるのを防ぐため．
- 活用の語幹を使用する方法が考えられるものの，表 1 で示すように終止形を使用しても適切な検索語と関連語が得られると判断したため．

一方で，各文から抽出した単語には「年」・「ある」・「それ」などの一般語が含まれる場合があり，一般語はウェブ検索に有用ではない．そこで，あらかじめ人手で作成した一般語辞書に含まれる単語を除き，残った単語を i 番目（ただし，$i = 1, 2, \ldots, N$）の文に対する検索語として使用する．また，抽出された検索語の個数を S_i とする．ここで，活用形のある一般語は終止形に直したものを，活用形のない一般語はそのものを一般語辞書に登録する．

ドライブが好きな私は大阪まで高速道路で行くことが多い。
札幌にいた時から日帰りで網走目指してドライブしたりしてたから、長距離運転はだいぶ慣れた方だと思う。
最近はガソリン代が高いのでちょっとどうしようかなと考え気味。
ドライブしたついでに温泉にも入りたい。
温泉で絶景を眺めながらゆっくりしたいし、色々美味しいもの食べたいし。
日帰りでドライブと温泉が楽しめる場所を開拓してみるか…。
最近、ゴルフを始めて色々本を買って打ちっぱなしで練習中。
近くに24時間営業のゴルフ打ちっぱなし場があるから、仕事帰りに練習してるんだけど…、低弾道のショットが多い気がする。
まあ、そんな簡単にうまく打てるようになるとは思っていないので、時間を見つけて打ちっぱなしで気長に練習をするか。

図 2　テキストの例
Fig. 2　An example of the input text.

2.3.2　関連語の抽出

抽出した検索語を用いてウェブ検索を行い関連語を取得する．関連語には検索語と同様の単語を使用する．具体的には以下の手順で各文に対して関連語を収集する．

(1)　抽出した検索語を用いてウェブ上で AND 検索を行う．
(2)　検索結果上位で参照されている $P_{\min}$ 件の関連記事に対して形態素解析を行い終止形に変換した単語を抽出する．
(3)　抽出した単語から一般語辞書に登録されている一般語を除く．
(4)　出現頻度の高い単語の順に関連語として選択する．

一方で，得られる関連語の個数は検索語や関連記事により変化する．そこで，あらかじめ設定した単語数 T に対し，$T - S_i$ 個（$i = 1, 2, \ldots, N$）の単語を関連語として使用し検索語と関連語の総数を T にする．

関連記事数が少ない場合ある限定的な内容に偏るため適切な関連語を得ることが難しい．その点から，本論文では得られた関連記事数 P が $P \geq P_{\min}$ の場合に関連語を抽出し，それ以外の場合は関連記事の再検索を行う．再検索の具体的な手順は以下のとおりである．

(1)　S_i 個の検索語から j 番目（$j = 1, 2, \ldots, S_i$）を除いた単語でウェブ検索を行う．その際，得られる検索件数を H_j とする．
(2)　H_j が最大となる $S_i - 1$ 個の単語を選択する．
(3)　$S_i \leftarrow S_i - 1$ とし関連記事を得る．
(4)　$P \geq P_{\min}$ を満たせば関連語を抽出して再検索を終了する．それ以外の場合は (5) に

進む．

(5)　$S_i > 1$ であれば (1) に戻る．それ以外の場合は，i 番目の文は検索語のない文として扱う．

表 1 に図 2 のテキストに対して得られた関連語を示す．ここで，表 1 の例は $T = 10$, $P_{\min} = 20$ として実行した結果である．

2.3.3　意味段落の境界候補の抽出

提案手法における意味段落の境界候補の抽出方法について説明する．初めに検索語と関連語を組にしたキーワード集合 K_i を作成する（ただし，$i = 1, 2, \ldots, N$）．キーワード集合の作成手順を**図 3** に示す．具体的には，各文において検索語の抽出と関連語の抽出を行い合計 T 個の単語を得る．検索語が存在しない場合には空のキーワード集合を作成する．検索語が合計個数 T を超えた場合には，検索語の中からランダムに T 個を選択する．また，ウェブ検索により関連記事を取得する際，関連記事数 P が $P < P_{\min}$ の場合は検索語を 1 単語ずつ減らし再検索を行う．再検索を繰り返しても関連記事数が $P \geq P_{\min}$ を満たさず，かつ $S_i < 1$ となった場合には空のキーワード集合を作成する．

キーワード集合作成後，j 番目の文を基準に K_{j+1-b} から K_j までのキーワード集合を含むブロック B1 と K_{j+1} から K_{j+b} までを含むブロック B2 を考え，単語 t の出現頻度を求める（ただし，$j = 1, 2, \ldots, N-1$ であり，b はブロック幅である）．そして，ブロック B1 と B2 に含まれる単語 t の出現頻度 w_t^{B1}，w_t^{B2} を用いて，j 番目と $j+1$ 番目の文の連結度 $C_{j,j+1}^b$ を式 (1) で算出する．

$$C_{j,j+1}^b = \frac{\sum_t w_t^{B1} w_t^{B2}}{\sqrt{\sum_t (w_t^{B1})^2 \sum_t (w_t^{B2})^2}} \tag{1}$$

連結度 $C_{j,j+1}^b$ は意味段落の境界付近では 0 に近い値をとる．したがって，意味段落の境界候補は連結度 $C_{j,j+1}^b$ の極小値（本論文では，これを連結度の谷と呼ぶ）を見つけることで求められる．本論文では連結度の谷を以下の式 (2) で定義する．

$$\begin{cases} C_{j-1,j}^b > C_{j,j+1}^b \\ C_{j,j+1}^b < C_{j+1,j+2}^b \end{cases} \tag{2}$$

一方で，話し言葉で書かれたテキストでは内容が局所的に変動する傾向にある．そのため，ブロック幅 b の値に応じて連結度の谷の検出位置が異なりやすいことや，小さい連結度の谷が複数検出されることが多い．局所的な内容の変化の影響を低減させるため，得られた連結度に対してスムージングを行う方法がある．しかし，スムージングにより連結度の値が平坦化される反面，意味段落の境界に対応する連結度の谷が不明瞭になるという問題がある（**図 4** (a)）．そこで，複数のブロック幅を用いて連結度を算出し，それらの平均値である平均連結度を用いて意味段落の境界を求める．一般的に，任意のブロック幅に対して意味段落の境界では連結度の谷は明瞭に検出される．それに対し，局所的な内容変化の影響を受ける箇所では小さい連結度の谷が複数検出されるものの，意味段落の境界の場合と比較して明瞭ではない．そこで，複数のブロック幅の連結度の平均値を用いることで，局所的な内容変

(1)　$i = 1$ とする．また，検索語と関連語の合計個数 T と関連記事数のしきい値 $P_{\min}$ を決定する．
(2)　i 番目の文から検索語を抽出する．ここで得られた検索語の個数を S_i とする．
(3)　$S_i \geq 1$ ならば次へ進む．それ以外の場合は空のキーワード集合を作成し（9）へ進む．
(4)　$T > S_i$ ならば次へ進む．それ以外の場合は，S_i 個の検索語からランダムに選択した T 個の単語をキーワード集合の要素とし（9）へ進む．
(5)　S_i 個の検索語を用いてウェブ検索を行い，得られた検索件数を P とする．
(6)　$P \geq P_{\min}$ ならば（8）へ進む．それ以外の場合は次へ進む．
(7)　再検索を行い最も検索件数が多くなる検索語の組合せを i 番目の検索語として使用する．また，検索語の個数を $S_i \leftarrow S_i - 1$ とし（3）に戻る．
(8)　検索結果上位 $P_{\min}$ 件の記事において文書頻度の高い順に単語を抽出し，S_i 個の検索語と $T - S_i$ 個の関連語を選択しキーワード集合の要素とする．
(9)　$i \leftarrow i + 1$ とする．
(10)　$i > N$ ならば終了する．それ以外の場合は（2）に戻る．

図 3　キーワード集合の作成手順
Fig. 3　Procedure of generating keyword sets.

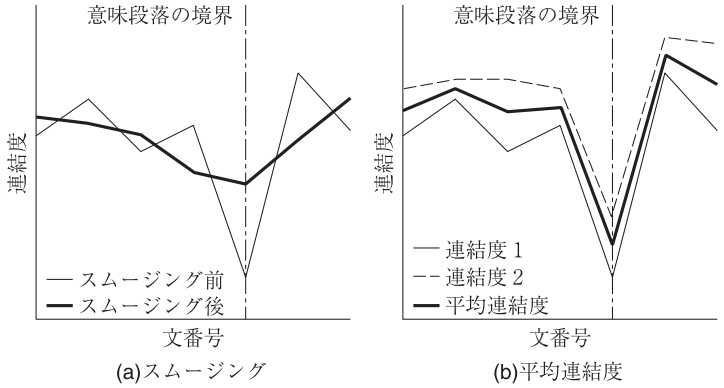

図 4　スムージングと平均連結度の相異点
Fig. 4　The difference between smoothed relation value and average relation value.

動の影響を吸収し，本来の意味段落の境界は正しく抽出できる（図 4 (b)）．平均連結度は，ブロック幅を $b = \{1, 2, \ldots, N/2\}$ とすると，以下の式 (3) で算出される．

$$C_{j,j+1} = \frac{C^1_{j,j+1} + C^2_{j,j+1} + \cdots + C^{N/2}_{j,j+1}}{N/2} \tag{3}$$

そして，提案手法では以下の式で定義する平均連結度の谷を意味段落の境界の候補位置とする（検出された候補位置の個数を r とする）．

$$\begin{cases} C_{j-1,j} > C_{j,j+1} \\ C_{j,j+1} < C_{j+1,j+2} \end{cases} \tag{4}$$

平均連結度を算出する際に，重複する文が存在しないように 2 つのブロックを決定する必要がある．そこで，本論文では平均連結度算出時のブロック幅を $b = \{1, 2, \ldots, N/2\}$ と設定する．また，空のキーワード集合は無視し，代わりにその前後にあるキーワード集合を用いる．実際に図 2 のテキストに対してスムージングを行った連結度と平均連結度（$T = 30$, $P_{\min} = 20$）を算出した結果を**図 5** に示す．スムージングはブロック幅 $b = 3$ を用いて求めた連結度に対して，ある基準点の連結度とその前後にある連結度の 3 つの平均値を用いて行った．図 5 の例において平均連結度に着目すると $j = 3$, 6 で平均連結度の谷が検出され，図 2 の内容の変化と一致していることが分かる．一方で，スムージングされた連結度では，本来の意味段落の境界位置において，連結度の谷の明瞭さが失われていることが分かる．

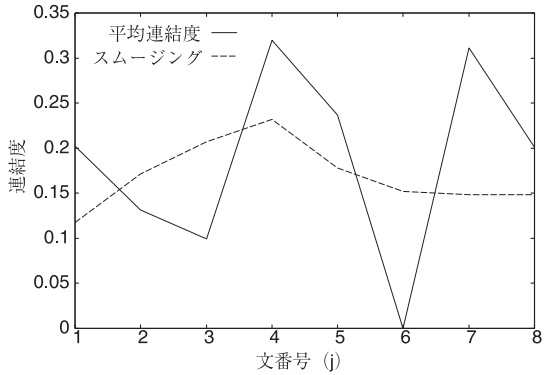

図 5　図 2 のテキストの例に対する平均連結度とスムージングによる連結度の比較
Fig. 5　A comparison between average relation value and smoothed relation value for the text shown in Fig.2.

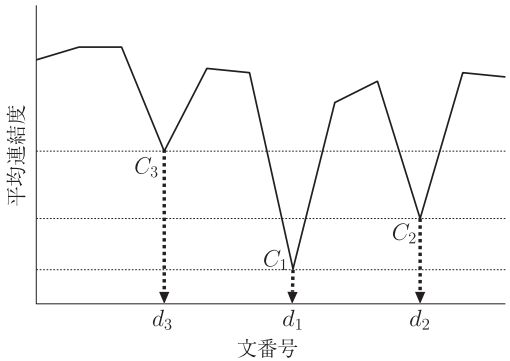

図 6　しきい値 C_m と分割位置 d_m の決定方法
Fig. 6　How to decide the threshold C_m and splitting position d_m.

2.3.4　しきい値選択

意味段落の境界候補の検出後，提案手法では分割位置と分割の順番を決定し，その順番に応じて複数の分割結果を作成する．そして，得られた分割結果に対して意味段落間の内容の異なり度合いを評価し，1 つの分割結果を選択する．初めに，r 個の意味段落の境界候補が検出されたときの平均連結度の値を小さい順に並べ替えたものを $C_1, C_2, \ldots, C_r$ とする．次に，それらに対応する順に検出された r 個の文番号 $d_1, d_2, \ldots, d_r$ とする（**図 6**）．そし

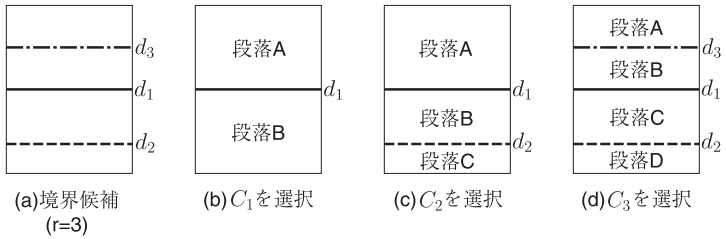

図 7　しきい値 C_m に基づく分割結果の作成方法
Fig. 7　How to split the text on the basis of the threshold C_m.

て，値の小さい順に並べ替えた $C_1, C_2, \ldots, C_r$ を分割箇所を決定するためのしきい値として用いる．具体的には，m 番目（$m = 1, 2, \ldots, r$）のしきい値 C_m を選択したとき，分割を行う文の位置は $d_1, d_2, \ldots, d_m$ とする（**図 7**）．ここで，m 番目のしきい値を選択した場合，分割箇所が m 個であるため得られる意味段落の個数は $m+1$ となる．

m 番目のしきい値 C_m を選択したときに得られる分割結果は以下の評価関数により評価する．

$$Q_1^m = \frac{1}{m+1} \sum_{l=1}^{m+1} \frac{\sum_t w_t^{all} w_t^l}{\sqrt{\sum_t (w_t^{all})^2 \sum_t (w_t^l)^2}} \tag{5}$$

$$Q_2^m = \frac{1}{m} \sum_{l=1}^{m} \frac{\sum_t w_t^l w_t^{l+1}}{\sqrt{\sum_t (w_t^l)^2 \sum_t (w_t^{l+1})^2}} \tag{6}$$

ここで，w_t^{all} はキーワード集合 K_i（$i = 1, 2, \ldots, N$）のすべてに含まれる単語 t の出現頻度である．また，w_t^l は m 番目のしきい値を選択したときの分割結果において l 番目の意味段落内のキーワード集合に含まれる単語 t の出現頻度である．Q_1^m は 0 以上 1 以下の値をとり，テキスト全体と各段落との共通する単語数を測る．分割が細かくなるほど，1 つの意味段落に含まれる単語の個数は小さくなるため，Q_1^m は小さい値をとるという性質を持つ．つまり，Q_1^m はキーワード集合に含まれる単語の個数で分割の細かさを測る指標であり，細かい分割が得られるほど Q_1^m のとる値は 0 に近くなる．Q_2^m も Q_1^m と同様に 0 以上 1 以下の値をとり，2 つの意味段落の内容の近さを測る指標である．すなわち，意味段落間の内容が異なるほど，2 つの意味段落で共通する単語の個数は少なく単語の出現パターンも異なるため，Q_2^m は 0 に近い値をとる．提案手法では，この 2 つの指標を用いて以下の式 (7) を満たす m（$m = 1, 2, \ldots, r$）番目のしきい値を選択する．

表 2　図 2 のテキストに対して提案手法を適用した結果
Table 2　Result of the proposed method for the text shown in Fig.2.

段落番号	文番号	出現単語（$w_t^l \geq 2$）
1	1，2，3	車，高速道路，価格，ドライブ，北海道
2	4，5，6	楽しめる，温泉，食べる，走る，ホテル，楽しむ，日帰り，観光，露天風呂，ドライブ
3	7，8，9	打つ，練習，スイング，上達，探す，打てる，ゴルフ，アイアン，打ちっぱなし，仕事，スコア，ドライバー，ボール，教える

$$\bar{m} = \arg\min_m (Q_1^m + Q_2^m) \tag{7}$$

$\bar{m}$ 番目に対応するしきい値 $C_{\bar{m}}$ を選択することで，内容ごとに細かく分割された結果が得られると考えられる．図 2 で示すテキストに対して提案手法を実行した結果（$\bar{m} = 2$）を**表 2** に示す．

3. 実　　験

本論文では，提案手法におけるパラメータの設定に関する予備実験と提案手法の有効性を確認するための実験を行った．予備実験では 2 つのパラメータ T と $P_{\min}$ に関して値を決定するために，ウェブ検索による検索結果を用いて調べた．本実験では，(1) 検索語と関連語の抽出方法と平均連結度の効果を調べる実験，(2) しきい値選択法の有効性を検証する実験，(3) 比較実験により提案手法の有効性を確認する実験，の 3 つの検証を行った．

3.1 予 備 実 験

提案手法では，ウェブ検索を行うことでテキストの内容に関連の高い単語を取得する．一方で，ウェブ検索により得られる関連記事が少ない場合，適切な関連語を得ることは難しい．また，関連語を過剰に取得すると，テキストの内容とは関係の弱い単語が含まれ，意味段落の境界を抽出する際に悪影響を与える．そこで，予備実験では**表 3** に示す 10 個の単語群を用いて，T と $P_{\min}$ の値を検討する．表 3 の単語群は，特定の内容に関するテキストから名詞を抽出したものである．

3.1.1 $P_{\min}$に関する予備実験

関連記事数 $P_{\min}$ に関する予備実験は以下の手順で行った．

(1)　単語群から S（$1 \leq S \leq 5$）個の単語をランダムに選ぶ．
(2)　S 個の単語を検索語として用いてウェブ検索を行い，P（$1 \leq P \leq 50$）件の関連記

表 3　予備実験で使用した単語群
Table 3　The list of word used in the preliminary experiment.

群	単語
1	道交法，飲酒運転，ひき逃げ，交通事故，自動車，責任，再犯，事故，社会，歩行者
2	厚生労働省，雇用，定年，高齢者，助成金，団塊世代，企業，職場，シンポジウム，保険料
3	宇宙，暗黒物質，銀河，重力，研究，空間，レンズ，ガス，望遠鏡，観測
4	津波，地震，気象庁，エネルギー，警報，沿岸，防災，プレート，太平洋，避難
5	感染，インフルエンザ，ワクチン，生活，高熱，薬，患者，重傷，自治体，医療
6	地球，エネルギー，温暖化，リサイクル，資源，化学物質，オゾン，大気汚染，二酸化炭素，工場
7	政治家，選挙，領収書，辞任，税金，政治団体，政治資金，疑惑，不正，公開
8	チーム，試合，日本代表，サポーター，勝利，決勝，選手，ホーム，監督，特典
9	外国為替，口座，手数料，通貨，取引，外為，セミナー，市場，資金，マージン
10	レシピ，スパイス，料理，野菜，辛口，ビーフ，ヘルシー，パン，ダイエット，レトルト

事を取得し W（$W = 10$，20，30，40，50）個の関連語を抽出する．

(3)　W 個の関連語を P 件分と $P+1$ 件分の間で比較し共通する関連語の個数を調べる．

(4)　上記 (1) から (3) をすべての単語群に対してそれぞれ行い，10 個の単語群での平均と標準偏差を求める．

予備実験の結果を**表 4** に示す．表 4 において，表中の上段の数字が共通単語数の平均，下段の数値が共通単語数の標準偏差である．予備実験の結果から，関連語数に関係なく関連記事数が 20 以上の場合において得られる関連語の 9 割程度が前後で一致することが分かった．この点から本論文の実験では関連記事数のしきい値 $P_{\min}$ は $P_{\min} = 20$ と設定する．

3.1.2　T に関する予備実験

次に，$P_{\min} = 20$ と固定し，検索語と関連語の総数を設定するしきい値 T について検討を行った．具体的には，表 3 の単語群を用いて以下の手順で予備実験を行った．

(1)　単語群から S（$1 \leq S \leq 5$）個の単語をランダムに選ぶ．

(2)　選択した S 個の単語を用いてウェブ検索を行い 20 件分の関連記事を得る．

表 4　関連記事数 P と関連語数 W を変化させたときの共通語数の平均（上段）と標準偏差（下段）
Table 4　The value of average (upper section) and standard deviation (lower section) of the number of common words with relation to the number of P and W.

関連語数	関連記事数				
	10	20	30	40	50
10	9.1	9.4	8.8	9.7	9.6
	0.700	0.663	0.872	0.458	0.663
20	18.0	18.8	19.0	19.7	19.5
	0.894	0.980	0.775	0.458	0.671
30	26.3	28.5	28.3	29.4	28.8
	1.616	1.204	1.269	0.800	0.980
40	34.4	38.0	38.0	39.4	38.1
	2.107	1.549	0.894	0.663	1.221
50	44.0	47.3	48.2	49.6	48.4
	2.608	1.616	1.077	0.490	1.356

(3)　20 件の関連記事から W（$W = 10$，20，30，40，50）個の関連語を抽出する．

(4)　(1) から (3) をすべての単語群で行う．

(5)　2 つの単語群間において W 個の関連語の中で共通している単語の個数を調べる．

(6)　(5) をすべての単語群の組合せに対して行い，平均と標準偏差を調べる．

実験結果を**表 5** に示す．実験の結果から，関連語の個数が 30 以下の場合において，すべての単語群の組合せの中で重複する関連語は平均 1 単語未満であることが分かった．関連語が多いほど文の内容を幅広く比較することが可能となる反面，異なる内容において重複する関連語は少ないことが望ましい．この点から，本実験では検索語と関連語の総数を決定するしきい値 T は $T = 30$ と設定する．

3.2　改善の効果を調べる実験

3.2.1　概　　要

この実験では，名詞検索法に対する変更点である，(1) 検索語と関連語の抽出方法，(2) 連結度の算出方法，の 2 つの変更に関して有効性を検証する．実験で使用するデータは実際のニューステキストとブログテキストを使用した．本実験では以下の**表 6** に示す 4 つの手法を用いてテキストセグメンテーションを行い，その精度を評価することで変更点の有効性を確認した．

また，意味段落の境界を抽出する際，隣接連結度や平均連結度の値に対してしきい値 D_T

表 5　10 の単語群間で重複する単語数の平均と標準偏差
Table 5　The average and standard deviation of the number of duplicating words.

関連語数	重複単語数	
	平均	標準偏差
10	0.022	0.147
20	0.200	0.542
30	0.600	0.879
40	1.044	1.299
50	1.644	1.594

表 6　名詞検索法に対する変更点の有効性を検証するために用いた 4 つの手法
Table 6　The list of methods to confirm the effectiveness of the modification to our previous approach.

	使用単語	連結度	特徴
手法 1	名詞	隣接連結度	名詞検索法
手法 2	すべて	隣接連結度	変更点（1）のみ
手法 3	名詞	平均連結度	変更点（2）のみ
手法 4	すべて	平均連結度	変更点（1）＋（2）

を設定し分割を行った（本実験では，固定しきい値法と呼ぶ）．具体的には，式 (2) と式 (4) の条件に $C^b_{j,j+1} < D_T$ と $C_{j,j+1} < D_T$ の条件をそれぞれ追加し分割位置を特定する方法である．ここで，しきい値 D_T は 0 以上 1 以下の値をとる．つまり，固定しきい値法では，j 番目の文において連結度の谷の条件を満たし，かつ j 番目の連結度の値が D_T 未満の場合に意味段落の境界が検出されたと判定し，j 番目と $j+1$ 番目の文の間で分割を行う．しきい値は $D_T = \{0.1, 0, 2, \ldots, 0.9\}$ と設定し，最も良いセグメンテーション結果が得られるパラメータを用いた．

3.2.2　使用したデータ

ニューステキストはオンラインで収集したものである．収集したニューステキストの詳細を**表 7** に示す．ニューステキストを用いた実験では，表 7 にある 7 つのジャンルから 3 つのジャンルを選択し，各ジャンルからテキストを 1 つずつランダムに抽出する．そして，選ばれた 3 つのテキストを 1 つに並べたものを 1 つのテストデータとして使用する．つまり，正解の分割位置は各ジャンルのテキストの結合位置である．実験ではテストデータを 100 個用意し，各手法を用いてテキストセグメンテーションを行った．

実験で使用するブログテキストは，3 人の評価者が複数の話題が含まれていると判断したものを使用した．具体的には以下の手順で収集した．

表 7　実験で使用したニューステキストの概要
Table 7　The detail of the news dataset used in the experiment.

ジャンル	記事数	記事内文章数		平均文章数
		最小	最大	
ライフ	68	1	61	10.6
国際	29	2	14	5.7
ビジネス	51	3	22	7.4
スポーツ	89	2	29	9.3
社会	62	2	37	7.9
エンタメ	111	1	39	7.9
政治	35	2	17	7.7

(1)　各テキストに対して各文に句点が抜けている場合にはあらかじめ付与する．
(2)　3 人の評価者が書かれている内容とその範囲を判断する．
(3)　3 人の評価者で 2 つ以上の内容が書かれていると判断し，かつそれぞれの範囲が一致したものを収集する．

収集されたブログテキスト数は 299 であり，文章数は最小で 7 文，最大で 71 文，平均 19.5 文である．また，ブログテキストにおける正解の分割位置は，3 人の評価者が判断した境界位置とする．

3.2.3　性能評価

手法の性能評価は出力された意味段落の境界位置と正解位置との比較により行う．具体的には，各テスト記事において式 (8) と式 (9) により適合率を再現率を算出する．

$$\text{適合率} = \frac{\text{手法で正しく検出された境界数}}{\text{手法で検出した境界数}} \tag{8}$$

$$\text{再現率} = \frac{\text{手法で正しく検出された境界数}}{\text{テストデータ内での境界数}} \tag{9}$$

次に，各テストデータで得られた適合率と再現率からテストデータ全体における平均適合率と平均再現率を求め，以下の式 (10) で示す F 値により性能を評価する．

$$\text{F 値} = \frac{2}{\frac{1}{\text{平均適合率}} + \frac{1}{\text{平均再現率}}} \tag{10}$$

F 値は平均適合率と平均再現率の調和平均であり，両方の値がともに大きいときに 1 に近

表 8　変更点の検証結果
Table 8　Result of the experiment to confirm the effectiveness of the modification to our previous approach.

テキスト	手法	選択しきい値	適合率（%）	再現率（%）	F 値
ニュース	手法 1	$D_T = 0.1$	63.1	68.5	65.7
	手法 2	$D_T = 0.1$	60.1	68.0	63.8
	手法 3	$D_T = 0.2$	87.3	84.5	85.9
	手法 4	$D_T = 0.2$	**89.8**	**86.0**	**87.9**
ブログ	手法 1	$D_T = 0.1$	18.7	23.8	20.9
	手法 2	$D_T = 0.1$	19.3	28.2	22.9
	手法 3	$D_T = 0.4$	25.1	34.8	29.2
	手法 4	$D_T = 0.3$	**28.4**	**37.0**	**32.1**

い値をとる．つまり，F 値が 1 に近い値をとるほどテキストセグメンテーションの精度が高いことを意味する．

3.2.4　実 験 結 果

検証実験の結果を**表 8** に示す．表 8 において，最も良い結果に対してボールド体で示している．ニューステキストとブログテキストともに，手法 4 が最も良いセグメンテーション結果を得ていることが分かった．使用単語の変更（手法 1 と手法 2）による効果を調べてみると，ニューステキストの場合，各文に内容を反映する名詞を多く含むことが多いため，すべての単語を用いた場合の F 値が 63.8 であるのに対し，名詞だけを用いる場合の F 値は 65.7 となり，手法 1 の方が分割精度は高かった．手法 1 と手法 2 では隣接文どうしの単語で連結度を算出するため，他の品詞を加えることで逆に連結度の値が低下することが原因であった．一方で，ブログテキストの場合では手法 2 の F 値は 22.9 となり，手法 1 の F 値（20.9）よりも高いことが確認できた．ブログテキストは話し言葉で書かれていることが多いため，形態素解析の失敗により名詞が正しく抽出できない場合がある．これに対し，助詞と助動詞を除くすべての単語を用いることで，ある意味段落内において共通する単語をより多く見つけられることから，セグメンテーション精度の向上につながった．平均連結度を用いた場合（手法 1 と手法 3），ニューステキストでは F 値で 20.2 ポイント，ブログテキストでは 8.3 ポイントともに改善されており，平均連結度の有効性を確認することができた．また，平均連結度とすべての単語を使用する方法を組み合わせる（手法 4）ことで，手法 1 と比較して名詞だけでなく動詞なども含めて複数文間の共通語を連結度算出時に考慮できることから，手法 4 が最も良い結果を得ることができた．

表 9　固定しきい値法としきい値選択法との比較
Table 9　Comparison between the automatic threshold selection and the modified method using fixed threshold.

手法	選択しきい値	適合率（%）	再現率（%）	F 値
固定しきい値法	$D_T = 0.3$	41.8	50.0	45.5
しきい値選択法	（n.a）	**42.9**	**52.3**	**47.1**

3.3　しきい値選択法の効果の検証

3.3.1　実験の概要

この実験では，固定しきい値法としきい値選択法と比較を行った．具体的には表 6 における手法 4 に対し，固定しきい値法としきい値選択法をそれぞれ適用し F 値を調べた．本実験では 100 個のニューステキストデータと 299 個のブログテキストデータを混合した 399 個のテストデータを使用した．また，固定しきい値法では，しきい値を $D_T = \{0.1, 0, 2, \ldots, 0.9\}$ と設定し，その中からテストデータに対して最も良い結果が得られるしきい値を決定した．そして，しきい値選択法を用いた場合と固定しきい値法を用いた場合を F 値により比較した．

3.3.2　実 験 結 果

実験結果を**表 9** に示す．実験の結果から固定しきい値法の F 値が 45.5 であるのに対してしきい値選択法の F 値は 47.1 となり，しきい値選択法の結果の方が良いことを確認できた．また，実験結果について符号検定を行ったところ，$P = 0.005544$ となり有意水準 1%であっても 2 つの手法の性能差は統計的に有意であることが確認できた．固定しきい値法が適切な分割結果を選択できない例としては主に以下の 2 点があった．

- $D_T = 0.1$ よりも小さい値をしきい値としなければ正しく境界位置を抽出できない．
- $D_T = 0.25$ のように，今回設定したしきい値の中間的な値を使用しないと抽出ができない，または過剰に境界位置を抽出する．

上記 2 つの問題に対しては，しきい値 D_T をより細かく設定する方法が解決策として考えられる．しかし，一般的にしきい値の設定に関する情報が既知であることは少ない．また，あらかじめ用意されたテキストに対して適切なしきい値を求めたとしても，他のテキストに対して適切であるとは限らない．これらの点から，与えられたテキストに対して適切なしきい値を自動的に選択する方法は望ましい．表 9 の結果から，本論文におけるしきい値選択法は上記の問題を解決する 1 つの方法として有効であることが確認できた．

3.4 比較実験

3.4.1 実験の概要

実験では比較手法として Hearst 法[16]，C99 法[17]，そして名詞検索法[18] を用いた．Hearst 法は単語単位で単語間の連結度を求め意味段落の境界を推定する手法である．そのため，Hearst 法をそのまま適用すると文と文の間とは異なる位置を境界として抽出することがある．そこで，提案手法との出力結果を統一するため，本実験では文単位で単語を抽出して連結度を算出するように Hearst 法を実行し意味段落の境界を求めた．また，Hearst 法では主に文間の結束度を求める際に必要な窓幅のパラメータと意味段落の境界を抽出する際に使用するしきい値の設定が必要である．そこで，本論文では窓幅としきい値を調整し，最もテキストセグメンテーション結果が良いパラメータの組合せを用いて性能の比較を行った．C99 法のパラメータの設定は参考文献 17) に基づいて行い，最も良いセグメンテーション結果が得られるパラメータを使用した．この実験で使用したテストデータは 3.3 節で使用したものと同一である．

3.4.2 実験結果

比較実験の結果を**表 10** に示す．表 10 の結果から提案手法の F 値は 47.1 となり最も高い精度が得られることを確認できた（Hearst 法は 23.8，C99 法は 28.7，名詞検索法は 32.2）．Hearst 法や C99 法では与えられたテキストから算出した単語の文書内頻度に基づいてテキストセグメンテーションを行う．そのため，分割精度は表記揺れや単語の出現の仕方に影響を受けやすい．これに対し，提案手法ではテキスト内の単語だけでなく，関連記事から求めた関連語も利用することで，意味段落内において共通する単語の個数を増やすことができ，意味段落の境界ではキーワード集合に含まれる単語の変化が明瞭になる．この点から，Hearst 法や C99 法と比べて提案手法の F 値が高くなったと考えられる．

また，名詞検索法と比較しても提案手法の方が F 値で 14.9 ポイント高いことが確認できた．ニューステキストを用いたテストデータに対しては名詞検索法でも十分に意味段落の境界を特定することができる．しかし，ブログテキストに対しては検索語が抽出できないことや固定しきい値により意味段落の境界を十分に特定できないという問題がある．これに対し，提案手法では名詞以外の単語も検索語や関連語として利用している．そのため，ブログテキストにおいて，意味段落の境界におけるキーワード集合の単語変化が名詞検索法よりもさらに明瞭になり，適切な連結度が得られ意味段落の境界位置の特定の精度向上につながった．さらに，ニューステキストとブログテキストでは内容や書き方が大きく異なるため，一般的には両テキストに有効なしきい値を見つけることは難しい．一方で，提案手法では与えられテキストに対して内容ごとに細かく分割を行うしきい値を自動的に選択できる．これらの理由から，提案手法のテキストセグメンテーション精度は名詞検索法よりも向上したと考えられる．

表 10　比較実験の結果
Table 10　Result of comparison experiment among four text segmentation methods.

手法	適合率（%）	再現率（%）	F 値
Hearst 法	20.3	28.7	23.8
C99 法	25.6	32.8	28.7
名詞検索法	29.8	35.0	32.2
提案手法	**42.9**	**52.3**	**47.1**

以上の実験結果から，ニューステキストとブログテキストの両方に対して，提案手法が最も有効であることが確認できた．

4. 考　　察

4.1 しきい値選択法について

本論文では検索語と関連語の単語の出現頻度に基づいて内容ごとに細かく分割を行うための評価関数を提案した．そして，複数の分割結果から評価関数の値を最小にするしきい値を選択し 1 つの分割結果を得た．実験の結果から固定しきい値を用いた場合よりも良いテキストセグメンテーション結果を得ることが確認できた．一般的に，与えられたテキストに対して適切なしきい値が既知であることは少ない．そのため，固定しきい値を用いる方法では複数の値から適切な結果を選択する方法が主である．一方で，その値が普遍的に有効であるかどうかは不明であり，与えられたテキストごとにしきい値が設定できる方法が好ましいといえる．その点から，本論文で提案する評価関数は適切なしきい値を自動的に選択する評価基準として一定の効果を示したと考えられる．さらに，しきい値選択法と平均連結度は単語の出現頻度のみに基づいて算出されるため，Hearst 法や名詞検索法にも適用が可能である．実際に Hearst 法と名詞検索法に対して平均連結度としきい値選択法を適用し，ニューステキストとブログテキストの混合テストデータを用いて F 値を求めた結果を**表 11** に示す．表 11 において，従来手法に平均連結度としきい値選択法の両方を適用した手法がそれぞれ拡張 Hearst 法と拡張名詞検索法である．表 11 の結果から，平均連結度やしきい値選択法が Hearst 法と名詞検索法の精度向上に寄与していることが確認できる．

表 11　Hearst 法と名詞検索法に平均連結度としきい値選択法を適用したときの F 値
Table 11　F-value of the previous methods that proposed criteria were applied.

手法	適合率 (%)	再現率 (%)	F 値
Hearst 法	20.3	28.7	23.8
拡張 Hearst 法	22.2	40.8	28.8
名詞検索法	29.8	35.0	32.7
拡張名詞検索法	41.4	46.1	43.6

しかし，実験結果から正しい分割よりも細かく分割を行うものが選択される傾向もあり，特にブログテキストにおいてその傾向が確認された．具体的には，1 つの意味段落中で文章数が大きい場合に，式 (7) による評価は過分割した結果を選択することが多かった．ブログテキストにおいて，文章数が増加するといくつかの話題に触れながら 1 つの内容が記述されることが多い．そのような場合において，式 (7) により評価を行うと細かい話題ごとに意味段落が生成されることが多かった．つまり，内容ごとに細かく分割する目的は達成されたものの，式 (7) により得られる意味段落は人が判断した意味段落に必ずしも一致しないことが分かった．キーワード集合に現れる単語の統計的な特徴を解析し，人間が判断する意味段落に近い結果が得られるよう評価関数について今後も検証する．また，統計的な指標の 1 つであるクラス内分散・クラス間分散比に相当する基準を検討し，分割の良さを測る指標を構築する工夫が必要である．

4.2　一般語辞書について

本論文では様々な分野において普遍的に存在する単語（たとえば,「年」「ある」「大きい」）を一般語として扱う．しかし，本論文における一般語辞書は人手で作成されたため，一般語の選定基準が曖昧であった．その結果，話題が変化したにもかかわらず共通単語（たとえば，代表的な人名や地名などの固有名詞）がいくつか存在する場合があり，意味段落の境界の判定に悪影響を与えた．また，テキストの内容に応じて一般語としての扱いが異なることが多いため，検索語として利用できる単語が一般語として除かれるという問題もあった．そこで，ウェブが様々な分野のテキストの集合であると考えられる点から，ウェブ全体に存在するような単語を一般語とすることで，上記の問題が解決できるかどうかを検証する必要がある．または，一般語辞書を用いることなく得られた関連記事から関連語を抽出する方法や，頻繁に現れる人名や地名に対して少ない重みを付与するなど平均連結度を算出する際の工夫が必要である．

4.3　単語の抽出方法について

本研究では検索語と関連語の抽出方法に関して以下の方法を用いた．

- 検索語は各文に含まれる単語から一般語を除いてすべて使用．
- 関連語は一般語を除いて関連記事内の文書頻度が高い単語を使用．

単語の選別方法としては TF-IDF や RIDF など単語に重みをつけて選択する方法が考えられる．しかし，単語の重みを計算する際，計算結果は算出の元になったコーパスに依存する問題がある．また，人名などの固有名詞が特徴語となりやすいため，検索語や関連語が固有名詞に偏る傾向が見られた．しかし，本研究では各文の内容に関連する単語を幅広く収集することが重要である．検索語の省略形やジャンル，上位概念に相当する単語をウェブ検索により取得することで表記揺れや単語の出現パターンの影響を低減させることができる．つまり，固有名詞などの特徴的な単語を収集することが主目的ではない．そこで，本論文では TF-IDF などの方法により単語に重みをつけず，検索語に関しては一般語を除きすべての単語を使用した．また，関連記事内の文書頻度が高い単語は検索語に共起しやすい単語であることから関連語として使用した．

提案手法では，関連記事をウェブ検索により取得する際，検索語が過剰であるときや検索語の組合せが正しくないときに関連記事が少なくなることがあった．少ない関連記事から関連語を抽出すると，関連語が局所的な内容に偏る傾向があり意味段落の境界の抽出に悪影響を与える．そこで，本論文では関連記事数が少ない（関連記事数が $P_{\min}$ 未満）ときには検索件数が最大となるように 1 つの単語を削除しながら再検索を行った．しかし，どの単語の組合せでも検索件数が 0 件となることがあり，どの単語を削除するのか判断できない場合があった．そのような場合において，実験ではランダムに 1 つの単語を削除することで対処したものの，文の内容を表す特徴的な単語が削除される場合もあった．そこで，文の内容を考慮し単語を削除する方法を検討するなどの改良が必要がある．

本論文では検索語が抽出されなかった文に対しては，その前後の文にあるキーワード集合を用いて分割を行うかを決定した．しかし，可能ならばその文と前後の文との関連性を比較したうえで判断できることが望ましい．実験の結果から，提案手法においてニューステキストでは検索語が抽出できなかった文が 1.9%（36/1,865 文）であったのに対し，ブログテキストでは 11.2%（652/5,830 文）であった．ブログテキストの場合，独り言・挨拶・擬声語だけなどの短い文や，意味的な内容は含まれているものの日本語表記の正しくない文が多く存在する．これらの文に対して単語の抽出方法や検索語の設定方法を工夫することで，すべての文に対してウェブ検索を行い関連語を取得できるよう検討する必要がある．

4.4　ウェブ検索による利点と問題点について

ウェブ検索を用いることで以下の利点が考えられる．

- ウェブ上には幅広い分野のテキストが蓄積されているため，事前に学習データを用意しなくても数多くの分野のテキストを含むコーパスとして利用できる．その結果，処理対象となるテキストへの制約が生じるという問題を解決できる．
- ウェブにはつねに新しいテキストが逐次追加されるため，学習データの更新を行う手間が省ける．また，学習パラメータの更新を行わなくても，検索結果の上位にある関連記事を使用することで直接的に関連語を得ることができる．

事前に学習データを用意する手法においても，大規模なコーパスを用意できれば精度の向上や制約の問題を低減させることができる可能性は高い．しかし，ウェブに匹敵するコーパスを作成することは一般的に困難である．一方で，ウェブは数多くの分野の記事を含む大規模コーパスの 1 つであることから，事前に大規模なコーパスを用意しなくても，ウェブを利用することで処理対象となるテキストに制約が生じるという問題を回避できると考えられる．また，学習データを用いる方法では新しく学習データを追加するたびにパラメータを更新する必要がある．これに対し，ウェブ検索により得られた関連記事から直接的に関連語を抽出することで，内容の新旧を問わず検索語に関連する単語が得られるという利点がある．さらに，学習データを利用する方法ではパラメータの推定に悪影響を与えるテキストを何らかの方法であらかじめ除く必要がある．一般的には人手で判断する方法が多く，学習データの作成には多くの労力が必要となる．一方で，本研究ではウェブ検索により得られた検索結果の上位に含まれる記事だけを使用することで上記の問題を回避している．ウェブ検索において複数の検索語で AND 検索を行った場合，検索結果の下位になるほど，検索語の一部しか含まれない記事が増加するという性質を持つ．つまり，検索結果の下位にある記事を使用することで，検索語に関連しない単語が得られるため，テキストセグメンテーションの精度に悪影響を与える．これに対し，ウェブ検索の検索結果上位にはすべての検索語を含む記事が多く，検索語に対して内容的に近い記事が得られる傾向にある．これらの点から，検索結果の上位から順に記事を収集し関連語を得ることで，検索語に対して関連性の低い記事を除外しつつ文の内容に関連する単語を得ることができる．以上の理由から，ウェブ検索を直接使用する方法が効率的に検索語と関連する単語を得る手段であると考えられる．

一方で，提案手法のセグメンテーション結果は利用する検索エンジンに依存する．たとえば,「事件」でウェブ検索を行った場合，事件に関連する単語だけでなく最近話題になっている地名や人名も関連語に多く含まれることがある．このことから，利用する検索エンジンのランキング手法によって関連語が変化することや，最近の話題になっている内容に関連語が偏るという傾向が考えられる．現段階では，ウェブ上から複数の記事を収集することで関連語が 1 つの限定的な内容に偏ることを防いでいる．今後は何らかの基準を用いて検索語に関連する記事を十分取得できたと判定できるまで，複数の検索エンジンを利用するなど，利用上の工夫が必要である．

5. おわりに

本論文では，ウェブ検索に基づくしきい値選択型テキストセグメンテーション手法を提案した．具体的には，初めに各文から検索語を抽出しウェブ検索により関連語を得た．次に，検索語と関連語を用いて局所的な内容変化の影響を低減しつつ意味段落の境界では明瞭な内容変化を示す平均連結度を算出した．そして，意味段落の境界候補を決定するしきい値を平均連結度から算出し，それぞれのしきい値に対応する分割結果を生成した．最後に，内容ごとに細かく分割を行う結果を選択する評価関数を用意し，評価関数の値を最小にするしきい値を選択して 1 つのテキストセグメンテーション結果を得た．

提案手法の有効性を検証するために，ニューステキストとブログテキストを用いた実験を行いテキストセグメンテーションの性能評価を行った．また，学習データを必要としない手法として Hearst 法，C99 法，名詞検索法，そして提案手法の合わせて 4 つの手法を用いてテキストセグメンテーションの性能を比較した．性能評価は手法により得られた出力境界と正解の境界の位置を比較し，平均適合率と平均再現率から算出した F 値を用いて行った．比較実験の結果，提案手法の F 値は 47.1（Hearst 法 23.8，C99 法 28.7，名詞検索法 32.2）となり，名詞検索法に対して 14.9 ポイント精度が向上した．以上の点から，ニューステキストとブログテキストの両方に対して，提案手法の方が名詞検索法よりも良いテキストセグメンテーション結果を得ることが確認でき，提案手法の有効性を確認することができた．

今後の課題として，人が判断する分割結果に近づけるために評価関数の検討が必要である．また，再検索を行う際に前後の文の内容を考慮した検索語の削除方法について検討する．さらに，学習データを用いる従来手法との比較により，テキストセグメンテーションの計算時間や精度の面で提案手法の長所や短所を明らかにする．

参　考　文　献

1) 別所克人：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，情報処理学会論文誌，Vol.47, No.3, pp.957–967 (2006).

2) 西脇正通，田中英輝：関連記事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，Vol.2002, No.104, pp.79–84 (2002).
3) 西澤信一郎，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，Vol.97, No.4, pp.145–152 (1997).
4) 松井祥峰，乾　伸雄，小谷善行：単語の結束度と文の表層情報を組み合わせたテキストセグメンテーション，情報処理学会研究報告，Vol.2004, No.73, pp.151–158 (2004).
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) 平尾　努，北内　啓，木谷　強：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，情報処理学会論文誌，Vol.41, No.3, pp.24–36 (2000).
7) 越仲孝文，奥村明俊，磯谷亮輔：HMM の変分ベイズ学習によるテキストセグメンテーション及びその映像インデキシングへの応用，電子情報通信学会論文誌，Vol.J89-D, No.9, pp.2113–2122 (2006).
8) 金寺　登，隅田飛鳥，池端孝夫，船田哲男：ビデオ教材作成支援を目的とした講義音声によるシーン分割，電子情報通信学会論文誌，Vol.J88-D-I, No.5, pp.977–984 (2005).
9) Beeferman, D., Berger, A. and Lafferty, J.D.: Statistical Models for Text Segmentation, *Machine Learning*, Vol.34, No.1–3, pp.177–210 (1999).
10) Morris, J. and Hirst, G.: Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text, *Computational Linguistics*, Vol.17, No.1, pp.21–48 (1991).
11) Kozima, H. and Furugori, T.: Segmenting Narrative Text into Coherent Scenes, *Literary and Linguistic Computing*, Vol.9, pp.13–19 (1994).
12) Ponte, J.M. and Croft, W.B.: Text Segmentation by Topic, *Proc. 1st European Conference on Research and Advanced Technology for Digital Libraries*, pp.113–125 (1997).
13) Utiyama, M. and Isahara, H.: A Statistical Model for Domain-Independent Text Segmentation, *Proc. 39th Annual Meeting on Association for Computational Linguistics*, pp.499–506 (2001).
14) Brants, T., Chen, F. and Tsochantaridis, I.: Topic-Based Document Segmentation with Probabilistic Latent Semantic Analysis, *Proc. 11th International Conference on Information and Knowledge Management*, pp.211–218 (2002).
15) Stokes, N., Carthy, J. and Smeaton, A.F.: SeLeCT: A Lexical Cohesion Based News Story Segmentation System, *AI Communications*, Vol.17, No.1, pp.3–12 (2004).
16) Hearst, M.: TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, *Computational Linguistics*, Vol.23, No.1, pp.33–64 (1997).
17) Choi, F.Y.Y.: Advances in Domain Independent Linear Text Segmentation, *Proc. 1st Conference on North American Chapter of the Association for Computational Linguistics*, pp.26–33 (2000).
18) 阿部直人，田邊勝義，奥田英範：ウェブ検索を利用したテキストセグメンテーション，電子情報通信学会論文誌，Vol.J91-D, No.3, pp.723–732 (2008).
19) 木田充洋，外池昌嗣，宇津呂武仁，佐藤理史：ウェブを利用した専門用語の分野判定，電子情報通信学会論文誌，Vol.J89-D, No.11, pp.2470–2482 (2006).
20) Fuchi, T. and Takagi, S.: Japanese Morphological Analyzer using Word Co-occurence – JTAG, *Proc. 17th International Conference on Computational Linguistics*, pp.409–413 (1998).

(平成 20 年 4 月 15 日受付)
(平成 20 年 9 月 10 日採録)

阿部　直人（正会員）

平成 11 年北海道大学工学部情報工学科卒業．平成 13 年同大学大学院修士課程修了．平成 18 年同大学院博士課程修了．同年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入所．現在，パターン認識，画像処理，ウェブマイニング，自然言語処理の研究に従事．博士（工学）．電子情報通信学会，日本データベース学会各会員．

内山　俊郎（正会員）

昭和 62 年東京工業大学工学部電気電子工学科卒業．平成 1 年同大学大学院修士課程修了．同年（株）NTT データ入社．平成 3〜5 年南カリフォルニア大学客員研究員．平成 11〜17 年通信・放送機構研究員，特別研究員．平成 18 より日本電信電話（株）サイバーソリューション研究所所属．Web データマイニングの研究に従事．博士（工学）．

内山　匡（正会員）

昭和 60 年名古屋大学理学部物理学科卒業．昭和 62 年同大学大学院修士課程修了．同年 NTT に入社．平成 10〜13 年 NTT コミュニケーションズ，平成 16〜18 年 NTT レゾナントにてポータルサービスの開発等に従事．平成 19 年より NTT サイバーソリューション研究所主幹研究員．ポータルサービスシステムの研究開発に従事．電子情報通信学会，日本応用数理学会各会員．

奥　　雅博（正会員）

昭和 57 年大阪府立大学工学部電子工学科卒業．昭和 59 年同大学大学院工学研究科博士前期課程修了．同年日本電信電話公社（現 NTT）に入社．日本語処理技術の研究開発に従事．現在，NTT サイバーソリューション研究所において，検索をはじめとするブロードバンドインターネットサービスに関する研究開発に従事．博士（工学）．電子情報通信学会，言語処理学会各会員．