自然言語処理 130－6
（1999．3．4）

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

平尾 努　北内 啓　木谷 強
(株)NTT データ 情報科学研究所
{hirao,kitauchi,tkitani}@lit.rd.nttdata.co.jp

あらまし
本稿では，単語重要度と語彙的結束性に基づきテキストを段落に分割する手法を示し，更に両手法を統合した精度向上手法を提案する．新聞記事を対象として実験を行い，記事の形式段落を正解としてセグメンテーションの精度を適合率・再現率で評価した．訓練セットを用いてパラメータを決定し評価した結果，語彙的結束性を用いた手法では適合率 39%，再現率 26%，単語重要度を用いた手法では適合率 40%，再現率 36%であった．両手法を統合した結果，適合率は 39%，再現率は 52%に向上した．
キーワード テキストセグメンテーション，語彙的結束性，単語重要度，自然言語処理

Text Segmentation Based on Word Importance and Lexical Cohesion

Tsutomu HIRAO　Akira KITAUCHI　Tsuyoshi KITANI
Laboratory for Information Technology
NTT DATA Corporation
{hirao,kitauchi,tkitani}@lit.rd.nttdata.co.jp

Abstract
In this paper, we first describe a text segmentation method based on lexical cohesion and word importance, and then propose merging results from the both methods to raise the segmentation accuracy. This proposed technique was evaluated using newspaper articles. Recall and precision rates were used to compare the paragraph borders extracted by the techique with those in newspaper articles. Segmentaion parameters were determined using training data results. Precision and recall using lexical cohesion was 39% and 26%, while those using word importance was 40% and 36%, respectively. For merged results, precision was 39% and recall jumped to 52%.
Keywords Text Segmentation, Lexical Cohesion, Word Importance,Natural Language Processing

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により，様々な形式の電子化文書が広く流通している．その中には章見出しや段落分けなどの構造化がされていない文書も多く，文書を理解するうえで読み手にとって大きな負担となることがある．

また，文書が意味的なまとまりで構造化されていれば，読み手の文書理解を支援するだけでなく，情報検索技術やテキスト要約技術への応用も期待できる [6, 10]．

本稿では，新聞記事を対象としてテキストを自動的に段落[1] 分割 (セグメンテーション) する手法

[1]ここでいう段落とはテキストの内容から考えて意味的にま

について述べる．ここでは，まず従来より提案されている語彙的結束性を用いたテキストセグメンテーション法を改良した手法を提案する．次に，単語重要度を用いたテキストセグメンテーション法を提案する．更に両手法を統合した精度向上手法を述べる．

以下，2 章ではテキストセグメンテーション技術の動向を述べ，3 章では語彙的結束性を用いたセグメンテーション法について説明する．4 章では単語重要度を用いたセグメンテーション法を説明し，5 章では両手法の評価実験を行い，その結果をもとに両手法を統合する精度向上手法について述べる．6 章では，5 章の結果をもとに提案手法の有効性を考察する．

2 関連研究

文書を段落に分割する手法としては，

(1) 文書を予めある単位に分割し，隣接する単位を結合する手法

(2) 分割されていない文書から段落分割のための境界を探索する手法

の 2 種類が考えられる．(1) の手法として，[7] では重要語の続く隣接単位を連結する手法を採用している．(2) のとして，[9] では異なり語の増減情報をもとに異なり語の増加率が極小となっている部分を段落の境界としている．[4] では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，その総和より段落の境界を推定する．[1] では，テキスト中のある基準点に対する左右の窓の類似度を求め，類似度が高い部分を語彙的連鎖とみなし，類似度が低くなる部分で連鎖が切れ話題の境界であると考えている．

上述したとおり段落分割に関する手法はいくつか提案されているが，どれも十分な精度を達成してはいない．評価対象とするテキストが異なるため一概には言えないが，特に適合率が低いという問題がある．上述の手法では，20%～30%程度の適合率である．

段落分割の精度が十分でない場合には，分割したセグメントを用いた検索や要約の精度に悪影響を与えることが考えられ，より高精度な段落分割手法が求められている．

本稿では，上述の (2) の手法を採用し，語彙的結束性および単語重要度を用いた段落分割法をそれぞれ提案し，更に両手法を統合して従来よりも高精度な段落分割を実現する．

とまったブロックであり，いわゆる意味段落に相当するものである．

3 語彙的結束性を利用したテキストセグメンテーション

2 章でも述べたが，語彙的結束性を利用して段落分割する手法はいくつか提案されている．ここでは [1] の手法を説明し，その問題点を改良した手法について述べる．

3.1 TextTiling 法

[1] では，意味的に関連の深い部分には，同一の語彙が繰り返し出現するという性質を利用して段落分割[2]を行っている．ここでは，ある基準点に対して左右に一定の単語幅を持った窓を設け，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，グラフの極小部分を段落の境界と推定する手法をとっている．窓間の類似度は，次に示す *cosine measure* で表される．

$$sim\,(b_1, b_2) = \frac{\sum_t w_t, b_1 w_t, b_2}{\sqrt{\sum_t w_t^2, b_1 \sum_t w_t^2, b_2}} \quad (1)$$

ここで，b_1 と b_2 は，それぞれ左窓と右窓であり，w_t, b_1 と w_t, b_2 は，それぞれ，単語 t の左窓，右窓における出現頻度である．

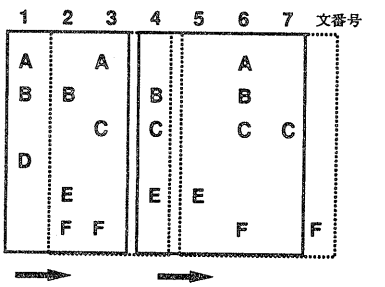

図 1: TextTiling 法

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考えると，基準点に対する左右の 9 単語の窓の類似度は，式 (1) を用いて，

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

となる．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

[2]TextTiling と名付けている

性が極小となる位置を境界とする．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 $depth\ score(d)$ を求め，

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

$depth\ score$ が以下の閾値 d_{th} を越えた場合に境界とする．ただし，d_{th} は類似度の平均，σ は類似度の分散である．

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

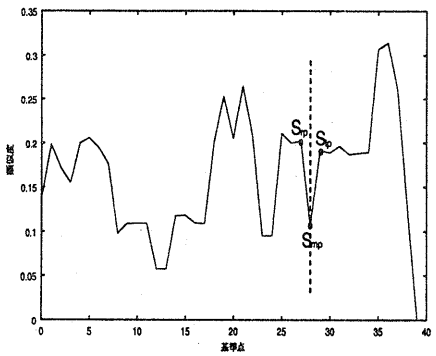

図 2: 基準点と類似度の関係

3.2 TextTiling 法の問題点

3.1 節では [1] の手法を説明したが，この手法には以下の 2 つの問題がある．

(1) 文書の先頭，末尾付近では左右同一の単語数を含む窓を設定できない．

(2) 窓幅が小さくなると，類似度が顕著に低くなる．

(1) 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に含まれる単語が少なくなることは避けることができない．(2) について，[11] のように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が，一般的に小さな範囲内で同一語が繰り返し出現することは稀である．よって，窓幅を小さくした場合には，左右の窓の類似性はみられなくなり，左右の窓の類似度が 0 となる基準点が多くなる．この場合は，語彙的連鎖があっても認定できず，話題境界の推定に悪影響を及ぼす．図 3 に窓が含む単語数を 17 語とした場合の基準点と類似度の関係を示す．類似度が 0 となる基準点があるため，類似度が極小となる基準点が認定できないことがわかる．

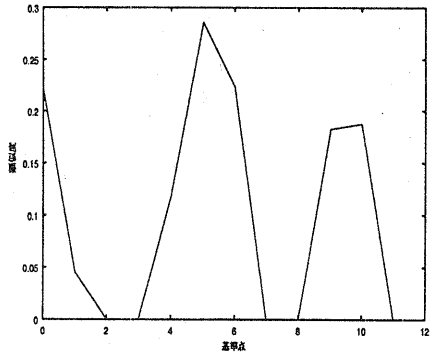

図 3: 窓幅を小さくした場合の基準点と類似度の関係

3.3 単語間の共起を考慮した語彙的結束性

3.2 節で述べたように小さな範囲内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間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 [6] ではシソーラスを用いている．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登録されているとは言えない．特に，「不況 → 倒産」のような連想関係は登録されていない．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間連語も併せて考慮する必要がある．

そこで，本稿では 3.2 節で述べた問題点の (2) 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．一般的に，ある語の近傍で頻繁に出現する語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 [3]．共起語は以下のように定義できる．

ある文書において，着目する単語 T が出現する全ての位置に対して，出現する文および前後 1 文に出現する T 以外の単語 W_j の総頻度 f_{W_j} が，

$$f_{W_j} \geq k \quad (k \text{ は任意の整数})$$

を満たす場合に，W_j を T の共起語とする．T が出現する毎に，関連語である W_j も同時に $\alpha(0 \leq \alpha \leq 1)$ 回出現したものとして，(1) 式での T の頻度に加算する．

図 4 で単語 C に着目すると，$k = 3$ とした場合には，B，F が共起語となり，C が 1 回出現すると B，F も同時に α 回出現したと考える．

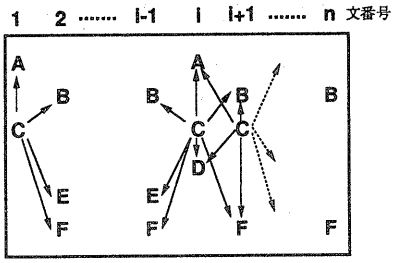

図 4: 単語の共起

上述のように，共起語を考慮して 3.1 節の式 (1) を用いて基準点と類似度の関係を表すと図 3 は図 5 のようになる．図 3 と比較すると類似度が 0 となる基準点が無くなり，値も大きく変わっていることがわかる．

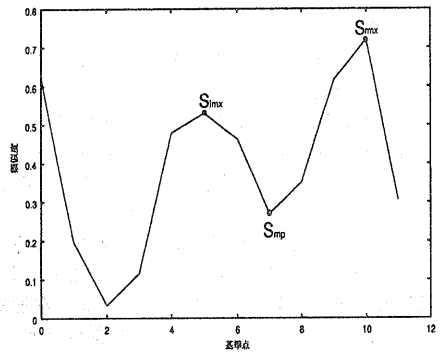

図 5: 共起を考慮した場合の基準点と類似度の関係

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 3.1 節での (2) 式，(3) 式は不適切となる．基本的には，極小点 mp において，グラフで大きく窪んでいる個所から順に段落境界とする手法をとる．

そこで本稿では，まず極小値 mp が含まれる範囲 (初期値は文書全体) において，左側の最大極大値 S_{lmx} と右側の最大極大値 S_{rmx} を用いて $depth\ score(d)$ を

$$d = min(S_{lmx}, S_{rmx}) - S_{mp}$$

と定義する．次に，以下の処理で $depth\ score$ の大きい mp から順に境界として認定していく．閾値 d_{th} は実験により最適値を決定する．

Step1 閾値 d_{th} に対して，$d \geq d_{th}$ を満たす d の最大値を与える mp を境界とする．

Step2 $d \leq d_{th}$ の場合は処理を終了する．そうで無い場合には Step1 で設定した境界による新たな範囲を対象に，d が最大となる mp を境界とする．

Step3 Step1，Step2 を繰り返す．

4 単語重要度を利用したテキストセグメンテーション

3 節では語彙的結束性，つまり，同一語や関連語の繰り返しという情報を用いて文書の段落分割を行う手法について述べた．ここでは，単語重要度に基づいた段落分割の手法について述べる．

4.1 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 [8] と文書集合全体から求める手法 [2] 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として後者を使用する．N 文書からなる文書集合中のある文書に出現する単語 T の重要度 $W(T)$ は，単語 T の出現文書数 df を用いて，

$$W(T) = \log\left(\frac{N}{df}\right) \tag{4}$$

と定義される[3]．

また，文書に含まれる任意の 1 文 L_k の重要度は，式 (4) を用いて，

$$W(L_k) = \sum_n W(T_n)/M \tag{5}$$

と定義する．これは，L_k に出現する単語の重要度の総和を L_k に出現する総単語数 M で正規化したものである．

4.2 単語重要度に基づく語彙的連鎖

名詞の連鎖に着目した [7] では，重要名詞の連続部分を語彙的連鎖として捉え，隣接する単位を結合するという手法をとっているが，本稿では語彙的連鎖を 4.1 節で定義した文の重要度を用いて認定し，連鎖の終端で段落に分割する手法を採る．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．

[3]式 (4) は一般的に IDF といわれる．

仮定 2 話題境界となる文は重要度が低い.

これはつまり，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．文の重要度の関係をグラフに表すと図 6 のようになり，グラフにおいて，重要度が極小値となる文を段落の最後の文とする．

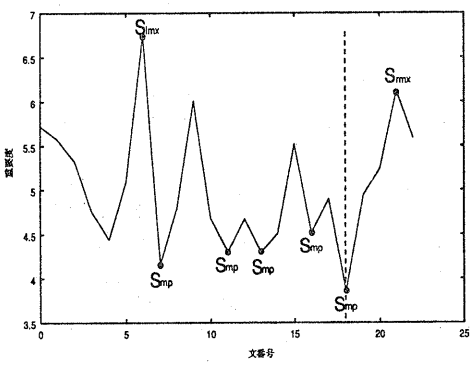

図 6: 文とその重要度の関係

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単調増加することによると考えられる．

なお，$depth\ score(d)$ については，3.3 節と同様に求め，閾値 d'_{th} は実験により最適値を決定する．

5 評価実験

3 章，4 章の手法を実装し，精度の評価実験を実施した．対象とする文書は，検索ベンチマーク用テストセット $BMIR-J2$[4] [5] に収録されている新聞記事から，文数が 20 文程度[5] のランダムに選んだ 60 件とした．なお，段落境界の正解は，新聞記事に付与されている形式段落とした．

評価指標として，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．

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

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

5.1 予備実験

語彙的結束性に基づく手法，単語重要度に基づく手法のそれぞれについて，上述した 60 文書を 30 文書づつの 2 つのセットに分け，一方の 30 文書を訓練セットとして、パラメータ (3.3 節，4.2 節での d_{th}，d'_{th}) の推定を行い，他方の 30 文書を評価用とした．なお，3.3 節での k，α は、経験的に $k=2$，$\alpha=0.1$ と設定した．また，窓幅は 17 単語と設定した．処理対象単語は名詞とした．表 1 に訓練セット 30 文書と評価セット 30 文書の平均文数と平均境界数を示す．両セットとも同様の傾向があることが分かる．

表 1: 訓練セットと評価セットの文書の傾向

	平均文数	平均境界数
訓練セット	21	6.3
評価セット	20	6

また，新聞記事の場合,「は」,「が」,「を」,「も」の助詞を含む文の次の文が段落境界にはならない場合が多いことが記事を分析した結果からわかった．そこで，これを経験則として採用した場合も併せて評価する．

語彙的結束性を利用した手法

訓練セットに対して d_{th} を変化させた場合の，適合率，再現率，出力境界数の変化を以下の表 2，表 3 に示す．

表 2: 語彙的結束性を用いた手法 (経験則なし)

d_{th}	適合率	再現率	出力境界数
0.01	0.39	0.28	4.6
0.02	0.39	0.28	4.5
0.03	0.39	0.28	4.5
0.04	0.39	0.28	4.4
0.05	0.39	0.27	4.3
0.06	0.39	0.27	4.3
0.07	0.40	0.27	4.2
0.08	0.38	0.25	4.1
0.09	0.38	0.24	4.0
0.10	0.37	0.23	3.9

[4](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.

[5]段落境界がある程度の数が存在することを推測して設定した

表 3: 語彙的結束性を用いた手法 (経験則あり)

d_{th}	適合率	再現率	出力境界数
0.01	0.40	0.27	4.2
0.02	0.40	0.26	4.1
0.03	0.40	0.26	4.1
0.04	0.41	0.26	4.0
0.05	0.41	0.26	4.0
0.06	0.42	0.26	3.9
0.07	0.40	0.24	3.8
0.08	0.40	0.24	3.7
0.09	0.41	0.23	3.6
0.10	0.39	0.21	3.5

単語重要を利用した手法

訓練セットに対して d'_{th} を変化させた場合の，適合率，再現率，出力境界数の変化を以下の表 4，表 5 に示す.

表 4: 単語重要度を用いた手法 (経験則なし)

d'_{th}	適合率	再現率	出力境界数
0.1	0.37	0.43	6.9
0.2	0.38	0.43	6.8
0.3	0.37	0.41	6.7
0.4	0.37	0.41	6.6
0.5	0.37	0.41	6.6
0.6	0.38	0.41	6.5
0.7	0.38	0.39	6.2
0.8	0.39	0.38	5.9
0.9	0.41	0.38	5.7
1.0	0.40	0.36	5.5

表 5: 単語重要度を用いた手法 (経験則あり)

d'_{th}	適合率	再現率	出力境界数
0.1	0.40	0.41	6.3
0.2	0.40	0.41	6.2
0.3	0.40	0.40	6.2
0.4	0.40	0.40	6.1
0.5	0.40	0.40	6.0
0.6	0.41	0.40	6.0
0.7	0.42	0.38	5.7
0.8	0.42	0.37	5.4
0.9	0.43	0.35	5.2
1.0	0.44	0.35	5.0

表 6 に語彙的結束性を用いた手法と単語重要度を用いた手法の正解出力結果の一部を示す．なお，表中の数字は文書中の境界となる文の番号を表す.

語彙的結束性を用いた手法では窓が含む単語数を 17 語に設定している．このため，文書の先頭，及び末尾から 3 ～ 5 文程度までは，片方の窓に含まれる単語数が 17 語に満たない．よって，文書の先頭，末尾の 3 ～ 5 文までの結束度は信頼性に欠ける．これは，表 6 において，出力した正解境界文番号が文書の半ばに多いことからも分かる．一方，単語重要度を用いた手法では，正解境界文番号が文書の先頭付近から末尾付近まで広く分布していることが分かる．これは，文の重要度の変化のみを手がかりとしているため，文書中の位置に係わらず段落境界の推定を行うことができるからである.

表 6: 出力境界文番号の異なり例

文書	語彙的結束性	単語重要度
1	10	2, 8
2	9	3, 7, 17
3	10, 12	2, 5
4	9, 16, 18	2, 5
5	3, 10, 13	3, 13
6	11, 15, 20	4, 9
7	5, 15	3, 5, 11, 19
8	3, 10, 13	3, 13
9	6	3, 11
10	11	4, 9, 11, 23

互いに出力する段落境界が異なった傾向であるならば，両手法の出力の和集合をとることで再現率の向上が期待できる．更に，両手法の出力の積集合が小さいので出力の和集合をとった場合にも適合率の低下は小さい．表 7に両手法の出力の和集合をとった場合の処理精度を示す.

表 7: 両手法の統合 (経験則あり)

d_{th}	d'_{th}	適合率	再現率	出力境界数
0.05	0.8	0.42	0.54	8.3
	0.9	0.42	0.54	8.1
	1.0	0.42	0.52	7.9
0.06	0.8	0.42	0.54	8.2
	0.9	0.43	0.54	8.1
	1.0	0.43	0.53	7.8
0.07	0.8	0.42	0.53	8.1
	0.9	0.42	0.52	8.0
	1.0	0.42	0.51	7.7

表 7 より，適合率を重視して 2 つのパラメータ d_{th}，d'_{th} について，$d_{th} = 0.06$，$d'_{th} = 0.9$ に設定した.

5.2 評価実験

評価セットを 5.1 節で設定したパラメータを用いて精度評価した結果を表 8 に示す.

表 8: 評価データに対する評価結果 (経験則あり)

d_{th}	d'_{th}	適合率	再現率	出力境界数
0.06	-	0.39	0.26	4.0
-	0.9	0.40	0.36	5.3
0.06	0.9	0.39	0.52	8.1

表 9: 適合率・再現率の理論値

d_{th}	d'_{th}	適合率の理論値	再現率の理論値
0.06	-	0.28	0.18
-	0.9	0.28	0.24
0.06	0.9	0.28	0.38

語彙的結束性を用いた手法，単語重要度を用いた手法共に適合率は40%程度となった．しかし，再現率を比較すると単語重要度を用いた手法の方が良い値となっている．両手法の和集合をとった場合には，適合率が語彙的結束性のみを用いた手法と同等，再現率は両手法と比較して大きく上回っている．

6 考察

5 章での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．

6.1 精度について

語彙的結束性を用いた手法である [1] の改良手法と単語重要度に基づく手法を提案した．両手法の精度を比較すると，単語重要度に基づく手法の方が適合率・再現率共に良い値となった．再現率については，語彙的結束性を用いた手法をより大きく上回っている．この原因としては，5.1 節でも述べたように，文書中の位置に係わらず段落境界を推定できるからである．単語重要度に基づく手法が語彙的結束性に基づく手法より精度が高いことから 4.2 節で導入した仮定も妥当であったと考える．また，表 7 と表 8 を比較すると評価セットを用いた場合が若干精度が低くなっているが，表 8とほぼ同様の傾向を示しており，設定したパラメータは妥当であったといえる．

一般に，l 文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，

$$E = \sum_{i=1}^{min(m,n)} \frac{i \times {}_mC_i \times {}_{l-1-m}C_{n-i}}{{}_{l-1}C_n}$$

となり，適合率・再現率はそれぞれ，

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

で計算できる．今回の評価実験において表 8 の出力境界数 n，正解境界数 m を用いて上式の適合率・再現率の理論値の平均を求めたところ，表 9 の結果を得た．

評価結果は，表 9 の値を全て上回っている．語彙的結束性を用いた手法の再現率以外は，理論値と10%以上の差がついている．両手法を統合した場合には，適合率では 11%，再現率では 14%理論値を上回っている．また，評価対象が異なるため，一概には言えないが従来の研究成果においては適合率が20%～ 30%程度であったことも考慮すると，提案手法は有効であるといえる．また，語彙的結束性に基づいた手法と単語重要度に基づいた手法の出力結果の和集合をとった結果，再現率は両手法と比べて向上し，適合率は単語重要度に基づく手法からわずかながら低下した．通常であれば，適合率は大幅に低下すると考えられるが，表 6 に示したとおり，出力する正解の傾向の相違が，適合率の大幅な低下を回避したと考えられる．ただし，セグメントを検索に用いる場合には，文書の重要セグメントを検索対象として適合率の向上を目的とし，要約に用いる場合には，文書の重要セグメントから要約文を抽出する処理を行う．よって，セグメントにある程度の文数が含まれていることが必要であり，出力境界数を増やして再現率を向上させることは適切ではなく，適合率を重視する必要がある．

また，今回は新聞記事を分析した結果として得られた経験則を導入したが，表 3 ，表 5 より効果があったといえる．このような言語の表層情報も有効な情報であり，精度向上には必要である．

評価対象とした文書は新聞記事であるが，書き方の特徴として話題毎に簡潔にまとめて書かれていることが挙げられる．このような構成がよく練られている文書では語彙的結束性や文の重要度の変化が顕著に現れるため，提案手法は有効であると考える．逆に文書の構成が練られていない場合には，提案手法は必ずしも有効であるとは限らない．

6.2 正解データについて

評価実験の正解データは新聞記事の形式段落としたが，[4] でも述べられているように形式段落と意味段落は必ずしも一致する訳ではない．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，段落分割に個人毎の揺れがあることが予想され，妥当性のある正解データを作成することが困難とな

る．新聞記事では簡潔に話題が述べられ，且つ，書き手も十分な訓練を受けているので，形式段落と意味段落を同一のものと考えた．実際の記事に目を通した限りでは特に問題が無かったと考える．

7 まとめ

語彙的結束性と単語重要度のそれぞれに基づいて，テキストを段落に分割する手法を提案した．更に，両手法の正解出力の傾向の違いに着目し，出力結果を併合して精度向上を図った．新聞記事を対象に記事の形式段落を正解として評価実験を行った結果，語彙的結束性を用いた手法では適合率39%，再現率26%，単語重要度を用いた手法では適合率40%，再現率36%であった．両手法を統合した結果，適合率は39%，再現率は52%に向上した．

参考文献

[1] Marti A. Hearst. TextTiling:Segmenting Text into Multi-paragraph Subtopic Passages. *Association for Conputational Linguistics*, Vol. 23, No. 1, 1997.

[2] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval.* McGrawHill, New York, 1983.

[3] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. *Proceedings of the 19th ACM SIGIR*, 1996.

[4] 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会, NLP-102-4,1994.

[5] 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会, DBS-114-3,1998.

[6] 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会, NLP-117-9,1997.

[7] 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会, NLP-117-20,1997.

[8] 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. 情報処理学会論文誌, Vol. 38, No. 2, 1997.

[9] 山崎誠. 文章の話題の展開を図る尺度. 計量国語学, Vol. 13, No. 8, 1983.

[10] 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会, FI-49-6,1998.

[11] 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, 1998.