複数の知識の組合せを用いたテキストセグメンテーション

望月 源, 本田 岳夫, 奥村 学
e-mail:{motizuki,honda,oku}@jaist.ac.jp
北陸先端科学技術大学院大学 情報科学研究科

[概要]

本稿では、我々はテキストのつながりを明示する語彙的結束性やその他の表層的な手がかりの情報を組み合わせて、テキストを段落に分割する手法について述べる。

本研究では、国語の現代文の問題集の段落分割の問題に対していくつかの段落分割の実験を行なう。実験ではパラメータとして、語彙的結束性、接続詞、修飾語などの表層的な情報を使用し、各パラメータの重み付けは人手と重回帰分析の両方で行なった。

[キーワード] 語彙的結束性, 表層情報, 重回帰分析, 段落分割

Text Segmentation used Combining Multiple Knowledge Sources

MOCHIDZUKI Hajime,HONDA Takeo,OKUMURA Manabu

School of Information Science, Japan Advanced Institute of Science and Technology
(Tatsunokuchi Ishikawa 923-12 Japan)

Abstract

In this paper, we present some methods of partitioning a text into segments with the aids of lexical cohesion and other surface clue informations.

The system's performances are judged by the comparision with segment boundaries marked as an attached model answer.

In this experiments, we take conclusivity of lexical cohesion and conjunction as well as modification,etc. as paramaters. The wait which multiplies to each paramater is determined by either hand or multiple regression analysis.

Key Words lexical cohesion, surface information, multiple regression analysis, segmentation

1　はじめに

テキストは単なる文の集まりではなく、テキスト中の文は互いに何らかのつながりを持つ。テキスト中にはこうしたつながりを捉える手がかりとなる表層上の情報が存在しており、それらを用いて談話構造の解析を行なう研究が多くなされている。本研究では、こうした手がかりを組み合わせて、段落分割を行なう手法を報告する。なお、段落には形式段落と意味段落という異なる想定ができるが、本研究では我々は　意味段落」を扱い、段落分割とは　意味段落」への分割のことをいう。

テキストのつながりを明示する表層的な情報として、語彙的結束性 [1] があり、我々は以前、この語彙的結束性を用いて、テキストを意味段落に分割する手法について報告した [15]。

しかし語彙的結束性の他にも、接続詞、修飾語、文のタイプなどのテキストのつながりを表す表層的情報が数多く存在する。こうした情報を抽出し、適切な重み付けを行なったものを段落分割に利用することは精度向上に有効であると考えられる。また、さまざまなテキストを一括して扱うよりも、テキストを分類し特徴的に近い分類毎に重みの計算などを推定した方が更に正確な段落分割が可能であると思われる。

こうした観点から、本研究では最初に語彙的結束性に基づく段落分割の実験を行ない。次に他の情報も重み付きのパラメータとして加えて、複数の情報を用いた段落分割の実験を行ない、複数パラメータの組合せによる段落分割の有効性を評価する。また、パラメータへの重みの推定方法として人手による経験的な決定方法の他に、重回帰分析による自動的な決定方法についても実験を行なう。次に、テキストを各パラメータの特徴から分類した上での段落分割について報告する。最後に重回帰分析の際に訓練データの数に対するパラメータ数の過多に起因すると思われる、重みの訓練データへの過適合を解消するための手法として、2 つのパラメータ除去基準をしめし、そのパラメータを用いて実験を行なう。

2　関連研究

語彙的情報やその他の表層的情報を用いた段落分割の研究がいくつか提案されている。

小嶋らは、異なり語に語彙的結束性の尺度を含めた LCP を提案している [8]。Hearst は、同語の反復による連鎖を用いた段落分割の手法を提案している [2]。

Litman and Passonneau は人間の被験者による段落分割と言語学的手がかりを用いた分割アルゴリズムの比較を行った [4]。彼らのアルゴリズムでは、言語学的手がかりとして参照名詞句、手がかり語、ポーズを用いて人手と機械学習の 2 つの分割規則の生成を行った。

また、段落分割以外の談話構造解析においても複数の知識を使用した研究が提案されている。

福本らは、主題、文タイプ、接続語句、同語反復などの情報を利用してテキストの構造化を行った [13]。

黒橋らは、手がかり語、同一 / 同義の語 / 句、2 文間の類似性の 3 つの表層的な情報を用いて科学技術文の文章構造の自動推定を行った [3]。

3　語彙的結束性による段落分割

ここでは語彙的結束性のみを用いた段落分割の手法について説明をする。

3.1　語彙的結束性

例えば、

例 1　膨張を続ける宇宙の中で数多くの星が誕生、消滅を繰り返しました。そして宇宙の誕生から約 100 億年後、他の星と同じ様にして、原始太陽を中心にして原始太陽系星雲と呼ばれるガスの円盤を作りました。

という文章には、意味的に関連のある単語の集まり { 星, 星, 星, 星雲 } が共起する。このような語の集まりを語彙的連鎖 [5] と呼ぶ。ここでは、シソーラス上の同一のカテゴリに属する語の集まりを語彙的連鎖として計算する。なお、本研究ではシソーラスに角川類語新辞典 [10] を用いる。

我々は、語義曖昧性を解消しつつ語彙的連鎖を生成する手法を提案している [15, 14, 6]。ここで、語彙的連鎖生成の手法を簡単に述べる。語彙的連鎖を漸進的に生成する過程で、語彙的連鎖を最近更新された語彙的連鎖が上位に来るようにスタック状に管理し、スタックの上位にある語彙的連鎖から順に、現在解析中の語との結束性を調べることで、その語の近傍の文脈が得ることができるので、語義曖昧性解消をしつつ、語彙的連鎖を生成することができる。

3.2　分割手法

ここでテキストの n 文目と n+1 文目の間の位置を $point(n, n+1)$ と呼ぶ。$point(n, n+1)$ を段落境界であるとみなす時、その語彙的連鎖の特徴により、点数 $scr_{chain_i}(n, n+1)$ を与える。テキスト中のすべての $point(n, n+1)$ で点数の総和 $scr(n, n+1) = \sum_i scr_{chain_i}$ を計算する。

この節での段落分割は、意味上のまとまりを示していると考えられる語彙的連鎖と、連鎖中で語の存在しない範囲で、我々がギャップと呼んでいる情報を用いて行なう。このギャップでは、別の話題について述べ、ギャップが終るとその連鎖に関する話題に戻っている可能性があると考えられる。

ここではこの連鎖とギャップの開始と終了の $point$ にそれぞれ点数を与える。

```
               |          text
 lexical chains|         1         2
      start-end|123456789012345678901234
chain a ( 1-24)|%%%%%%% %%# #%%%%%% ##%#
chain b ( 4-13)|   ##    %#%#
chain c (14-16)|              %%%
chain d ( 8- 9)|       %#
```

図 1: 語彙的連鎖の例

図 1 では横一列が一つの連鎖にあたり、#または%で連鎖中の語がどの文に含まれるかを示している。例えば、連鎖 d では、8 文目と 9 文目に連鎖中の語が含まれていることを表す。また、連鎖 b では、6 文目から 9 文目までの範囲に語が存在しない。このような連鎖中で語の存在しない範囲が連鎖のギャップである。従って、連鎖 d では $point(7,8)$(連鎖の開始)、$point(8,9)$(連鎖の終了) に、連鎖 b では $point(3,4)$(連鎖の開始)、$point(5,6)$(ギャップの開始)、$point(9,10)$(ギャップの終了)、$point(13,14)$(連鎖の終了) に、それぞれ点数を与える。

こうした連鎖とギャップの点数の総和が高い $point$ ほど段落境界になりやいことになる。

なおこの節での段落分割の計算では、連鎖、ギャップの開始、終了位置に与える点数は 1 点とし、各点数への重みも 1 として計算した。

4 他の表層的情報を加えた段落分割

前節では、語彙的結束性のみに基づいた段落分割について説明したが、テキスト内には、これ以外にも談話構造解析に役立つと考えられる表層的な情報がいくつか存在している。ここではそれらを含めてパラメータとし、各パラメータに人手による重み付けを行い段落分割をする方法について述べる。

4.1 使用するパラメータ

パラメータとして使用する表層的情報は具体的には次のものを用いた。

テキストの n 文目と n+1 文目の間 $point(n,n+1)$ に以下の表層的情報が出現した時対応するパラメータ P_i に点数を与える。

- グループ G_1:助詞の情報 $(P_1 \sim P_4)$
  n+1 文目に助詞「は」が出現したら、$point(n,n+1)$ のパラメータ P_1 に 1 点
  n 文目に助詞「は」が出現したら、$point(n,n+1)$ のパラメータ P_2 に 1 点
  n+1 文目に助詞「が」が出現したら、$point(n,n+1)$ のパラメータ P_3 に 1 点
  n 文目に助詞「が」が出現したら、$point(n,n+1)$ のパラメータ P_4 に 1 点
- グループ G_2:接続詞の情報 $(P_5 \sim P_{10})$
  n+1 文目の文頭に以下の接続詞が出現したら $point(n,n+1)$ の該当するパラメータ番号 P_x に 1 点
  - 「添加」の接続詞 (ex. しかも、そして)、P_5
  - 「強調」の接続詞 (ex. むしろ、とにかく)、P_6
  - 「説明」の接続詞 (ex. 例えば、つまり)、P_7
  - 「順接」の接続詞 (ex. ゆえに、だから)、P_8
  - 「逆接」の接続詞 (ex. しかし、だが)、P_9
  - 「転換」の接続詞 (ex. ところで、それでは)、P_{10}
- グループ G_3:前方照応の情報 $(P_{11} \sim P_{13})$
  n+1 文目の文頭に以下の前方照応が出現したら $point(n,n+1)$ の該当するパラメータ番号 P_x に 1 点
  - 「あ」型の前方照応詞 (ex. あの、あんな)、P_{11}
  - 「こ」型の前方照応詞 (ex. この、こんな)、P_{12}
  - 「そ」型の前方照応詞 (ex. その、そんな)、P_{13}
- グループ G_4:主語の有無 (P_{14})
  n+1 文目の文に主語がなければ $point(n,n+1)$ の P_{14} に 1 点
- グループ G_5:同一タイプの文の連続 $(P_{15} \sim P_{18})$
  n 文目と n+1 文目がどちらも以下のタイプの文である時、$point(n,n+1)$ の該当するパラメータ番号 P_x に 1 点
  - 叙述文 (ex. ~ている、ません)、P_{15}
  - 判断文 (ex. ~に違いない、と判断する)、P_{16}
  - 断定文 (ex. ~のである、なのだ)、P_{17}
  - その他、P_{18}
- グループ G_6:修飾語の情報 (P_{19})
  n+1 文目で修飾語の変わる連鎖がある時、$point(n,n+1)$ の P_{19} に 1 点を加算する。

P	重み	P	重み	P	重み
1	$0.25 \times$ 文数	2	$0.15 \times$ 文数	3	$-0.25 \times$ 文数
4	$0.15 \times$ 文数	5	$-3.5 \times$ 文数	6	$-3.5 \times$ 文数
7	$-0.1 \times$ 文数	8	$-0.5 \times$ 文数	9	$-0.5 \times$ 文数
10	$0.1 \times$ 文数	11	$-0.15 \times$ 文数	12	$-0.15 \times$ 文数
13	$-0.15 \times$ 文数	14	$-0.1 \times$ 文数	15	$-0.15 \times$ 文数
16	$-0.15 \times$ 文数	17	$-0.25 \times$ 文数	18	$-0.3 \times$ 文数
19	$0.20 \times$ 文数	20	$1.0 \times$ 文数	21	$1.0 \times$ 文数
22	$0.1 \times$ 文数	23	$0.1 \times$ 文数		

表 1: 人手により決定された重み

- グループ G_7:語彙的連鎖の情報 (P_{20} ~ P_{23})
  前節の場合と同様に、
  n+1 文目から始まる連鎖が存在する時、P_{20}
  n 文目で終る連鎖が存在する時、P_{21}
  n 文目から始まるギャップが存在する時、P_{22}
  n+1 文目で終るギャップが存在する時、P_{23}
  にそれぞれ 1 点を加算する。

G6,G7(P_{19},P_{20} ~ P_{23}) の情報は$point(n, n+1)$ に出現する数だけ加算していく。G1 ~ G5(P_1 ~ P_{18}) の各情報は出現したら 1 点、出現しなければ 0 点として点数を付ける。

なお、接続詞は [7, 9] を参照し、その機能により筆者が分類を行った。文のタイプは文末表現を手がかりにして 9 つに分類した。また、各パラメータとしてあげた情報には、$point(n, n+1)$ で切れ易い / 切れ難い両方の場合が含まれている。例えば同一タイプの文の連続は一般にその $point$ では境界になり難いが、逆接の接続詞の出現は境界になり易いと考えられる。このことは各パラメータへの重み付けの際に考慮する。

4.2 パラメータの重み付け

上で述べた各パラメータの点数に人手による重みづけを試みた。ある $point$ での分割され易さを表すと思われるパラメータには正の、分割され難さを表すと考えられるパラメータには負の重み付けを行っている。表 1に我々が決定した各パラメータの重みを示す。

ある $point(n, n+1)$ での切れ易さ $scr(n, n+1)$、そこでの各パラメータに重みをかけた値の総和によって計算される。

5 重回帰分析による重みの推定

前節では各パラメータの重みは人手により経験的に決定していたが、この節では、同様のパラメータに対して、その重みを重回帰分析の手法を用いて自動的に推定し、段落分割を行なう手法について述べる。重回帰分析によるパラメータの重みの推定に注目した研究としては、新聞記事の要約手法を提案した渡辺の研究がある [12]。渡辺は重回帰分析を計算するために与える S の値に被験者による結果を使用している。

5.1 重みの推定手法

$point(n, n+1)$ の各パラメータ $P1$ ~ $P23$ についての次式

$$S = a + \sum_{i=1}^{p} w_i \times P_i$$

(a は定数項、P_i はパラメータ i の点数、p はパラメータの数)

を考えると、訓練テキスト中の $point$ の数だけ得ることができ、S に何らかのを値を与えて重回帰分析を行うことで自動的に各パラメータの重み w_iを計算することができる。

本研究では、この式の S の値として以下の 4 通りを与えて実験を行なっている。ここで境界とは問題集における解答の $point$ を指す。

S_1 境界の $S = ($テキスト長 $- 1)/$境界数
非境界の $S = -1$

S_2 境界の $S = 10$
非境界の $S = -1$

S_{11} 境界の $S = ($テキスト長 $- 1)/$境界数
境界 $\pm i$ に位置する非境界の $S = ($境界の Sの値$)/2^i, (i = 1, 2, 3, ...)$

S_{12} 境界の $S_i = 10$
境界 $\pm i$ に位置する非境界の $S = ($境界の Sの値$)/2^i, (i = 1, 2, 3, ...)$

S_{11} と S_{12}の非境界では S の値が 1 以下になった時点で-1 とし、同じ位置で複数の値が計算される場合には値の高い方をその位置の S の値とする。

ここでの段落分割は人手による重みの代わりに、重回帰分析によって推定した重みを用いて$point(n, n+1)$ での境界になり易さ $scr(n, n+1)$ を計算する。

6 テキストの分類による段落分割

前節では、訓練テキスト全体を使用して重回帰分析を行い重みの推定を行ったが、ここでは、各テキストをパラメータの特徴から分類し、各分類ごとに重回帰分析による重みの推定を行う方法を述べる。

テキスト間の類似度の測定、クラスタ間の類似度測定には、それぞれ代表的な距離である内積と、平均距離を用いた [11]。

6.1 分類手続き

テキストの分類の手続きは、まず訓練テキストを分類し、次に各評価テキストがどのクラスに属するかを判定することによって行なう。

訓練テキストの分類の手続きは次のようになる。1. 各テキスト間の類似度を計算する。2.1 テキスト 1 クラスタとしてクラスタを作る。3. 各クラスタ間の類似度を計算する。4. 類似度が閾値に満たなくなるまで、最も類似度が高いクラスタどうしを融合し 3 に戻る。

最終的に決定されたクラスタ毎に重回帰分析による重み値の推定を行なっておく。

次に、各評価テキストの属するクラスの判定手続きは次のようになる。1. 各クラスタと評価テキストの類似度を計算する。2. 最も高い類似度が閾値以上ならば、そのクラスに属するものとし、閾値に満たなければどのクラスにも属さないものとする。

テキスト間の類似度計算には、各パラメータの値から計算されるベクトル[1] どうしの内積を用いた。

テキスト i と jの類似度はそれぞれのベクトル $V_i = (v_{i1}, v_{i2}, \cdots, v_{ip})$,$V_j = (v_{j1}, v_{j2}, \cdots, v_{jp})$ の内積を求める次式によって求めた。

$$Sim(V_i, V_j) = \sum_{k=1}^{p} v_{ik} v_{jk}$$

任意の 2 クラスタ間の類似度は平均距離を用いて計算した。

任意のクラスタ C_i, C_j間の類似度 D_{ij} は、

$$D_{ij} = \frac{1}{n_i n_j} \sum_{\alpha \in C_i} \sum_{\beta \in C_j} d\alpha\beta$$

(ここで n はテキストの数を表す。)

クラスタの融合基準となる閾値は各テキスト間の類似度の平均値を基本として平均の 0.6 倍から 1.4 倍の間で変化させて計算した。

段落分割は、評価テキストが属すると判定されたクラスで計算された重みを使用して $point(n, n+1)$ での切れ易さ $scr(n, n+1)$ を計算することにより行なう。ただし、どの分類にも属さないと判定された評価テキストには、分類を行なう前の全訓練テキストから推定した重みを用いて段落分割を行った。

7 パラメータの除去

5、 6節で訓練テキストから重回帰分析による重みの推定に用いるデータとして取得した式は平均 385 個であり、各式のパラメータの数は 23 個であった。

8節で後述するように、実験結果、表 6、 7ともに訓練テキストでの結果の割に、評価テキストでの結果があまり向上してない。

この原因として、データ数に対するパラメータ数が多過ぎるために訓練テキストに対して、過適合を起こしていることが考えられる。そこで、2 つの基準からパラメータを除去した上で、重回帰分析を試みた。

7.1 除去方法 1

この方法は、あるパラメータが 1 度も出現しないテキストの数のテキスト全体に対する割合 (0.1 ~ 0.9) によって、そのパラメータを除去するものである。

各割合に応じた除去されるパラメータは表 2のようになった。

割合	除去パラメータ番号
0.1	5,6,7,8,9,10,11,12,13,14,15,16,17,18
0.2	5,6,7,8,9,10,11,12,13,16,17,18
0.3	5,6,7,8,10,11,12,13,16,17,18
0.4	6,7,8,10,11,13,16,17,18
0.5	6,7,8,10,11,13,16,18
0.6	6,7,10,11
0.7	6,7,10,11
0.8	6,10,11
0.9	6,11

表 2: 除去パラメータ 1

7.2 除去方法 2

もう 1 つの除去方法は、パラメータを情報の種類からグループ化して、除去するパラメータをグループ単位で組み合わせていくものである。このグループは 4 節で説明したパラメータのグループ G_x に対応している。表 3の組合せで除去した。

8 実験

ここまで説明してきた一連の段落分割の手法について、実際のテキストを用いて実験を行なった。

段落分割の評価には、現代文の受験参考書の問題「本文を N段落 (段落境界数 $N-1$) に分けるとすると、どこで切るのがもっとも適当か」の解答との一致を用いる。

実験では出力境界数 M を scrの上位から順に境界候補数 $\times \frac{1}{3}$個まで出力させる。評価には情報検索の分野で用いられる再現率 (Recall)、適合率 (Precision) と

[1] ベクトルの各要素は、テキストの同一パラメータの点数の総和をそのテキストの境界候補数 ($point$ 数) で割ったものである。

グループ	除去パラメータ番号
G:2	5 ~ 10
G:3	11 ~ 13
G:5	15 ~ 18
G:4	14
G:2+3	5 ~ 10,11 ~ 13
G:2+5	5 ~ 10,15 ~ 18
G:2+4	5 ~ 10,14
G:2+3+5	5 ~ 10,11 ~ 13,15 ~ 18
G:2+3+4	5 ~ 10,11 ~ 13,14
G:2+4+5	5 ~ 10,14,15 ~ 18
G:3+5	11 ~ 13,15 ~ 18
G:3+4	11 ~ 13,14
G:3+4+5	11 ~ 13,14,15 ~ 18
G:4+5	14,15 ~ 18
G:2+3+4+5	5 ~ 10,11 ~ 13,14,15 ~ 18

表 3: 除去パラメータ 2

いう評価指標を用いる。再現率、適合率は次の式で与えられる。

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

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

我々は国語現代文の問題集から 24 のテキストを採集した。このテキストに対して上記の各実験を行なった。

表 4 に語彙的結束性に基づく段落分割の結果を示し、表 5 に他の表層的情報を加えた段落分割の結果を示す。$point(n, n+1)$ の scr は、パラメータの各点数に表 1の重みをかけた値の合計として計算する。いずれの値も 24 テキストの平均値である。

5節から 7節までの重回帰分析による実験では次のように訓練テキストと評価テキストを分けて使用した。

24 のテキストを 8 テキストづつ 3 組にわけ、そのうち 2 組 16 テキストを訓練用、残りの 1 組 8 テキストを評価用のデータとした。この組合せを変えることにより 3 回の cross-validation の平均によって評価を行なった。

表 6に重回帰分析による重みを使用した段落分割の結果を示す。

表 7に内積によるクラスタリングの実験結果を示す[2]。

表 8に除去方法 1 によってパラメータを減少させた結果を示し、表 9に除去方法 2 によってパラメータを減少させた結果を示す。

なお、表 7、表 8、表 9では 4 通りの S の与え方の内、最も良かった S_2についてのみの結果を示した。他の S はこれらよりも良くなかった。

[2]表中の a 欄は作成されたクラスタに分類された評価テキストの数を表し、b 欄はいずれのクラスタにも分類されなかった評価テキストの数を表す。

再現率	適合率
0.422917	0.198923

表 4: 語彙的結束性に基づく段落分割の結果

再現率	適合率
0.492361	0.243690

表 5: 他の表層的情報を加えた段落分割の結果

	訓練		評価	
S	再現率	適合率	再現率	適合率
S_1	0.642708	0.273611	0.363889	0.167484
S_2	0.653819	0.290501	0.506250	0.228428
S_{11}	0.487500	0.195426	0.404167	0.177533
S_{12}	0.542708	0.215464	0.449306	0.211402

表 6: 重回帰分析による重みを使用した分割結果

訓練テキスト				評価テキスト	
再現率	適合率	a	b	再現率	適合率
閾値=平均*0.6					
0.653819	0.290501	24	0	0.506250	0.228428
閾値=平均*0.7					
0.677083	0.308438	20	4	0.582639	0.256495
閾値=平均*0.8					
0.729167	0.314741	18	6	0.593056	0.261623
閾値=平均*0.9					
0.807292	0.334458	18	6	0.513194	0.235034
閾値=平均*1					
0.810764	0.338436	16	8	0.374306	0.187852
閾値=平均*1.1					
0.984375	0.427572	18	6	0.313889	0.164289
閾値=平均*1.2					
0.996528	0.435053	18	6	0.350000	0.180658
閾値=平均*1.3					
1.000000	0.436947	16	8	0.391667	0.204460
閾値=平均*1.4					
1.000000	0.436947	15	9	0.398611	0.202079

表 7: 内積によるクラスタリングの結果

	訓練テキスト		評価テキスト	
割合	再現率	適合率	再現率	適合率
0.1	0.592014	0.267224	0.514583	0.227765
0.2	0.597222	0.266654	0.507639	0.224257
0.3	0.597222	0.265612	0.486806	0.225516
0.4	0.610417	0.272838	0.513194	0.233185
0.5	0.615625	0.277004	0.513194	0.233185
0.6	0.626042	0.284582	0.516667	0.232229
0.7	0.626042	0.284582	0.516667	0.232229
0.8	0.636458	0.289790	0.520139	0.228285
0.9	0.650347	0.286334	0.506250	0.228428

表 8: 除去方法 1 の結果

	訓練テキスト		評価テキスト	
グループ	再現率	適合率	再現率	適合率
G:2	0.647569	0.293808	0.521528	0.231634
G:3	0.626042	0.275442	0.485417	0.218011
G:5	0.600694	0.269332	0.523611	0.240643
G:4	0.639931	0.281426	0.506250	0.228428
G:2,3	0.633681	0.284854	0.500694	0.221217
G:2,5	0.614583	0.276029	0.563194	0.236276
G:2,4	0.619792	0.280905	0.521528	0.231634
G:2,3,5	0.614583	0.280427	0.486806	0.231119
G:2,3,5	0.595486	0.273765	0.500694	0.227004
G:2,4,5	0.604167	0.273053	0.570139	0.239023
G:3,5	0.589583	0.261569	0.454167	0.222041
G:3,4	0.615625	0.270234	0.485417	0.218011
G:3,4,5	0.574653	0.259039	0.488889	0.227021
G:4,5	0.579167	0.262078	0.523611	0.240643
G:2,3,4,5	0.592014	0.267224	0.514583	0.227765

表 9: 除去方法 2 の結果

9 考察

最初の実験で行なった語彙的結束性による段落分割の実験結果よりも、2 番目の他の表層的情報を加えた場合の方が良い結果を得ている。これは複数の情報を組み合わせて使用した方がより適切な段落分割が可能になることを示している。

また、3 番目の実験では各パラメータへの重み付けを行なう手法として重回帰分析による重みの推定を行なった。しかし、テキストの組合せによって精度にばらつきがある。これは異なる特徴を持つテキストを一律に使用して重回帰分析を行なったために、特徴が合わないテキストで得られた重みによって、段落分割を試みた場合に極端に精度が落ちたためだと考えられる。

そこで、4 番目の実験では、訓練テキストをパラメータを用いたベクトルの内積をもとに分類し、各分類毎に重みを推定した上で、各評価テキストを閾値以上の類似度で最も類似した分類の重みを用いて分割を行なう手法をとった。この手法によると、閾値を類似度の平均 $\times 0.8$ に設定した時、最も良い結果を得た。これは他のどの手法よりもかなり良いものであった。しかし、閾値をこの値に設定する基準は今のところ明らかでない。

また、重回帰分析による訓練結果に対して評価結果があまり向上せず、推定された重みが過適合する傾向が見られる。原因としてパラメータの数が訓練データの数に対して多過ぎることが考えられる。そのため最後の実験では 2 つの基準によって、パラメータの数を減少させて重回帰分析を行なった。除去方法 1 では割合が $0.6 \sim 0.7$ の時に最も良い結果を得た。除去方法 2 では接続詞、主語の有無、同一文タイプの連続 (G2,4,5) の情報に関するパラメータを除去した時に最も良い結果を得た。どちらの場合もすべてのパラメータを用いた重回帰分析による場合よりも良い結果を得た。このため一定の基準を設けてパラメータを減少させることは段落分割にとって有効であると考えられる。

10 おわりに

本研究ではまず、語彙的結束性に基づく段落分割の実験を行ない、次にその他の表層的な情報もパラメータとして組み合わせて用いる段落分割の実験を行ない、次に訓練テキストをパラメータの特徴から分類し、各分類毎に重みの推定をした上で、各評価テキストを段落分割する実験を行なった。いずれの実験も実際の国語現代文の問題に対し実験を行なった。

結果として、訓練テキストをパラメータの特徴から分類し、各分類毎に重みの推定をした上で、各評価テキストを閾値以上の類似度で最も類似した分類の重みを用いて分割を行なう手法が最も良い結果を得た。この時の分類に用いる最も良い閾値は類似度の平均 $\times 0.8$ であったが、この値に設定する基準がないため今後この基準を明らかにしなければならない。

また、重回帰分析の際に、データの数に対してパラメータの数が多いことから起こると思われる、推定された重みの訓練テキストへの過適合に対処するため、2 つの基準で除去した後のパラメータを用いて重回帰分析による段落分割の実験を行なった。

除去方法 1、2 どちらの方法も設定した全てのパラメータを用いた重回帰分析による場合よりも良い結果を得た。このため、一定の基準を設けてパラメータ数を減らすことは重回帰分析を用いた段落分割に有効であることがわかった。なお、除去方法 2 で結果の良

かった除去パラメータに接続詞が含まれていたが、これは談話構造における一般的な接続詞の捉え方と異なる結果となった。このことについてはもう少し詳しく調査する必要があると思われる。

今後の課題として、最後に実験を行なった除去後のパラメータを用いて、訓練テキストを分類する手法による段落分割の実験を行ない、どの程度段落分割の精度をあげられるかを確かめることがあげられる。

参考文献

[1] Michael Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, 1976.

[2] Marti A. Hearst. Multi-Paragraph Segmentation of Expository Texts. Technical report, UC Berkeley Science Technical Report, 1994.

[3] Sadao Kurohashi and Makoto Nagao. Automatic Detection of Discourse Structure by Checking Surfce Information in Sentence. In *COLING'94*, pp. 1123–1127, 1994.

[4] Diane J. Litman and Rebecca J. Passonneau. Combining Multiple Knowledge Sourse for Discourse. In *the 33rd ACL*, 1995.

[5] Jame Morris and Graeme Hirst. Lexical Cohesion Computed by Thesaurual Relations as an Indicator of the Structure of Text. *Computational Linguistics*, Vol. 17, pp. 21–48, 1991.

[6] Manabu Okumura and Takeo Honda. Word sense disambiguation and text segmentation based on lexical cohesion. In *COLING94*, pp. 755–761, 1994.

[7] 所一哉. 現代文レトリック読解法. 匠出版, 1987.

[8] 小嶋秀樹, 古郡廷治. 単語の結束性にもとづいてテキストを場面に分割する試み. 情報処理学会研究会資料 NL95-7, pp. 49–56, 1993.

[9] 成田杢之助. 口語文法表覧. 共文社.

[10] 大野晋, 浜西正人. 角川類語新辞典. 角川書店, 1981.

[11] 鳥脇純一郎. 認識工学-パターン認識とその応用-. コロナ社, 1993.

[12] 渡辺日出雄. 新聞記事の要約のための一手法. 言語処理学会第 1 回年次大会, pp. 293–296, 1995.

[13] 福本淳一, 安原宏. 文の連接関係解析に基づく文章構造解析. 情報処理学会研究会資料 NL88-2, pp. 9–16, 1990.

[14] 本田岳夫. 日本語文章の語彙的結束性に関する基礎的研究. Master's thesis, 北陸先端科学技術大学院大学, 1994.

[15] 本田岳夫, 奥村学. 語義曖昧性を考慮した有意な語彙連鎖の生成. 情報処理学会研究会資料 NL97-14, pp. 95–102, 1993.