自然言語処理 117－20
（1997. 1. 21）

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

西澤 信一郎　中川 裕志
横浜国立大学 工学部 電子情報工学科
E-Mail: {shin, nakagawa}@naklab.dnj.ynu.ac.jp

概要: 計算機可読な文書が増大しつつある現在，利用者がこれらから効率良く情報を得るための支援システム・ツールが必要不可欠である．特に，話し言葉からの書き起こし文書 (座談会など) では，段落などの構造や章見出しなど基本的な情報が存在しないため，文書検索などのためにもこれらを自動的に得ることが重要となる．本研究では，このような書き起こし文書を対象とし，文書解析の重要な要素である段落分割について検討する．手法としては，情報検索で良く用いられる *tf* や *idf* という数値を用い，シソーラス情報を用いずとも比較的良好な結果を得ることが出来ることを確認した．

Bottom-up Discourse Segmentation based on Word Frequency

Shin'ichiro Nishizawa　Hiroshi Nakagawa
Yokohama National University
E-Mail: {shin, nakagawa}@naklab.dnj.ynu.ac.jp

Abstract: Now, we need computer aided systems to get useful informations such as keywords, abstracts, subtopic structures, etc. from documents, because there are a lot of documents which are available on computer. We discuss here a method of Japanese discourse segmentation, which is one of the most important part for text analysis. Our target documents are transcriptions of a discussion among several people, and we pay our attention to word frequency in the document. As a result of our tests for that method, we show better performance than already proposed methods. Numerically the results are about 70% recall and about 20% precision compared with human judgments about discourse boundaries.

1　はじめに

章見出しや段落などによる構造化がなされていない文書を読むという行為は，その文書の利用者にとって大きな負担となる．これらの構造は，話のまとまりや話題の流れ，アウトラインなどを利用者に与えるものであり，利用者が文書を利用するたびにこれらをゼロから構築することは非常に手間がかかり，効率が悪い．このような，いわゆる「ベタ書き」の文書の例として，話し言葉からの書き起こし文書が挙げられる．この中には，座談会，対談や会議の記録など重要な情報を含むものも多く，これらの文書へ利用者が手軽にアクセスし，必要な情報を得るためには，話題に応じた段落分け，大まかな章見出しの設定，各段落におけるキーワード設定などといった処理をあらかじめ行なう必要がある．計算機可読な文書が増大しつつある現在，これらの作業を自動的に行なう支援ツールの必要性もまた大きくなっていると考えられる．本研究では，話し言葉の書き起こし文書を対象とした段落分割 (セグメンテーション) について検討する．ここでいう段落とは，テキストの内容から見てひとまとまりとなっているようなブロックということであり，いわゆる意味段落に相当する．ここでは，意味段落を認識するための要素として，主に「名詞の出現頻度」を利用する段落分割の手法について述べる．また，実際の書き起こし文書を対象として，人間による段落分割の結果との比較実験を行なった．

2 関連研究

段落分割の研究としては，シソーラスから得られる語彙的結束性(語の類縁性)を主な情報として利用するものがいくつか提案されている．

[本田 94] では，文書中に現れる，意味的に関連がある語の集まりを語彙的連鎖としてとらえ，この開始，終了位置，また語が現れないギャップの位置などに得点を与え，その総和から段落境界を推定する．[Koz93] では，語の類似性の尺度として，テキスト中のある単語とその周辺に存在する単語間の相互の結束性を示す LCP を提案し，文書全体でこの値が谷になる部分を段落境界とみなす方法を述べている．[山本 92] では，文書に含まれる語間の結束の強さをシソーラスでのカテゴリーのレベルに合わせて三種類の値とし，これによる評価関数を設定している．段落分割の手順は，評価関数の値が最も改善されるように隣接段落を連結していき，改善がそれ以上なされなくなった時点で段落分割が完了，としている．また，この手法では，「手がかり語」として接続詞，副詞などの情報も併せて用いている．同様に，語彙的結束性や手がかり語の情報など複数の知識を用いるものとして [望月 95，望月 96] がある．さらに，同一語の文書中での出現頻度を用いる [Hea94] の手法がある．ここでは，隣接ブロック間の類似度を cosine measure で計算し，この値の変化から段落境界を推定している．

以上の研究に対し，本研究ではシソーラス情報を用いず，文書中に出現する同一名詞の出現の状態に着目するものである．ただし，[Hea94] とは異なり，隣接ブロックとの類似度を計算するのではなく，名詞が文書中の連続ブロックで出現するかどうかを判断材料とする．また，全名詞を対象とするのではなく，文書中で重要語と考えられる名詞を出現頻度などを参考に取り出し，それらの名詞の出現状態に着目するものである．

なお，自然会話コーパスの段落分割を目的とし，名詞の照応，手がかり語，ポーズなどの情報を用いる [PL93] の研究がある．本研究でも座談会の書き起こしなどある程度話し言葉の特徴を残していると考えられる文書を対象とするが，あくまでも書き起こされたテキストを対象とするものである．よって，ポーズなど語彙情報以外の情報は用いない．

3 名詞の出現頻度に基づく段落分割

前述したように，本研究では段落分割のために文書中の名詞の出現の状態(出現頻度)を利用する．ここではその方法について説明する．なお，段落分割の手法としては，(1) 対象とする文書をある単位 (文，段落など) であらかじめ分割しておき，段落として分かれない隣接単位を連結することで段落分割を行なう方法，(2) 対象とする文書をひとまとまりとみなし，段落として分割可能な位置を探索する方法，の二通りが考えられるが，本研究では (1) の手法をとる．また，文書は初期状態として一文毎に分割されているものとする．本稿では，この一文毎および処理が進むに従ってこれらが連結されて生成されるものをブロックと呼ぶ．すなわち，文書の初期状態では “一文 = 一ブロック” であり，処理の進行につれて一ブロックあたりに含まれる文数が増加していく．これによって段落が形成されることになる．

3.1 tf.idf による重要語のランキングを利用した手法

検索を目的とした文書のランクづけを行なう際に，文書の構造 (章，節など) によらずに文書全体を同サイズの領域に分割し，各領域内の単語で算出される単語の重みを利用する，という手法がある [HP93]．これを参考とし，ここでは，文書内の名詞の *tf.idf* を用い，前記の方法で各領域毎に上位にランクされる名詞を抽出し，それらが連続して出現する隣接ブロックを文書全体で連結していくことにより段落分割を行なう手法を以下のように示す[1]．

手順 1 (tf.idf 連結法)

1. 対象とする文書 (全体が N ブロックから成る) を，先頭から k ブロック毎の領域に分割する．
2. 各領域毎に，そこに含まれるすべての名詞について，名詞毎の重み $w_{i,j}$ を次のように求める [FBY92]．

$$w_{i,j} = freq_{i,j} \times idf_i$$

$$freq_{i,j} = \frac{\text{領域 j における 名詞 i の出現回数}}{\text{文書全体における 名詞 i の出現回数}}$$

[1] 以降で **tf.idf 連結法** と記述する．

$$idf_i = \log_2 \frac{\text{全領域数}}{\text{名詞 i を含む領域数}} + 1$$

3. $w_{i,j}$ が閾値 w_{th} 以上である名詞 i を各領域毎に選び，それらの和集合を重要語集合 W とする (図 **1** 参照).

4. 文書の先頭より，W に含まれる名詞が連続して出現するブロックを探索し，それらを連結して一つの新しいブロックとする.

5. 連結作業の終了した文書に対して，繰り返し上記の 1. からの手順を実行する．繰り返しの終了条件は，連結作業の前後で文書全体のブロック数が変化しない場合とする.

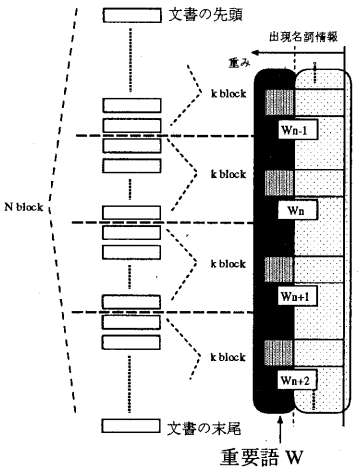

図 1: 等幅領域による重要語の抽出

3.2 idf を利用した手法

文書全体における，ある語についての idf (inverted document frequency) は，一般に $\frac{\text{全文書数}}{\text{その語を含む文書数}}$ で決まる．つまり，文書をいくつかの決まった数のブロックに分割した時，ある語が少ないブロックにまとまっているほど idf の値が大きいこととなる．これを利用した段落分割の手法を以下に示す[2].

手順 2 (idf 連結法)

[2] 以降で，**idf 連結法**と記述する.

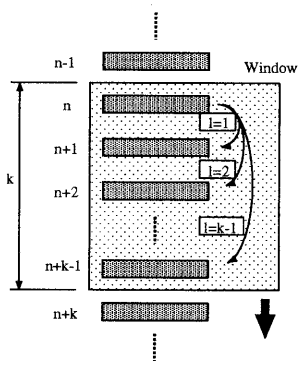

図 2: idf の値の変化を利用した段落分割

1. 対象とする文書 (全体が N ブロックから成る) について，文書の n ブロック目から k ブロック分をカバーするような窓を想定する.

2. 窓の内部について，先頭のブロックより一ブロックずつ順に連結したと仮定し，文書中に出現する全名詞について，$IDF_{n,l}$ を次のように求める (図 **2** 参照).

$$IDF_{n,l} = \sum_i idf_{i,l} \quad \begin{cases} n = 1, 2, \cdots, N \\ l = 1, 2, \cdots, k-1 \end{cases}$$

$$idf_{i,l} = \left(\log_2 \frac{AllBlock}{iBlock} + 1 \right) / W_{num}$$

ただし，

AllBlock ... 連結を仮定した時の全ブロック数，つまり *(N-l)*

iBlock ... 連結を仮定した時の名詞 i を含むブロック数

W_{num} ... 文書中の名詞種類数

3. 窓を移動させながら，文書全体について $IDF_{n,l}$ を求め，それが最大値をとる n, l を得る．これに従って，文書の n ブロック目より $n+l$ ブロック目までを連結する.

4. 以上の処理を繰り返しながら，$IDF_{n,l}$ の最大値と文書のブロック数とをグラフにすると図 **3** のようになる．このグラフがピークとなる時点での段落分割の様子を最終的な出力とする.

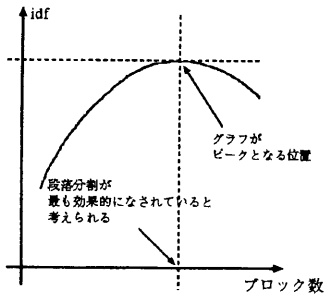

図 3: idf の変化による段落分割位置の決定

4 実験

tf.idf 連結法および idf 連結法は，共に文書内に出現する名詞に着目している．これらの名詞を大きく分けると，(1) 文書全体に渡って頻出する名詞，(2) 文書の特定の位置に集中的に出現する名詞，となると考えられるが，特に idf 連結法をそのまま文書に適用した場合，(1) のような名詞がノイズとなってしまう可能性がある．一方，tf.idf 連結法では，連結処理毎に「重要語」を抽出していることから，これを利用して，(1) のような名詞を除去するようにすれば，idf 連結法適用時での影響が少なくなるものと考えられる．以上より，ここでは両手法を次のように併用することによる段落分割の実験を行なう．

1. 対象となる文書について，まず tf.idf 連結法を適用し，結果を得る．この時，手順 1 の各繰り返し毎において，「重要語」ではない名詞の出現情報を各ブロックより除去する．

2. 上記の出力結果に対して，idf 連結法を適用し，最終結果を得る．

なお，各実験では，文書中の (a) 話題の変化を表わす語句 (接続詞など)，(b) 話題の連続を表わす語句 (前方照応詞など) がどのように結果に影響を及ぼすかについても着目する．また，文書内における名詞の連鎖の状態を用いて段落分割を行なう手法との比較実験も併せて行ない，その結果について検討する．

4.1 実験の対象データ

実験用の文書としては，雑誌に掲載されている座談会の記事など，話し言葉からの書き起こしがもとになっている文書を用いることとした．この時，文書を電子化したうえで句点にしたがって一文を一ブロックとして分割し，形態素解析を行なうことによって，各ブロックに含まれる名詞情報などを抽出し，その情報を用いている[3]．また，各文書についての人間による段落分割の結果を正解として用いた．段落分割のサンプル数は各文書 12 であり，6 以上のサンプルで段落境界と判断されている結果を正解とし，実験で得られる段落境界と比較することとする．**表 1** に，実際に用いた文書について示す[4]．

表 1: 実験に用いた文書

データ名	全文数	名詞種類	正解段落境界数
slpnp	509	1019	70
saigai	374	912	56
mt	325	564	45

4.2 tf.idf 連結法による実験

ここでは，**表 1** に示した文書に対し，次のような条件を組み合わせた実験を行なった．

1. 等分割の幅 k を 5 ブロック刻みに全ブロック数の 1/10 まで増加させ，各々の k の場合について調べる．

2. 以下の語句がブロックの先頭に出現する場合，そのブロックを段落の開始位置とする場合 (**Cue1**)，もしくはしない場合 (**Cue2**) の各々について調べる．

   - 転換の接続詞 (「ところで」など)
   - 「最初に」「次に」「それから」「最後に」など話の起承転結を示すと考えられる語句．

[3] 形態素解析には，JUMAN3.1[松本 96] を用いた．

[4] **slpnp** は，1995 年 5 月に行なわれたパネルディスカッションの録音からの書き起こしである．また，**saigai** は，**bit** 誌 1995.8,Vol.27,No.8 の記事，**mt** は，人工知能学会誌 1989.11,Vol.4,No.6 の記事である．

表 2: tf.idf 連結法による実験結果

Cue1(接続詞などの手がかり語を考慮する)

文書	連結	分割幅	再現率	適合率	E(b=1.0)	段落境界数の変化	名詞種類数の変化
slpnp	**Ca**	20	0.786	0.190	0.694	508 → 289	1019 → 85
saigai	**Ca**	5	0.714	0.192	0.697	373 → 208	912 → 189
mt	**Cc**	25	0.800	0.167	0.724	324 → 216	564 → 37

Cue2(接続詞などの手がかり語を考慮しない)

文書	連結	分割幅	再現率	適合率	E(b=1.0)	段落境界数の変化	名詞種類数の変化
slpnp	**Ca**	20	0.700	0.178	0.716	508 → 275	1019 → 101
saigai	**Ca**	15	0.804	0.174	0.714	373 → 259	912 → 114
mt	**Cc**	15	0.800	0.164	0.728	324 → 220	564 → 27

- 「一つめ」「二つめ」など事柄の列挙を示すと考えられる語句.

3. 手順 1 において連結対象とする連続ブロックの種類を次の三種類の方法について調べる.
   - 重要語集合 W に含まれる名詞が連続して出現するブロックを連結する (**Ca**).
   - W に含まれる名詞を各ブロックから削除し，その結果名詞情報を含まなくなった連続ブロックを連結する (**Cb**).
   - **Cb** による最終出力に対し，更に W に含まれる名詞が連続して出現するブロックを連結する (**Cc**).

なお，手順 1 での閾値 w_{th} は，各ブロック毎での $w_{i,j}$ の平均値とする．また，実験の評価には，再現率，適合率および Rijsbergen の E を用いる[5].

この実験の結果，すべての文書において，上記の 2. に関しては **Cue1** の場合に E が最小となる結果を出力した．また，1. については，これらと各文書で E が最小となる場合との関係についてばらつきがあった．さらに，3. についても，**slpnp** および **saigai** が **Ca** で E が最小となるのに対し，**mt** では **Cc** で E が

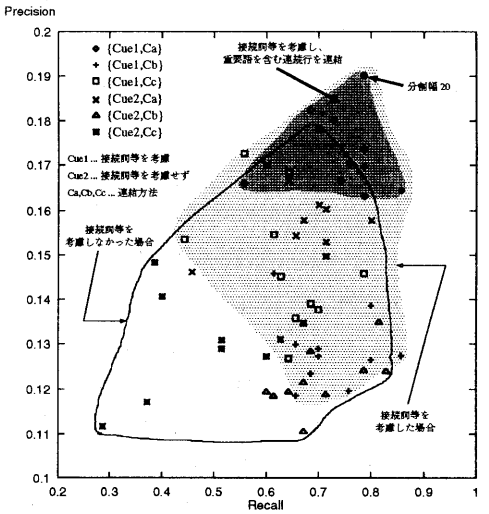

図 4: **slpnp** における最終出力結果の再現率と適合率

[5] これらは次のように定義される [FBY92].

$$\text{再現率}(R) = \frac{\text{出力結果に含まれる正解数}}{\text{全正解数}}$$

$$\text{適合率}(P) = \frac{\text{出力結果に含まれる正解数}}{\text{全出力結果数}}$$

$$E = 1 - \frac{(1+b^2)PR}{b^2P+R}$$

最小となる，というように，文書によってばらつきがあった．以上の実験について，**表 2** に，各文書の結果として最良だった (E が最小であった) 場合を示す．なお，**図 4** は，**slpnp** を対象とした 1～ 3のすべての組合せによる実験結果についての，最終出力に関する再現率 – 適合率のグラフである．

4.3 idf連結法による実験

表 3: idf 連結法の実験対象の文書

データ名	全段落境界数	名詞種類	正解段落境界数
slpnp	289	85	70
saigai	208	189	56
mt	216	37	45

4.2節の実験による出力結果を用いて，更に idf 連結法についての実験を行なう．ここでは，**表 2** の **Cue1** の結果を対象とした．これらの文書のデータを**表 3** に示す．また，実験の際の条件は次の 1. ～ 3. の各々について組み合わせたものとした．

1. 窓幅を 10 ブロック刻みに 50 ブロックまで増加させる．
2. 以下の語句がブロックの先頭に出現する場合，そのブロックを段落の開始位置とする (**Cue1**)，もしくはしない (**Cue2**) こととする．
   - 転換の接続詞 (「ところで」など)
   - 「最初に」「次に」「それから」「最後に」など話の起承転結を示すと考えられる語句．
   - 「一つめ」「二つめ」など事柄の列挙を示すと考えられる語句．
3. 以下の語句がブロックの先頭に出現する場合，そのブロックと直前のブロックとを連結するものとみなす (**An1**)，もしくはしない (**An2**) こととする．
   - 「それは」「それが」など「そ」型の前方照応詞．
   - 「これは」「これが」など「こ」型の前方照応詞．

実験の評価には，4.2節と同様に再現率，適合率および Rijsbergen の E を用いる．この実験の結果として，**表 4** に，各文書毎に (1){**Cue1,An1**} の条件 (接続詞等手がかり語を考慮する) 場合，(2){**Cue2,An2**} の条件 (手がかり語を考慮しない) 場合，それぞれの最良の結果について示す[6]．

この結果から，実験対象の文書すべてにおいて，{**Cue1,An1**}(接続詞などの手がかり語を考慮する) という条件の方が結果が良いことがわかる．また，窓幅については，文書により最適と考えられる値にはばらつきがみられた．なお，**slpnp** を対象とした {**Cue1,An1**} の場合の段落境界数と idf の値の関係を**図 5** に示す．

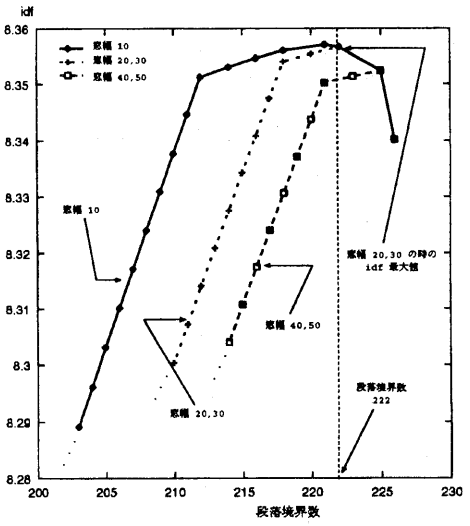

図 5: idf の値の変化の様子

さらにこの実験では，文書内における名詞の連鎖の状態を用いて段落分割を行なう方法との比較を行なった．これは，大まかにいって以下のような手順で段落分割を行なうものである [本田 94][7]．

1. 文書を一文が一ブロックになるように分割し，その文書内に出現するある名詞について，それが連続して出現するような範囲 (*chain*) を文書全体に

[6] 窓幅で 'all' とは，すべての窓幅の場合において同じ結果を得たということである．

[7] ただし，ここではシソーラス情報は用いていない．

表 4: idf 連結法による実験結果

{Cue1,An1}(接続詞などの手がかり語を考慮する)

文書	窓幅	再現率	適合率	E(b=1.0)	段落境界数の変化
slpnp	20,30	0.686	0.216	0.671	289 → 222
saigai	all	0.679	0.204	0.686	208 → 186
mt	10	0.756	0.180	0.709	216 → 213

{Cue2,An2}(接続詞などの手がかり語を考慮しない)

文書	窓幅	再現率	適合率	E(b=1.0)	段落境界数の変化
slpnp	all	0.671	0.187	0.693	289 → 276
saigai	all	0.679	0.189	0.704	208 → 201
mt	all	0.756	0.171	0.721	216 → 199

おいて探し，その先頭のブロックと末尾のブロックに 1 点を加える．これを全名詞について行なう．

2. 文書内に出現するある名詞について，それが連続して出現しないような範囲 (gap) を文書全体において探し，その先頭のブロックと末尾のブロックに，gapの長さに比例した得点を加える．これを全名詞について行なう．

3. 各ブロックについて，1.，2. の得点の総和を出し，それが閾値[8]以上であるようなブロックが段落の先頭であるとする．

この比較の結果を図 6 に再現率 – 適合率のグラフで示す．このように，どの文書に対しても，[本田 94] の名詞の連鎖の状態を利用する手法に対して，“tf.idf 連結法 +idf 連結法” による段落分割 (段落境界決定) の結果の方が，再現率，適合率共に良好であるといえる．ただし，[本田 94] の手法では，シソーラス情報が重要な役割を果たしているが，ここではそれを用いていない．これについては後述する．

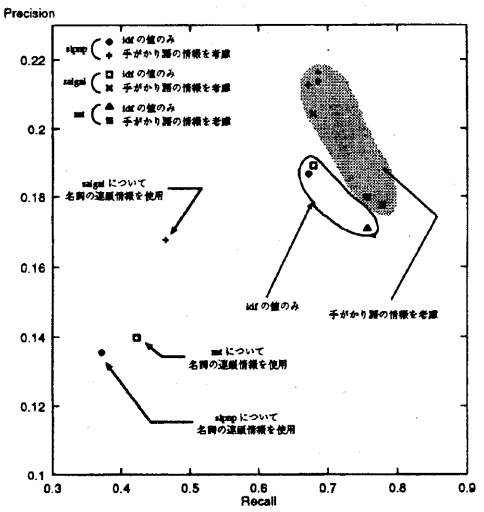

図 6: tf.idf 連結法 + idf 連結法の場合の再現率と適合率

5 おわりに

実験の結果，三種類の文書について手順 1 および手順 2 による段落境界決定のパフォーマンスは，平均して再現率 70% 程度，適合率 20% 程度であった．ここでは，その結果について考察する．

[8] ここでは，$\frac{\sum_{\text{全ブロック}} \text{得点総和}}{\text{全ブロック数}-1}$ を閾値とする．

まず手順 1 における分割幅，手順 2 における窓幅についてであるが，これらは比較的重要なパラメータであると考えられるものの，名詞毎の平均出現回数など対象文書における他の数量との関係を見い出すことが出来なかった．処理の自動化を図る上で，これは重要な要素であり，今後検討の必要がある．

また，接続詞などの手がかり語を考慮した効果が両手法において見られることから，この要素が段落分割時に与える影響は大きいと考えられる．これについては，手がかり語として用いるべき語がどのようなものかをさらに検討する必要がある．また，今回の実験では，これらの語が存在した場合に，無条件で段落境界とみなしたり直前ブロックとの連結を行なうなどしているが，何らかの重みを用いた処理を行なう必要もあると考えられる．

ところで，**表 2** からわかるように，tf.idf 連結法によって文書中の名詞種類数は大幅に絞り込まれる．この処理を行なわずに idf 連結法のみを文書に用いた場合，適合率は 10% 程度 (E=0.8 程度) となってしまうことから，今回の実験における "tf.idf 連結法→ idf 連結法" という適用順序は妥当なものであったと考えられる．

さらに，**図 6** のように，名詞の連鎖の状態を用いる手法と比較して，本研究での手法はかなり良好な結果を示す．これは本来，名詞の連鎖をとらえる場合にはシソーラス情報を利用する [本田 94] ところを，これを用いていないためだと考えられる．そこで，この種の手法の例として，たとえば [望月 96] では，本研究とは異なる性質の文書 (国語問題集の問題) を対象としているため直接比較はできないものの，再現率 55% 程度，適合率 25% 程度という結果を出している．これと比較しても，**表 4** に示した実験の最終結果はさほど劣るわけではなく，このことから，シソーラス情報の利用による本研究での手法のパフォーマンス向上が期待できる．ただし，そのためには，および本研究でのシソーラス情報の利用形態などを検討する必要があると考えられる．

参考文献

[FBY92] William B. Frakes and Ricardo Baeza-Yates, editors. *Information Retrieval Data Structures & Algorithms*. P T R Prentice-Hall, Inc., 1992.

[Hea94] Marti A. Hearst. Multi-paragraph segmentation of expository text. In *ACL '94 Proceedings*, pp. 9–16, 1994.

[HP93] Marti A. Hearst and Christian Plaunt. Subtopic structuring for full-length document access. In *SIGIR '93*, pp. 59–68, 1993.

[Koz93] Hideki Kozima. Text segmentation based on similarity between words. In *Proceedings of ACL-93*, pp. 286–288, 1993.

[PL93] Rebecca J. Passonneau and Diane J. Litman. Intention-based segmentation: Human reliability and correlation with linguistic use. In *Proceedings of the 31rd ACL*, pp. 148–155, 1993.

[本田 94] 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 94-NL-102, pp. 25–32. 情報処理学会, 1994.

[松本 96] 松本裕治, 黒橋禎夫, 山地治, 妙木裕, 長尾真. 日本語形態素解析解析システム JUMAN version 3.1 使用説明書. 京都大学長尾研究室, 奈良先端科学技術大学院大学松本研究室, Nov 1996.

[望月 95] 望月源, 本田岳夫, 奥村学. 複数の知識の組み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 95-NL-109, pp. 47–54. 情報処理学会, 1995.

[望月 96] 望月源, 本田岳夫, 奥村学. 重回帰分析とクラスタ分析を用いたテキストセグメンテーション. 言語処理学会 第 2 回年次大会発表論文集, pp. 325–328. 言語処理学会, 1996.

[山本 92] 山本和英, 増山繁, 内藤昭三. 手がかり語および語の類縁性を併用した段落分け. 情報処理学会研究報告 92-NL-92, pp. 41–48. 情報処理学会, 1992.