文書内の名詞の出現頻度を用いた段落分割

西澤 信一郎　　森 辰則　　中川 裕志
横浜国立大学工学部

1 はじめに

計算機ネットワークの発達にともない，電子化された大量のテキストが入手可能となっている現在，それらを利用者が効率よく利用するため，段落分割，キーワード抽出，自動抄録作成などのテキスト処理技術が求められている．この時対象となるテキストは，マニュアル，新聞記事などのように，章・節などの構造をあらかじめ持つものばかりでなく，会議の議事録からの書き起こしなど，そのような構造を表層に持たないものも考えられる．このような，いわゆる話し言葉からの書き起こしテキストを対象とした文書処理を行なう際には，章や節，段落などに相当するテキストの構造を把握することが最重要課題の一つとなる．

このような背景を基に，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．同様の研究としては，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として，[1, 2] など，また，接続詞や副詞などの「手がかり語」の情報を併せて利用するものに [3, 4] などがある．これらと比較して，本論文で述べる手法では，シソーラスを用いず，文書中の同一名詞の出現頻度に着目する点が異なる．さらに，同一語の出現頻度を利用する研究として [5] がある．これは隣接ブロック間の類似度を cosine measure で計算し，この値を利用するものであるが，本論文ではこのような類似度計算は行なわない．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 [6] の研究があるが，本研究では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り，よって，ポーズなど語彙情報以外の情報は用いない．

2 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頻度を利用する．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2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1文毎に分割されているものとし，この1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 “1文＝1ブロック” であり，処理の進行につれて1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．

2.1 tf.idf による重要語のランキングを利用した手法

2.1.1 アルゴリズム

以下で述べる **tf.idf 連結法**は，対象とする文書中の全名詞から重要語を抽出し，それらの出現状況に基づいて段落境界を決定する手法である．重要語は，文書を同サイズの領域に等分割し [7]，*tf.idf* による重みを利用して決定する．

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 m ブロック毎の領域に分割する．なお，m はあらかじめ与える正整数である．
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 [8]．ここで，$w_{i,j}$ は「先頭から j 番目の領域における名詞 i の重み」である．

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

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

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

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．なおここでは，$w_{i,j}$ の平均値を w_{th} とする．
4. 文書全体で，W に含まれる名詞の出現状況に従い，ある条件に該当する隣接ブロックを連結して一つの新しいブロックとする．なお，連結の条件については，いくつかの実験を行なう (後述)．
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．

2.1.2 実験

ここでは，表 1 に示す文書を対象とした実験を行なう[1]．この時，すべての文書は初期状態として，句点に従って1文を1ブロックとして分割されているものとする．また，大学生12人を対象とした人手での段落分割の

表 1: 実験に用いた文書

文書名	全文数	名詞種類	正解段落境界数	初期適合率
slpnp	509	1019	70	0.138
saigai	374	912	56	0.150
mt	325	564	45	0.139

実験を各文書についてそれぞれ行ない，過半数が段落の境界であると判定した結果を正解として，計算機による実験結果との比較を行なう．各文書の初期状態を段落分割の結果とし，この正解と比較した場合，再現率は100%となり，適合率は表 1 の初期適合率の数値となる．

なお，手順1の実験の際の条件として，以下の1.～3.の項目を組み合わせることとする．

1. 等分割の幅 m を $5, 10, 15, \cdots, 50$ と変化させ，各々の場合について調べる．
2. 表 2 に挙げるような「手がかり語」の影響を考慮する場合 (**Cue1**)，もしくはしない場合 (**Cue2**) の各々について調べる．
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二種類の方法について調べる．
   - 重要語集合 W に含まれるある名詞が一種類以上連続して出現する隣接ブロックを連結する (**Ca**).
   - 重要語集合 W に含まれる名詞が全く存在しない隣接ブロックを連結する (**Cb**).

各実験の結果は再現率，適合率および Rijsbergen's E により評価し[2]，E の値が最小の場合を最良の結果とした．この結果，隣接ブロックの連結法については **Ca** の方法を用いた時，さらに加えて，手がかり語を考慮した

[1] 文書の出典は以下のとおりであり，これらを JUMAN3.1(京都大学長尾研究室などで開発) で形態素解析した結果を利用している．

- [slpnp]：情報処理学会 SIGNL パネルディスカッションからの書き起こし．1995.
- [saigai]：bit 誌 *Vol.27,No.8* pp.29–43. 1995.
- [mt]：人工知能学会誌 *Vol.4,No.6* pp.671–680. 1989.

[2] これらは次のように定義され [8]，ここでは，E の式中において $b = 1$ とした数値を用いた．

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

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

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

表 2: 「手がかり語」について

段落の開始位置を示す	転換の接続詞 (「ところで」など)
	「最初に」「次に」「それから」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
段落の継続を示す	「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

表 3: tf.idf 連結法による実験結果

Cue1(手がかり語を考慮する)

文書名	m	再現率	適合率	E	段落境界数
slpnp	50	0.857	0.196	0.681	508 → 306
saigai	50	0.893	0.193	0.683	373 → 259
mt	35	0.844	0.184	0.698	324 → 207

Cue2(手がかり語を考慮しない)

文書名	m	再現率	適合率	E	段落境界数
slpnp	15	0.957	0.167	0.715	508 → 401
saigai	25	0.928	0.179	0.699	373 → 290
mt	10	0.911	0.177	0.703	324 → 231

場合に最良の結果が得られた．**Ca** の方法による結果を表 3 に示す．この結果の適合率を表 1 の初期適合率と比較すると，**Cue1** の場合，平均して14%から19%へと改善される結果となった．また，領域幅 m と出力結果との間には，この実験でははっきりとした関係は認められなかったものの，全体的に，m の値を大きめ (30以上) に取る方が良いのではないかと考えられる．なお，図 1 は，文書 "saigai" 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．

2.2 idf の変化に基づく段落分割

2.2.1 アルゴリズム

idf の定義 (手順 1 参照) より，文書をいくつかの決まった数の段落に分割した時，ある名詞が少ない段落にまとまっているほど idf の値が大きい．この性質を利用した段落分割の手法として，**idf 連結法**の手順を以下に示す．

手順 2 (idf 連結法)

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 ブロック目に着目する．
2. n～$n+1$ ブロック，n～$n+2$ ブロック，…，n～$n+l$ ブロックを連結したと仮定し，それぞれの場合

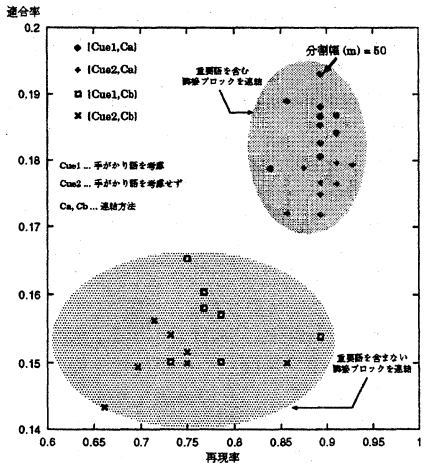

図 1: "saigai" における最終出力結果の再現率と適合率

について，文書中に出現する全名詞に関する $IDF_{n,l}$ を次のように求める．なお，k はあらかじめ与える正整数である．

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

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

ただし，
$iBlock$ ... 名詞 i を含むブロック数
W_{num} ... 文書中の名詞種類数

3. 文書全体について得られる $IDF_{n,l}$ が最大値をとる n_{max}, l_{max} を得る．これに従って，文書の $n_{max} \sim n_{max} + l_{max}$ ブロックを連結する．
4. 以上の処理を繰り返すことにより，段落分割を行なう．なお，今回は，各文書の正解段落境界数まで段落分割を行ない，その際の再現率および適合率，Rijsbergen's E から最適な段落分割位置を判断するものとする．

2.2.2 実験

ここでは，2.1.2節での実験結果として得られる，表 4 に示す各文書を用いて idf 連結法の実験を行なう．これは，idf 連結法において予想される，文書全体にわたって出現するような名詞の idf がノイズとなるような影響が tf.idf 連結法による重要語抽出処理によって除去できると考えられるからである．また，2.1.2節での実験と同様に，実験の際の条件は，以下の 1. ～ 2. の項目をそれぞれ組み合わせたものとする．

表 4: idf 連結法の実験対象の文書

文書名	全段落境界数	名詞種類	正解段落境界数
slpnp	306	321	70
saigai	259	286	56
mt	207	199	45

1. 窓幅 k を $10, 20, \cdots, 50$ と変化させ，各々の場合について調べる．
2. 表 2 に挙げるような「手がかり語」の影響を考慮する場合 (**Cue1**)，もしくはしない場合 (**Cue2**) の各々について調べる．

なお，手順 2 で述べたとおり，段落分割は各文書の正解段落境界数に到達するまで行ない，全出力結果のうち Rijsbergen's E の値が最良の段落分割の様子を最終結果とする．この結果，各文書についての最良の結果として表 5 に示す数値を得た[3]．

表 5: idf 連結法による実験結果

Cue1 (手がかり語を考慮する)

文書名	k	再現率	適合率	E	段落境界数
slpnp	all	0.829	0.204	0.673	306 → 284
saigai	all	0.518	0.367	0.570	259 → 78
mt	all	0.644	0.212	0.681	207 → 136

Cue2 (手がかり語を考慮しない)

文書名	k	再現率	適合率	E	段落境界数
slpnp	all	0.857	0.2	0.676	306 → 299
saigai	all	0.339	0.339	0.660	259 → 55
mt	all	0.756	0.198	0.687	207 → 171

この結果から，2.1.2節での実験と同様，手がかり語の影響を併せて考慮すべきであると考えられる．また，適合率を表 1 と比較すると，以上の段落分割処理により，最終的に平均して 14% から 25% へと改善される結果となった．また，今回の実験で用いた k の範囲においては，どの値を用いても結果に差異が見られなかった．なお，文書 "saigai" を対象とした実験結果において，出力される段落境界数と再現率および適合率の関係を図 2 に示す．

さらに本研究では，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 (1) 文書中の全名詞について，それらが連続して出現する範囲 (*chain*) の先頭と末尾の位置に得点を与え，(2) 同じく全名詞について，それらが出現しない範囲 (*gap*) の先頭と末尾の位置に得点を与

[3] 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．

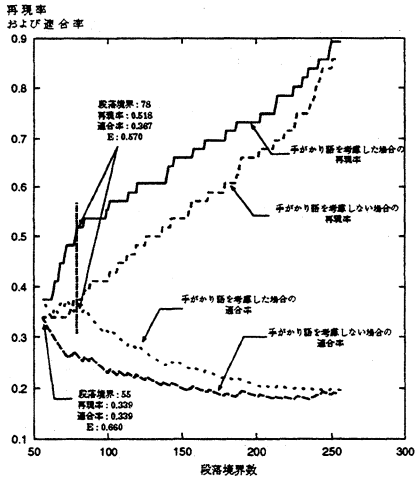

図 2: 段落境界数と再現率，適合率との関係

え，(3)*chain*と *gap*に関する得点の総和を元に段落の境界位置を決定する，という手順で段落分割を行なうものである [1][4]．この比較の結果を表 **6** に示す．このよう

表 6: 他方式との比較実験

文書名	手法	再現率	適合率	E
slpnp	名詞の連鎖	0.371	0.135	0.901
	本研究での手法	0.829	0.204	0.673
saigai	名詞の連鎖	0.464	0.168	0.877
	本研究での手法	0.518	0.367	0.570
mt	名詞の連鎖	0.422	0.140	0.895
	本研究での手法	0.644	0.212	0.681

に，どの文書に対しても，前述したように名詞の連鎖の状態を単純に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．

3 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 25% 程度であった．ここでは，その結果について考察する．

まず手順 1 における領域幅 m，手順 2 における窓幅 k 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，会議などでは「司会者」の発話が談話の流れ (意味段落) に大きな影響を与えると考えられるため，この情報を一種の「手がかり語」として利用できるのではないかと考えられる．

さらに，表 **6** のように，名詞の単純な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 [1] ところを，これを用いていないためだと考えられる．この種の手法の例として，たとえば [4] 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 **5** 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．

[4] ただし，ここではシソーラス情報は用いていない．

参考文献

[1] 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.

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

[3] 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.

[4] 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.

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

[6] 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.

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

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