文書内における名詞の出現頻度を用いたテキストセグメンテーション

6B－1

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

1 はじめに

本研究は，対談の書き起こしなどタグのない話し言葉コーパスを対象とした段落分割の手法について述べるものである．同様の目的の研究としては，手がかり語，ポーズなどの情報を用いる [1] の研究があるが，これとは異なりここでは文書中の名詞の出現頻度を利用する手法について述べる．この手法は [2, 3] などと異なり，シソーラス情報を用いない．また，同一名詞の出現頻度に着目する点では [4] と同様であるが，本研究では cosine measure による隣接ブロックの類似度を用いるのではなく，着目する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ものである．

2 tf.idf連結法による段落分割

2.1 アルゴリズム

ここで述べる手法 (**tf.idf 連結法**) は，文書内の名詞の *tf.idf* を用いて文書中の名詞の重要度を計算し，この値に基づいて連結されるべき隣接ブロックを決定する，というものである[1]．以下にその手順を述べる．

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 [5]．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った (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．

2.2 実験

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
slpnp	509	1019	70
saigai	374	912	56
mt	325	564	45

ここでは，表 1 に示す文書を対象とした実験を行なった[2]．なお，大学生 12 人を対象とした段落分割の実験を行ない，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用いた．まず，手順 1 において，(a): 重要語集合 W に含まれるある名詞が一つ以上連続して現れる隣接ブロックを連結，(b): 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，(c):(b) → (a) の順に適用，それぞれについて実験を行なった結果，(a) もしくは (c) の方法を適用すべきであるという結果を得た．各文書において，最も良い結果を表 2 に示す．なお，評価には再現率，適合率および Rijsbergen's E を用いた[3]．この実験の結果，tf.idf 連結法と併せて，転換の接続詞 (「ところで」など) のように段落の開始位置を表すような，いわゆる「手がかり語」の影響を考慮すべきであるという結果を得た．

3 idf の変化に基づく段落分割

3.1 アルゴリズム

本研究で用いている 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

*Bottom-up Discourse Segmentation based on Word Frequency
by Shin'ichiro Nishizawa and Hiroshi Nakagawa,
Yokohama National University, 79-5 Tokiwadai, Hodogayaku, Yokohama 240, Japan.

[1]対象とする文書は初期状態として 1 文毎に分割されているものとする．ここでは，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，これによって段落が形成されることになる．

[2]文書の出典は以下のとおり．
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 *Vol.27,No.8* pp.29–43. 1995.
mt 人工知能学会誌 *Vol.4,No.6* pp.671–680. 1989.

[3]再現率を R，適合率を P とした時，$E = 1 - \frac{2PR}{P+R}$ である [5].

表 2: tf.idf 連結法による実験結果

文書	連結	m	再現率	適合率	E	段落境界数
slpnp	(a)	20	0.786	0.190	0.694	508 → 289
saigai	(a)	5	0.714	0.192	0.697	373 → 208
mt	(c)	25	0.800	0.167	0.724	324 → 216

い．この性質を利用した段落分割の手法 (**idf 連結法**) 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. $n \sim n+1$ ブロック，$n \sim n+2$ ブロック，…，$n \sim n+l$ ブロックを連結したと仮定し，それぞれの場合について，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と，文書のブロック数に対して $IDF_{n,l}$ の最大値がピークとなるような場合がある．この時の段落分割を最終的な段落分割結果とする．

3.2 実験

表 3: idf 連結法による実験結果

文書	k	再現率	適合率	E	段落境界数
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

ここでは，表 2 に示した実験結果として得られた文書に対し，さらに idf 連結法を適用する実験を行なった．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．この実験の結果，2.2節での実験と同様に，手がかり語の影響を併せて考慮すべきであるという結果を得た．各文書において，最も良い結果を表 3 に示す[4]．なお，slpnp を対象とした実験での段落境界数と idf の値の関係を図 1 に示す．

[4] k の値が 'all' とは，実験で用いたすべての k(10,20,...,50) の場合において同じ結果を得たということである．

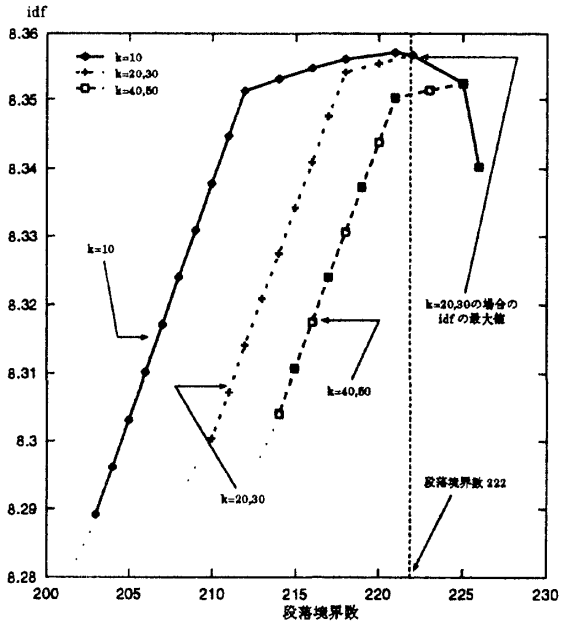

図 1: 段落境界数と idf の値の対応

4 おわりに

実験の結果から，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．なお，同じ文書を対象として，語彙連鎖を用いる手法 [2] による実験を行なった結果，平均して再現率 40% 程度，適合率 15% 程度であった．この手法で重要な要素であるシソーラス情報をここでは用いていないことなどのため直接比較は難しい[5]が，この比較から，本研究での手法が段落分割に際して有効な手段であり，シソーラス情報の利用によりさらなるパフォーマンス向上が期待できると考えられる．

参考文献

[1] 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.

[2] 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.

[3] 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.

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

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

[5] 例えば [3] ではシソーラス情報を用いた同様の手法で再現率 55% 程度，適合率 25% 程度という結果を出している．