テキスト分割のための統計的モデル

内山 将夫　　井佐原 均
通信総合研究所

1　はじめに

複数トピックからなる文章を切り分けて，それぞれの切り分けた部分が一つのトピックになるようにすることを，テキスト分割と呼ぶ．

テキスト分割は，情報検索や要約などにおいて重要である．まず，情報検索においては，文書全体ではなく，ユーザの検索要求を満す部分(トピック)だけを検索した方が効果的である(Salton, Singhal, Buckley, and Mitra 1996)．また，要約においては，長い文書をトピックに分ければ，それぞれのトピックごとに要約を作成することにより，文書全体の要約を作成できるし，重要なトピックだけを選んで要約を作成することもできる (Nakao 2000)．

情報検索や要約等が対象とする文書は，分野を限定しない文書であるので，それらを分割する手法も分野を限定しないものである必要がある．本稿で述べる手法は，テキスト内の単語分布のみを利用してテキストを分割する．そのため，訓練データが存在する分野に限られることなく，どんな分野のテキストでも分割できる．

提案手法は，テキストの分割確率が最大となるような分割を選択するというものである．このようなアプローチは，分野を限定しないテキスト分割としては，新しいアプローチである．なお，従来の研究で，分野を限定しないテキスト分割の研究では，主に，語彙的な結束性を利用してテキストを分割している．その例としては，意味ネットワーク上での活性伝播に基づく結束性を利用するもの (Kozima 1993)や，単語分布の類似度(コサイン)を結束性としたもの (Hearst 1994)や，単語の繰り返し状況に基づいて結束性を計るもの (Reynar 1994)や，文間の類似度としてコサインを直接使うのではなくコサインの順位を結束性の指標とするもの (Choi 2000) などがある．

2　統計的モデル

本節では，テキストの分割結果の確率を定義し，それを用いて最大確率であるような分割を定義する．そして，次節で，最大確率であるような分割を選ぶアルゴリズムを示す．

まず，n 個の延べ単語からなるテキスト $W = w_1 w_2 \ldots w_n$ が与えられたとき，m 個の区間からなる分割 $S = S_1 S_2 \ldots S_m$ の確率は，

$$\Pr(S|W) = \frac{\Pr(W|S)\Pr(S)}{\Pr(W)} \qquad (1)$$

である．$\Pr(W)$ は，W が与えられたときには定数なので，最大確率の分割 (最適分割)$\hat{S}$ は，

$$\hat{S} = \arg\max_S \Pr(W|S)\Pr(S). \qquad (2)$$

$\Pr(W|S)$ の定義　区間 S_i に n_i 個の延べ単語があるとして，S_i 中の j 番目の単語を w_j^i とし，$W_i = w_1^i w_2^i \ldots w_{n_i}^i$ とする．つまり，S_i と W_i とを一対一に対応させる．このようにすると，$n = \sum_{i=1}^{m} n_i$，$W = W_1 W_2 \ldots W_m$ である．

このとき，ある区間に属する単語列は，その他の区間には独立に生起し，更に，同一区間に属する単語も，区間が与えられているという条件下では確率的に独立であるとすると，

$$\begin{aligned}\Pr(W|S) &= \Pr(W_1 W_2 \ldots W_m|S) \\ &= \prod_{i=1}^{m} Pr(W_i|S) \\ &= \prod_{i=1}^{m} Pr(W_i|S_i) \\ &= \prod_{i=1}^{m}\prod_{j=1}^{n_i} \Pr(w_j^i|S_i). \end{aligned} \qquad (3)$$

次に，W 中における異なり単語の数を k，W_i において w_j^i と同じ単語の数を $f_i(w_j^i)$ とし，$\Pr(w_j^i|S_i)$ を定義する．

$$\Pr(w_j^i|S_i) \equiv \frac{f_i(w_j^i)+1}{n_i+k}. \qquad (4)$$

なお，$f_i(w_j^i)$は，厳密には，次式で定義される．

$$f_i(w_j^i) \equiv g(w_j^i|w_1^i w_2^i \ldots w_{n_i}^i) \tag{5}$$

$$g(w_j^i|w_1^i w_2^i \ldots w_{n_i}^i) \equiv \sum_{k=1}^{n_i} \delta(w_k^i, w_j^i). \tag{6}$$

ただし，δについては，単語aと単語bとが同じとき$\delta(a,b)=1$，そうでないとき，$\delta(a,b)=0$である．

$\Pr(S)$の定義　$\Pr(S)$の定義に関しては，任意性が大きい．我々は，予備実験の結果，および，MDL(Minimum Description Length)原理[1](山西 韓 1992)などを考慮した結果として，以下の式を採用した．

$$\Pr(S) \equiv \left(\sqrt{\frac{1}{n}}\right)^m. \tag{7}$$

3　最適分割選択アルゴリズム

本章では，分割Sのコスト$C(S)$を，

$$C(S) \equiv -\log \Pr(W|S) \Pr(S) \tag{8}$$

と定義し，このコストが最小となる分割$\hat{S} = \arg\min_S C(S)$を選択することにより，最大確率である分割$\hat{S}$を選択する．ここで，$C(S)$は以下のように展開できる．

$$\begin{aligned} C(S) &= -\log \Pr(W|S) \Pr(S) \\ &= -\sum_{i=1}^{m} \sum_{j=1}^{n_i} \log \Pr(w_j^i|S_i) + \frac{m}{2} \log n \\ &= \sum_{i=1}^{m} c(w_1^i w_2^i \ldots w_{n_i}^i|n,k). \end{aligned} \tag{9}$$

ただし，

$$\begin{aligned} & c(w_1^i w_2^i \ldots w_{n_i}^i|n,k) \\ &\equiv \sum_{j=1}^{n_i} \log \frac{n_i + k}{f_i(w_j^i) + 1} + \frac{1}{2} \log n \\ &= \sum_{j=1}^{\#(w_1^i w_2^i \ldots w_{n_i}^i)} \log \frac{\#(w_1^i w_2^i \ldots w_{n_i}^i) + k}{g(w_j^i|w_1^i w_2^i \ldots w_{n_i}^i) + 1} \\ &\quad + \frac{1}{2} \log n, \end{aligned} \tag{10}$$

ここで，$\#(\cdots)$は，その引数である単語列の長さ(延べ単語数)である．

3.1　アルゴリズム

まず，用語を定義する．延べ語数nのテキスト$W = w_1 w_2 \ldots w_n$において，i番目の分割候補点g_iとは，単語w_iとw_{i+1}の間を言う．ただし，g_0はw_1の直前，g_nはw_nの直後である．このとき，分割候補点は$g_0, g_1, \ldots, g_n$の$n+1$個ある．また，分割候補点の集合をノード集合とするグラフを考えるとき，$e_{ij}(0 \le i < j \le n)$は$g_i$から$g_j$への有向辺である．このように定義されたグラフの例を，図1に示す．

このとき，e_{ij}は，単語列$w_{i+1} w_{i+2} \ldots w_j$をカバーするという．$e_{ij}$は，テキストの，ある一区間$w_{i+1} w_{i+2} \ldots w_j$を表現している．そのため，$e_{ij}$のコスト$c_{ij}$を，(10)式を利用することにより，次式で定義する．

$$c_{ij} = c(w_{i+1} w_{i+2} \ldots w_j|n,k) \tag{11}$$

ただし，kは，W中の異なり単語数である．

以上の準備の下で，最小コストを与える最適分割を求める手順は以下の通りである．

Step 1. 有向辺e_{ij}のコストc_{ij}を(11)式により計算する．$(0 \le i < j \le n)$.

Step 2. g_0からg_nまでの最小コストパスを求める．

ここで，Step 2を効率的に解くアルゴリズムは良く知られている[2]．なお，Step 2は，全て

[1]MDL原理は，確率モデルを選択する際に，その記述長が最小のモデルを選ぶという原理である．このとき，モデルの記述長の一部として，パラメタの記述長を含むが，この記述長は，自由パラメタの数をr，観測事象の数をnとしたとき，$\frac{r}{2}\log n$とすれば良いことが分かっている．我々の場合，テキストが与えられているとすると，分割Sには，区間数mと，各区間に属する単語の数$n_1, n_2, \ldots, n_m$の，全部で$m+1$個のパラメタがある．ただし，$n_m = n - \sum_{i=1}^{m-1} n_i$であり，かつ，テキスト中の延べ語数$n$は与えられているので，自由パラメタの数は，$m$個である．そのため，パラメタの記述長は，$L = \frac{m}{2}\log n$である．ここで，記述長と確率との対応を考え，(7)式において，$\Pr(S) = \exp(-L)$と定義した．なお，logは自然対数である．

[2]本稿で述べた手法を実装したプログラムが第1著者より入手できる(`http://www2.crl.go.jp/khn/nlp/members/mutiyama/index.html`)．なお，DPを用いてテキストを分割する研究としては，(Ponte and Croft 1997; Heinonen 1998)がある．

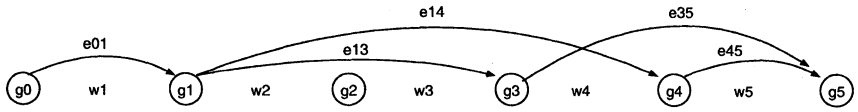

図 1: 分割候補点をノードとするグラフ

の可能なパスの中での最小コストパスを求めるものであるが，その代りに，パスの長さを指定した最小コストパスを求めることもできる．その場合は，区間数を指定した場合の最適分割に対応している．

このようにして求めた最小コストパスについて，その各辺にカバーされる単語列を，それぞれ一つの区間とすると，それは最適分割である．たとえば，図 1で，$e_{01}e_{13}e_{35}$ が最小コストパスであるとすると，最適分割は，$[w_1][w_2w_3][w_4w_5]$ である．

なお，実際にテキストを分割するときには，全ての分割候補点を考慮するのではなく，たとえば，文と文の間でのみテキストを分割したい場合がある．その場合には，分割位置として許される分割候補点の間にのみ有向辺を張るようにすれば良い．そして，そのグラフ上での最小コストパスを探索すれば良い．

4 実験

実験には，(Choi 2000) により，各種テキスト分割手法の比較に用いられたデータ[3]を用いた．Choi は，彼の提案手法 $C99$ では，従来の手法と比較し，誤り確率 (Beeferman, Berger, and Lafferty 1999) が半減されたと述べている．

この実験データは，700 個のテキストからなり，個々のテキストは，10 個のテキスト断片を連結したものである．そして，それぞれのテキスト断片は，Brown Corpus からランダムサンプリングされたテキストの最初の s 行である．表 1には，実験データの諸元を示す．

各テキストは，Choi のパッケージにあるライブラリを利用した stemmer により正規化され，提案手法により分割された．ただし，分割可能な位置は，(Choi 2000) と同様に，文間のみである．誤り確率は，Choi のパッケージにある評価プログラムにより計算された．

表 1: 実験データの諸元 (Choi 2000)

s の範囲	3 – 11	3 – 5	6 – 8	9 – 11
テキスト数	400	100	100	100

その評価結果を表 2と表 3に示す．これらの表において，$U00$ は，提案手法において，大域的な最小コスト分割を求めたときの評価結果であり，$U00_{(b)}$ は，提案手法において，区間数を 10 に指定したときの評価結果である．また，$C99$ は，Choi のアルゴリズムによる最適分割の評価結果であり，$C99_{(b)}$ は，Choi のアルゴリズムにおいて区間数を 10 に指定した場合の評価結果である[4]．また，二つの表において，$*$ は，比較対象であるアルゴリズムの誤り確率が t 検定により，有意水準 5%で有意差があることを示し，$**$ は，有意水準 1%で有意差があることを示す．なお,「3 – 11」などの列の数字は，それに該当するテキストにおける誤り確率の平均であり,「全体」は，全部のテキストについての誤り確率の平均である．

これらの表から，提案手法が，$C99$ あるいは $C99_{(b)}$ と，同等あるいは，より精度良くテキストを分割できると言える．そして，$C99$ あるいは $C99_{(b)}$ は，分野非依存のテキスト分割

[3] http://www.cs.man.ac.uk/~choif/software/C99-1.2-release.tgz

[4] $C99_{(b)}$ の行にある数値は，(Choi 2000) の Table 6 のもとの若干異なる．その理由は，元々の数値は 500 のサンプルテキストに基づいたものであるのに対して，この表のものは，700 のサンプルに基づいて我々が再実験した結果だからである (Choi, personal communication).

表 2: 分割数が無指定の場合の分割精度

	3 − 11	3 − 5	6 − 8	9 − 11	全体
$U00$	12%*	9%**	10%	11%	11%**
$C99$	13%	18%	10%	10%	13%

表 3: 分割数が指定された場合の分割精度

	3 − 11	3 − 5	6 − 8	9 − 11	全体
$U00_{(b)}$	10%**	9%	7%**	5%**	9%**
$C99_{(b)}$	12%	11%	10%	9%	11%

手法のなかでは，その他の従来手法よりも精度良くテキストを分割できるので，我々の提案手法が，従来手法よりも精度良くテキストを分割できることが言える．

5 提案手法の特徴

提案手法のテキスト分割における特徴としては，長い文章でも短かい文章でも，分割数が，大幅には変動しないというものがある．この理由は，(9)式の，$\frac{m}{2}\log n$における$\log n$が，長い文章ほど大きくなるので，長い文章においては，短かい文章よりも分割が抑制されやいからである．この性質は，我々がテキスト分割をする目的が要約のため，という観点からは適した性質である．なぜなら，要約では，文章の長さに関わらず，それを適当に少ないトピックにまとめる必要があるので，分割の結果得られる区間数は，文章の長さに，それほど影響されない方が望ましいからである．しかし，応用によっては，任意に指定した粒度の分割が望ましい場合もあると考えられる．そのために，再帰的な分割が適当であることがこれまでの実験から判っているが，より有効な分割方法を考えることは今後の課題としたい．

提案手法は，テキスト分割のために，テキストの各区間における単語の確率 (密度) を求めている．このような密度は，重要単語の抽出 (Bookstein, Klein, and Raita 1995) や，重要説明箇所の特定 (黒橋, 白木, 長尾 1997) に有用であることが知られている．提案手法を，このようなアプリケーションに対して適用することも興味深い．

6 おわりに

我々は，分割確率最大化という観点からテキストを分割する手法を提案した．提案手法は，従来手法と比べて，同等以上の精度でテキストを分割することができた．このことは提案手法がテキストの分割に有用であることを示している．我々は，今後，実際の応用でのテキスト分割の有効性を調べたいと考えている．

参考文献

Beeferman, D., Berger, A., and Lafferty, J. (1999). "Statistical Models for Text Segmentation." *Machine Learning*, **34** (1-3), 177–210.

Bookstein, A., Klein, S. T., and Raita, T. (1995). "Detecting Content-bearing Words by Serial Clustering – Extended Abstract." In *Proc. of SIGIR '95*, pp. 319–327.

Choi, F. Y. Y. (2000). "Advances in domain independent linear text segmentation." In *Proc. of NAACL-2000*.

Hearst, M. A. (1994). "Multi-Paragraph Segmentation of Expository Text." In *Proc. of ACL'94*.

Heinonen, O. (1998). "Optimal Multi-Paragraph Text Segmentation by Dynamic Programming." In *Proc. of COLING-ACL'98*.

Kozima, H. (1993). "Text Segmentation Based on Similarity between Words." In *Proc. of ACL'93*.

黒橋禎夫, 白木伸征, 長尾眞 (1997). "出現密度分布を用いた語の重要説明箇所の特定." 情報処理学会誌, **38** (4), 845–854.

Nakao, Y. (2000). "An Algorithm for One-page Summarization of a Long Text Based on Thematic Hierarchy Detection." In *Proc. of ACL'2000*, pp. 302–309.

Ponte, J. M. and Croft, W. B. (1997). "Text Segmentation by Topic." In *Proc. of the First European Conference on Research and Advanced Technology for Digital Libraries*, pp. 120–129.

Reynar, J. C. (1994). "An Automatic Method of Finding Topic Boundaries." In *Proc. of ACL-94*.

Salton, G., Singhal, A., Buckley, C., and Mitra, M. (1996). "Automatic Text Decomposition Using Text Segments and Text Themes." In *Proc. of Hypertext'96*.

山西健司 韓太舜 (1992). "MDL 入門：情報理論の立場から." 人工知能学会誌, **7** (3).