単語分布の偏りに基づく文書分割

吉川拓哉
東京大学学際情報学府*

田中久美子
東京大学情報理工学系研究科†

1 はじめに

Web 上の検索などにおいて、まとまった意味のある領域を抽出する技術への要求は大きい。その意味では従来、この問題に対して Hearst が提案した TextTiling に代表される手法が数多く提案されている [2, 4]。しかし、これらの従来手法は

- 局所的な処理により文脈の断絶点を発見することを基本とし、
- さらに精度向上のために単語辞書や品詞など言語的知識を用いている

といった特徴がある。しかし、これでは web などの現代用例を多く含み、かつ多言語で記述された雑多な文書の区分けは難しい。なぜなら、品詞解析ソフトウエアがさまざまな言語で用意できるとは限らない上、言語的知識を多様な文書に対して用意できるとも限らないためである。

そこで本稿では、与えられた文書から得られる統計だけをもとに文書分割を行う手法を提案する。具体的には、大域的な単語の偏りを捉えることでその断絶点を定式化することによる。

2 関連研究

テキスト分割問題の解法としては Hearst[2] の TextTiling が代表的である。この手法は各単語の出現頻度を成分とするベクトルを利用するベクトル空間モデルを基礎としている。具体的には、テキスト上の各位置においてその左右に一定幅の窓を用意し、その左右の窓内における各単語の出現頻度を求め、それによって決まるベクトルの類似度を二つのベクトルの余弦として計算する。こうして各位置に対して両側の窓内の単語分布の類似度が計算された後に、その両脇に比べ相対的に鋭く類似度が落ち込んでいる点を文脈の変化する位置として判断し、テキストを分割する。

TextTiling には、これを基礎として単語に Salton[3] らによる tf.idf 指標を利用した重み付けをするなど様々な改良版が存在する。最近では Dias、Alves[4] らが、単語密集度をスコアとして定量化したものも考慮に入れた重み付けを提案している。

しかし TextTiling の固定長ウィンドウによる各位置での変化の考察では、ウィンドウサイズに依存した局所的な特徴しか解析できない。そのため目的に応じて大域的な特徴を掴むべく拡張することは難しい。段落のサイズが不明な場合、ウィンドウサイズの大きさが本当に適切であるのかどうかといった判断も難しい問題である。

3 手法

以上のような観点から、本手法では分割する対象テキストに含まれる情報から得られる統計のみを用い、トピックを形成する単語が最適に分けられるような指標を大域的に計算する。手法の流れとしては、

1. トピック形成に寄与する単語を抽出し、
2. 各位置でテキストが分割されるかどうかを判断するための指標を計算し、
3. それに基づいて分割する点を選ぶ、

となる。

3.1 分割指標計算

本手法の基本的な考え方は、含まれるトピックに依存して分布に偏りのある重要語が分割後もできる限りその偏りを保つようにテキストを分割するというところにある。

* 東京大学大学院学際情報学府学際情報学専攻
Yoshikawa.Takuya@iii.u-tokyo.ac.jp
† 東京大学大学院情報理工学系研究科創造情報学専攻
kumiko@i.u-tokyo.ac.jp

そこでまず一連のテキストをある点 i で二つに分割したときを考える。この分割に従って各単語 w の各出現は分割点の左右どちらに属するかが決まるので、左側に出現する回数を n_l、右側に出現する回数を n_r とする。もしこの分割がこの単語の偏りをできる限り保っているのであれば、この二つの数の差は大きくなっているはずである。そこで値 $(n_l - n_r)^2$ をこの単語 w の位置 i での分割に対応した偏り指標として $U_w(i)$ で表すことにする。この位置での分割の指標は、この値に重要度による重み α_w(3.3 節で後述) を掛けたものをすべての単語について足し合わせたものとして定める。

$$U(i) = \sum_{w \in W} \alpha_w U_w(i). \tag{1}$$

3.2 分割

U の定義から、テキストの連結部分においては周辺に比べ相対的に U の値が大きく極大となる傾向がある。そこで実際に分割する点を決めるために、

$$S(i) = \sum_{j=1}^{L} (2U(i) - U(i-j) - U(i+j)) \tag{2}$$

を各分割候補点 i に対し計算し、この値の大きいものを分割点として選ぶ。これは U をプロットしたグラフ上において U が描く曲線と高さ $(i, U(i))$ を通る水平線が近傍 $[i-L, i+L]$ で囲む領域の符号付面積に相当する。ここで L は分割点の近傍を決めるための適当な値である。

3.3 重要語抽出

与えられたテキストの表面的情報のみからトピック形成に大きく寄与する単語を選出するために長尾、水谷、池田 [1] らが開発した手法を応用する。この手法は一言でいうと、各単語について「テキスト中での出現分布が一様分布からどの程度乖離しているか」を定量的に表現したものをその単語の重要度として採用する、ということになる。ここで定量に用いるのは統計学の適合度検定で用いられる χ^2 値である。

[1] と違い、本稿は単一のテキストを対象として単語分布の偏りを調べる必要がある。そこで対象となるテキストを適当な粒度で分割しその部分テキスト単位への単語の分布がどのように偏っているかを利用する。

まず分割対象となるテキスト T を単語単位で分割し、位置のインデクス i $(1 \leq i \leq N)$ を割り当てる。すると各単語 $w \in W$ にはその出現位置の分布が定まる。次にテキストをある定まった数 M 個に等分割し、それらを T_j $(1 \leq j \leq M)$ とする。ここまできたらテキスト T 全体と各部分テキスト T_j での単語 w の出現回数を n、n_j として

$$\chi^2_w = \sum_{j=1}^{M} \frac{(n_j - n/M)^2}{n/M}, \tag{3}$$

を計算し、そのテキスト中でのトピック形成における重要度と解釈する。

ここで計算された値は、単語の分布が独立性と一様性を満たすという仮定のもとで漸近的に自由度 $M-1$ の χ^2 分布に従うので、もし $\chi^2(M-1)$ 分布において有意に棄却される域に入るならば、その単語がある特徴を持って分布していると推論できる。事実 [1] ではこの指標による重要語自動抽出法がかなり有効なものであることを実験により示している。

この結果を重みとして計算に用いるために、実際には $\chi^2(M-1)$ において上側確率がある十分低い値になる点での χ^2 値を引いた値を重要度として用いる。ここで値が負になった場合にはトピック形成に有意に寄与しない単語として重要度 0 を一律に付与することにする。

4 実験

TextTiling と本手法を文章分割実験により比較する。本来 TextTiling の問題は文章の意味段落を発見することであるが、できる限り客観的な比較を可能とするために、今回は適当な数の新聞記事を無作為に選びそれらを連結したものに両手法を適用する事で比較する。

またその後に、一般の文章に対して本手法を適用した結果のグラフを提示することで、本手法が単一の文書内で意味のまとまりをどのように捉えるのかを観察する。

各種パラメータ

本手法において用いるパラメータは、単語の重み計算でのテキスト分割数 $M = 10$、判定で考慮する近傍 $L = 100$ とする。

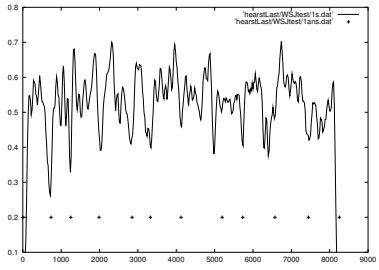

図 1　WSJ: TextTiling(縦: 類似度、横: 位置、極小点に分割点が現れる)

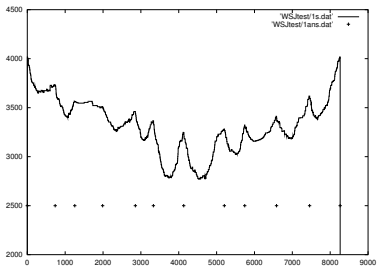

図 2　WSJ: 本手法 (縦:U, 横:位置、極大点に分割点が現れる)

テストデータ

1987 年の Wall Street Journal の記事の中から無作為に 11 個抽出したものを連結することで一つのテストセットとする。見つけるべき連結点は各セット毎に 10 個所存在することになる。これを 10 セット用意したものが今回のテストデータとなる。

問題設定

両手法とも、各セットについて分割すべき点を判断するのに対象テキスト以外の情報は利用しないものとする。ただし正解数だけは事前に与えることにする。よって比較すべきは各セットに対する正解 10 個のうち、いくつを適切に選出できたかということになる。適合率、再現率はこの設定ではこの意味での正解率として等しくなる。

結果

あるテストセットに対し TextTiling、本手法のグラフをプロットしたものは図 1、2 のようになる。[*1]

もとのテキスト中での境界点から 40 語分[*2]ずれまでの範囲で分割している場合を正解としたときと、60 語分ずれまでを正解とした場合とですべてのセットに対する平均の正解率を表 1 に示す。表中の重み付 TextTiling とは、TextTiling に本手法で用いた重みを組み入れた結果である。

表 1　WSJ でのテスト結果

	本手法	TextTiling	重み付 TextTiling
40 語	0.74	0.48	0.69
60 語	0.76	0.51	0.73

本手法は、通常の TextTiling よりも精度が 25% ほど高かった。TextTiling において、本手法と同じ χ^2 の単語選別を行っても、5% ほど精度が高かった。40 語と 60 語の二つの正解基準においてあまり精度に差がないことから、本手法の方が分割点をよりピンポイントで捉えていることが分かる。

同じ重みを付けた場合でも本手法が少し精度において上回った理由としては、本手法が大域的な偏りに比重をおいているために、時々出現する意味の切れ目とは無縁の局所的パターンの変化に TextTiling 程は影響を受けなかったということが考えられる。

ここで得た結果から、言語知識やコーパスを用いないという条件を考慮すると、本稿で用いた偏りに基づく文章単位のモデルが比較的妥当なものであるといえるであろう。

一般の文章への適用

ここではダンテの戯曲へ本手法を適用した結果を用いてその特徴を観察する。具体的にはダンテの戯曲の冒頭から始まる 5 つの章を用いて U を計算し、その結果のグラフを観察する。単語の重要度を求めるときに今回はテキストを 10 等分にした。よって単語の分布がどれくらい有意に一様分布からはずれているかは自由度 9 の χ^2 分布の値から判断される。今回はある程度大域的な傾向を掴むために χ^2 値が 25 以上のものだけをトピック形成に有意[*3]に寄与していると仮定して採用する。

結果は図 3 のようになった。一つ目の段落境界は別として、残りの三つにおいては U の値が盛り上がり、全体としての大きな傾向を比較的よく捉えて

[*1] 図中、グラフの下にある点は記事の連結点を現す。

[*2] およそ二文の長さ

[*3] $\chi^2_{0.005}(9) = 23.589$ であるから極端に有意。

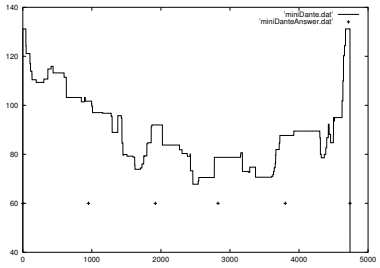

図 3　ダンテ: 本手法 (縦=U, 横=位置)

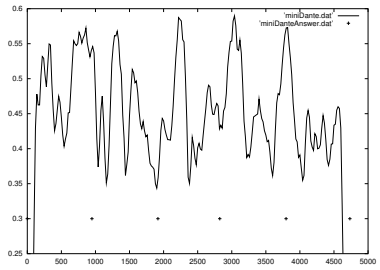

図 4　ダンテ: TextTiling (縦=類似度, 横=位置)

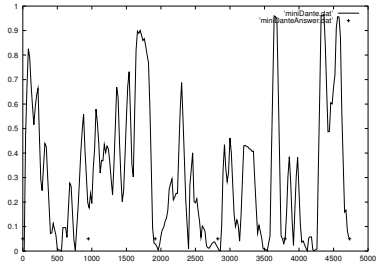

図 5　ダンテ: 重み付 TextTiling (縦=類似度, 横=位置)

いることが分かる。

同じテキストに対し TextTiling を用いると図 4 のようになる。本手法と違い、局所的な変化に敏感に反応しているために大域的な情報が埋もれてしまっていることが分かる。また本手法と同じ重みを付けた結果は図 5 である。重みを付けても局所的解析に起因する細かい変化は取り除かれていない。

この二つの比較から、本手法が内部を変えることなく比較的意味の明白な統計量をパラメタとして調節するだけで、対象とする文章のスケールにあった大域的な解析が行えるという利点を持つことが分かる。

5　考察

今回は分割を考える上である位置を切れ目に、全体を二つに分けることで指標 U を計算した。これは、重要語の選出に一様な分布からのずれを尺度としたのと整合性を持つような大域的な指標である。これにより、文章内での単語の大域的な出現パターンを探ることができる。

しかし、実際に分割点を判断するときに必要となるより詳しい解析に関しては本手法では十分ではない。この点に関しては、大域的な分割に続き、分割されたそれぞれの区間を再帰的により細かく分析するなどの工夫が考えられる。こういった工夫によって、どのような応用が可能となるのかといったことを実証するのが今後の課題である。

参考文献

[1] 長尾真、水谷幹男、池田浩之 (1976): 日本語文献における重要語の自動抽出, 情報処理, Vol. 17, No. 2.

[2] M.A.Hearst(1994): Multi-paragraph segmentation of expository text, In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics.

[3] Salton, G., Yang, C.S., and Yu, C.T.(1975): A theory of term importance in automatic text analysis. Amer. Soc. Inf. Sc 26, 1

[4] Dias, G., Alves, E.(2005): Discovering Topic Boundaries for Text Summarization based on Word Co-occurrence. International Conference on Recent Advances in Natural Language Processing.