社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DBS－145 (16)
2008－ＦＩ－ 91 (16)
2008／6／20

ウェブ検索を利用したしきい値選択型テキストセグメンテーション

阿部直人 内山俊郎 内山 匡 奥 雅博
日本電信電話株式会社 NTT サイバーソリューション研究所
abe.naoto@lab.ntt.co.jp

Text Segmentation using Web Search with Threshold Selection

Naoto Abe, Toshio Uchiyama, Tadasu Uchiyama and Masahiro Oku
NTT Cyber Solutions Laboratories, NTT Corporation

1. はじめに

テキストセグメンテーションは与えられたテキストを内容的なまとまりである意味段落に分割する処理を行うことである。発表者らはウェブ検索を利用したテキストセグメンテーション法(名詞検索法)を検討した[1]。しかし、局所的な内容の変動の影響を受け易く、意味段落の境界を決定するために使用した固定しきい値では意味段落の境界を十分に抽出できない問題があった。そこで、本発表では意味段落の境界を決定するしきい値を自動的に選択する方法を提案する。また、実際のニュース記事やブログ記事を用いて実験を行い提案手法の有効性を検証する。

2. 提案手法

名詞検索法と提案手法の基本概念を図1に示す。提案手法では助詞以外の全ての単語を使用し、活用形のある単語は終止形に変換したものを使用する。次に、検索語と関連語を用いて平均連結度[2]を算出し、平均連結度の極小値と検出箇所を調べる。そして、極小値を値の小さい順に並び替えたものを $l_1,l_2,\cdots,l_M$ とし、それに対応する順に並び替えた検出箇所を $d_1,d_2,\cdots,d_M$ とする。

並び替え後、提案手法では $l_1,l_2,\cdots,l_M$ をしきい値として用い、評価関数を最小にするしきい値を選択する。具体的には、しきい値 l_i(i=1,2,…,M)を選択したとき、d_j(j=1,2,…,i)が意味段落の境界としてテキストを分割し、i+1 個の意味段落を生成する。そして、式(1)の評価関数の値を計算する。

$$Q_i = Q_i^1 + Q_i^2 \tag{1}$$

$$Q_i^1 = \frac{1}{i+1}\sum_{k=1}^{i+1}\frac{\sum_t w_t^{all} w_t^k}{\sqrt{\sum_t \left(w_t^{all}\right)^2 \sum_t \left(w_t^k\right)^2}}$$

$$Q_i^2 = \frac{1}{i}\sum_{k=1}^{i}\frac{\sum_t w_t^k w_t^{k+1}}{\sqrt{\sum_t \left(w_t^k\right)^2 \sum_t \left(w_t^{k+1}\right)^2}}$$

ここで、w_t^{all} はテキスト全体における単語 t の出現頻度、w_t^k は k 番目の意味段落における単語 t の出現頻度である。提案手法では、式(1)を最小にするしきい値 l_{i^*}を選択し、d_j(j=1,2,…,i*)で分割した結果を提案手法の結果とする。

3. 実験

実際のニュース記事とブログ記事を使用して実験を行った。

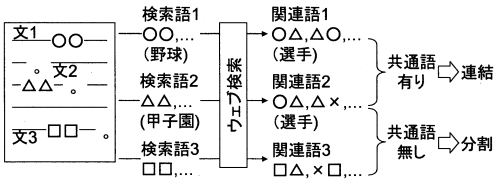

図1 ウェブ検索を利用したテキストセグメンテーションの概要

また、比較手法として名詞検索法と Hearst 法[3]を用いた。ニュース記事を用いた実験では、複数の記事を一つに連結したテスト記事を100個作成し、連結箇所を正しく検出できるかどうか調べた。ブログ記事は3人の評価者が記述内容とその範囲を判定し、結果が一致するもの(83 記事)を使用した。 実験結果を表1に示す。

表1 テキストセグメンテーション実験結果

テキスト	手法	適合率	再現率	F 値
ニュース記事	Hearst 法	50.7	47.0	48.8
	名詞検索法	60.6	74.5	66.8
	提案手法	83.2	87.5	85.3
ブログ記事	Hearst 法	8.6	12.6	10.2
	名詞検索法	23.1	36.1	28.2
	提案手法	45.2	69.8	54.9

4. まとめ

本発表では、意味段落間の異なり度合いを考慮したしきい値選択型テキストセグメンテーション方法を提案した。また、実際のニュース記事やブログ記事を用いた実験を行った。その結果、名詞検索法と比較してニュース記事ではＦ値で18.5ポイント、ブログ記事では 26.7 ポイントの改善が見られ、提案手法の有効性を確認できた。

参考文献

1) 阿部直人, 田邊勝義, 奥田英範 : ウェブ検索を利用したテキストセグメンテーション, 電子情報通信学会論文誌, Vol.J91-D, No.3, pp. 723-732, 2008.
2) 阿部直人, 内山俊郎, 内山匡, 奥雅博 : ウェブ検索を利用したブログテキストセグメンテーション法, 電子情報通信学会第19回データ工学ワークショップ, 2008.
3) Hearst, M.: TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, *Computational Linguistics*, Vol.23, No.1, pp.33-64, 1997.