

語の類縁性を用いた日本語文章の段落分けの試み

6G-8

山本 和英*

増山 繁*

内藤 昭三**

{yamamoto@toki, masuyama@}tutkie.tut.ac.jp

naito@atom.ntt.jp

*豊橋技術科学大学知識情報工学系

**NTT基礎研究所

1 はじめに

談話における文間の結束性の表現形態には、照応・省略、接続的な語句による明示的な表現[1]などの他に、類縁性のある語の使用、すなわち語彙的結束性がある[4]。我々はすでに、語の類縁性に基づく談話の結束性の評価尺度について考察した[2]。本報告では、この評価尺度を利用した段落分けの方法を提案する。

まず、対象とする文章から文間の語彙的結束性を表現するグラフ(結束グラフと呼ぶ)を構成する。次に、結束グラフに対して、最適な段落分けの評価関数を提案し、定式化する。この評価関数を実際の文章に適用して、計算機による自動段落分けの実験を行ない、その結果について考察する。

2 文章の結束グラフ

2.1 シソーラス

語彙的結束性の判断基準として、角川類語新辞典[3]をシソーラスとして使用した。また本研究では、この類語辞典を形態素解析用辞書としても使用した。また、例えば、「蒲公英」など、一般には使用されない漢字語が見出しになっている場合は、「たんぽぽ」のようにひらがな語も追加登録している。

2.2 結束グラフの構成

以下の手順により、文章から結束グラフを構成する。

- 一文に一節点を対応させる。文中の単語がシソーラスにあれば、その語を、対応する節点の語彙要素とする。ただし、出現回数も考慮に入れる。
- 節点对、すなわち辺には、以下の式による結束度を与える。

$$\begin{aligned} & \text{2文の結束度} \\ &= f(w_1 \cdot \text{同じ単語の組数} + w_2 \cdot \text{小分類一致の組数} \\ &+ w_3 \cdot \text{中分類一致の組数}) \\ & \quad (\text{ただし, } w_1 > w_2 > w_3 > 0 \text{ は定数}), \\ & \quad f(x) = \exp(-\lambda \cdot x), \lambda > 0 \end{aligned}$$

Basic study on paragraphing Japanese sentences using lexical cohesion

Kazuhide YAMAMOTO*, Sigeru MASUYAMA*, Shozo NAITO**

*Toyohashi Univ. of Tech., **NTT Basic Research Labs

関数 $f(x)$ は一般に単調非増加関数であり、文間の距離を結束度に反映させる。今回の実験では、指数関数を採用した。

3 段落分けの評価関数

今回使用する段落分けの評価関数は以下のとおりである。

$$\begin{aligned} & \text{評価関数 (最大化)} \\ &= \alpha \cdot \sum_{\text{段落}} \left(\frac{\text{段落内2文の結束度}}{\text{段落内文数}} \right. \\ & \quad \left. - \frac{\text{段落内と段落外の2文の結束度}}{\text{段落内文数}} \right) \\ & \quad - \beta \cdot \sum_{\text{段落}} \left(\frac{\text{段落内文字数} - \frac{\text{全文字数}}{\text{全段落数}}}{\text{段落内文数}} \right)^2 \\ & \quad (\alpha > 0, \beta > 0 \text{ は定数}) \end{aligned}$$

評価関数は、結束度と段落長の二要素から構成される。さらに、結束度の要素は、段落内結束度と段落間結束度から構成される。段落の性質から前者は大きい方が、後者は小さい方が望ましい。段落内の文数が増えると、一般に結束度の単純和は増加するので、段落内の文数で割ることで正規化を行なっている。

段落長の要素は、極端に長い、あるいは短い段落ができることを避けるためのものである。これら二つの要素は、定数 α 、 β を上下させることによって、結束度と段落長の重視する割合を変化させることができる。

4 段落分け実験

上記モデルに基づき、実際に段落分けを行なうプログラムを作成した。プログラムは、形態素解析、結束度の算出、段落分け、の3つのモジュールから成っている。

4.1 形態素解析

動詞、形容詞などの語尾変化の考慮を行なった後で、文節数最小法によって形態素解析を行なう。助動詞の変化や、接続助詞などは考慮しない、簡易版の形態素解析である。本研究では、形態素解析はシソーラス掲載の語を切り出すことが目的であり、そのために必要な範囲での簡易な解析を行なっている。また、シソーラスには、固有名詞は基本的には掲載されていないので、別に固有名詞辞書を追加した。

4.2 結束度の算出

形態素解析の終了後、すべての文の組合せについて、前述した計算式で結束度を計算する。今回の実験では、

3種類の枝の重みの比 $w_1:w_2:w_3$ を、 $10:8:3$ としている。

4.3 段落分け

現在、Off-lineによる方法、つまり文章全体がすでに入力されている状態を仮定して段落分けを行なっている。具体的な段落分け手順は以下の通りである。

1. 初期設定として、1文1段落とする。
2. 隣接する段落を統合するときの評価関数の減少量が最も小さい（あるいは最も増加する）位置で段落を統合する。
3. この操作を繰り返し行ない、目的関数が最大となった段落の分け方を最も自然な段落分けの候補とする。

Off-lineによる方法には、文章全体を1段落として初期値とし、それを分割していくことにより、一つずつ段落数を増やしていくという、上記の方法とは双対な方法が考えられる。しかしこの分割法では、例えば第1ステップで、文章の（語数に基づく）中央から大きくはずれるような位置での段落の分割は行なわれず、段落長がちょうど二分されるような位置で分割される傾向がある。このため、数段落に分割されるまでは、段落長の影響が強く出てしまい、結束度を考慮した自然な段落分けを行なえないと判断したためである。

5 実行結果および考察

日本経済新聞の社説33編を用いて実験を行なった。実験の結果、原文に対して55%程度、出力結果に対して42%程度の段落の一致が見られた。

パラメータ α 、 β は、一致度には影響のないことがわかった。また、ある一定値に達するまでは、パラメータ λ を大きくするほど一致度が大きくなった。これは、局所的な結合度の大小が段落の構成に影響していることを示す。ただし、 λ が一定値以上になると、隣接文の結束度も、0に近くなってしまうので、一致度は極端に悪くなる。

一致の割合 (%)

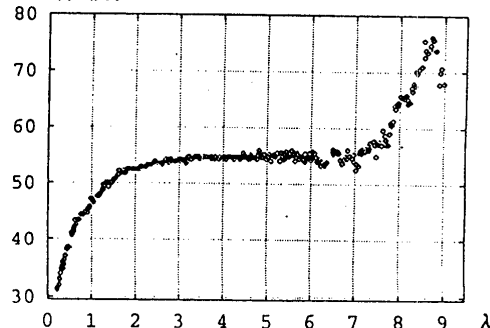


図1：原文に対する一致の割合

一致の割合 (%)

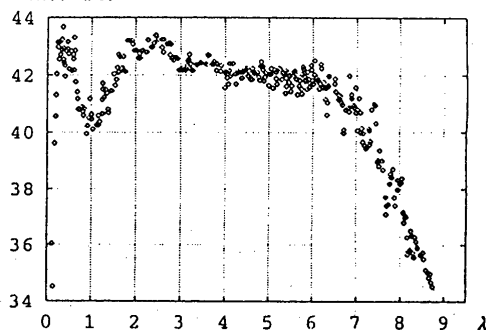


図2：出力結果に対する一致の割合

語彙的な結束性を考慮して段落分けを行なうと、文間の意味的なつながりを考慮していることになるので、ある程度自然な段落分けが可能である。ただ、「手がかり語」[1]的な要素が文中に含まれることが多いので、手がかり語を考慮していない現在の段落分けの出力結果では、不自然に感じられる部分がある。

しかし、そのような段落に対しては、最小限の変更を行なうか、手がかり語をも考慮して段落分けを行なうことによって、自然な段落分けが得られるものと考えられる。

謝辞

本研究で、シソーラスに使用した「角川類語新辞典」を機械可読辞書の形で提供いただき、その使用許可をいただいた(株)角川書店に感謝する。

参考文献

- [1] 山本 和英, 増山 繁, 内藤 昭三: 手がかり語を用いた日本語文章の段落分けに関する実証的考察, 情報処理学会 NL 84-9, pp. 65-72 (1991).
- [2] 山本 和英, 増山 繁, 内藤 昭三: 語の類縁性に基づく談話の結束性の評価尺度について, 電気関係学会東海支部連合大会 (1992).
- [3] 大野 晋, 浜西 正人: 角川類語新辞典, 角川書店 (1981).
- [4] Halliday, M. and Hasan, R.: *Cohesion in English.*, Longman Group (1976).