

分割と統合による新聞社説の接続構造解析

4 J-4

奈良雅雄 比留間正樹 田村直良

横浜国立大学 工学部 電子情報工学科

{masao,masaki,tam}@tamlab.dnj.ynu.ac.jp

1 はじめに

本研究¹では、分割と統合による文章解析手法 [2] を拡張し、解析率の向上を図る。

近年のインターネットや電子媒体の発達により大量の電子化された文書があふれてきており、これらを自動的に処理する手法の必要性が増している。文章の構造化は文章理解、要約等の処理の前提となる過程であるが、大量の文書を高速に処理するためには、なるべく深い意味解析に立ち入らない「表層的」な処理により行なうことが求められる。

文末表現から文章構造を組み立てる手法、表層的な表現から構造化する手法もいくつか提案されているが、大域的構造、局所的構造、両者ともに良好に解析する手法は少ない。

我々は、トップダウン的解析 (分割) とボトムアップ的解析 (統合) の双方の利点を生かし、文章の木構造を根から葉、葉から根へと同時に生成してゆく文章解析手法を提案した [2]。しかし、この手法において分割のパラメータである「名詞の連鎖情報」の重みがあまり反映されてなかった。

そこで本研究では、情報検索の分野で用いられる tf.idf 法を用いて重要語を抽出、その重要語の連鎖情報を利用するように分割部を拡張し、新聞社説への応用により評価する。

2 分割と統合による文章解析アルゴリズム

2.1 トップダウン的構造化 (分割)

望月ら [1] のテキスト・セグメンテーションの手法は、文と文の境界について以下の判別式で、

$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + \cdots + a_px_p \quad (1)$$

(x_i : パラメータ i の点数、 a_i : パラメータ i の重み)

閾値を越えた $\hat{y}$ によりテキスト分割の可不可を判定するものである。

我々は、(1) の評価値がテキストの「非連続性の強さ」と相関があると仮定し、この値をもとに文章の構造化を行なう。すなわち、すべての文間について、評価値を求め、評価値の高い順にセグメントの分割を行い、二分木を作る。

2.2 分割のパラメータと訓練

パラメータを以下のような観点から選択する [1]。

- 助詞は「は」と「が」の出現
主題、主語の存在が判断できると考えられる。
- 接続語句の有無
接続語句は文間の接続関係を表層的に明示していると考えられる。
- 指示語 (こそあど) の有無
指示語の存在より、その前後の数文との密接な関係があると考えられる。
- 時制の情報
着目している境界の前後の文の時制の変化について調べる。
- 文末のタイプの情報
段落内の構造に文末のモードタイプが非常に良く反映している。
- 名詞の連鎖情報
名詞の連鎖より、文章中の焦点の変化を見ることが出来る。

訓練ではテキスト中のすべての文と文の境界についてパラメータを評価し、正解としてその境界が形式段落と一致するとき $y = 10$ 、しないとき $y = -1$ を与える。

日本経済新聞から 80 編の社説を用いて、重回帰分析によりパラメータの重みを求めた。

2.2.1 tf.idf 法による重要語の抽出

田村 [2] の手法において、分割のパラメータとして使用している「名詞の連鎖情報」は、出現頻度の低い名詞も利用していたので、重みとしては小さいものであった。

そこで本研究では、パラメータの「名詞の連鎖の情報」として重要語のみの連鎖を使用するように拡張した。

情報検索の分野でよく用いられる tf.idf 法より名詞の重みを算出し、重みが閾値以上の名詞を重要語とする。

ある文書 j における名詞 i の重み $w_{i,j}$ は次のように求める。

$$\begin{aligned} w_{i,j} &= tf_{i,j} \times idf_i \\ tf_{i,j} &= \frac{\text{文書 } j \text{ における名詞 } i \text{ の出現回数}}{\text{全文書における名詞 } i \text{ の出現回数}} \\ idf_i &= \log \frac{\text{全文書数}}{\text{名詞 } i \text{ を含む文書数}} \end{aligned}$$

¹Text Structuring for Editorial Texts by Composition and Decomposition of Segments

Masao Nara, Masaki Hiruma, Naoyoshi Tamura
Department of Electrical and Computer Engineering,
Yokohama National University

2.3 ボトムアップ的構造化 (統合)

連続する4つのセグメント S_1 、 S_2 、 S_3 、 S_4 において、隣合う2つのセグメントの結束性の強さ(次節参照)をそれぞれ R_1 、 R_2 、 R_3 とすると、

$$R_1 < R_2 > R_3$$

の場合のみ、セグメント S_2 、 S_3 を統合して新しいセグメント S_{23} を作る(図1)。

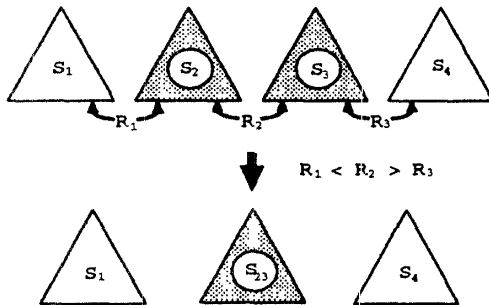


図1: セグメント統合

以上の操作を、文の先頭から繰り返し適用し、セグメントを統合してゆく。

2.4 結束性の強さ

本研究では接続構造関係の観点から見た「接続」のみを扱い、以下の「結束性の強さ」という尺度を導入する。

- 形式段落をまたぐ結び付きより形式段落内の結び付きの方が結束性が強い
- 接続表現のあるものの間の結び付きの方が無いものの間より結束性が強い
- 構造木の上部の接続構造関係より下部の接続構造関係の方が結束性が強い。
- 接続構造関係において、並列型、直列型、転換型の順で結束性が強い。

3 分割と統合の融合

本研究における解析アルゴリズムは、前節で述べたトップダウン解析とボトムアップ解析の良いところのみを採り入れたアルゴリズムで、次の2つの手順が相互に呼び合う。

topdown

1. 処理範囲が1セグメントなら終了。
2. (1)式より、セグメント列において最大の分割箇所を求め、二分割する。
3. それぞれのセグメント列を **bottomup** により構造化する。

bottomup

1. 処理範囲が1セグメントなら終了
2. セグメント列上で統合できうるセグメントを統合する。
3. 得られたセグメント列を **topdown** により構造化する。

4 解析結果の評価

解析結果の木構造の評価では、正解をどのように設定するのかという問題がある。そこで今回は拡張を行なったトップダウン解析部(分割)について評価を行なう。

新聞社説² 20編を tf.idf 法を用いた改良をする前後の解析し、その結果より再現率³、適合率⁴を求めた(図2)。その際、正解は元の社説の形式段落とした。

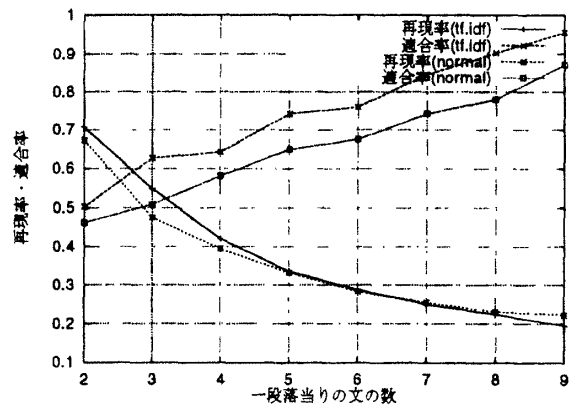


図2: 再現率・適合率

この結果から、tf.idf法を用いて拡張した手法の方が全体的に再現率・適合率ともに高くなった。新聞社説の一段落当たりの平均文数である3付近で、最も顕著に現れている。以上より、tf.idf法を用いた重要語を名詞の連鎖として利用することにより、解析の精度が向上することがわかった。

5 まとめ

トップダウン的に解析してゆくアプローチ、ボトムアップ的に解析してゆくアプローチで文章を解析する手法を示した。また分割のパラメータとして、tf.idf法を導入することで重要語の連鎖を使用するよう拡張し、実験で解析の精度向上に役に立っていることを示した。

本手法の中心的なアルゴリズムは非常に単純であるため、柔軟に拡張が可能である。今後の課題としては、格情報などの意味情報の利用および本手法の応用が考えられる。

参考文献

- [1] 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, 1996.
- [2] 田村直良, 和田啓二. 統合と分割による文章の構造解析. 言語処理学会 第3回年次大会 発表論文集, pp. 385-388, 1997.

² 日本経済新聞 94年1月・2月の社説からいくつかを使用

³ 再現率 = $\frac{\text{形式段落と一致した境界数}}{\text{形式段落の境界の数}}$

⁴ 適合率 = $\frac{\text{形式段落と一致した境界数}}{\text{本手法により検出された境界の数}}$