統合と分割による文章の構造解析

田村直良，和田啓二
横浜国立大学 工学部 電子情報工学科
{tam,keiji}@tamlab.dnj.ynu.ac.jp

1 はじめに

本研究では、論説文の文章構造についてモデル化し、それに基づいた文章解析について論じる。

近年のインターネット等、電子媒体の発達により大量の電子化された文書がアクセス可能になってきており、文書理解、自動要約等、これらを自動的に処理する手法も必要性が増している。文章の構造化はそれらの処理の前提となる過程であるが、人間がその作業を行なう場合を思えば容易に分かるように、元来非常に知的な処理である。しかし、大量の文書を高速に処理するためには、記述されている領域に依存した知識を前提とせず、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、画一的な観点からの文章の構造化では、大域的構造、局所的構造、両者をともに良好に解析する手法は少ない。我々の手法は、相互に再帰的なトップダウン的、ボトムアップ的アルゴリズムにより文章を構造化するものである。

以下、第2章では前提となる文章構造のモデルを提案し、第3章では解析アルゴリズムについて説明し、最後に第4章で実験結果を述べる。

2 文章の論説モデル

2.1 文末表現と論説文の構造

2.1.1 文のムードタイプ

文末表現はその文の性格を示す重要な情報であるといえる。我々は、福本 [1]、仁田 [5] の分類に基づいて、述語の文末表現に注目し、筆者の意見の強さを表す度合によって大きく3つのムードタイプに分類した。さらに、本研究ではこれを利用することにより論説文の構造の解析を行なう。

以下にムードタイプの分類を示す。

意見 筆者の願望や疑問などの意見が含まれる文
これらは仁田 [5] において、表出、希望、問いかけの発話・伝達のモダリティにあたる。

意見　問掛　推量

断定 筆者の判断が含まれる文
これらは仁田 [5] において、述べ立ての発話・伝達のモダリティにあたり、また判断・推量の言表事態めあてのモダリティをとる。

断定　推量　理由

叙述 事実を述べている文
これらは仁田 [5] において、述べ立ての発話伝達のモダリティを持ち、現象描写文である。

叙述　可能　伝聞　様態　存在
継続　状態　使役　例示

2.1.2 ムードタイプと論説文構造の特性

以上のようなムードタイプと論説文構造の関係を調べた。図1は文の位置と各ムードタイプの出現頻度の関係を、304個の社説[1]で調べた結果である。各文章はそれぞれ文数が異なるので、文の位置は0～1に規格化してある。

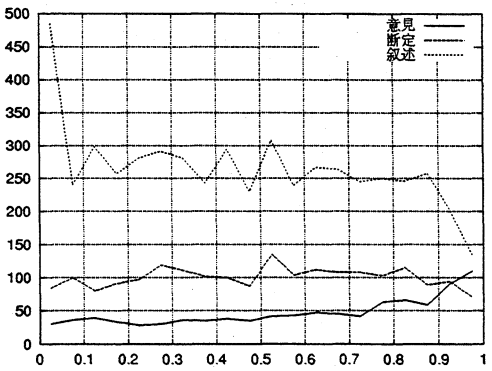

図 1: 叙述文の出現頻度

これによると、まず最初に導入があり、そこで簡単に社会一般の事実・出来事を叙述し、それに対する筆者の主張を述べる。その後、展開部分では本格的に話題の内容を示し、それについてある判断を加えていきながら、その判断に基づく筆者の意見を述べる。そして、最後に結論部分では、これまでの展開部分で述べられた筆者の意見を総括し、それを一文ないしは二文で述べる。

2.2 論説文の修辞レベル

論説文の構成を図2のように考える。本研究では、この階層的な文章構造の構築を目標に、文章の解析手法を考える。文章の修辞レベルとは、以下の通りである。

- 論証レベル：論説文章の再上位のレベルである。このレベルの構造は、固定的に「導入」、「展開1」、「展開2」、...、「展開n」、「結論」をノードとして，これ以下の構造を統括する。

[1] 日本経済新聞 94年1月から6月までの社説

- 話題レベル：このレベルでは、主に名詞の分布、連鎖に着目して文章全体における話題の構造を扱う。
- 思考レベル：[2] を参考に、思考レベル、言明レベルを導入する。これらの構造は、修辞構造理論の基づいており、表 1のように分類される。表中で、n、n1、n2 は核 (nucleus) を、s は衛星 (satellite) を表わす。
- 言明レベル：ノードは、一文あるいは言明レベルの修辞関係 (表 1参照) に対応する。各文は、命題と文末情報に相当するモダリティからなるとするが，本研究では、命題部分からは名詞の出現を，モダリティ部分からはムードタイプのみを扱う。

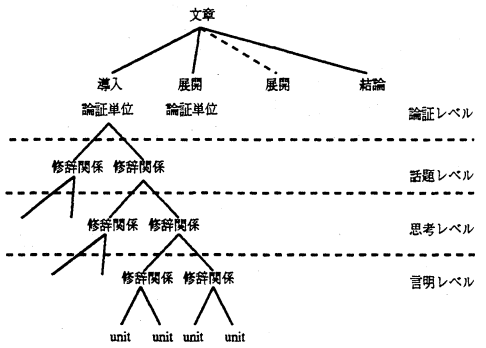

図 2: 論説文の修辞レベル

思考レベル				
直列型			並列型	転換型
n1 → n2	n ← s	s → n	n1 → n2	n → n
順接 逆接 換言	添加	条件 結論 一般化 相反 因果	並列 選択 対比	転換
言明レベル				
n ← s				
説明, 強調, 例示				

表 1: 修辞関係の分類

3 文章解析のアルゴリズム

3.1 文章解析のトップダウン的アプローチ

3.1.1 文章のセグメンテーション

望月ら [3] のテキスト・セグメンテーションの手法は、各文の文間について

$$\hat{y} = a_o + a_1 x_1 + a_2 x_2 + \cdots + a_p x_p \qquad (1)$$

(x_i : パラメータiの点数、a_i : パラメータiの重み)

なる式で、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) 式の評価値がテキストの「非連続性の強さ」と相関性があると仮定して、この値の大きさをもとに文章の構造化を行う。すなわち、

構造化のトップダウン・アルゴリズム

1. テキスト中のすべての文間について (1) により評価値を求める。
2. 評価値の高い順に分割を行い、二分木を作る。

3.1.2 セグメンテーションのパラメタと訓練

パラメータは以下のような観点から選択した。各パラメータの事象が成立する場合に値 1 を与える。

- 助詞は「は」と「が」の出現
  着目している境界の前後の文について調べる。これにより主題、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も明示していると考えられる。
- 指示語 (こそあど) の有無
  指示語が用いられているということは、その前後の数文と密接な関係があると考えられる [4]。
- 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。以前の調査によると、過去形となるのは叙述文のみで、過去形の叙述文は導入に用いられる。
- 文末のムードタイプの情報
  論説文などの文章は導入、展開、結論のような構造があると考えられる。また、それらの構造に文末のムードタイプが非常に良く反映している。
- 名詞の連鎖の情報
  名詞の連鎖を評価することにより、文章中で焦点がどのように変化するかを見ることができると考えられる。

訓練は、日本経済新聞からの 30 編の社説を用いた[2]。社説中の形式段落の位置を $y = 10$、それ以外の文間を $y = -1$ として、訓練を行い、(1) 式の重みを求めた。

3.2 文章解析のボトムアップ的アプローチ

3.2.1 セグメントの隣接関係

いま二つのセグメントが隣接しているとして、それについらの修辞関係の同定は、次のような手順で行なう。

1. 右セグメントの左端が形式段落の切れ目で接続表現があれば、それにより同定する。

[2]セグメンテーションに関する別の実験から、訓練データの社説は 50 編程度で十分であることがわかっている。

2. セグメントが部分木になっている場合、評価は核の規則を用いるが、これは基本的には、修辞構造の内の核を優先するものである (次節参照)。

3. 接続表現があればそれにより同定する。

4. 左右のセグメントが両方とも一文であるならば、言明レベルの修辞関係を優先する。

5. 文末表現の連接関係によって同定する。

6. デフォルトは "順接" とする。

3.2.2 核の規則

一つのセグメントを評価するとき、基本的に核だけでそのセグメントの評価をできると仮定する。よってそのセグメントの代表となる文は、基本的に核とする。以下のような例外処理を加えて決定する。

- 前文と後文が両方とも核になるもの
  - 前文が主 ··· 換言・並列・選択・対比
  - 後文が主 ··· 順接・逆接・転換
- 前文が核となり、後文が衛星となるもの ··· 説明・強調・例示・添加
- 前文が衛星となり、後文が核となるもの ··· 条件・結論・相反・一般化・因果・提起

3.2.3 結束性の良さの指標

前節で述べた観点に基づき、「結束性の良さ」の指標を導入し、これを「結束性の強さ」と呼ぶことにする。

この結束性の強さの評価は以下のような規則を1から順に参照することにより行なう。

1. 形式段落間より形式段落内の方が結束性が強い。
2. 接続表現のあるものの方がないものより結束性が強い。
3. 言明レベルの修辞関係より思考レベルの修辞関係の方が結束性が強い。
4. 思考レベルにおいては並列型、直列型、転換型の順で結束性が強い。
5. 思考レベルにおいて同型同士ならば先に出たものが結束性が強い。

3.2.4 セグメント統合のアルゴリズム

前節の結束性の強さに基づいて、次のアルゴリズムにより文章の構造をボトムアップ的に解析していく。

構造化のボトムアップ・アルゴリズム

1. セグメント統合のアルゴリズム
   連続する4個のセグメント S_1、S_2、S_3、S_4 において、S_1 と S_2、S_2 と S_3、S_3 と S_4 の結束性の強さをそれぞれ 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 と S_3 を統合して新しいセグメント S_{23} を作る (図3参照)。

2. 文の並びから始めて、セグメント統合のアルゴリズムを繰り返し適用し、セグメントを統合していく。

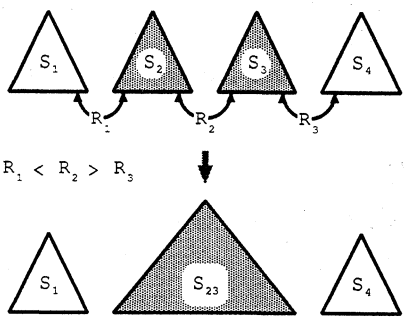

図 3: セグメント統合のアルゴリズム

3.3 トップダウン的アプローチとボトムアップ的アプローチの融合

3.3.1 トップダウン vs. ボトムアップ

ここで、トップダウン的なアプローチとボトムアップ的なアプローチについて比較する。

トップダウン的アプローチ

- パラメータにより、(表層的に) 明確に指標が現れている箇所ほど早い段階で分割が行われている。
- 構造木の葉にあたる下部に近付くにつれ、接続表現などの文間の関係から考えると適当でない分割が行なわれる。これは評価関数による判定では二文間の関係という細かい関係まで正しく評価できないことによる。

ボトムアップ的アプローチ

- 対象とする構造が小さいほど、結束性の強さや修辞関係は正しく判定される。
- 反面、大きい構造同士の修辞関係の判定は困難である。

3.3.2 統合アルゴリズム

本研究で提案する解析アルゴリズムは、次の二つの手順からなる。

topdown

1. 範囲が1セグメントなら終了。
2. (1) 式により、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. 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るセグメントのみを「セグメント統合のアルゴリズム」により統合する。
3. そのセグメント列を topdown により構造化する。

4 解析システムと実験

4.1 構造木の生成例

本研究のアルゴリズムに基づいて解析された結果を図4に示す。

4.2 解析結果の評価

アルゴリズムにより30の社説[3] (総段落数323、総文数899) を解析した結果、生成された木構造が明らかに誤りであるところが31箇所見つかった。そのうち17箇所は修辞関係の同定誤りによるもので、スパン生成の間違いは14箇所であった。

全体的な評価は表2のようになる。人間が見て「誤りがない」とすることができたのは、30の社説のうちの8文章であった。これらの文章はどれも相対的に他の文章より短く、セグメントの分割がうまくいったと考えられる。逆に許容範囲外にあるとするものも11文章であった。これらの文章は相対的に文章自体が長く、セグメントの分割誤りが目立った。なお、許容範囲外としたのはセグメントの分割誤りが二つ以上あるか、またはセグメントの分割誤りが一つで修辞関係同定誤りが一つ以上あるものとした。

誤りなし	8
許容範囲内	11
許容範囲外	11

表 2: 解析結果の全体評価

5 まとめ

本研究では、いくつかの観点から論説文の構造を階層化し文章構造のモデルを提案した。これに基づき、重回帰解析によるテキストセグメンテーションの手法を応用して文章のトップダウン構造解析のアルゴリズムを提案した。また、セグメント間の結束性の強さ、修辞関係解析からボトムアップ構造解析のアルゴリズムを提案した。さらに両者の長所、短所の検討により、両者を組み合わせる手法を提案し、実験により有効性を確認した。

謝辞

本実験で利用したコーパスは、日本経済新聞 CD-ROM '93 ～ '94 版から得ている。同社、および利用に関して尽力された方々に深く感謝します。

参考文献

[1] 福本淳一, 安原宏. 文の連接関係解析に基づく文章構造解析. 情報処理学会研究報告, Vol. 88, No. 2, 1992.

[2] 小野顕司, 浮田輝彦, 天野真家. 文脈構造の解析. 情報処理学会研究報告, Vol. 70, No. 2, Jan. 1989.

[3] 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.

[4] 高山浩史. 構造的連接性による必須格省略の復活に関する研究. 横浜国立大学, 卒業論文, 1994.

[5] 仁田義雄. 日本語のモダリティと人称. ひつじ書房, 1991.


```
[ ]
 |-[[(1,1), 順接,(1,2)], 転換,[(2,1), 並列,(2,2)]]
 |- 結論
 |-[ ]
    |-[[(3,1), 転換,(3,2)], 逆接,(4,1)]
    |- 転換
    |-[ ]
       |-[ ]
       |  |-[ ]
       |  |  |-[ ]
       |  |  |  |-[ ]
       |  |  |  |  |-[(5,1), 順接,(5,2)]
       |  |  |  |  |- 順接
       |  |  |  |  |-[(5,3), 並列,(5,4)]
       |  |  |  |- 順接
       |  |  |  |-[(6,1), 並列,(6,2)]
       |  |  |- 順接
       |  |  |-[ ]
       |  |     |-[(7,1), 順接,(7,2)]
       |  |     |- 順接
       |  |     |-[[(8,1), 順接,(8,2)], 順接,(8,3)]
       |  |- 結論
       |  |-[ ]
       |     |-[(9,1), 転換,[(9,2), 対比,(9,3)]]
       |     |- 順接
       |     |-(10,1)
       |- 結論
       |-[(11,1), 転換,(11,2)]
```

図 4: 構造木の生成例

[3] 日本経済新聞 94 年 1 月の社説からいくつかを使用