2003－NL－155 (10)
2003／5／26

文章のセグメント間関係解析に基づく文章構造解析

春日 隆緒 † 田村 直良 ††

†(株) 日立製作所 インターネットプラットホーム事業部
†† 横浜国立大学大学院 環境情報研究院
{takao,tam}@tamlab.eis.ynu.ac.jp

修辞構造理論では、20 あまりの修辞関係を定義し、この修辞関係によって文章を関係づけ、それらを階層的に表現しているが、修辞構造木の根の付近では、大きな単位（セグメント）で修辞構造を同定することは困難であり、また用意された修辞関係が適切であるとは限らない。本稿では、小さな意味段落内を修辞構造で扱いつつ意味段落間の関係付けをすることにより、２段階的な文章構造解析を示す。実現した文章解析器は、新聞社説記事を対象に、漸進的な処理により修辞構造解析を行い、トピック関連語彙を元に意味段落間の関係付け、論説文の定型的な構造をふまえた上で、ある話題から筆者の主張を導き出すまでの論旨の展開の過程を考慮にいれた解析を行う。

Text Structure Analysis based on Inter-segment Relation Analysis

Takao Kasuga† Naoyoshi Tamura††

†Internet Platform Division
Hitachi Ltd.

†† Graduate School of Environment and Information Sciences
Yokohama National University

{takao,tam}@tamlab.eis.ynu.ac.jp

In this paper, we present text structure analysis, which produces semantic segments according to the rhetorical structure theory at the first stage and analyze inter-relationship of the segments at the second stage. In the rhetorical structure theory, relations between sentences and segments are classified with pre-defined 20 rhetorical relations, and upper level segments are produced hierarchically with sentences and lower segments recursively. However, structuring such as the identification of relation is getting more difficult as coming near to the top level of the structure. We show two-stage representation. In the implementation, stack-based incremental analysis are done for rhetorical structure analysis and then relations between produced semantic segments are analyzed according to the topic words chain and the schema-based assertion flow analysis.

1　はじめに

本稿では、論説文におけるセグメント間の関係解析による文章構造解析について述べる。

文章とはテキスト表現による伝達の手段・方法であり、伝達するためには、伝える内容であるところの、考えや主張を具体的な形にして文章表現のメカニズムに組み込まなければならない。そして、この文章表現のメカニズムを理解することにより、読み手は書き手の考えや主張を論法を含めて理解することができる。このように文章に論証性をもたせるものが修辞構造をはじめとする文章構造である。

文章を構成する文、段落などの各ブロックは、“理由”と“結論”、“説明”と“具体例”などの意味的な関係でつながっている。これらの意味的関係の種類は無数にあるわけではなく、ある程度決まった関係しか用いられないと考えられる。そのような文章の意味的関係を記述するための枠組の一つに、修辞構造理論 (RST:Rhetorical Structure Theory)[5] がある。Mann らは広範囲の数百パラグラフの文章を調査し、20 あまりの修辞関係を定義した。RST では、この修辞関係によって文章を関係づけ、それらを階層的に表現している。

修辞構造理論を利用した文章の構造解析を実現する研究がなされてきた。小野ら [3] は日本語の論説文を対象に、接続詞に着目して文脈表現を形式化している。福本 [2] は文末表現に着目し、論説文を筆者の主張という観点から捉えて、文章を構造化している。我々は、これまで接続詞や文末表現による解析、トップダウン的な解析とボトムアップ的な解析を組み合わせた手法など [1, 7] の解析手法を検討してきた。

しかし、この修辞構造理論は、記述理論としては有力であるが、形式的な定義や定式化については述べられていない。そのため、そのままの形で文章解析に応用するには不十分なところがある。これまでの修辞構造解析には次のような問題点があると言える。

- 修辞構造木の根の付近では、大きな単位（セグメント）で修辞構造を同定することは困難である。また、根の付近では提案された修辞関係が適切であるとは限らない。
- 筆者がどのように論旨を展開しているのかを把握することができない。

これまでの文章解析は修辞構造理論に基づいての木構造解析が多かったが、上記の問題を考えると、単純な木構造ではなく筆者の論旨の展開レベルでの構造も表現できるような文章構造が必要であると考えられる。

そこで本研究では、筆者が論旨を展開する際の論証単位を“意味段落”とし、意味段落内は修辞構造解析し、意味段落を論旨の展開の観点より関係付けることにより、論旨の展開構造を表現し、より抽象性の高いレベルでの文章解析を実現することを目的とする。

2　論説文の構造

2.1　論説文における傾向

論旨の展開について、社説 30 記事（日本経済新聞 93 年）の筆者の論旨の展開の傾向について調査を行ったところ、以下の傾向が見られた。

- 記事には見出しがついており、これを**トピック**とすると、まず第一段落でトピックについての概要が述べられている。
- その後、トピックに関連した事例が、筆者の主張を交えて述べられる。その事例に関する話題が続く場合もある。
- ある事例について論旨が展開され、本題に戻る場合が多い。
- 最後のいくつかの段落において、トピックに関しての結論が述べられている。

2.2　意味段落の機能

上記のような論旨の展開において、意味段落には大きく分けて 3 つの役割があると考えられる。**トピックの導入**、**事例の展開**、**結論**である。以下にそれぞれの役割を具体例（日本経済新聞社説 93 年 1 月 26 日）と共に述べる。

- **トピックの導入**

(1)　ミスター・チャド・ローウェン。
(2)　曙太郎関。
(3)　横綱昇進を心からおめでとうと言いたい。
(4)　日本の相撲の歴史に画期的なことが起きたと思う。

まず導入部では、トピックについての大まかな概要が述べられ、それについての筆者の意見が述べられている。これにより、広い現実世界の中からこれから述べる特定の範囲を限定するとともに、論証の方向性をおおまかに示す。例外として、導入部が叙述文のみで構成される場合があるが、その場合は前者の機能のみをはたすと考えられる。ここでもし導

入がなかったとしても論証の内容は解釈することはできるが、非常に唐突でまとまりのない感じを受けるだろう。このように導入は文章の内容のおおまかな範囲と方向性を示すことによって、後に続く論証での筆者の意見主張を読者にスムーズに受け入れさせる役目があるといえる。

- **事例の展開**

(16) 相撲人気は、世界中で高まっている。
(17) 曙を生んだハワイでは日本の「大相撲ダイジェスト」をテレビが放送し、三大ネットワーク系のニュース番組にぶつかっても、高い視聴率を得ている。
(18) 地元紙では、フットボールや野球と並ぶ扱いだという。

展開部は、導入で提示された分野に関することから、定められた方向性にしたがって本格的に論旨の展開を行う。

導入での大まかなアウトラインを受け継いで、展開ではさらに詳しい論証が述べられる。ここで述べられる内容は導入で述べられた話題に関する内容である。導入部でのトピックに展開部のトピックが、意味的に内包されているといえる。

- **結論**

(25) 私たちは、米国で生まれた野球というスポーツ文化を日本に根付かせた。
(26) 英国で生まれたサッカーやゴルフも日本人の好きなスポーツにすっかりなっている。
(27) 日本で生まれたものが世界に広がっていく。
(28) 世界に通用する文化を、自分たちのなかに発見する。
(29) そうしたことが多ければ多いほど、私たちは自分の国を誇りに思うことができるのではないだろうか。

結論部では、展開部で導かれた結論を総括し、導入部に対する結論を述べる。これがないと、展開部の各論証単位で導かれた結論はまとまりを見い出せず、読者は中途半端な理解のまま文章を読み終えてしまうであろう。導入部とは逆に、結論部の例外としては意見文のみのものがあるが、結論の機能を考えると叙述部はさほど重要ではないといえる。

3 文章構造解析

3.1 文章構造解析手法の全体と修辞構造解析

文章構造解析は、

- 意味段落を論証単位とし、意味段落間は論旨展開の観点で関係付ける。
- 意味段落内は、修辞構造による。

修辞構造解析は、 Marcu[6] を基にしている。解析器は、修辞構造の部分木を要素とするスタックと文（ルックアヘッド）を先読みする機構をもち、増進的にルックアヘッドを進めつつスタックの上部の要素から修辞構造木を生成する還元操作とルックアヘッドをスタックにプッシュするシフト操作を選択し木の生成を進める。 Marcu[6] では、各時点で還元操作とシフト操作のどちらをとるかを、事例からの機械学習による判定機構によっているが、われわれの手法では、パラメータに優先順位を設け、その値により動作を決めている。

使用したパラメータを表 1に示す。接続詞の分類は、 [4] による。

文のパラメータ	
文のムード	意見 (意見, 問掛, 推量), 断定 (断定, 推定, 理由), 叙述 (叙述, 可能, 伝聞, 様態, 存在, 継続, 状態, 使役, 例示)
時制	現在, 過去
接続詞	敷衍, 拡張, 増強

文間の関係についてのパラメータ	
主題による結束性	主題維持, 主題省略, 主題変移
語彙結束性	有, 無

表 1: 修辞構造解析で使用するパラメータ

3.2 論旨展開構造

本節では、意味段落間を論旨の展開の観点で関係付けた、論旨展開構造について述べる。

前で述べたような、３つの意味段落の機能と、社説にみられた論旨展開の傾向を考慮にいれた、２次元的な構造とする。図 1に論旨展開構造を示す。

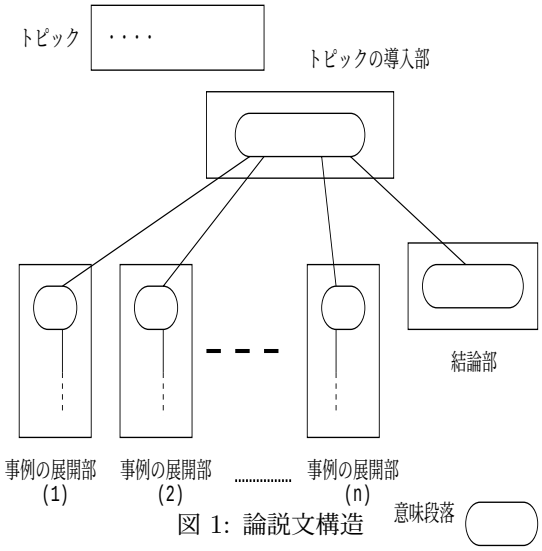

図 1: 論説文構造

２節で述べた意味段落の３つの機能と、社説に見られる傾向を考慮にいれ、図 1のような構造を考える。全体として、大きな３つのセグメントに分れている。

- トピック導入部
  第一意味段落がこれにあてはまり、ここでトピックの概要が述べられる。
- 事例の展開部
  ここでは、トピック導入部に関連した事例について述べられる。
- 結論部
  最後の意味段落がこれにあてはまり、ここでトピック導入部に対する結論が述べられる。**結論**の関係で第一意味段落と関係づく。

さらに、事例の展開部ではいくつかの意味段落が２次元的に結合しており、結び付く関係の種類としては、以下のものを考える。

- 事例
  関係先の意味段落の内容に関する事柄に関して、述べられている。
- トピック事例
  論旨の展開がトピックの本題に戻ってきた場合がこれにあたる。
- 接続詞による関係
  関係先の意味段落と、接続詞に応じた意味的な関係を持つ。
- 順接
  直前の意味段落と意味的に順接の関係にある。

図 2に事例の展開部構造を示す。

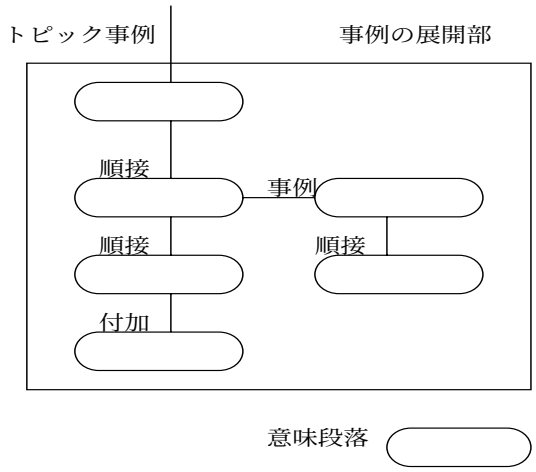

図 2: 事例の展開部構造

3.3 トピック関連語彙の抽出

意味段落間の関係付けにあたって、その記事における話題の中心、または関連する語句という意味で、**トピック関連語彙**の抽出を行う。

まず、トピック関連語彙候補として以下を抽出する。

- トピック中の全ての名詞
  その社説記事のタイトルであり、そこに出現する名詞は話題の中心であると考えられる。
- 第一段落中の全ての名詞
  社説に見られる傾向として、第一段落はトピックの導入であることを述べた。よって、第一段落に出現する名詞はトピックに関連した語句として重要度が高いと考えられる。

その後、トピック関連語彙候補に対して重み付けを行い、重みの高い上位５割の単語を**トピック関連語彙**とする。重み付けは、$tf * idf$ を用いる。

3.4 関係同定

ここでは、トピック関連語彙を用いて意味段落についての関係評価について述べる。

1. 社説記事では、第一意味段落でトピックについての概要が述べられており、第一意味段落

を中心に論旨の展開が行われると考え、これを**トピック導入** とする。

2. 意味段落の先頭に接続詞が存在する場合、その接続詞により関係が決まる。関係付けの対象は、直前の意味段落とする。

3. 注目している意味段落（Aとする）について、Aの全ての文の主題からなる集合を求める。べつの意味段落（Bとする）中の全ての文の主題・題術部分とマッチングをとり、一致した場合にAはBの**事例**とする。

   対象意味段落（B）は、直前の意味段落から始め、第一意味段落まで関係をたどりながら評価を行う。

4. 事例としての対象意味段落が存在しない場合、注目している意味段落中の全ての文における主題の集合について、トピック関連語彙とマッチングをとり、一致した場合**トピック事例**とする。関係付けの対象は、第一意味段落とする。

5. 以上の関係がみられない場合は、デフォルトとして**順接**とする。関係付けの対象は直前の意味段落とする。

6. 最後の意味段落については、これを**結論**とする。関係付けの対象は、第一意味段落とする。

4 結果とその評価

4.1 修辞関係について

本研究では、新聞社説記事１０記事に対して、文章構造解析を行った。表2に、各記事における修辞解析結果を示す。表において、「修辞関係数」は各記事においてスパンを結び付けている修辞関係数の総計である。「誤った関係」は、人手により修辞関係を評価し明らかに誤りであるとみなされた関係数の総計である。評価の際には、「〜という意味関係にもとれないことはない」といったような修辞関係は解析結果として許容範囲とし、誤りとはしない。

解析された修辞関係の中で、明らかな誤りは１６個所であり、全体の約５%となった。ここでは、各誤りについて述べ、検討する。

- 並列的関係
  ユニット間に並列的な関係が存在するときは、ほとんどの場合ユニット間に接続詞が存在していた。しかし、３個以上のユニットが並列的であった場合、接続詞表現が存在しない場合がある。その場合、並列的な関係は判定することができなかった。

記事番号	文数	修辞関係数	誤った関係
1	65	53	2
2	45	38	2
3	61	48	1
4	43	32	1
5	50	36	2
6	19	12	1
7	25	19	2
8	18	11	0
9	51	40	3
10	31	24	2
計	408	313	16

表 2: 記事毎の修辞解析結果

  並列的関係を判定するには、様々な定型パターンを用意しておく方法が考えられるが、表現方法には限りがなく、全て網羅することは困難であるといえる。

- 対比、対応
  ユニット同士に存在する語句が対比・対応している場合、その対応関係を判定することができない場合があった。

- 倒置的な表現
  倒置的な表現が用いられた場合、意味的に正しい判定を行うことができない場合があった。

- 定型的な表現
  定型的な表現が用いられた場合、その意味関係を判定することができない場合があった。

- 文末のムード
  文末のムードのタイプを意味的に同定できる場合においても、ムードタイプの分類で網羅しきれていない文末表現があった。

  これは、社説には定型的な言い回しが多いことから、さらにムードタイプの分類を検討・細分化すること等により再分類を行う必要がある。

4.2 意味段落間関係について

表3に、各記事における意味段落間の解析結果を示す。

解析された意味段落間関係の明らかな誤りは９個所であり、全体の約１０%となった。明らかな誤りの認められない記事が半分ほどあり、概ね良い結果が得られたといえる。接続詞による関係同定はほぼ

記事番号	意味段落間関係数	誤った関係
1	12	2
2	9	0
3	13	1
4	11	0
5	15	4
6	6	0
7	5	0
8	6	1
9	8	0
10	7	1
計	92	9

表 3: 記事毎の意味段落間関係解析結果

問題なく行えた。しかし、一つの意味段落間関係を誤ってしまったために、連鎖的に誤ってしまった場合などもあった。

- 意味段落分け
  修辞構造解析の部分での誤りにより、結果的に意味段落間の関係が誤ってしまうケースがあった。

  – ・・・な必要がある。一つの有力な方法は、・・・。

  この２文の間は、形式段落の境となっている訳であるが、修辞関係の評価でも述べたように、例示的な関係を判定できずに、意味段落分けされてしまった。

- ３つの機能としての関係
  論旨展開構造において、トピックの導入部、事例の展開部、結論部の３つに機能を分類した。本研究では、社説１０記事に対して解析を行った訳だが、３つの機能としての分類はほぼ問題なく行えたと言える。事例の展開部においても、「特定の事例について論旨が展開し、本題に戻る」といったような論旨展開構造を正しく表現できた場合が多かった。

  しかし、結論部と事例の展開部との境界が、社説記事としてあいまいな場合もあった。これは、筆者によってある程度論旨の展開方法が異なっているためであると考えられる。よって、論旨展開構造における機能の分類の細分化等が必要であると考えられる。

5 おわりに

本研究では、新聞社説記事を対象に、漸進的な処理を用いて修辞構造解析を行い、トピック関連語彙を元に意味段落間の関係付けをすることにより、２段階的な文章構造解析を行った。また、構造解析において、論説文の定型的な構造をふまえた上で、ある話題から筆者の主張を導き出すまでの論旨の過程を考慮にいれた解析を試みた。

社説のような定型的な表現・構成を持つ文章は、約１割程度の誤り以外は、許容範囲の解析を行うことが可能であった。修辞構造木の根の付近の構造について、論旨の展開構造解析を行うことによって、概ね良い関係付けが行えた。また、論旨の展開の観点から意味段落の機能を３つに分類し、修辞構造木では把握できなかった論旨展開を表現した。

謝辞

本研究では、日本経済新聞社による９３年新聞記事 CD-ROM を用いた。

参考文献

[1] 有道啓史. 表層情報による文章の構造解析に関する研究. Master's thesis, 横浜国立大学大学院, 1992.

[2] 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. 情報処理学会自然言語処理研究会, Vol. 88, No. 2, 1992.

[3] 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.

[4] M.A.K.Halliday. *An Introduction to Functional Grammar*. Edward Arnold(Publishers) Ltd., 2001. 邦訳：機能文法概説, 山口他訳, くろしお出版,2001 年.

[5] W. C. Mann. Rhetorical structure theory : Description and construction of text structure. In G. Kempen, editor, *Natural Language Generation*, pp. 279–300. Martinus Nijhjoff Publishers, 1987.

[6] Daniel Marcu. *The theory and practice of discourse parsing and summerization*. MIT Press, 2000.

[7] 田村直良, 和田啓二. セグメントの分割と統合による文章構造解析. 自然言語処理, Vol. 5, No. 1, pp. 59–78, 1998.