文章の分割と文脈構造の解析

小野　顕司　　住田　一男　　浮田　輝彦　　天野　真家

（株）東芝　総合研究所

1．はじめに

我々は文脈構造を文と文，さらにそれらが結合されたものの間にある接続関係の総体であると捉え，「思考の流れ」という観点から，その記述方法及びテキストから自動抽出する方法について考察してきた（小野[89]，木下[89]）．文脈構造は，文を最小単位とし，それらが2項接続関係（例示関係，並列関係，順接関係等２４種）で互いにつながれて構成される２分木として表現する．図１にその概略を示す．これは，いわば文章中の各接続詞の２項オペレータとしての"スコープ"を明確にしたものといえる．この文脈構造を実テキストから抽出するため，従来は「思考制約規則」と呼ぶ，隣接する２接続関係間に存する文脈構造に対するプリファレンスをルール化したもの約400個を用いていた．

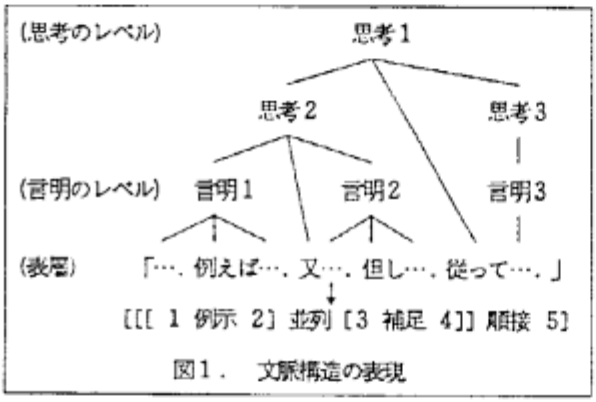

図１．　文脈構造の表現

しかし，この方法では
（１）接続表現の無い部分については解析ができない，
（２）大局的な構造が規定できず，構造候補を十分絞りきれない，という２つの問題点があった．

（１）の部分を調べてみると，前後に接続関係を示す修辞的表現が存在するため，その部分に接続表現を用いていないケースがかなりあることがわかった．

また（２）については，話題表現や反復語句の分布といった情報によって構造を絞り込める可能性が高いことが判った．

これらの点に対処するため，接続表現にかぎらず，文脈構造を示唆するような個別の修辞表現や表層的特徴一般を扱えるようにシステムを拡張した．具体的には，各表現と，各表現が規定する前後の文脈の構造（文章の部分的なまとまりや，それらの間の相対関係を示す機能）とを if-then ルール（セグメンテーションルールと以降呼ぶ）の形で準備し，入力されたテキスト中に該表現が存在する場合，その表現が規定する構造にそぐうように文脈構造候補を生成するようにする．

本報告ではその概要を説明する．

Text Segmentation and Discourse Analysis
Kenji Ono, Kazuo Sumita, Teruhiko Ukita, and Sin'ya Amano
R&D Center, Toshiba Corp.

＊本研究は，ＩＣＯＴからの委託により第５世代コンピュータプロジェクトの一環として行っている．

2．処理概要

図２に処理全体の概要を示す．入力されたテキストはまず形態素・構文解析され，その結果から各文の間の接続関係が抽出される．抽出された接続関係の並びを接続系列と呼ぶ．次のセグメンテーション処理部では，入力されたテキスト中にセグメンテーションルールとして登録されている表現がないか調べ，存在する場合は，該表現が持つ構造規定を"制約"として接続系列中に付加する．この処理については次節で詳しく述べる．次の構造候補生成処理では，構造制約が付加された接続系列を元に，制約を満たすような文脈構造を総て生成する．
生成された各構造候補に対し，従来通りの思考制約規則による評価をおこない，最終的に構造を絞り込む．

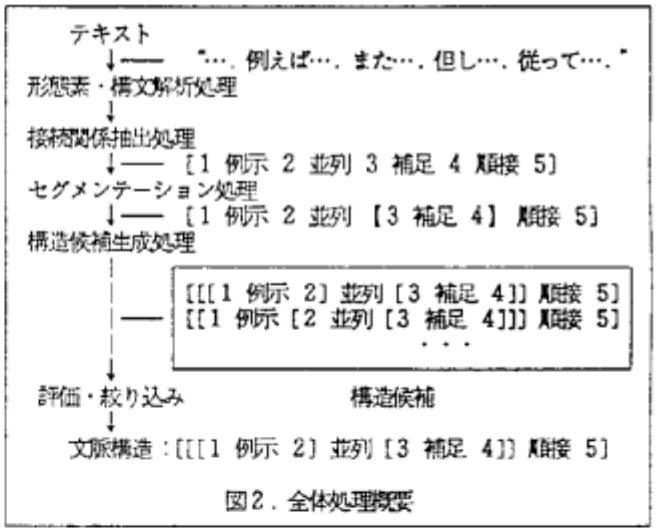

図２．全体処理概要

3．セグメンテーション処理

3．1　セグメンテーション処理で扱う修辞表現の例

図３に示すような文章を考えてみる．記号 'ε' は，文頭に明示的な接続表現が無いことを示す．また，A，B，…は，適当な名詞句を示すものとする．

このような文章に於いては，（１）文 N+1〜文 M-1 までが一つのまとまりを構成していること，（２）文 M からの何文かで，またひとつのまとまりを構成していること，（３）上述の２つのまとまりは，ともに文 N の内容を詳細説明するものであり，対等（並列）の関係であること，またその意味で，一つのまとまりを構成すること，といった構造が自明に認識される．

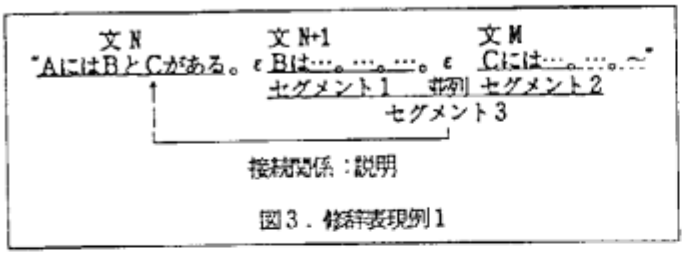

図3．修辞表現例1

以下，この（1）や（2），（3）のような文のまとまりをセグメントと呼ぶこととする．セグメント1については，その開始位置（開始文）も終了位置（終了文）もはっきりしているが，セグメント2については，終了位置（終了文）が判っていない．

別の例を図4に示す．ここで"同列"とは，接続詞"すなわち"などに代表させるような接続関係である．

どちらの例においても，表層的に接続表現のない部分（図中'ε'で示された部分）の修辞的な接続関係は自明である．

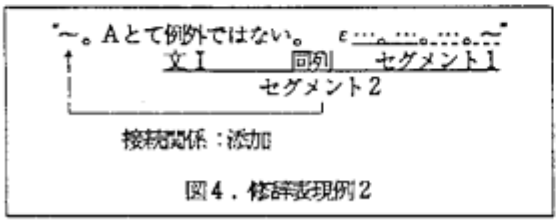

図4．修辞表現例2

このような，1文ないし数文にわたる表層的特徴（同語反復や話題表現，接続表現，特定の修辞的表現等）が持つ構造規定性を文脈構造解析に反映させる処理がセグメンテーション処理である．

3．2　セグメンテーションルール

セグメンテーション処理は，前節で示したような個別の修辞表現毎に用意されたセグメンテーションルールを，入力テキストが該表現を含む場合適用する処理である．セグメンテーションルールの適用は，接続系列中に判明した接続関係と特定の構造制約記号を挿入するという形で行う．

この構造制約記号は以下に示す3つである．

（1）　'［' および '］'

'［' と '］' で囲まれた部分が1つの部分構造を構成することを示す．開始，終了位置の判っているセグメントに対して用いる．すなわち，文章の部分的なまとまりを示す制約である．

（2）　'(' ，および ')'

'(' は，生成される構造中にその箇所から始まる部分構造が存在することを示す．')' は，その箇所で終わる部分構造が存在することを示す．開始位置，終了位置のどちらかが判らないセグメントに対して用いる．

（3）　'@'

その箇所で終わるような部分構造が存在しないことを示す．後続する文ないしセグメントが，直前の文，ないしは前接するセグメントに内容的に直接係っていることが自明な場合，用いる．

図3に示した修辞表現に対するセグメンテーションルールは，以下のように表現できる．

●If sentence N has such expressions as
"AにはBとCがある" and
If sentence N+1 has such expressions as
"Bは…" and
If sentence M(> N) has such expressions as
"Cは…" then
●insert the relations and the constraints
as follows:
[… N @ 説明 【N+1 … M-1】@ 並列 M …]

図5．セグメンテーションルール例

このルールを図6に示す5文からなる文章に適用した例を図7に示す．従来14構造候補生成されるところが本例では適切な構造制約によって2候補しか生成されていないことがわかる．

文1　生き物は動物と植物からなる．
文2　動物は人間とけものからなる．
文3　この場合の動物はけものの意味である．
文4　一方，人間を含むものとして動物という場合もある．
文5　植物は陸生植物と水棲植物からなる．

図6．テキスト例

接続系列
［1 ε 2 継続 3 対比 4 ε 5］

ルール適用 ↓

［1@ 説明 【2 継続 3 対比 4】@ 並列 5］

構造候補生成 ↓

［1 説明 [[[2 継続 3] 対比 4] 並列 5]]
［1 説明 [[2 継続[3 対比 4]] 並列 5]]

評価・絞り込み ↓

［1 説明 [[[2 継続 3] 対比 4] 並列 5]]

図7．図6のテキストに対する処理例

4．まとめ

文脈構造解析において個別的な修辞表現や複数の文にわたる表層的特徴を扱うための処理であるセグメンテーション処理について述べた．今後この方式にのっとってセグメンテーションルールを拡充してゆき，構造抽出精度を高めていく予定である．

5．参考文献

・小野他：文脈構造の分析，情報処理学会自然言語処理研究会資料　70-2，1989.
・木下他：日本語テキスト理解における文脈構造抽出法，情報処理学会「談話理解モデルとその応用」シンポジウム，pp.125-136，1989.