日本語論説文の自動抄録のための文脈構造解析

7B−10

小野　顕司　住田　一男　知野　哲朗　浮田　輝彦

(株)東芝　研究開発センター　(株)東芝　関西研究所

1．はじめに

論説文などには，内容の構成や論述に構造的なものが感じられる．その構造性は，読者が持つ文章の話題に関する知識に依る部分もあるが，文章の修辞的な特徴から把握されている部分も多い．特に読者の知識を前提としない一般解説記事等では，何が何だから何なのかという論旨の構造を読者に明示するため接続詞等を多用するので，この傾向は著しくなる．

我々は，このような修辞的な手がかりをもとに，知識を用いずに文章の構造（文脈構造）を解析するシステムを開発してきた(1),2),3),4))．本稿ではこれら修辞的表現の処理に関してどのような問題があり，それにどう対処したかについて，列挙表現を例として述べる．

2．文脈構造解析の概要

文脈構造とは文と文，さらにそれらが結合された単位の間に存在する接続関係を記述したものである．これを，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３５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．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．この構造を抽出するのが文脈構造解析である．

文脈構造解析部は話題解析部，セグメンテーション処理部と文脈構造候補生成・絞り込み部の３つから成り，自動抄録のための文脈構造を単文解析結果から抽出する．図１にその概略を示す．

単文解析部は，各文に含まれる接続詞などの接続表現を抽出し，各文の接続関係を決定する．解析には，正規表現*で記述された接続表現約１５００からなる辞書（単文解析用辞書）を用いる．

話題解析部は，各文中の話題表現（助詞‘は’，‘も’，‘こそ’などで話題化されている名詞句）を抽出する．そして，抽出された名詞句が前文あるいは前方の文章中に現れていないか分析する．こうして，話題の構造や推移に関する情報，例えば，“ＡにはＢとＣがある。Ｂは…。…。Ｃは…。”，“…としてはＡなどがある。Ａは…。”といった文章に対する構造を検出する．

セグメンテーション処理部は，複数の文にわたる修辞的な表現を手がかりにして，文脈構造に関する情報を抽出する．譲歩的な表現（“確かに…。…。しかし…。”のような文章）や，列挙表現（“…は２つある。一つは…。…。もう一つは…。”）に対処する規則を中心に約１５０のルールを用いている．

構造候補生成・絞り込み部は，抽出された接続関係や文脈構造に関する情報をもとに，可能なすべての文脈構造候補を生成した後，ｐ／ｎルールと呼ぶ隣接する接続関係の間の局所的な構造に関するプリファレンス規則で評価し，上位規定個を出力する．

次に，各処理部に於けるデータ構造とその流れについて述べる．

単文解析部で各文の接続関係が取り出された後，文番号と接続関係とを交互に並べたリストが作られる．これを接続関係系列と呼ぶ．

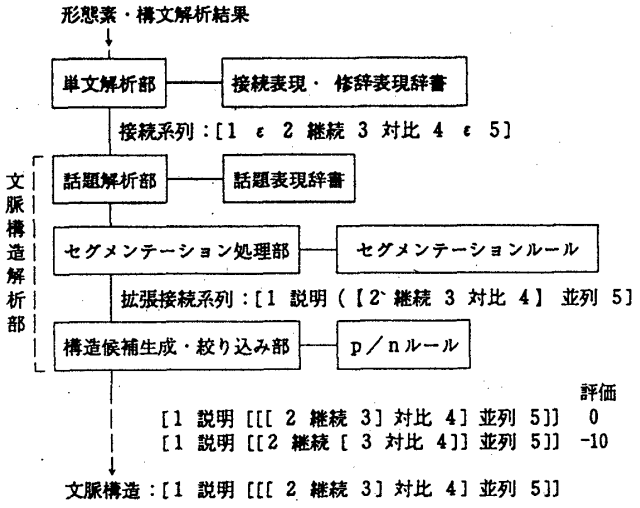

図１　文脈構造解析の概要

この接続系列の中に，話題解析部やセグメンテーション処理部が後述する“構造制約記号”を適宜挿入する．このような構造制約記号が挿入された接続系列を，拡張接続系列と呼ぶ．

構造候補生成・絞り込み部は，挿入された構造制約条件を満たす範囲で，その拡張接続系列に対して可能なすべての文脈構造候補を生成し，思考制約規則と呼ぶ局所的な構造に関する規則で評価した後，上位規定個（１個）を出力する．

3．列挙表現を解析するときの問題点

列挙表現とは，箇条書きなど幾つかの事柄を並べて述べる表現一般のことである．典型的な列挙表現は，以下の５種類である．

① (1) (2) (3)，一二三，(a) (b) (c)等の記号による呈示
② ‘・’等，単一記号の連用によるマーキング
③ “…．第一に…。第二に…。第三に…。”
④ “…．まず…。次に…。さらに…。…も…。最後に…。”
⑤ “…には，Ａ，Ｂ，Ｃがある。Ａは…。Ｂは…。”
（Ａ，Ｂ，Ｃは名詞句）

①～④で用いられている修辞的表現を以下シーケンシング表現と呼ぶことにする．処理上問題となるのは次の点である．

a) ①に関して，章や節の見出しと間違える．
b) ②に関して，文中に２カ所以上この種の列挙表現があった場合，混同してしまう．
c) ③に関して，シーケンシング表現の多様性が問題となる．例えば“第３に”という代わりに“また，第３の問題は”といった言い方をすることもある．
d) ④に関して，“次に…を述べる。”といった列挙表現でない表現と区別しないといけない．また“最後に”という表現は，文章全体の結語として使われている場合もあるので，区別しなくてはならない．

Discourse Structure Analysis for Automatic Text Skimming System for Japanese Expanatory Texts, by K. Ono, K. Sumita (Toshiba Research and Development Center), T. Chino, and T. Ukita (Toshiba Kansai Research Lab.)

e) ⑤に関して，列挙表現を示すマーカが他と違って固定でないので，話題解析結果を利用した特別な処理が必要である．
f) 最後の列挙要素の叙述範囲（スコープ）の決定
g) 各列挙要素のスコープの中の文脈構造に関して：
通常最初の文が独立していて，その説明が以降続く．この構造的な特徴を接続系列中に反映しなければならない．
h) "…には次の3つがある。第一は…。第二は…。"のような，列挙文の前に予告的な文がある場合の文脈構造上の措置．
i) "図＊に…を示す。…。表＊に…を示す。"といった図表の参照表現は，列挙的な場合がある．段落の先頭にそういった表現がある場合は以降にその図表の説明が続くが，段落の最後にあった場合はそれまでの説明の補足であることが多い．このような構造的な違いを処理しなくてはならない．このことは，"第1章では…について述べる。…。第2章では…について述べる。…"といった前書きの中の表現にもあてはまる．
j) 字下げ，改段落されている部分の影響：
これは，最終要素のスコープ決定や，図表の参照表現のスコープ決定への影響が大きい．

4．対処方法

c)，d)の問題は単文解析部で対処する．単文解析用辞書登録の例を表1に示す．

表の第2フィールドは，登録された表現の文中の出現場所に関する条件を示す．文頭(BUNTOP)，文末(BUNEND)，文の先頭の節の末尾(TERMEND)，などが指定できる．第3フィールドは形態素列に対する条件である．記述がある場合のみ，チェックされる．

表中3行目の登録は，"第一に…"，"第一の…"，"第一は…"といった表現にマッチする．もしこのテンプレートが助詞 'に' 'の' 'は' を含まないものであると，"…。第一彼は未成年だ。"といった，「そもそも」という意味の表現と誤マッチしてしまう．6行目の登録には品詞条件が書かれているので，"まずいのは，…"といった表現に誤マッチすることがない．

最後の登録は，不定部分 ".＊" を含むものである．"最初に話したいことは，…" "最初に問題になるのは，…"といった表現を検出するためのものである．文中の同じ箇所に複数の登録がマッチしたときは長い方を優先するので，2番目の登録よりも優先される．

単文解析辞書には "…を述べる。"といった文末表現を捉える登録も存在する．複数の登録がマッチした場合，優先順序は，マッチした位置に従って，文頭→文末→文の先頭の節の末尾→その他，となっているので，d)のような問題は自動的に回避される．

このような工夫により，c)やd)の問題に対処している．

次に，f)の対処を例として，列挙表現を処理するセグメンテーションルールについて説明する．以下の3ケースに分けて処理される．

1) シーケンシング表現を含む文で段落が終わっている場合：
 1-a) 次の文（次の段落の先頭文）に "以上" など概括的な表現がある場合 → スコープはその文まで：
 1-b) それ以外 → スコープは次の段落の最後まで
2) 1)以外の場合
 2-a) "以上" のような概括的な表現を含む文が段落内にある場合 → スコープはその文の前まで．
 2-b) 2-a)以外の場合 → スコープはその段落の終わりまで．

同様のヒューリスティックにより，b)，g)〜j)の問題を処理している．このように，セグメンテーション処理では，段落の境界を目印とした文の相対的位置関係が重要である．これらの指定をしやすいように，ルールの記述方法は工夫されている．

表1 単文解析用辞書

表層	位置条件	形態素品詞列
(?[1 a I])	BUNTOP	
(さいしょ|最初)[にのは]	BUNTOP	
第[1一][にのは]	BUNTOP	
(初め|はじめ)[にのは]	BUNTOP	
(一|ひと)つに?は	BUNTOP	
まず	BUNTOP	/まず<副>
まず第[一1][にのは]	BUNTOP	
最初.＊(問題|こと|の)は	TERMEND	

a)は，本稿で説明しなかったが，単文解析部の前処理である書式解析部で処理される．e)は，話題解析結果を利用したセグメンテーションルールで処理される．

5．解析結果の評価

文脈構造解析結果の評価は，3人の被験者に対し，新聞社説および教科書から採取した4文ないし5文からなるパラグラフ40個について文脈構造を記述してもらい，その結果とシステムの解析結果とを比較するというやり方で行った．構造の比較は，構造の形状（ツリーの形）に関してのみで，接続関係については見ていない．

まず被験者間の一致度であるが，完全一致は35％であった．構造の部分的な一致も認めると，一致率は58％となる．その部分に対するシステムの解析結果の一致度は41％であった．（システムが完全にランダムに構造を作るとすると，一致率は1％以下である．）

文脈構造の評価方法には定まったものはなく，この評価方法も，接続関係を見ていない点を含め，いろいろ問題がある．現在よりよい評価方法を検討中である．

6．終わりに

列挙表現の対処を例として，単文解析処理と文脈構造解析処理の実際について述べた．生成された構造に基づいて抄録文が生成される．列挙表現の場合，3節で述べたg)の処理結果が特に抄録文に影響する．

現在のルール数は，接続表現・修辞表現辞書1350，セグメンテーションルール150，p／nルール約600である．

今後更にルールを追加してゆく．典型的な修辞表現をすべて網羅するだけでも，ある程度のルール数は必要であると思われる．また，その後は，特定の事柄を述べるときの典型的な文章構成に関するルールを増やすことが必要であると思われる．これは文章の内容に依存するので，どこまで記述できるかについても検討しなければならない．しかし，門外漢の人間が専門文献を読んだ場合でも文章の構造が掴めるということを考えれば，一般的なルールだけでもある程度の解析はできると考えている．

7．参考文献

1) 小野，浮田，天野：文脈構造の分析，情報処理学会研究会報告 NL70-2，1989.
2) 小野，住田，浮田，天野：文章の分割と文脈構造の解析，情報処理学会第43回全国大会講演論文集3，pp.251-252，1991.
3) Sumita,K.,Ono,K.,Chino,T.,Ukita,T., and Amano,S.:A Discourse Structure Analyzer for Japanese Text, Proceedings of International Conference on Fifth Generation Computer Systems, Vol.2, pp.1133-1140, 1992.