TM-1198

日本語文章の構造解析システム

福本　淳一、安原　宏（沖）

August, 1992

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03)3456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

日本語文章の構造解析システム

Japanese Text Structure Analysis System

福本　淳一　　安原　宏

Junichi Fukumoto　　Hiroshi Yasuhara

要旨

論旨の展開構造を文章の構造ととらえ、書き手の主張に基づく文章構造解析システムを開発した。本システムは、文章の形態素解析結果から抽出された文タイプ情報、主題提示情報、接続語句情報、指示語情報、同語反復情報の各情報を入力とし、文脈現象の解析に基づいて記述した約400規則を用いて文章の構造解析を行う。文章の構造解析は、まず文章中の連接2文間の関係を解析し、その関係の強さにより文のグループ化を行い、最後に文章構造要素を認識することにより文章全体の構造を得るものである。

1 まえがき

従来まで行われてきた1文レベルの自然言語処理では、文章全体の意味するものや文の流れを判断することは困難である。このような1文レベルの処理を越える文脈処理技術は、文章要約などの自然言語応用システムや高度な機械翻訳システムの実現に不可欠な要素技術であると考えられる。

これまで文脈処理に関して幾つかの研究[1)2)3)]が行われてきたが、これらは限られた範囲の問題について予め準備された知識を用いたものであり広い範囲の文章に対応できないという問題があった。また、文章の要約処理の実現を考えた場合、与えられた文章全体の構造を認識し、そこから重要部分を抽出する必要がある。本研究は、大量の文章の文脈現象の解析から得られた知識に基づいて記述した文脈規則を用い、広い範囲の文章の構造解析を行うシステムの実現を目指している。

論説文等の文章は、ある事柄についての書き手の考えや意見などの書き手の主張を述べることを目的とした文章である。このような文章においては、書き手の主張が読み手である読者に伝わるような論旨が展開されている。この論旨の展開構造が文章の構造であると考えられる。一つの文の中では、語及び文節間に明白な形態的指標が存在するのと同様に、複数の文からなる文章においても形態的指標が存在すると考えられる[4)]。我々は、広い範囲の文章の各文間の形態的表現の解析から得られた知識を規則として記述し、それを用いることにより文章の構造化を行った[5)6)]。

論説文などの文章を構成する各文には、書き手の主張を表わす文や主張のために必要な客観的な事実を述べた文が存在する。日本語においてはこのような書き手の態度は主に文末の表現として現われ、特に文末の助動詞と動詞の機能によって決まる[4)7)]。我々は、各文間の関係を決定するための形態的な情報として、書き手の主張に基づいて分類した文タイプ情報を用いた[8)]。また、助詞「は」等により提示される主題提示語、文章の流れを明示的に決定する機能のある接続語句、文章中の他の部分を参照する機能のある指示語、文章中で繰り返し用いられる同一名

詞も文間の関係を決定するための形態的な情報として用いた[6]。

以上の情報を用いて文章構造を解析するため、我々は新聞社説記事中に現われる文脈現象の解析から得られた知識を基に、文章の構造化のためのモデル及びそのモデルに基づく言語仕様を定義した。これを用いて記述した規則を適用することにより文章の構造解析を行うシステムを開発した。本システムでは、まず、文章の形態素解析結果から各文の文タイプ、主題提示語、接続詞語句、指示語、同語反復の各情報を抽出し、これらを用いて文章中の連接2文間の関係の解析を行う。そして、その関係に基づき文のグループ化を行い、各生成されたグループ間の関係の解析から文章構造要素を認識することにより文章全体の構造を得る。

以下、2章で文脈現象の調査結果について述べた後、3章で文章構造のモデル及び記述言語について述べる。次に、4章で本システムの文解析及び文章構造解析手法について述べ、5章で新聞社説記事を用いた文章構造解析の評価について述べる。

2 文脈現象の調査分析

文章中の各文間の関係を決定するため、新聞社説記事を用いて文間関係と照応現象の調査を行った。

2.1 文間関係の調査

文章を構成する各文は、書き手の考えや意見などの書き手の主張が現われている文（主張文）とそのような意見を表わすために必要な文（叙述文）とに分類できる。文タイプの連鎖からどのように文間関係が決定できるかを調査するため、朝日新聞社説記事70編より文間関係が存在すると考えられる2文を抽出し、2145件の文間関係データを収集した[8]。そして、各文間関係データの2文に文タイプ情報を付与することにより、文タイプの連鎖と文間関係との関連を分析した。

文タイプ情報としては、朝日新聞社説記事の各文末表現を調査することにより、主張文について [問掛文, 断定文, 推量文, 要望文, 判断文, 意見文, 理由文, 義務文] の8つに、叙述文について [現在, 過去, 可能, 伝聞, 様態, 叙述, 存在, 継続, 状態, 使役] の10に再分類した[8]。以下にそれぞれの文末表現の例を示す。

○主張文

・問掛文	のではあるまいか	どう～のか	
	恐れはないか	どう～だろうか	
	ありはしないか	あるのだろうか	
・断定文	である	ことである	のである
	明らかである	のではない	のだ
・推量文	だろう	ではあるまい	
	とはいえまい	べきだろう	
	かもしれない	やむをえまい	
・要望文	～たい	期待する	～したい
	てもらう	てもらいたい	てほしい
・判断文	はずだ	はずはない	と考える
	といえる	とは言いがたい	
	ているようだ		
・意見文	てはならない	必要がある	望ましい
	必要である	当然だ	大切だ
・理由文	ているからである	からだ	わけだ
	わけではない	ためだ	
・義務文	なければならない	べきだ	
	ねばならない	べきである	

○叙述文

・現在	「る」で終わる形		
・過去	「た」で終わる形		
・可能	できない	いえない	
・伝聞	と聞く	という	
・様態	強い	厳しい	多い
・叙述	名詞＋だ		
・存在	がある	はない	にある
・継続	つつある		
・状態	ている	てくる	てしまう
・使役	させる	せる	

文間関係としては、永野[4]や市川[9]による分類がある。我々は、これをもとに図1に示す20種類の文間関係を設定した。

図1 文間関係名

文間関係データの分析から、主張文－主張文、主張文－叙述文の連鎖については、文タイプの連鎖によりいくつかの文間関係を決定することができたが、叙述文－叙述文、叙述文－主張文の連鎖に関しては、文タイプ情報のみでは明確な関係を決定することができなかった。そこで、これらについては、さらに主題提示情報を用いることによりいくつかの関係を決定できた。

2.2 照応現象の調査

文章中の指示語及び同一名詞の反復からどのような文間の関係が決定できるかを調査するため、文間関係の調査に用いた文章を用いて文章中の照応現象の調査を行った。この調査では、照応語とそれが指示参照するものである先行詞とこれらの間の照応関係及びその文間の距離を照応現象データとして、3003件を収集した[8]。

照応現象データの分析から以下のことが明らかになった。まず、指示語の照応に関しては約8割がその前文を指しているため、指示語が用いられている文は、その前文の内容の一部を指し示すことにより話題が展開されていると考えられる。また、文間関係データとの関連を調べると、指示語については累加、背景関係が比較的多く現われ、転換関係が少ないという傾向があった。これからも、指示語が用いられた場合、その指示語の参照している文に関しては、前文から引き続いた内容が述べられる傾向があることがわかる。

同一名詞については、並列、転換関係が多く現われる傾向があった。並列関係は同じ内容が並列的に述べられているという関係であり、このような場合には同じ語句が用いられる傾向があることが分かる。しかし、同じ名詞が使われた場合でも転換関係が多く現われているため、同一名詞句の反復情報のみでは単純に話題の転換と判定することはできず、その他の情報が必要であると考えられる。

3 文章構造のモデル化

3.1 文章構造

書き手の主張の観点から文章を構造化するためには、書き手の主張を表わす文とその前後の文がどのような関係で結ばれているかを捉える必要がある。また、そのような関係によって結ばれた文のまとまり間の関係も捉える必要がある。この関係を一方が主で他方が従である2項関係と捉えることで文章全体を木構造として構造化できる。しかし、実際の文章中には、主従を決定しにくいものがいくつか存在する。図2に示す例のように、文章中である事柄について述べられた文が順に列挙される場合、これらの文間に主従の関係はないと考えられる。

図2 文脈現象の例 (1)

また、図3に示す例のように、「第1に」「第2に」「第3に」のような表現を用いて、いくつかの文を序列的に並べるものもある。これらの文間にも主従の関係はないと考えられる。

図3 文脈現象の例 (2)

なお、図2、図3において数字はパラグラフ番号－文番号を示す。

このような文間の関係を主従の関係のないn項関係と捉えることにより、これらの文を1つのまとまりとして表現できる。また、主従の関係のあるものについても、同じ話題について述べられているものを1つのまとまりとして表現できる。このようないくつかの文をまとめる操作をグループ化と呼ぶことにする。文章中の各文間の関係のうち主従の関係があるものについては同じ話題について述べられているものをグループ化し、また、主従の関係がないものについてはそれらをグループ化することにより得られたものを文章の構造とする。

3.2 文章構造モデル

文章構造をモデル化するため「。」で区切られた1文を1つのノードとして表わし、これらの文間に主従の関係が認められる時、その関係をノード間のアークで表わす。このとき、アークには従となるノードから主となるノードの方向へ矢印をつける。また、いくつかのノードをまとめてグループ化したものも1つ

のノードとして扱う。本文章構造モデルで扱うノードには、文を表わすSノード (sentence node) とグループを表わすGノード (group node) の2種類がある。各ノードの性質はノードの属性として表わす。Gノードには、以下の3つのタイプがある。

○連続タイプ

ある事柄が連続して述べられているものをグループとして表現したノード

○並列タイプ

並列的な内容や序列的に述べられているものをグループとして表現したノード

○スコープタイプ

文間関係が存在するノードのうち同じ話題に関するものをまとめてグループとして表現したノード

Gノードの各タイプを表現したものを図4に示す。図中、ノードは「○」で、アークは「←」で表わされている。また、連続タイプでは、ノードは事柄が連続している順に「⇢」で結ばれて表わされている。

図4 Gノードの種類

文章構造は、ノード (Sノード, Gノード) 間にアークを張ったり、ノードをグループ化することにより表現される。アークを張るための条件は、各ノードに付与された属性やそのノードから張られているアークの情報により、また、Gノードの場合にはその要素ノードの属性によっても決まる。

3.3 記述言語

文脈現象の分析結果を文章構造解析のための知識とみなし、これを規則として記述するため、前節で述べた文章構造モデルに基づく記述言語を定義した。

本言語で記述される規則は、以下のようにルール名、ルールグループ名、パターン部、条件部、実行部から構成されている。

<ルール名>:<ルールグループ名>{
<パターン部> <条件部> <実行部> }

規則は、ルール名の後にそのルールグループ名を付与すること

により、いくつかをまとめてグループとしてモジュール化できる。パターン部では、文章構造のノード及びアークに関する条件を記述する。条件部では、ファクト形式の情報とのパターンマッチ及び任意の述語を記述することができる。パターン部及び条件部に記述された全ての条件が満たされた場合その規則は適用され、実行部の記述にしたがって、ノード、アーク及びファクト形式の情報の追加・修正・削除を行う。これによって文章の構造解析を行う。

本言語仕様に基づいて記述した文章構造解析規則の記述例を図5に示す。この規則のパターン部では、連続する2つのノードがそれぞれ主張文、叙述文であり、叙述文に主題提示があり、それらの間にアークがないことを表わす。条件部では、叙述文の主題提示語が主張文に現われないことを示す。実行部では、それらの間に転換関係のアークを張ることを表わす。

図5 文章構造解析規則記述例

4 文章構造解析システム

文章構造解析システムのシステム構成を図6に示す。本システムでは、まず、文解析部において入力された文章の形態素解析を行い、その結果から構造解析に必要な情報を抽出し、ワーキングメモリにセットする。次に、トランスレータにより文章構造解析規則を変換し、ルールメモリにセットする。文章構造解析部では、ルールメモリにセットされた文章構造解析規則とワーキングメモリにセットされた文解析情報とのマッチングを繰り返し、その内容を書き換えていくことにより文章の構造解析を行う。

図6 システム構成

4.1 文解析部

文解析部では、文章中の各文についての形態素解析結果をもとに、文章の構造解析のための入力となる情報として、主題提

示語、文タイプ、接続語句、指示語、及び同語反復の各情報を抽出する。

1) 主題提示語情報抽出

主題提示語情報の抽出処理においては、主題提示語として、助詞「は」「も」が付属する語句の抽出を行う。一般に、主題提示語は、自立語に助詞「は」等が付属する形を持つが、この形の中には、「特には」「～とはいえない」のように主題提示の機能を持たないものも存在する。そこで、新聞社説記事を用いた調査から、以下のものを主題提示語として抽出した（「も」も同様）。

名詞+は	名詞+においては	名詞+では
名詞+とは	名詞+にとっては	名詞+としては
名詞+には	名詞+については	
名詞相当語+は	名詞相当語+には	

2) 文タイプ情報抽出

文タイプ情報の抽出処理においては、文章中の各文の文末表現を解析することにより、主張文については[問掛文, 断定文, 推量文, 要望文, 判断文, 意見文, 理由文, 義務文] の8種類の文タイプを、叙述文については[現在, 過去, 可能, 伝聞, 様態, 叙述, 存在, 継続, 状態, 使役]の10の文タイプを抽出する。

3) 接続語句情報抽出

接続語句情報の抽出処理においては、文中で接続詞等の文間の関係を決定する語句、及び「第1に」「1つは」のように文章の流れを決定する機能を持つ語句を抽出する。接続詞情報としては、品詞が接続詞であるものをとり、接続的な機能を持つ語句は、新聞社説記事の分析で得られた結果をもとに決定した。

4) 指示語情報抽出

指示語情報の抽出処理においては、文中の指示代名詞及び指示連体詞を含む文節を抽出する。但し、人称代名詞の1人称、2人称のものや、「その半面」、「このところ」等の直接指示参照するものがないものは抽出しない。後者の抽出しないものは、新聞社説記事の分析で得られた結果をもとに決定した。

5) 同語反復情報抽出

同語反復情報の抽出処理においては、文章中で繰り返し用いられている名詞語句を抽出する。また、反復語に対して修飾語

が係っているかどうかの情報も抽出する。

4.2 文章構造解析部

文章構造解析部では、文解析部で抽出された情報と文章構造解析規則との適用を繰り返しながら文章構造の解析を行う。この解析では、まず、文章中の隣接文間の関係の解析を行い、次に、解析された文間関係を基にノードのグループ化を行い、文章全体をいくつかのGノードとして構造化する。最後に、文章を構成するために必要な要素を生成されたGノードより認識し、それぞれの要素として認識されたGノードをさらにグループ化し、文章全体の構造を得る。

1) 隣接文間解析処理

隣接文間の関係は、主として新聞社説記事の解析結果から得られた文のタイプ間関係によって決定する。但し、文間に接続詞情報が存在した場合、また、指示語情報や主題提示情報などを用いて記述された規則に適用した場合、そちらを優先して文間関係を決定する。文間関係としては、文脈現象の調査で用いた20の関係を分類・整理し、次の17の関係を設定した。これらの関係は、ノードの主従の順序と結び付きの強さにより以下の3つのクラスに分類されている。

○クラス1（前文が主で結び付きの強い関係）

例示・補足・呼応

○クラス2（後文が主で結び付きの強い関係）

背景・前提・根拠・結果

○クラス3（結び付きの弱い関係）

逆接・転換・序列・累加・反復

並列・対比・継続・展開・連係

2) ノードグループ化処理

ノードグループ化処理では、隣接文間解析処理で得た文間関係を基に、まず、継続、連係関係で結ばれたノードを連続タイプのGノードとしてグループ化する。次に、結び付きの強い文間関係で結ばれたノードをスコープタイプのGノードとしてグループ化する。そして、以上の処理でグループ化されなかったノードについては、文間関係に強度を設定し、強度の大きい関係で結ばれているノードを先にグループ化する。連続する3つのノード間の関係をそれぞれR1, R2とし、それらの強度関係

を以下に示す。（R1 > R2 のとき、R1が強い強度を持つとする。）

(a) 連続する3つのノードが主張文の場合

クラス1の関係 > クラス2の関係 > クラス3の関係

(b) 連続する3つのノードが順に主張文, 叙述文, 叙述文の場合

クラス1の関係 > クラス2の関係 > クラス3の関係

(c) それ以外の場合

展開 > クラス1の関係 > クラス2の関係 > クラス3の関係

また、グループ化により生成されたGノード間の関係についても、その主ノードの情報等を用いることでグループ間関係の解析処理を行い、さらに上位のGノードとしてまとめる。

3) 文章構造認識処理

文章構造認識処理においては、文章を構成するために必要な要素となるGノードを調べ、構成要素として認識されたノードを1つのGノードとしてまとめたうえで文章構造の要素名を認識する。現在、文章全体の構造の要素として認識しているものには、[事例紹介, 問題提起, 序論, 本論, 結論] がある。構造解析された文章構造の概略を図7に示す。

図7 文章構造の概略

5 文章構造解析の評価

本システムによる文章構造解析結果の評価のため、約400の文章構造解析規則を用いて朝日新聞社説記事13編の構造解析結果の評価を行った。評価は、

・ノード間の関係の妥当性

・生成されたGノードの要素のノードの妥当性

・文章構造要素名の妥当性

の3項目について○△× (各2, 1, 0点) の3段階で行い、それらの合計値をすべてが○であった場合に100となるように規格化したものを文章構造解析の評価値とした。1名の評価者により行った評価の結果を表1に示す。

表1 文章構造解析の評価結果

新聞社説記事を用いた評価結果のうち、いくつかのグループ化処理が正しく行われないという問題点があった。これは、本システムで主題提示語として扱うものが助詞「は」の付与する名詞語句のみであるため、これだけでは文章中の話題の流れを十分に扱えないためであると考えられる。また、Gノードが生成された場合、先頭ノードの主題提示情報と中心となるノードの文タイプ情報をGノードに付与しているが、これについても検討する必要がある。

6 あとがき

本稿では、新聞社説記事中の文脈現象の分析から得られた知識を基に、文章の構造化のためのモデル及びそのモデルに基づく言語仕様について述べた。そして、これを用いて記述した規則を適用することにより文章の構造解析を行うシステムおよびそのシステムを用いた文章構造解析の評価結果について述べた。

本システムでは、文の表層情報として現われた書き手の主張を述べた文を中心として文章の構造解析を行っている。そのため、新聞社説記事等の論説などのタイプの文章の構造化については約7割という評価結果がえられた。しかし、一般の新聞記事のように単に事実が羅列してあるタイプの文章の構造化は困難である。このような問題を解決するためには、文章中で用いられている言葉の連鎖情報や表層情報だけでなく意味的な情報も用いる必要があると考えられる。

本文脈処理技術の応用としては、文章の構造情報を用いた文章の要約処理の実現や文章における省略部分の補完による機械翻訳システムの訳質の向上が考えられる。要約処理の実現のためには、得られた文章構造からどこを重要部分として抽出すべきかを検討する必要がある。また、省略語の補完のためには、どのような文間関係の場合に省略現象が起こるかの関連を調査する必要がある。

本研究は第5世代コンピュータプロジェクトの一環としてICOTからの委託で行われたものである。研究の機会を与えて頂いたICOT淵所長、内田部長、有益な助言を頂いたICOT第6研究室田中室長、及びICOT自然言語理解ワーキンググループの

皆様に感謝します。また、研究協力頂いた(株)沖テクノシステムズラボラトリの皆様にも感謝します。

参考文献・引用文献

(1) Schank, R.C. and Abelson, R.P. : "Scripts, Plan, Goals and Understanding", John Wiley and Sons, 1977.

(2) Grosz,B.J. : "Attention, Intention, and the structures of Discourse", Computational Linguistics, Vol.12, No.3, pp.175-204, 1986.

(3) Sidner, C.L. : "Focusing in the Comprehension of Definite Anaphora", in Brady, M. and Berwick, R.C. (eds.), Computational Models of Discourse, MIT Press, pp.267-330, 1983.

(4) 永野 賢 : "文章論総説－文法論的考察－",朝倉書店, 1986.

(5) 福本 淳一 : "筆者の主張に基づく日本語文章の構造化", 情報処理学会自然言語処理研究会 78-15, pp.113-120, 1990.

(6) 福本, 安原 : "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 88-2, pp.9-16, 1992.

(7) 山梨 正明 : "発話行為",大修館書店, 1986.

(8) 福本, 安原 : "日本語文章の構造化解析", 情報処理学会自然言語処理研究会 85-11, pp.81-88, 1991.

(9) 市川 孝 : "国語教育のための文章論概説",教育出版, 1978.

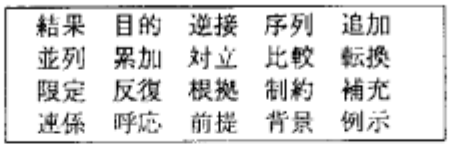

図1 文間関係名

Figure 1 Relationships between Sentences

1-1 ジョギング中の突然死や社長の急死などの不幸な事件が、このところ目立っている。
1-2 産業構造が変わり技術革新が進んで、働く人のストレスもつのってきた。
1-3 高齢化への歩みが速まるなかで、働き盛りの中高年の健康管理が特に重要な問題になっている。
（昭和62年10月 2日付け朝日新聞社説記事より）

図2 文脈現象の例 (1)

Figure 2 An Example of Context Phenomena (1)

5-1 第1に、水の節約が、水道料金の値上げにつながらぬようにしなければならない。
8-1 第2に、給水配管の漏水対策を積極的に進めてほしい。
9-1 第3に、最も重要なのは、今回の水不足が、東京集中によってひき起こされた構造的な問題だ、という認識を持つことである。

（昭和62年 8月30日付け朝日新聞社説記事より）

図3 文脈現象の例 (2)
Figure 3 An Example of Context Phenomena (2)

(a) 連続タイプ　　(b) 並列タイプ　　(c) スコープタイプ

図4 Gノードのタイプ
Figure 4 Three Types of G-nodes

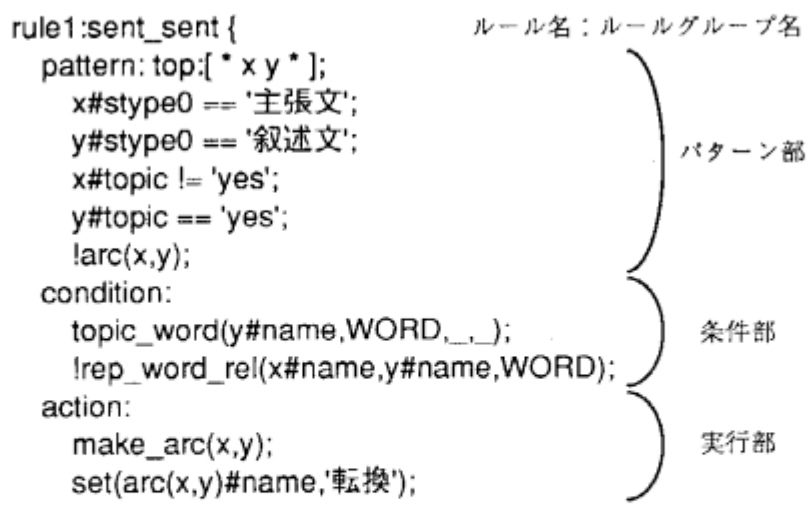

図5 文章構造解析規則の記述例

Figure 5 : An Example of Text Structure Analysis Rules

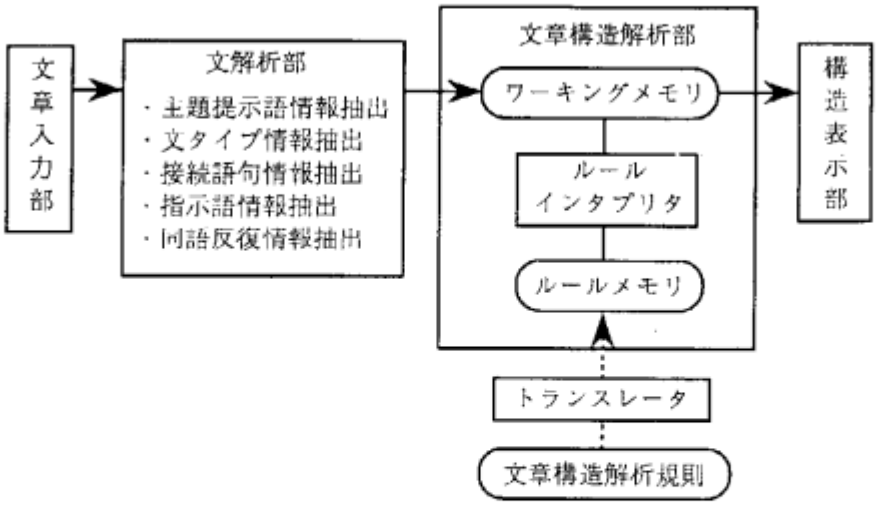

図6 システム構成
Figure 6 An Overview of the System

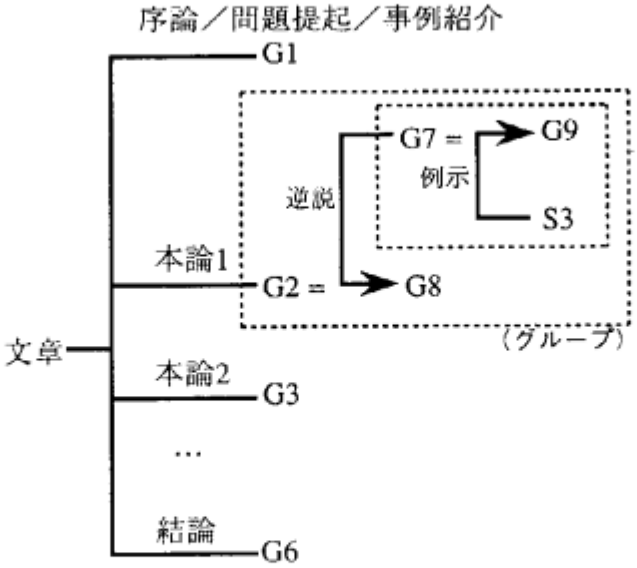

図 7　文章構造の概略
Figure 7 : An Overview of Text Structure

文章名	評価値
社説 ('87/10/1 No.1)	81.0
社説 ('87/10/2 No.1)	76.1
社説 ('87/10/2 No.2)	83.3
社説 ('87/10/4 No.1)	81.5
社説 ('87/10/5 No.1)	80.8
社説 ('87/10/7 No.2)	68.3
社説 ('87/10/8 No.1)	73.0
社説 ('87/10/10 No.2)	71.2
社説 ('87/10/11 No.1)	66.4
社説 ('87/10/12 No.1)	74.5
社説 ('87/10/13 No.2)	63.3
社説 ('87/10/16 No.2)	65.6
社説 ('87/10/30 No.1)	79.8
平均値	74.2

表1 文章構造解析の評価結果

Table 1 Evaluation of Text Structures Analysis