自然言語処理 67－5
（1988. 7. 22）

意味連結パターンを用いた係り受け解析

稲垣博人、壁谷喜義、小橋史彦

ＮＴＴヒューマンインタフェース研究所

本報告では、自然言語処理の基本技術である係り受け解析処理の高精度化手法について述べる。従来の係り受け解析処理が、高い解析精度を得るために、対象分野を限定しているのに対し、本解析手法では、対象文章の文章構造情報に着目した意味連結パターンを活用しているため、世界知識やフレームなどの分野に依存した知識を必要とせずに高い解析精度を得ることが可能となる。最短距離の係り受け候補を正解とする処理が８４％程度の正解率しか得られないのに対し、本係り受け解析手法を用いることにより正解率として９７％の高い解析精度を得ることができた。

Modification Analysis using Semantic Pattern

Hirohito INAGAKI, Kiyoshi KABEYA,and Fumihiko OBASHI

NTT Human Interface Laboratories

1-2356 Take, Yokosuka-Shi, Kanagawa, 238-03, Japan

We propose new modification analysis method which decreases ambiguity using modification semantic pattern of target sentence. The semantic pattern is composed of semantic relation of modifier and modificand and is extracted from target sentence, not given externally, so that we can obtain high modification accuracy to any fields of document.

Compared to the primitive modification method which modificand is decided from the point of view that modifier often depends nearest modificand, using our modification method, we can obtain high modification accuracy of 97%.

1．　はじめに

オフィスにおいては、必要な情報を迅速かつ的確に収集することが要求されている。特に、情報ソースとして文書を対象とする場合、文書の探索支援として、索引抽出、要約・抄録、文書分類などが必要となる。しかしながら、これらの文書探索支援技術は、ほとんど人手に頼らざるを得ないのが現状であり、膨大な文書処理を可能とする機械処理化が強く望まれている。

文書の自動索引抽出、要約・抄録生成などに関する研究は、数々行われている［1，2，3，4］が、質の高い索引や要約などを実現するには、文章構造を把握するめの基本的な言語処理技術、つまり、係り受け解析を高い精度で行なうことが必須条件となってくる。

プリミティブな係り受け解析手法として、各文節の品詞属性から係り受け候補を決定し、係りに最も近い候補を受けとする方法がある。この手法は、係りに最も近い候補が受けとなる確率が高いという日本語の特性を利用しており、処理は簡略化できるものの、精度が十分に上がらないという問題がある。この手法の欠点を補うべく、世界知識［5］や格情報［6］を用いることにより係り受けのあいまいさを解消しようという試みがなされている。前者の手法は、特定分野の上位－下位関係、全体－部分の関係などを知識として登録し、世界知識の条件を満たす格関係候補を正解とする処理を行なっている。また、後者の手法では、意味情報を付与した動詞の格関係の文型表を用い、文型表の中で結合が許される格関係のみ正解とする処理を行い、係り受けを決定する。しかし、どちらの手法も構築に膨大な労力を必要とする世界知識や格辞書などを用いているため、広範囲な分野の文章に対してこれらの手法を適用することは難しい。

本係り受け解析手法は、文章中に頻出する言い替え表現、繰り返し表現、複合語表現などに着目し、これらの表現が持つ意味関係を意味連結パターンとして抽出し、その意味連結パターンにより、世界知識を用いずに、曖昧な係り受け関係を一義に決定する手法である。

2．　意味連結パターンを用いた係り受け解析

本係り受け解析では、文章中の一義に決定される係りと受けの意味的な連結関係を意味連結パターンとして登録し、この意味連結パターンを用いて曖昧な係り受け関係を一義に決定することを特徴とする。このため、分野に依存した大規模な世界知識を用いずに、係り受けの曖昧さを解消することが可能となる。

係り受け解析処理の流れを図１に示す。形態素解析、基本的係り受け解析、並列構文解析、読点文節解析、意味連結パターン解析の５つの処理に分かれており、以下で個々の処理について述べる。対象文章としては特許請求範囲文を使用した。

2．1　形態素解析

漢字かな混じり文を文節単位、単語単位に分割し、各単語の品詞、用言なら活用形を付与する。同時に、各単語にその単語の意味を表わす意味カテゴリ番号［7］を付与する。

本形態素解析では、文節数最小法により候補を絞り、さらに、文節内単語数最小法、単語の出現頻度評価、連語関係評価を適用することにより、最尤候補を選出している［8］。

自立語の辞書規模は約８万語であり、解析精度としては、一般文書（新聞記事、社説、特許請求範囲文など）

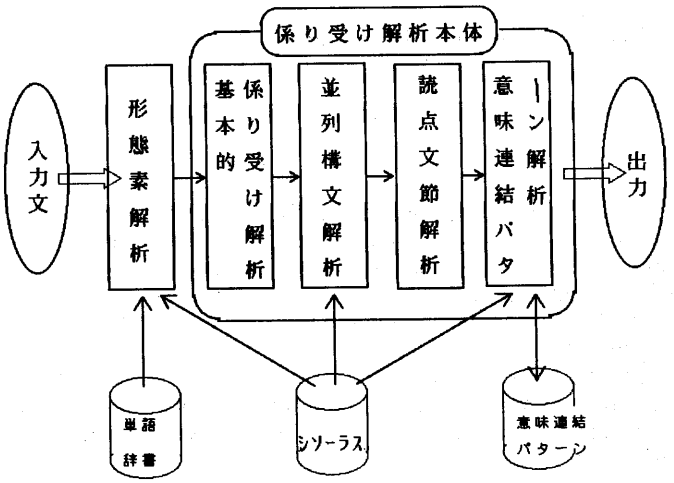


図１．　係り受け解析処理の流れ図

に対して、約９８％の精度で正しい形態素情報（単語切り出し、読み、品詞、活用形）を得ることができた。以降の係り受け解析処理では形態素解析結果の正しいもののみ対象とした。

２．２　基本的係り受け解析

形態素解析の結果を用いて、以下に示す（１）～（７）までの手順に従って基本的係り受け解析を行う。

（１）文節を構成する単語から係りおよび受け文節の属性を求める。

（２）表１の係り文節と受け文節の係り受け関係表より、係り受けの可能性があるか、ないかの判断を行ない、係り受け候補を選出する。係り受け関係の曖昧さを減少させるため、受け文節の属性として、名詞、動詞、形容（動）詞の３分類にわけると共に、用言については活用形も考慮した。

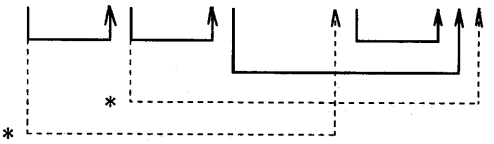
（３）係り受けが他の候補と交叉する候補を排除する。

（４）係り受けにおいて、格が重複している場合、最も用言に近い候補のみ係るとする。

（５）複数の係り受け候補が存在する場合、句読点内の候補を優先的に処理する。

例１）句読点内候補優先処理例

光線の／反射光を／受光して、／運搬車を／制御する

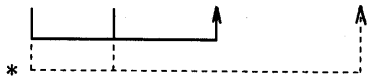


（＊は、不適切な係り受けを示す。）

（６）係り受けの範囲を限定する表現（「場合」、「とき」など）が存在する場合、係り受け範囲限定表現までの係り受け関係を優先処理する。

例２）係り受け限定表現の例

Ｃ値を／１２に／設定したとき／／出力する．



（７）後置表現（「～に対し」など）内の文節の係り受けは一義に決定する。

表１．　係り受け関係表

係り文節＼受け文節		名詞	動詞終止形	形容(動)詞終止形	動詞連体形	形容(動)詞連体形	動詞連用形	形容(動)詞連用形	動詞仮定形	形容(動)詞仮定形
格助詞	が,に,で,より		○	○	○	○	○	○	○	○
	の	○			○	○				
	と	○	○		○		○		○	
	を、へ、から		○		○		○		○	
係助詞	は		○	○			○	○		
	も		○	○	○		○		○	
副助詞	か,くらい		○	○	○	○	○	○	○	○
	のみ		○	○			○	○	○	○
用言	動詞(連体)	○								
	形容詞(連体)	○								
	動詞(連用)		○		○		○		○	
	形容詞(連用)		○	○	○		○	○	○	○
その他	連体詞	○								
	副詞		○	○	○	○	○	○	○	○

注）主な係り受け関係のみを示した。

この基本的係り受け解析の解析結果から、ベースとなる意味連結パターンを抽出する。詳細については、２．５節で述べる。

２．３　並列構文解析

並列構文は曖昧性が高く、解析が難しいものの１つであるが、文章の構造を決定する上で最も重要な位置をしめるため、正確な解析が必要となる。

基本的な並列構文の分析は、首藤ら［９］や長尾ら［１０］によって行なわれており、並列構文を決定付ける重要な手がかりを得ているが、決定的な並列構文解析手法を得るには至っていない。

並列構文解析では、並列構文を名詞句の並列構文と述部の並列構文とに分けて処理を行い、並列要素（並列構文を構成している文節群）、並列範囲（並列を構成する各名詞句または述部全体）を決定した後、係り受け適正化処理を行う。

２．３．１　名詞句の並列構文解析

名詞句の並列構文解析手法としては、

①表層的並列構文解析法［１１］

②並列要素間の意味類似度解析法［１２］

の２つがある。ここでは、両者を併用した処理により名詞句の並列構文解析を行う。

表２は、特許請求範囲文に出現する名詞句の並列構文を分類したものである。解析においては、２項型の並列構文を基本文型とした。多項型並列構文は２項型の特殊文型としてとらえ、２項型並列構文の解析を完了した後、処理を実行する。

例３）２項型並列構文例（括弧は並列範囲を示す。）

（（電流と）（電圧とを））測定する

例４）多項型並列構文例

（（本体内に．．．　テレビと、）

（本体前面に．．　鍵盤と、）

（本体内部に．．　演算装置とを））備えたパソコン。

以下に具体的な処理手順を示す。

（１）表層的処理により抽出できる並列情報、つまり、並列接続詞および並列接続詞の前に位置する単語（並列要素）の抽出を行う。

（２）最後尾の並列要素の候補（名詞）を選出する。

（３）（１）で抽出した並列要素と（２）で選出した候補との意味類似度を算出する。この場合、単語の文字列の類似度だけでなく、意味カテゴリ番号を用いて、意味類似度を計算する。意味カテゴリ番号は、３桁の数値からなり、１桁目が大分類、２桁目が中分類、３桁目が小分類を示すため、上位の桁ほど分類が大ざっぱとなり、意味的類似性が弱くなる。

つまり、意味類似度は、以下のように与えることができる。

（単語の文字列が一致）

＞（意味カテゴリ番号の一致）

＞（意味カテゴリ番号の上位２桁一致）

＞（意味カテゴリ番号の上位１桁一致）

（４）以上のようにして求めた意味類似度を基にして最も意味類似度が高い候補を最尤候補とする処理を行う。但し、最尤候補が多数存在する場合には、最も並列接続詞に近い候補を最尤候補とする。

（５）並列の範囲は、２項型と多項型では、多少異なり、多項型並列構文では、最後尾の並列範囲は２項型と同じく並列接続詞から最後尾の並列要素までとし、他の並列範囲は、句読点から接続詞までとする。

例５をもちいて、意味カテゴリを用いた並列構文の解析例を示す。

例５）

[1]下限 予測 信号 と／[2]基準 信号 に／[3]貯水 量 信号 を
194 411 820 191 820 395 123 820

I

／[4]加えた／[5]上限 予測 信号 を／[6]求め、
262 194 411 820 742

II

表２．　名詞句の並列構文

２項型並列構文		
並列構文	出現数	出現比率
NP と（、） NP	46	32%
NP および NP	27	19%
NP，（or 、 ・） NP	12	8%
NP または NP	5	3%
NP も NP も	3	2%
NP もしくは NP	2	2%
NP．あるいは NP	2	2%
NP （、）かつ NP	1	1%
多項型並列構文		
並列構文	出現数	出現比率
NPと、NPと、・・・ NP	37	25%
NP、NP、 ・・・NP（、）およびNP	8	6%
NP、NP、 ・・・ NP	3	2%
合計	146	100%

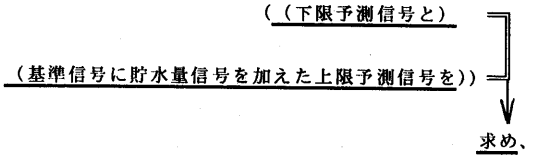
（対象データ：特許請求範囲文４０篇）

（左上の数値は文節番号を示し、２重線は並列関係を示す。）

例５で、第１文節に並列接続詞「と」があるため、並列構文解析ルーチンが起動される。まず、並列接続詞「と」を抽出すると共に、並列要素「下限予測信号」を抽出する。次に（２）の処理で並列要素の候補、第２文節「基準信号」と第５文節「上限予測信号」を抽出する。（３）の処理で、第２文節と第５文節におけるそれぞれの意味類似度を算出する。この場合、第５文節の方が第２文節に比べ、意味類似度が高いので第１文節と第５文節が並列要素であると決定する。

後半の並列範囲は第２文節から第５文節までの「基準信号に貯水量信号を加えた上限予測信号」となる。前半は並列要素だけなので「下限予測信号」が並列範囲である。この並列範囲で示された２つの名詞句が用言「求め、」の並列目的語となっている。

つまり、以下の解析結果が得られる。



２．３．２　述部の並列構文

述部の並列構文は、連用中止法による並列構文と並列接続詞（「、かつ」など）を介した並列構文との２種類に分類できる。

表３に特許請求範囲文で用いられている述部の並列構文を示す。約７割以上が連用中止法であり、残りの３割は、並列接続詞を介した並列構文であることがわかる。

表３．　述部の並列構文

並列構文		出現数	出現比率
[用言・連用形]	[文]	８６	７５％
[用言・終止形]　と共に、	[文]	１３	１１％
[用言・連用形]　、かつ、	[文]	１２	１１％
[用言・連用形]　、しかも、	[文]	１	１％
[用言・連用形]　、又、	[文]	１	１％
[用言・連用形]　、さらに、	[文]	１	１％
合計		１１４	１００％

（対象データ：特許請求範囲文４０篇）

述部の並列では、以下の手順で解析を行う。

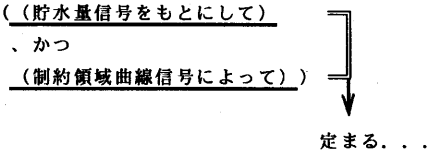
（１）連用中止法は、用言間の意味カテゴリの類似性、文型の類似性を用いて並列要素を決定する。連用中止法により複数個の並列が構成される場合、他動詞同志または、自動詞同志のみが並列要素であるとする。

（２）並列接続詞を介した並列構文には、短文を並列する「、かつ」型並列構文と長文を並列する「と共に」型並列構文がある。前者の場合、並列の接続詞に近い候補を並列要素とし、後者の場合、並列接続詞に最も遠い候補を並列要素とする。

「、かつ」型並列構文の並列接続詞には「、かつ」以外に、「、または」があり、「と共に」型並列構文の並列接続詞には「と共に」以外に「さらに」、「しかも」などがある。

（３）前半の並列範囲は、句読点から並列の接続詞までとし、後半の並列範囲は、並列の接続詞から最後の並列要素までとする。

例６）「、かつ」型並列構文例

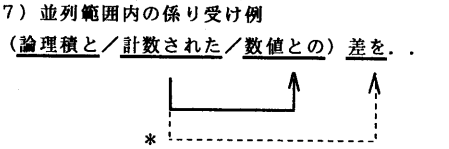


２．３．３　係り受け適正化処理

係り受け適正化処理では、並列構文により決定された文章構造に着目して、不適切な係り受け関係を排除する。処理は（１）、（２）からなっている。

（１）係り受けが並列範囲内に位置する場合、並列範囲内の係り受けを優先的に処理する。

例７）並列範囲内の係り受け例



（２）受け候補が並列範囲内にある場合、並列要素全体が受け候補であるとする。

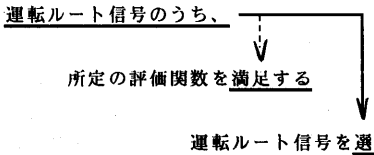
２．４　読点文節解析

読点文節の解析では、並列構文解析で解析できなかった読点文節の係り受け解析を行う。解析手順は以下の通り。

（１）受け文節候補の選出。

（２）読点は、受け文節が遠くにあることを明示することが多いので、受け文節候補の中で、係りから最も遠い候補を正解とする。

例８）読点文節解析例



この例では、「運転ルート信号のうち、」という文節の受け候補は、「満足する」と「選択し、」の２つであるが、読点文節処理により、「運転ルート信号のうち、」は係りから遠い文節「選択し、」に係ると判断する。

２．５　意味連結パターン解析

並列構文解析、読点文節解析をおこなっても曖昧さが残る文節の係り受け解析を意味連結パターンを用いて解析する。ここで意味連結パターンとは、係りと受けの関係を単語と単語の共起的関係として捕らえ、その単語間の共起を記述したものである。その場合、係りと受けの関係は、単語の意味の共起的関係であるため、単語の意味カテゴリの共起関係も意味連結パターンとして扱う。以下に意味連結パターン解析の手順を示す。

（１）係りと受けの関係が一義に決定される係り受け関係から意味連結パターンを抽出し、保存する。

例９の文から抽出した意味連結パターンを表４に示す。

例９）

走行路を複数のブロックに分割し、退出点で減速信号を発信する指令装置。

（２）先に抽出した意味連結パターンを用いて、曖昧な係り受け関係の判定を行う。

まず、すべての係り受け候補の意味連結パターンを作成する。次に、作成した意味連結パターンと、保存してある意味連結パターンとのマッチングを行い、係り受け候補の中で意味連結パターンとの一致度が最短候補に比べ、最も高い候補を正解とする。

同様な意味連結パターンが頻出する場合、その意味連結パターンを持つ文節列を複合表現的に扱う。例えば、「第１の誘導線」、「第２の誘導線」という表現が頻出する場合、これらの表現を複合的に扱い、係り受けの曖昧さを減少させる。

係り受け候補を一義に決定できた場合、その係り受け関係から抽出される意味連結パターンも順次保存する。

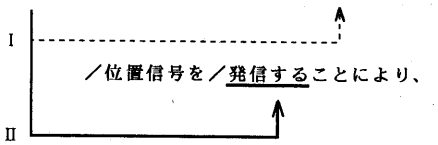
表４．　意味連結パターン抽出例

係り		受け		意味連結パターン	
係り文節	意味ｶﾃｺﾞﾘ番号	受け文節	意味ｶﾃｺﾞﾘ番号	文節構成単語	意味カテゴリ番号
走行路を	０４７	分割し	２６７	("走行路" ,"分割")	(０４７, ２６７)
				("分割" ,"走行路")	(２６７, ０４７)
複数の	１２８	ﾌﾞﾛｯｸに	７００	("複数" ,"ブロック")	(１２８, ７００)
				("ブロック","複数")	(７００, １２８)
ﾌﾞﾛｯｸに	７００	分割し	２６７	("ブロック","分割")	(７００, ２６７)
				("分割" ,"ブロック")	(２６７, ７００)
退出点で	３１３	発信する	７５６	("退出点" ,"発信")	(３１３, ７５６)
				("発信" ,"退出点")	(７５６, ３１３)
減速信号を	８２０	発信する	７５６	("減速信号","発信")	(８２０, ７５６)
				("発信" ,"減速信号")	(７５６, ８２０)
減速信号を	２６２	発信する	７５６		(２６２, ７５６)
					(７５６, ２６２)
発信する	７５６	指令装置	３８３	("発信" ,"指令装置")	(７５６, ３８３)
				("指令装置","発信")	(３８３, ７５６)
発信する	７５６	指令装置	４５２		(７５６, ４５２)
					(４５２, ７５６)

意味連結パターンを用いた係り受け解析例を次に示す。

例１０）

指令装置から／壁面への／接触を／検出する



例１０において、文節「指令装置から」の係り受けを解析した場合、係り受け候補が２つ存在する（Ⅰ、Ⅱ）。この場合、係り文節に最も近い候補を正解とすると、係り受けは誤りとなる。表４の意味連結パターンがすでに保存されていたとすると、本手法では、表４の意味連結パターンと係り受け候補（Ⅰ，Ⅱ）の意味連結パターンとのマッチングをとる。第１候補「検出する」は一致するパターンは存在せず、第２候補「発信する」は（"指令装置"，"発信"）のパターンと一致がとれ、第２候補が正解であると判断する。

また、この例では、単語の文字列同志により意味連結パターンとの一致がとれたが、"発信"（意味カテゴリ番号７５６）が"送信"（意味カテゴリ番号７５６）となっていても、意味連結パターンの意味カテゴリ番号が一致しているため、先の結果と同様に、係り受け関係があると判定できる。

（３）（２）の処理で決定できなかった係り受け関係については、候補として出現する頻度が最も多い候補を正解とする。

（４）最尤候補処理では、（３）で決定できない係り受け候補について最も係りに近い受け候補を正解とする処理を行う。

３．係り受け解析結果と評価

２章で述べた係り受け解析手法を、特許請求範囲文１０篇（公告番号　昭62-20561～昭62-20570）に適用した結果について述べる。各処理の平均正解率を図２に示す。正解率は以下のように定義する。

$$\text{正解率} = \frac{\text{係り受け関係を正しく一義に決定できた文節数}}{\text{対象文章の総文節数}}$$

（１）基本的係り受け解析では、全文節の４０％程度しか一義に決定できず、高精度の係り受け解析処理を用いて係り受けの曖昧さを減少させる必要があることがわかる。

（２）一義に決定できなかった係り受けを最短距離の候補に係るとして強制的に決定した場合の正解率を最短候補正解率とすると、基本的係り受け解析における最短候補正解率は８４％となり（図中、点線で示す。）、基本的係り受け解析だけでは、文章構造を十分把握できないことがわかる。

（３）並列構文解析および読点文節解析は、係り受け解析の正解率を大幅に向上させるわけではないが、基本的係り受け解析に比べ、これらの処理により係り受け候補数を半分近くに減少させる効果がある。

（４）最終的に係り受けの候補を絞る意味連結パターン解析では、係り受けの正解率として平均９３％を得ることができた。読点文節における正解率が平均５２％であることから、意味連結パターン解析により、文章中の全係り受け関係の半分近くを一義にかつ正しく解析していることになる。これは、文章の係り受け解析において、意味連結パターンが非常に有効な手段であることを示している。

（５）意味連結パターンによる係り受け解析実行後、最尤候補処理を用いて、係り受け関係を半強制的に決定しても、係り受け解析の正解率は向上し、最終的には平均正解率９７％を得た。

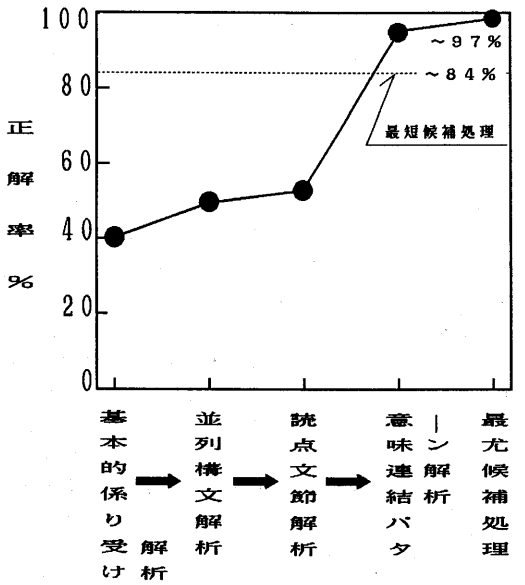
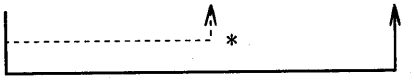


図２．　係り受け解析結果

（６）表５に本係り受け解析手法における誤り傾向を示す。誤りの多くは、並列構文解析部と意味連結パターン解析部で生じており、解析ルールの拡充および格情報の併用などによる対処法が有効であると考えられる。

例１１）係り受け誤り例

車両感知器を車両走行路に沿って、所定間隔で配設し、



（＊は係り受け誤り）

表５．　係り受け解析の誤り傾向

処理項目	誤り率
基本的係り受け解析	０．２％
並列構文解析	０．６％
読点文節解析	０．１％
意味連結パターン解析	１．７％
最尤候補処理	０．３％
合計	２．９％

４．　おわりに

本稿では、対象文章中の係り受け関係から意味連結パターンを抽出し、その意味連結パターンをもちいて、曖昧な係り受けの中で最も適切な係り受け候補を抽出する手法について報告した。特許請求範囲文を対象に、本係り受け解析手法を適用したところ、係り受けの正解率として、平均９７％を得た。最短候補を用いた係り受け解析処理が正解率８４％であるのに対し、本手法が係り受け関係の決定に非常に効果的であることを確認した。

今後は、現在の意味連結パターンに加え、格情報を併用した係り受け解析手法を構築し、対象文章を拡大すると共に、文章の索引抽出、抄録・要約などの応用分野への適用を考える。

謝辞　本研究の機会を与えて下さり、かつ熱心な討論をしていただいた当研究所の石川浩一郎主幹研究員、清水明宏研究主任、基礎研究所の内藤昭三主任研究員に感謝するとともに、実験を進めるにあたり多大な協力を頂いたＮＴＴ技術移転（株）の清末三恵子嬢に感謝します。

参考文献

［１］斉藤ほか：日本語文解析によるキーワード抽出，電子通信学会技術研究報告,vol.81,no.90,1981.

［２］北ほか：要約支援システムCOGITO，情報処理学会研究会資料,NL 58-7.

［３］鈴木ほか：高頻度隣接語を利用した科学技術文献の自動抄録，情報処理学会第３２回全国大会，4T-11.

［４］細野ほか：漢字の出現頻度情報を用いた日本語文献の自動分類，情報処理学会研究会資料，NL 47-7.

［５］高松ほか：技術抄録文からの関係情報の自動抽出，情報処理学会論文誌，vol.25,no.2,1984.

［６］絹川ほか：日本語文構造解析による自動インデクシング方式，情報処理学会論文誌，vol.21,no.3,1980.

［７］大野ほか：角川類語新辞典，角川書店，1981.

［８］本間ほか：連語解析を用いたべた書きかな漢字変換，情報処理学会論文誌，vol.27,no.11,1986.

［９］首藤ほか：日本語技術文における並列構造，情報処理学会論文誌，vol.27.no.2,1986.

［１０］長尾ほか：科学技術論文における並列句とその解析，情報処理学会研究会資料，NL 36-4.

［１１］押金：パターンを用いた日本語の名詞並列句解析，情報処理学会第３６回全国大会，3T-2.

［１２］田村ほか：意味解析に基づく並列名詞句の構造解析，情報処理学会研究会資料，NL 59-2.