パターン・マッチングによる重要語の自動抽出

細野公男，後藤智範（慶大・文），諸橋正幸（日本ＩＢＭ）

1.はじめに

自動索引の研究は，1950年代後半にIBMのLuhnによってはじめられて以来，今日まで様々な角度から研究されてきた．この研究は，情報検索の学問対象としてだけではなく，現在の科学技術分野の文献データベース作成に要する多大なコストの削減を可能にする，という実用的な側面からも，重要な研究課題として期待されている．

自動索引の目的は具体的には，

1．個々の文献に対する索引語の付与

2．各主題分野における索引語リストの作成

の２点に集中している．この２つの問題について，現在までに行なわれてきた研究のアプローチは概して，

1．数量的側面からの研究

2．言語処理的側面からの研究

に，大別される．どちらの側面からの研究も，実際の商用レベルでの大規模なデータベース作成において，実用化された例は極めて少ない．

上記の研究の前段階として，文章中から，重要語－free key word／phraseを抽出するという研究が，70年代に盛んに行なわれるようになってきた．この問題については，特に言語処理的側面からの研究が成果をあげている．その代表的な例が，70年代前半からDDC（Defence Documentation Center）で実用化されているMAI[1),2)]である．MAIは巨大な用語辞書と，key phraseを構成している単語の結合パターンを記述したパターン辞書の２つを効果的に利用している．重要語は，単語，句の区別なく処理できるが，２つの辞書の作成に多大なコストを要する，という欠点を持っている．

一方，MAIとは異なったアプローチをとったシステムで，MAIと同時期に開発されたLockheedのPHRASE[3),4)]がある．PHRASEは60年代に入って行なわれた自然言語処理の方向に沿ったシステムである．入力文に対し完全な統語解析（parsing）をすることによって重要語を抽出する．PHRASEはparsingだけで抽出するのでsyntactic ambiguityを避けることはできない．また統語解析を行なうために入力文中の単語の品詞をすべて知っている必要がある．従って，parseが不可能になり，重要語抽出に支障をきたすことが起こりうる．大量の文に対し，このような方法を適用することは現実的ではない．

科学技術文献に含まれる文が，言語処理上，日常使われる文とは異なった特色を持ち，また重要語がある種の特徴を持っていることは明らかである．しかしながら，両システムともそのような特徴を，重要語抽出のために有効に利用してはいない．本論文は，科学技術文献を構成している文，および重要語の持つ諸特性を明らかにし，それらを有効に活用した重要語抽出の方法を説明する．さらに，この方法を実現したプログラムを作成し，いくつかのデータに対して行なった実験結果について報告する．

2.パターン・マッチングによる重要語の抽出

2.1　アプローチ

2.1.1　科学技術文献の特徴

次の文は科学技術分野における代表的なデータベースであるINSPEC（Information Services in Physics, Electrotechnology, Computer and Control）の抄録に含まれている文である．例２－１は引用付で囲まれた句，例２－２は関数表現を含んでいる．

（例　２－１）

THE APPLICATION OF THIS REAL IMAGE AS AN 'ANALOGUE 3-D PICTURE STORAGE' TO DIGITAL IMAGE PROCESSING HAS BEEN SUCCESSFULLY APPLIED TO THE COMPUTER CONTROLLED ANALYSIS OF FAST MOVING PARTICLE.

（例　２－２）

SHOWS THE EXISTENCE OF AN OPTIMAL MESH T WHICH MINIMISES THE ERROR | U - U (T) | /SUB 1, OMEGA//SUP 2/, WHERE U STANDS FOR THE SOLUTION OF A SECOND ORDER CONTINUOUS ELIPTIC PROBLEM AND U (T) IS THE SOLUTION OF THE APPROXIMATE PROBLEM, SOLVED IN A FINITE ELEMENT SPACE V (T).

これらの例からわかるように，科学技術文献に含まれる文は，日常使われる文とは趣を異にしている．科学技術文献を構成している文には，次のような特徴がある．

1．複雑な構造を持つ文が多い．

2. 一般に，名詞句は長く，構造も複雑である．
3. 関数表現など，日常使われる文にはない句を含む．

従って，特定の主題分野に限定した，完全な統語解析を行なうことは，重要語を抽出するという目的からはあまり得策とはいえない．

2.1.2 重要語の諸特性

一方，科学技術文献に含まれる重要語には，次のような統語論的，語用論的特性が見られる．

1. 単語であるよりも句であるものが多く，それらの多くは名詞句である．
   （例 2－3）
   INITIAL VALUE PROBLEM
   LINEAR DIFFERENCIAL EQUATIONS
2. 前置詞（OF, BY, etc.），および接続詞（AND）を含んでいるものもある．
   （例 2－4）
   PLASMA DIAGNOSTICS BY LASER BEAM
   CONVERGENCE OF NUMERICAL METHODS
   SPEECH ANALYSIS AND PROCESSING
3. 重要語は，科学技術文献中に，頻繁に使用される次のような慣用表現の後に含まれることが多い．
   （例 2－5）
   THE PROBLEM OF－－－－－－
   THE BASIC METHOD OF－－－－－
4. また，次のように慣用的に頻繁に使用される文の目的語に含まれることも多い．
   （例 2－6）
   THIS PAPER DESCRIBES－－－－－．
   THE AUTHOR DISCUSSES－－－－－．

2.1.1，および 2.1.2にあげた特徴を持つことから重要語抽出には，統語論的アプローチ，あるいは統計的アプローチをとるよりも，以下に述べるパターン・マッチングによるアプローチが現実的である．

2.2 パターン・マッチングによる重要語の抽出

2.2.1 重要語抽出の方法

前節で述べたように，科学技術文献に含まれる文の構造，および名詞句の構造は複雑であるが，動詞句は日常使われる文と比較して，複雑な構造を持つものは少ない．動詞句の統語パターンを記述し，文中の動詞句を同定することは，困難なことではない．文の統語構造が複雑であっても，木構造的な方法をとらず，左から右に動詞句を照合し，照合した部分とその前の部分を分離すれば名詞句が抽出できる．

名詞句からの重要語の抽出は，前節で述べた慣用表現を，統語パターンとして記述することにより可能である．抽出された名詞句に対し慣用表現の統語パターンを照合し，照合した部分を除去すれば，重要語が抽出される．

従って，図2－1に示されるように，上記の過程を再帰的に行なうことにより，文の統語構造に左右されずに，重要語を順次抽出することが可能である．

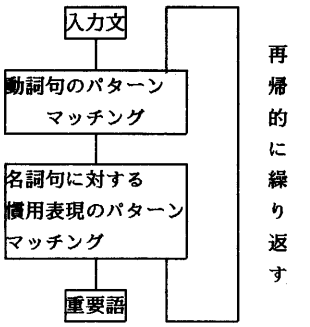

図2－1 重要語抽出の方法

2.2.2 ILIS

本研究は前節で述べた，パターン・マッチングによる重要語の抽出方法を実現するため，IBMで開発した ILIS（Interactive Language Implementation System）[5] というプログラミング言語を用いた．ILISは，自動バック・トラッキングの機能を持ったrule－drivenな言語で，単語をベースとしたパターン・マッチングの機能を持つ．また，stack automatonの動作が実現できるので，コンパイラーの作成，自然言語処理，および問題解決等の記述言語として使われている．

ILISの詳細な機能についてここでは触れないが，次にあげるような言語上の機能を持っている．

1. Pattern matching
   入力文を処理するために作成されたプログラム－規則（rule）－に従って，入力文中の語と，規則中の語との照合をする． さらに，各語に対し，内部ビット列を持ち，内部ビット列の照合をすることも可能である．
2. Action
   照合した後の様々な処理を，スタックを用いて行なう．
3. Application
   ILIS本体は，PL／1で書かれており，利用者がPL／1で書いたプログラムをILISの規則

の中で，サブルーチンとして呼び出すことができる．

2.2.3　抽出システムの構造

システムは，図２－１に示される処理を実現するために，５つの主要な規則から構成されている．図２－２にこれを示す．

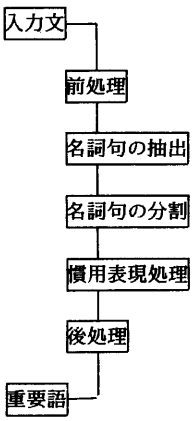

図２－２　抽出システムの構成

５つの主要な規則は，パターン・マッチングを効果敵に行なうための制御機構で，照合すべきパターンは多くの別の規則として記述される．これらの規則は，５つの主要な規則から，必要に応じて呼び出され，照合される．規則の数を少なくし，より一般化するために，入力stringを構成する単語には品詞が付与され，事実上，入力stringは品詞列と見なされる．

前処理では，入力STRING中の語は，辞書を用いて品詞が付与される．これより後の処理を効果的に行なうために，特殊文字，および熟語について処理がなされる．

２－２で述べた名詞句の抽出は，実際のシステムでは２段階で行なわれる．

名詞句抽出規則では，動詞句を中心としたパターン・マッチングが行なわれ，照合した動詞句の前の部分が抽出される．

この段階で抽出された名詞句は，複雑な構造を持つものがあるため，名詞句分割規則で前置詞を中心に，名詞句を分割する．

慣用表現処理規則は，分割された名詞句から慣用表現を取り除く．

この処理を経た段階で，重要語は抽出されているが，冠詞や 'other', 'each' 等の不要な語が，重要語の前後についている場合がある．後処理規則は，主にこれらの不要な語を除去する．

次の章で各規則の処理構造を説明する．

3.各規則における処理構造

3.1　前処理

前処理は次の３つの処理を含む．

1. 熟語の処理
2. 特殊記号の処理
3. 品詞の付与

3.1.1　熟語の処理

熟語の処理は次の２種類が考えられる．

1. 熟語そのものをパターンとして登録し，入力文中の語句と照合する．
2. 熟語を単語化し，品詞を付与し，単語と同様に処理する．

本システムは，規則の数を少なくし，汎用化するために，後者の方法を採用している．前処理規則では，入力文中の熟語は，それを構成する全単語を，次の例のように，単一の単語として扱う．

（例　３－１）

THE AUTHOR LOOKS AT THE DOUBLE－POLYSILICON APPROACH USED IN THIS DEVICE, AS WELL AS OTHER ASPECTS OF ITS FABRICATION.

↓

THE AUTHOR LOOKS AT THE DOUBLE－POLYSILICON APPROACH USED IN THIS DEVICE, ASWELLAS OTHER ASPECTS OF ITS FABRICATION.

3.1.2　特殊文字の処理

ILISでは，単語は，一般に空白文字で区切られた単位である．従って，'WITH：' は１つの単語であるが，'WITH ：' は，２つの単語である．この不都合を避けるため，英語の文に含まれる個々の特殊記号について，個別に処理する必要がある．

1.ハイフン　：　'－'
2.スラッシュ　：　'／'
3.カッコ　：　'（', '）'

以上の３つの特殊文字については，原文のとおりとする．カッコで囲まれた句は，カッコの前の語句についての説明である場合が多いため，本システムでは処理の対象としていない．

4.コロン　：　'：'
5.セミコロン　：　'；'

以上の２つの特殊文字については，原文で単語の

後に直接ついているものは，空白を挿入する．

（例３－２）

THIS PAPER DEALT WITH: LINEAR CIRCUIT---.

THIS PAPER DEALT WITH : LINEAR CIRCUIT---.

6.引用付　　：　　'''

引用付で囲まれた句も処理対象としているので，引用付は削除する．

（例３－３）

THE　'DATA　AVAILABLE'　SIGNAL－－－.

THE　DATA　AVAILABLE　SIGNAL－－－.

3.1.3　品詞の付与

本システムでは，図３－１に示されるように16種類の品詞を用いている．

PARTS OF SPEECH	SYM	POS
BE VERB	BEV	5
VERB	VER	6
NOUN	NOU	7
ADJECTIVE	ADJ	8
ADVERB	ADV	9
PAST PARTICIPLE	PAP	10
PREPOSITION	PRE	11
PRESENT PARTICIPLE	PRP	12
ARTICLE	ART	13
AUXILIARY VERB	AUX	14
CONJUNCTION	CON	15
NEGATIVE	NEG	16
DEMONSTRATIVE PRONOUN	DPR	17
INTERROGATIVE PRONOUN	IPR	18
RELATIVE PRONOUN	RPR	19

図３－１　品詞リスト

図中央の記号は，辞書の中で品詞を示す記号として各単語に付与されている．ILISは各パターンに対し内部に 32bit長のビット列を持っている．このうち第５ビットから第32ビットは，利用者が自由に使える領域となっている（本システムで使用したILISの版では，内部ビット列の長さは32であるが，最新の版は64ビットである）．

本システムでは，Thompsonらの REL system[6]と同様に，この内部ビット列を品詞の識別に利用している．通常，第５ビットから第32ビットは，すべて'０'が立ててあり，特定の位置に'１'を立てることによって品詞を表わす．図３－１の右側の数字は，内部ビット列中に'１'を立てる位置を示している．

ILISには辞書と呼ばれる機能がある．これは，uniqueな単語をキーとして，任意の情報を格納し，呼び出す機能で，本来の辞書的な利用のほか，複数の文の間の文脈の受渡し等をねらったものである．本システムでは，この機能を本来の辞書として用い，キーとなる単語に対し，品詞を受渡し情報とする．

図３－２は本システムに使われている辞書の一部である．システムは辞書を読み込み，入力文中の単語を辞書と照合する．照合すると，品詞記号に従って，内部ビット列に，図３－１に示された位置に'１'を立てる．入力文中の単語で，辞書の中にないものは第25ビットに'１'を立てる．このような語は，未登録語として名詞句分割プログラムまでは，名詞と同等に処理される．複数の品詞を持つ語は，内部ビット列に，複数の位置に'１'が立てられる．たとえば，'CONTROL'が名詞，動詞の２つの品詞を持っていると仮定すると，第５および第６ビットの位置に'１'が立てられる．

```
R    CHEAP            NOU.
R    CHECKING         NOU PRP.
R    CLASS            NOU.
R    CLASSICAL        ADJ.
R    CLASSIFICATION   NOU.
```

図３－２　辞書の一部

3.2　名詞句の処理

電電公社のLUTE－EJシステム[7]では，名詞句の構造を次のように把握している．

<NP> ::=（<NHD>）｛<NP>｜NOUN｝（<NMP>）

NP　：　名詞句

NHD　：　前方修飾句'premodification'

NMP　：　後方修飾句'postmodification'

さらに NHD，および NMPの構造を厳密に記述している．しかしながら，前章で述べたように科学技術文献は，引用付で囲まれた句，関数表現，さらにカッコで囲まれた句を含んでいるため，上の NHD，およびNMPについて，統語構造を厳密に記述することは，重要語抽出という目的には，効率的とはいえない．本システムでは，名詞句の構造を厳密に記述することは避け，次に説明する名詞句抽出規則を経て抽出された句が，重要語を含む名詞句であると，把握している．

3.2.1　名詞句の抽出

前章で述べたように，入力文中から不要なパター

ンを削除することによって名詞句を抽出する．この過程で，動詞句を中心とした不要な句は，２つの過程を経て除去される．

第１は，文頭に来る不要なパターンの除去である．科学技術文献に含まれる文は，名詞句で始まらないものが数多く見られ，そのため以下の図に示されるパターンを削除する．

＜副詞句＞
＜接続視＞
＜IT－THAT句＞
＜前置詞句＞,※
＜動詞句＞,※
＜不定詞句＞,※
＜分詞構文＞,※

図　３－３　文頭に来る不要句のパターン

上の表の中で，※のついたパターンの最後に'，'があることに注意したい．前置詞句，不定詞句でも，次の名詞句との間に'，'を含まない，次の例文に代表される句はこの段階では処理されない．

（例　３－５）

IN THIS PAPER THE AUTHOR PROPOSED－－－－－－－－－－－．

第２は，通常の宣言文に対する不要なパターンの除去である．次の図にあるパターンを照合し，マッチしたパターンは文中から除去される．

＜接続詞＞
＜副詞句＞
＜IT－THAT句＞
，　＜前置詞＞
，　＜不定詞句＞
，　＜現在分詞＞
＜関係代名詞＞　＜動詞句＞
＜関係代名詞＞
AND　＜動詞句＞
＜動詞句＞　＜関係代名詞句＞　＜動詞句＞
＜動詞句＞　＜関係代名詞句＞
＜動詞句＞

図３－４　文中に現われる不要句のパターン

この過程は，前章で述べたパターン・マッチングの基本構造そのものである．文中の語に対し，左から右にカーソルを１語ずつ移動させ，第２表に含まれるパターンにマッチしないと，その語がスタックにpushされる．マッチするとそのパターンを入力stringから除去し，stackにある個々のパターン（単語）を連結して１つの入力stringとし，次の名詞句分割プログラムにわたされ，処理される．例３－４の文に対し，名詞句抽出プログラムは，'IS ESTABLISHED USING'が動詞句パターンにマッチし，下線部分を出力する．

（例３－４　入力文）

THE EXISTENCE OF STABLE PERIODIC OSCILLATORY SOLUTIONS IN A TWO SPECIES COMPETITION MODEL WITH TIME DELAYS IS ESTABLISHED USING A COMBINATION OF HOPF-BIFURCATION THEORY AND THE ASYMPTOTIC METHOD OF KRYLOV, BOGOLIUBOFF AND MITROPOLSKY.

3.2.2　名詞句の分割

例３－４の文に代表されるように，名詞句抽出プログラムは前置詞句等を含む名詞句を出力する．科学技術文献には，この例のように，複数の前置詞句から構成されている名詞句を含む文が，数多く見られる．

名詞句分割プログラムは，次表にあげるパターンを照合し，このような長い名詞句を分割する．3－４の例では，最初に前置詞'IN'がマッチし，

THE EXISTENCE OF STABLE PERIODIC OSCILLATORY SOLUTIONS

が抽出され，慣用表現処理プログラムにわたされる．慣用表現処理プログラム以降のすべてのプログラムの処理が終わると，'A TWO SPECIES－－－'が新たな入力stringとなり，'WITH'が照合し，

A TWO SPECIES COMPETITION MODEL

が出力される．

名詞句抽出プログラムで処理されなかった'IN THIS PAPER THE AUTHOR'のような名詞句はここで分割される．＜名詞句＞＜冠詞｜指示代名詞＞というパターンを照合し，＜冠詞｜指示代名詞＞より前の部分をスタックにpushし，次の慣用表現処理プログラムにわたす．

3.3 慣用表現の処理

科学技術文献には２種類の慣用表現が見られる．

1. 'THE PROBLEM OF－－－－－'のように，'＜冠詞＞　＜形容詞＞　＜名詞＞　OF'，に代表される統語構造を持つ不要語で，点線の位置に重要語が来ることが多い．（2.1.2を参照）
2. 'THIS PAPER'，または'THE AUTHOR'のように，'＜冠詞＞　＜指示代名詞＞　＜名

詞＞'に，代表される統語構造を持つ句．

1のケースには，点線の位置に2のケースが来ることがあるので，最初に1のパターンを処理して点線の部分を抽出し，2のパターンを除去し，重要語だけを抽出する．

例3－4では名詞句分割プログラムが

THE EXISTENCE OF STABLE PERIODIC OSCILLATORY SOLUTIONS

を，慣用表現処理プログラムにわたす．最初に'THE EXISTENCE OF'が慣用表現のパターンに照合し，

STABLE PERIODIC OSCILLATORY SOLUTIONS

が，2のケースを処理するプログラムにわたされる．この語は2のケースの統語パターンにマッチしないので重要語として抽出され，後処理プログラムにわたされる．

慣用表現処理プログラムでは，辞書にない語は，名詞とは別の品詞として扱われる．たとえば，1のケースの結果として，'ASTABLE MULTIBIBLATOR'が抽出されても，2のケースの処理で，'＜形容詞＞＜名詞＞'のパターンにマッチすることはない．

3.4 後処理

後処理プログラムは 次の3つの処理を行なう．

1. 'AND'を含む重要語の処理
2. 重要語の前後についている不要語の除去
3. 誤って抽出された重要語ではない句（ノイズ・パターン）の処理

'AND' を含む重要語の処理：

多くの自然言語処理システムにおいて，等位接続詞 'AND'に対し様々な処理の方法が存在する．本システムでは ，'AND'の前後の単語の品詞を調べ，等しければ，'AND'で分割せずに，そのまま出力する．相違している場合は ，'AND'の前後の名詞句は独立したものとみなし，分割する．

次の例3－5で ，'TERMINALS'と 'JOB'は共に名詞なので（と仮定する），1つの名詞句としてそのまま，次の不要語を処理するプログラムにわたされる．

（例 3－5）

CHARACTER TERMINALS AND JOB TRANSFER

一方，例 3－6では 'RESISTANCE'は未登録語，'A' は冠詞で，品詞は異なっているので，最初に'A RESISTANCE'が，不要語を処理するプログラムにわたされ，次に'A ZENER DIODE'がわたされる．

（例 3－6）

A RESISTANCE AND A ZENERDIODE

不要語の処理：

例3－6に代表されるように ，'AND'の処理の段階では，重要語の前後に冠詞や，'SOME'，'EACH'，'OTHER' のような形容詞がついている場合がある．不要語処理では，これらの語の持つ品詞を照合するのではなく，これらの語をパターンとして，直接マッチングさせて除去する．

ノイズ・パターンの処理：

上記の不要語処理プログラムを経た段階で，重要語は抽出される．しかし，本システムを構成している最初の4つのプログラムにおいてミス・マッチを生じる可能性がある．従って，誤って抽出された語を除去する必要がある．ノイズ・パターン処理では動詞，関係代名詞等，重要語に含まれてはならないパターンを照合し，除去する．

4.実験結果と考察

4.1 実験対象

2章で提案され，3章でILISプログラムを用いて実現したパターン・マッチングによる重要語抽出システムの能力を調べるために ，小規模な 実験を行なった．

対象分野としてコンピュータ科学を選び，この分野の代表的なデータベースであるINSPECデータベースの抄録を用いた．100件の抄録中に含まれる 344の文を 実験対象とした .100件の 抄録は ，81年度のINSPECテスト・テープの，抄録を含む最初の 100レコードであり ，344の文は 100件の抄録に含まれるすべての文で，人為的な判断によって選ばれたものではない．次の文は実験対象に含まれる文の例である．カギカッコの中の番号は文番号を示している．

（例 4－1）

<<< NO. 00008 >>>.
THE RECORDED DATA CAN BE ACCUMULATED AND ANALYZED AT A LATER TIME WITHOUT CONSTANT PERSONNEL MONITORING.

<<< NO. 00009 >>>.
IN THIS PAPER THE EXISTENCE OF A POLYNOMIAL APPROXIMATION OF INPUT-OUTPUT MAPS FOR DISTRIBUTED SYSTEMS IS PROVED.

本システムで使用した辞書は，この 344の文に含まれる日常的な単語と，熟語，およびそれらの品詞からなっている．単語の総数は1300，熟語の総数は

48である．図３－２に，本システムで使われている辞書が示されている．

本システムの中でILISのプログラムとして記述されている動詞句を中心とする，文中から削除すべきパターン，および慣用表現等の，名詞句から削除すべきパターンは，344の文を調べて作成されたものである．図４－２，および図４－３は，本システムの中でプログラムとして記述されている，動詞句のパターン，慣用表現のパターンの一部を，それぞれBNF で記述し直したものである．

```
< BEV > < ADV > < ADJ > < PRE >
< BEV > < ADV > < ADJ >  TO
< BEV > < PAP > < ADV > < PRP >
< BEV > < PAP >  AND  < PAP >
< BEV > < ADJ >  OR  < ADJ >
```

図４－２　動詞句パターンの一部

```
< ART > < ADJ > < NOU >  AND  < NOU >  OF
< ART > < NOU >  AND  < ART > < NOU >  OF
< ART > < ADJ > < NOU >  OF  < PRP >
< ART > < ADJ > < NOU >  OF
```

図４－３　慣用表現パターンの一部

4.2. 実験結果

344の文について，抽出されるべき重要語に対しシステムは80％を的確に抽出した．ただし，システムが抽出した重要語には，いくつかのノイズが含まれる場合がある．

例４－２，および例４－３は，共に本システムの出力結果の一例である．例４－２は，抽出が成功した例，例４－３は，抽出の誤りを含んだ例を示している．

（例　４－２　抽出が成功した例）

```
<<< NO. 00019 >>>
A FAST NUMERICAL METHOD IS PRESENTED FOR THE
SOLUTION OF NONLINEAR ALGEBRAIC SYSTEMS WHICH
ARISE FROM DISCRETIZATIONS OF ELLIPTIC
BOUNDARY VALUE PROBLEMS
 *****KEY PHRASE *****
 FAST NUMERICAL METHOD
 NONLINEAR ALGEBRAIC SYSTEMS
 DISCRETIZATIONS OF ELLIPTIC BOUNDARY VALUE
 PROBLEMS
```

（例　４－３　抽出の誤りを含んだ例）

```
<<< NO. 000191 >>>
THE CONTROL SYSTEM MANAGES THE NETWORK AT
FIRST BY INTERACTING WITH THE TELEPROCESSING
MONITOR, THEN BY MEASURING THE NETWORK
COMPONENTS WITH THE GOAL OF GIVING THE
LOCALIZATION OF IRRECOVERABLE FAILURES AND THE
LIST OF THE NECESSARY MANUAL ACTIONS
 *****KEY PHRASE *****
 CONTROL SYSTEM
 FIRST BY INTERACTING WITH THE TELEPROCESSING
 MEASURING THE NETWORK COMPONENTS
 GIVING
 IRRECOVERABLE FAILURES
 NECESSARY MANUAL ACTIONS
```

4.3　考察

実験を繰返して行く過程で，名詞句抽出において直接動詞句のパターンを照合すると，'CONTROL', 'SET' に代表される名詞と動詞の複数の品詞を持つ単語について，ミス・マッチが生じる場合が多いことが明らかになった．このようなミス・マッチを減らすために，動詞句の直前の名詞句を次のように定義した．

＜NP＞＝＜NOU＞（＜PAP＞）｜＜DPR＞

最終的には，名詞句抽出規則では，最初に上記の名詞句パターンを照合した後,動詞句を照合する，という方法をとった．

次に，例４－３に代表される抽出の誤りを分析すると，抽出の誤りは，次に示される３つのケースに分類される．

1. 抽出された重要語の前後，あるいは中に不要な単語がついている．
2. 除去すべき語が誤って抽出されている．
3. 抽出されるはずの語句の一部，または全部が抽出されていない．

上記の誤りは，次にあげるような原因によるものと考えられる．

1. 動詞句，不定詞句，前置詞句，などの統語パターンが不十分である場合．
2. パターン・マッチングの実行順序による場合．
3. 品詞認定に誤りを生じる場合．

１については，必要な統語パターンを追加することによって改善されるであろう．

２，および３については，次のような理由が考えられる．規則が文としての構造を求めずに，句構造

で処理するため，誤った解析をした場合でも，システムは成功とみなし，バックトラックをしないことがある．そのため，実行順序が処理に大きな影響を与えてしまう．

5.おわりに

パターン・マッチングを用いた重要語の抽出の方法を提案し，ILISで，重要語抽出プログラムを作成した．本システムの利点を次に示す．

1. 大規模な辞書を必要としない．
2. システムの能力は文の統語構造の複雑さに影響されない．
3. システムの能力は文の意味構造にも左右されない．

一方，本システムは次のような欠点を持っている．

1. 形態素解析機構を持っていないため，名詞については１つの単語で，単／複それぞれについて，辞書に登録する必要がある．動詞については，さらに，各時制ごとに登録する必要がある．
2. 名詞と動詞の複数の品詞を持つ単語については，ミス・マッチが生じる場合がある．

以上のことから，次のような改良を加えることによって，重要語抽出の精度の向上が見られると考えられる．

1. 形態素解析機構を持たせる．

これにより，

A）名詞については，単複それぞれの単語，動詞については，さらに，各時制に関してそれぞれの単語を，辞書に登録する必要がなくなる．

B）単数／複数，および時制が識別できるので，動詞と名詞の複数の品詞を持つことによって生じるミス・マッチを減らすことができる．

2. より多くのデータを用いることによりパターンおよび辞書を充実させる．

最近，重要語抽出システムに関して，英語，および日本語について，格文法[8),9)]およびシナリオ[10)]を利用したシステムが開発されている．科学技術文献中の慣用表現をこのような方向から研究することにより，自動索引システムに近づくことができると思われる．

引用文献

1）Klingbiel, P.H. "Machine-Aided Indexing of Technical Literature" Information Storage and Retrieval, vol.9, no.2, 1973, p.477-494.

2）Klingbiel, P.H. "Technique for Machine-aided Indexing" Information Storage and Retrieval, vol.9, no.9, p.477-494. 1973.

3）Earl, L.L. "The Resolution of Syntactic Ambiguity in Automatic Language Processing" Information Storage & Retrieval, vol.9, p.277-308. 1973.

4）Earl, L.L. "Use of Word Government in Resolving Syntactic and Semantic Ambiguities" Information Storage & Retrieval, vol.9, p.639-664. 1973.

5）Sowa, J.F. "Interactive Language Implementation System" 1982.

6）Thompson, F.B. & Thompson, F.H. "Practical Natural Language Processing: The REL System as Prototype" in M. Rubinoff & M.C. Yorbits, eds., Advances in Computers. vol.13, Academic Press, New York, p.110-168. 1976.

7）飯田仁，他 "ATNと各解析を融合した英文名詞句解析" 自然言語処理研究会資料,vol.34, p.45－50. 1982.

8）絹川博之 "情報検索のための日本語解析" 情報処理，vol.20, no.10, p.907－910. 1979.

9）絹川博之，木村睦子 "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.24, no.3, p.200－207. 1980.

10）猪瀬博，他 "シナリオを用いる論文抄録理解・作成援助システム" 情報処理学会論文誌，vol.24, no.1, p.22－29. 1983