文献標題を利用した索引作製の自動化に関する考察*

安部憲広** 豊田順一*** 田中幸吉***

Abstract

The analysis of subjects contained in the Documents is one of the most important problems in the Document Retrieval System. There were methods in which role-indicators or links or weights are attached to the Documents in order to raise retrieval efficiency. From the experimental data in the past, it is well known that the system with role-indicators is at a disadvantage in the recall-factor. But the cause of such faults are the following three; 1) the property of role-indicator itself. 2) the ambiguity of rules which assign role-indicator to words in the Documents. 3) wrong assignment of role-indicator by the intereference of the man.

We organize the role-indicator system which is quite different from usual one, and design the algorithm of automatically indexing documents by words with role-indicators and tried a few experiments and also considered the retrieval system suitable to this automatic subjects recognition system.

あらまし

情報の索引作成を中心とする内容の分析は，文献検索などにおいて最も重要な点である．そうした主題分析において，文献に role-indicator, link, weight などを附加する方法がある．このうち，role-indicator を附加する方法は，過去の検証によれば，呼び出し率が悪いとされている．しかし，そうした欠点は，role-indicator そのものの特性と，その指定規準の不明確さ，および人間の介在に基づく誤りなどから生じるものであると考えられる．筆者らは，従来の role-indicator とは全く異なった概念を導入し，かつ自然言語の一般的処理を考察したうえで，role-indicator を自動的に指定する Algorithm を構成，実験し，さらに，こうした方式に基づく検索システムをも考察した．

1. まえがき

文献検索などにおいて，個々の文献の主題の表現方法は重要な問題である．従来，行なわれてきた方法は，検索効率向上のために，role-indicator, link, weight などの文献への附加が行なわれている．しかし，このうち role-indicator による方法では，呼び出し率が著しく低いとされている．その原因としては，role-indicator の指定が詳細すぎるため，role-indicator の指定において，その用法を規定することができず，主題分析時に誤差を生むことが考えられる．さらに，従来のように人手による主題分析作業では，知識などの面から個人差による情報の損失，誤差の生じることが考えられる．そこで，筆者らは，従来の role-indicator とは全く異なる概念で role-indicator を設定し，電子計算機で自動的に主題分析†1 による索引を作成する Algorithm を構成した．

Algorithm 構成においては，従来の単なる統計処理ではなく，自然言語の一般的特性を考慮したうえで，一種の統辞論的立場をとっている．こうした立場は，現在行なわれている自動化の方法よりも積極的なものである．そして，実験により，こうした方法でかなりの結果が得られうることを示した．

本考察の対象は，英文で記述された論文標題である．

まず，本考察の意図に関して解説し，つぎに標題文の構成上の特徴と主題との関連を記述し，さらに主題分析の自動化の Algorithm を示し，それに従がった

* Some Considarations on an Automatic Indexing System by Use of a Title of the Document, by Norihiro Abe, Junichi Toyoda and Kokichi Tanaka (Osaka University)

** 大阪大学・基礎工学研究科
*** 大阪大学・基礎工学部・情報工学科

†1 本文でいう主題分析は通常の意味より狭義のものであり，標題に表現される主題を抽出し，索引として使用することを意味する．

実験，および結果例を挙げて結果についての考察を行ない，最後には，本方式に適合した検索システムの一例を示した．

2. 考察意図

本節において，筆者らが本方式を考察した意図を解説する．要約すれば

（1） 自然言語処理による結果の統計的妥当性．

（2） 自動化に要求される機械の能力の限界．

（3） 文献検索システムの効率向上と，実用面での合目性との妥協．

に基づいて考察を行なった．それぞれに関して以下で解説する．

（1） 自然言語はその多様性のために，論理的性格や構文などを正確に記述することは困難である．種々の表現形式のため，たとえば，機械翻訳のような語句の表面的な“意味”の変換ですら，すべての表現に対して有利な方法はない．まして，自然言語の表現する“意味”はあいまいな場合が少なくなく，解析によってその“意味”，“内容”を完全に推測するAlgorithmの構成は不可能である．

しかし，自然言語の表現領域と，その表現される“内容”の値域の制限によって，自然言語の持つあいまいさや多様性を一意的解釈に規定することが，かなりの範囲において，可能であり，そのような条件をみたす領域では，十分使用に耐えうる解釈が行ないうると考えられる．本考察では，標題文という自然言語の部分集合を対象とし，そこに表現される“内容”を主題事項に制限し，その解釈の妥当性を結果の統計的妥当性に求めた．そのような規定に従がいえぬ例も存在するが，それらは例外と考え，例外が少ないならば，その解釈の規定は有意であると考えた．

（2） 現在のcomputerにとって，可能な言語の処理は，その言語を構成する系列の認識，すなわち，シンボルの認識とその系列のsyntaxの認識に限られる．したがって，主題分析の自動化に際しても，与えられた系列中の語とそのsyntaxしか利用できない．単語の“意味内容”，あるいは系列全体の“意味内容”などはcomputerは知り得ない．このような観点から筆者らは科学論文標題の持つ主題の分析を試みた．標題はあまり複雑な構文を持たず，比較的あいまい度も低いと考えられ，computerにとって取り扱いやすいと考えられ，また実用上の入力面からも標題を取り扱うべきであると考える．標題に示される主題を表示するMeta-言語を従来と異なる上位の概念に対して設定することにより，標題文の句構造，慣用句に基づいて，標題中に示されている主題をMeta-言語上で表現できるAlgorithmが作成可能であると考えられる．

（3） 文献情報検索を構成する立場には2つの立場が考えられる．1つは，索引の構造を精密にし，たとえばrole-indicatorを索引語に付けて，文献集合を細分化しようとするものである．

このためには，自然言語Lで書かれた文献Dの主題を表示するMeta-言語 L^* の構造を微細に規定し，

$$f;\ D \text{ over } L \rightarrow L^*$$

なるAlgorithmを作成する必要がある．この方法は適合率100%，雑音率0%を目的とするが，このようなAlgorithmの作成は，前記した自然言語の性質から困難であり，また客観性にも欠け，実際には呼び出し率も低い．

これに対して，他の立場として，索引の構造を簡単にし，たとえばuniterm systemなどにより，文献集合をあらく分割し，合目的であると判断される文献集合を求める方法がある．当然，このシステムでは，適合率は下がるが，呼び出し率は上昇するであろう．この場合，先に記述した意味でのAlgorithmは，前者に比して設定面および客観性の面においても有利である．

実用に供する場合，利用者の要求する文献の条件などが格納文献のそれと一致しないことが少なくない．このような場合には適合率が低下したとしても，呼出し率を上げておくほうが，確実に必要文献が入手できるであろう．

標題文は論文における論議事項の概要を記述したものであり，この標題文を変換し索引として利用するのは，後者の立場に近いものであろう．

以上で，本システム考察意図の解説とする．

従来，労力に比して効率の低いrole-indicatorによる主題分析に対して，自動化可能な部分を対象として，実用に耐えうる程度の効率をもつ検索システムの構成を意図したものである．

3. 標題文の特徴

3.1 構成上の特徴

標題文の特徴は，その構文的特徴と語句的特徴に区別される．それぞれに関して解説する．

（1） 構文的特徴

標題文の大部分は句構造であり，一般の文のような

主部，述部といった構成形態をとらない．句構造中のそれぞれの句の間に存在する関連は前置詞を媒体として表示されている．もともと，前置詞は，作用の条件状況のカテゴライズを明示する機能を有しているが，本例のような句構造をとる文においては，句の表現する事象を因果関係づけるものとして，前置詞の用法を規定することが可能となると考えられる．したがって標題文をその構成要素である句に分割し，それに続く前置詞の機能によって，それぞれの句の表現する事象を与えることができると考えられる．

（2） 語句的特徴

構文的特徴に付随して生じるものであるが，標題文には特有の表現形式がある．そこで使われる語は，その語を含む句や前後の句の表現事象を，前置詞とともに，規定する機能をもつと考えることが可能である．たとえば

An application to ……
An approach to ……
An example of ……
A method of ……

などの波線を施した語それ自体は標題中において主要な主題を表現するものではなく，主要な主題の明示のための指標として，また主要な主題の説明的事項としての機能をもっていると解釈される．

したがって，句を構成する語の特徴と，前置詞およびその系列の機能により，句の表現する事象を指定することが可能である．

3.2 標題文の表現する主題項目

標題に表示される主題は，通常の論文内容と比較して，かなり情報の総括をうけている．いいかえれば，標題においては，主題はきわめて詳細な限定，条件，相互関係などのもとで規定されているのでなく，すでに抽出された形での主題となっている．標題には，論文記述者の論議事項と，その関連事項とが記述されている．

筆者らは，前述の考察意図，およびこうした点より，主題表現としてつぎに記述する記号とその役割を設定した．これらの記号を構文記号と呼ぶ．このような主題表現事項の役割は，従来の役割とは異なったものであり，かつ本考察の対象外にも利用分野は多いと考える．たとえば，特許文書における特許の索引などに応用される．

M；論文の中心事項．
P；中心事項の属性，特性，部分，説明事項．
R；中心事項の関連，包括事項．
O；中心事項の対象とする目的事項．
D；中心事項を限定する属性事項．
C；論文の考察域に関する条件，使用機器などの表示事項．

これらの記号より構成される言語の意味記号の定義を以下のように設定すれば，主題表示の Meta-言語 L* が構成される．

〔定義 2-1〕 意味記号は，標題文を構成すを句である．
〔定義 2-2〕 意味記号は，構文記号に対応づけられる．
〔定義 2-3〕 主題表示の Meta-言語 L* は
{n−Np} により表現される．
ただし，n∈{M, P, R, O, D, C}
Np は名詞句である．

つぎに，標題文の主題をメタ言語上で表現するための変換規則の構成を行なう．変換規則は，標題文を構成する各句に，ロールを指定する機能をもつものである．そのために必要な条件などの考察を次節で行なう．

4. 変換規則の設定

4.1 設定のための諸考察

前記の標題文の特徴および資料調査によって，以下に示す諸手順から標題文の各句にロールを指定する変換規則を構成した．

（1） 前置詞の機能

標題文中での前置詞の基本的機能を以下のように仮定した．

OF；主要事項とその属性および関連事項を結合する．
FOR；中心事項とその目的事項を結合する．
ON；中心事項とその属性事項または条件事項を結合する．
BY；中心事項とその条件事項を結合する．
IN；中心事項とその条件事項を結合する．
TO；中心事項と目的事項または属性事項を結合する．
WITH；中心事項と限定条件事項を結合する．

しかし，こうした機能は基本的なものであり，実際の標題文中での各句の表わす事象を決定するには不十分である．実例を示して，具体的な処理を示す．

以下，構造の浅いものから，深い構造へと，さらに種々の変形された構造のものへと，考察をすすめる．

A Design of Automata.　　(1)

A Principle of Automata. (2)

A Covering of Automata. (3)

のような "of" のみが出現する標題文中での "of" の機能を考察する．(1), (2)では Automata が中心事項であることは明らかである．これに対して，(3)では，Automata は重要な事項ではあるが，covering は Automaton 理論の一分野であって，(3) のような標題では，一般的な Automaton の理論よりも，むしろ covering 理論についての記述を意味することは自明である．この差の認定は，Design, Principle, Covering の語によって行ないうる．すなわち，Design, Principle は "常用語" であり，Covering は "常用語" でないことから容易に認定が行なえることになろう．

同様な点を，他の前置詞に対して，考察することによって，構造の浅い標題文については，完全な処理が可能となり，また構造の深いものに対する基本的処理方法を示すものとなる．基本的処理を示すとは，深い構造のものも，適当な方法で浅い構造の組み合わせとして処理可能であることをいう．

したがって，前置詞の系列の出現する標題文に関する処理—組み合わせの方法—を考察する必要がある．もちろん，組み合わせのみでは不十分であり，新たに前置詞系列に対しての機能の設定も必要である．

たとえば

Application of Boolean Algebra to Switching Circuit Design. (4)

では，Np of Np, Np to Np の組み合せと考えるか，Np of Np to Np に対する (of, to) の系列としての機能を設定するかの決定を要する．中心事項の指定から行なう方針をとれば，"TO" の基本的機能から考えて，"OF" との連続系列としての機能は，資料調査の結果から考えても，不要である．なぜならば，"TO" には目的事項を限定する作用が強く，これに対して "OF" には，中心事項を限定する作用が強いため，(OF, TO) の系列では，中心事項の限定を "OF" によって行なうことができるためである．(4) の場合まず Np of Np によって各句の表現事象を決定し次に to Np の名詞句の表現事象を決定する．しかし，

Design of Comb Filter for a Servo Controller. (5)

などでは，Np of Np, Np for Np の組み合わせでは不十分である．これは "OF" "TO" の基本的機能に比して，"OF" "FOR" とは独立に句の表現事象を決定できるものではないことによる．いいかえれば，"OF" "FOR" においては，中心事項限定の作用力が，

SEQUENCE	%	SEQUENCE	%
—on—.	2.1	—of—by—of—of—.	0.1
—on—of—.	2.6	—of—on—.	0.7
—on—of—of—.	0.6	—of—for—.	1.3
—on—of—in—.	0.8	—of—for—of—.	0.7
—on—of—by—.	0.2	—of—for—with—.	0.1
—on—of—to—.	0.1	—of—for—on—.	0.3
—on—by—.	0.5	—of—with—.	1.7
—on—for—.	0.8	—of—to—.	1.1
—on—in—.	0.5	—of—to—with—.	0.2
—. (OR—and—.)	27.6	—of—to—to—.	0.1
—for—.	8.5	—of—of—.	2.8
—for—of—.	3.3	—of—of—in—.	0.1
—for—with—.	0.6	—of—of—of—.	0.6
—for—in—.	1.3	—of—of—for—.	0.2
—for—by—of—.	0.2	—of—of—to—.	0.3
—for—of—of—.	0.1	—of—of—by—.	0.1
—in—.	5.2	—to—.	1.4
—in—of—.	0.7	—to—of—.	1.6
—of—.	21.5	—with—.	1.5
—of—in—.	4.8	—with—of—.	0.3
—of—in—of—.	0.1	—with—on—.	0.6
—of—by—.	0.9	OTHER	1.4
—of—by—of—.	0.1		

Fig. 1 Patterns of Preposition Sequence in the Title

"TO" に対して "OF" のもっている作用力に比して，"FOR" に対しては弱いからである．"OF" で中心事項を決定し，"FOR" で目的事項を決定する妥当性が，"TO" に比して少ないことが，資料分析により推測されたことによる．このため，新たに Np of Np for Np の形態に対して，系列としての機能の設定を行なう必要がある．この設定は，資料調査に基づいて行なった．他の前置詞系列に対しても，同様な考察を行ない，必要な系列に対する機能を設定した．こうした機能の設定は，**Fig. 1** に示すように，標題文を標成する前置詞系列の出現頻度に基づいて，必要最少数にとどめた．機能の設定数は，多いほど質が向上するが，Algorithm の構成においてあいまい度が増すことになる．現時点では，20 の変換規則を設定している．

以上で，大部分の手順は記述されたが，つぎのような場合に，新しい処理が必要となる．たとえば，

The Construction of Minimum Redundancy. (6)

General Recursive Functions of Natual Number. (7)

では，(1), (2), (3) のような決定が行なえない．なぜなら，"常用語" の出現が無いため，中心事項の指定が行ないえないのである．筆者らは，このような

(WORDS OF NO SIGNIFICANCE)	(IDIOMATIC WORD)
A	ANALISIS
ANABNORMAL	APPLICATION
ABOUT	APPROACH
.....	APPROXIMATION
HIGH	ASPECT
ITS	.....
LARGE	CLASS
MANY	COMPARISON
MORE	DEVELOPMENT
NEW	EFFECT
.....	ESTIMATION
PRACTICAL	EXPRESSION
SOME	EXTENSION
THE	FOUNDAION
THEIR	.....
THREE	IMPROVEMENT
.....	INTRODUCTION
	INVESTIGATION
	.....
	METHOD
	NOTE
	PRINCIPLE
	PROBLEM
	PROCEDURE
	PROPERTY
	.....

Fig. 2 Examples of Φ Word and U-Word.

場合，主要な事項はより有意な説明を加えられているとの仮定を置いて，修飾部の長い句を中心事項と決定した．有意な説明とは，冠詞，一般の形容詞（many, some, ……）および副詞などの修飾語を除いた語による修飾を意味する．こうした語群を“空語”と呼ぶことにし，こうした語群を登録して，参照する．“空語”は，いわゆる“不要語”よりも小さな規模のものであり，“常用語”とともに，学術分野，学術の進展にとって変化を受けないものである点，更新の手続きが不必要である点，システム構成に有利である．こうした“常用語”，“空語”の一部を **Fig. 2** に示す．

以上で，変換規則の設定に関する基本的部分の考察は終了したが，他の変形的処理の考察を加える．

（2） 他の処理

標題文は，大部分句構造であるが，複文型，現在分詞型，受動体型も含まれている．このような型は，文中に“AND”“～ING”“～ED”を含むことより呼ぶことにする．このうち“AND”は，Fig. 1 に示すように，出現頻度も高く重要である．それぞれに関して，処理の方法を示す．

〈i〉 “AND”の処理

“AND”の用法としては，語の並列記述と文の接続記述がある．前者は，修飾語の接続と語の等位接続を意味し，後者は意味の分離を意味するものとする．標題文中の“AND”用法の規定は，構文中心の解析のため，文意の切れ目の確認などは容易ではなく，現時点では，誤まりの最少化に重点をおき，出現変度の高い用法に規定している．

筆者らは，“AND”の文形として

名詞句 and 名詞句 (8)

名詞句 and 名詞 (9)

名詞 and 名詞 (10)

形容詞 and 名詞句 (11)

を考えた．ここに名詞句とは，2語以上からなる句をさし，名詞とは区別する†2．例をあげておけば

series and parallel system は (11) の例

analysis and synthesis は (10) の例

などである．上記の中には，“名詞 and 名詞句”は含まれていない．したがって，たとえばつぎの例

Games and Statistical Design

では，Games は形容詞として処理されてしまうが，こうした例は非常に少ない．この誤りを防止するには，品詞リストを用意すれば十分であるが，辞書規模に問題があるため，そうした方法は考慮していない．しかし，上記4つの型でかなりの処理が可能である．

文意の切れ目を意味する“AND”としては

a）前置詞のない標題中の“AND”．

b）前置詞を含む一般の系列において，“AND”のつぎに続く語が“ITS”“THEIR”……などの所有代名詞，または冠詞の場合の“AND”．

と規定した．たとえば，

A Generalization of Algol and Its Formal Definition. (12)

では，“Formal Definition”は切り離して考える．この際問題となるのは“Its”の内容であるが，その指示物は考慮せず，前文の中心事項を後文に補足することで“Its”の内容を補なうこととした．

〈ii〉 “～ING”型の処理

Meta—文法のレベルで，形容詞や名詞に使用される“～ING”型の語は問題はないが，たとえば

An Algorithm for Assigning Role-Indicator to The Title. (13)

において，assigning は role-indicator の修飾語ではないため，assigning を Role-indicator から切り離して考える必要がある．そのためには，前置詞のつぎに続く“～ING”型の語のうちで，修飾語として働く語

†2 4.2 では名詞句 NP は，名詞のみから成る語を含んでおり，ここでの名詞句とは意味が異なる．

と，そうでない語とを区別しておく必要がある．筆者らは，このような切り離しを受けない語を登録し，参照した．

switching, covering, programing, sorting, …… などの語が，その一例である．

〈ⅲ〉 "～ED" 型の処理

前置詞の直前の語は一般に名詞（または，名詞としての働きをする語）であるが，たとえば

A Class of Language Recognized by Finite Automata. (14)

において，Recognized は Language と切り離す必要のある語である場合がある．このような語としては，他に applicable to などの語もあり，こうした語も "～ED" 型の語に含めて登録しておく．

以上で，現時点で可能と考えられる処理の大部分を考察した．これより，手順の形式化†3 を行なう．

4.2 諸定義

変換規則構成に必要な記号を定義する．

〔定義 4-1〕

（1） $\Delta\#$ は標題文の文頭を示す．

（2） NP_i は文頭より i 番目の名詞句である．ただし，以後名詞のみからなるものも名詞句として含むとする．

（3） $W(NP_i)$ は，i 番目の名詞句の末尾の単語とする．ただし，末尾語が名詞でない場合は，その直前の単語とする．

（4） $N(NP_i)$ は，i 番目の名詞句を構成する単語数である．

（5） NP_i が，$(\varphi)^1\cdot N$ であるとき，$NP_i \Rightarrow N$ とかく．ただし，$\varphi \in \Phi$, $\Phi=\{$a, an, the, its, their, ……$\}$ なる "空語" であり，$N(\varphi)=0$ とする．また，$(\varphi)^1=\varphi$ or Λ (null word) であり，$\cdot$ は concatenation である．

（6） NP_i の末尾語 $W(NP_i)$ が，"常用語" の集合Uの要素であるとき，$W(NP_i)\in U$ とかく．

〔定義 4-2〕

（1） 標題文の部分系列

$\#\ P_i\ NP_i\ \#P_{i+1}\ NP_{i+1}\cdots\cdots\#P_{i+n}\ NP_{i+n}$

が条件 α のもとで，それぞれの名詞句が変換規則によってロール $X_{i1}, X_{i2}, \cdots\cdots X_{in}$ を指定されるとき，変換規則を

$\langle m\rangle\langle\!\langle \sim\#P_i NP_i\cdots\#P_{i+n} NP_{i+n}\rangle\!\rangle;\ \alpha, \langle X_{i1},\cdots X_{in}\rangle^K$

とかく．ここに $\#P_j(i\leq j\leq i+n)$ は前置詞である．また，$X_{ij}(1\leq j\leq n)$ は，ロール $\{M, P, R, O, D, C\}$ の要素であり，$\langle m\rangle$ は，変換規則のラベルである．

（2） $i=1$ のとき

$\langle m\rangle\langle\!\langle \#\Delta NP_1\#P_2NP_2\cdots\#P_n NP_n\rangle\!\rangle;\ \alpha, \langle X_1,\cdots, X_n\rangle^K$

とかき，NP_{i+n} が文末の名詞句であるとき，

$\langle m\rangle\langle\!\langle \sim\#P_i NP_i\cdots\#P_{i+n} NP_{i+n}\#\rangle\!\rangle;\ \alpha, \langle X_{i1}\cdots X_{1n}\rangle^K$

とかく．とくに，$NP_1=\Lambda$ のとき，$X_1=\Lambda$ とかく．

$\langle m\rangle\langle\!\langle \sim\#P_i NP_i\cdots\#P_{i+n} NP_{i+n}\rangle\!\rangle;$

$\alpha, \langle X_{i1},\cdots X_{ik}, n, X_{il},,\cdots X_{in}\rangle^K$

のとき，句 $NP_{i(k+1)},\cdots NP_{il}$ のロール指定は，ラベル n の変換規則に従がうことを意味する．

（3） $X_{ik}=X_{i(k+1)}$ であるとき，X_{ik}' としるす．

（4） 上記の変換規則の定義において

K=0; それに続く系列によらず，ロール指定の追加を許さない．

K=1; 続く前置詞が {in, on, by} に含まれていれば，ロール指定を続く系列に従って追加する．

K=2; 続く前置詞が {with, for, to, in, on, by} に含まれていれば，ロール指定を続く系列に従って追加する．

4.3 変換規則

〔1〕 $\langle\!\langle \#\Delta\ NP_1\ \#\rangle\!\rangle;$ $\langle M\rangle^0$

〔2〕 $\langle\!\langle \#\Delta\ NP_1\ \#OF\ NP_2\rangle\!\rangle;$

（i） $W(NP_1)\in U$, $\langle P, M\rangle^2$

（ii） $\neg(i)\wedge(N(NP_1)>N(NP_2))$, $\langle M, R\rangle^1$

（iii） $\neg(i)\wedge\neg(ii)$, $\langle P, M\rangle^2$

〔3〕 $\langle\!\langle \#\Delta\ NP_1\ \#OF\ NP_2\ \#OF\ NP_3\rangle\!\rangle;$

（i） $(W(NP_2)\in U_V(W(NP_2)\notin U_\wedge NP_2\Rightarrow N))_\wedge NP_3\not\Rightarrow N$, $\langle P, P, M\rangle^2$

（ii） $(W(NP_2)\in U_V(W(NP_2)\notin U_\wedge NP_2\Rightarrow N))_\wedge NP_3\Rightarrow N$, $\langle P, M'\rangle^2$

（iii） $\neg(i)\wedge\neg(ii)$, $\langle P, M, R\rangle^1$

〔4〕 $\langle\!\langle \#\Delta\ NP_1\ \#OF\ NP_2\ \#FOR\ NP_3\rangle\!\rangle;$

（i） $NP_1\Rightarrow N_\wedge NP_2\Rightarrow N_\wedge(W(NP_1)\in U, W(NP_2)\in U)$, $\langle P, P, M\rangle^2$

（ii） $NP_1\Rightarrow N_\wedge NP_2\Rightarrow N$, $\langle M', O\rangle^1$

（iii） $\neg(i)\wedge\neg(ii)$, $\langle 2, O\rangle^1$

〔5〕 $\langle\!\langle \Delta\#\ NP_1\ \#ON\ NP_2\ \#\rangle\!\rangle;$

（i） $W(NP_1)\in U$, $\langle P, M\rangle^2$

（ii） $\neg(i)$, $\langle M, C\rangle^0$

〔6〕 $\langle\!\langle \#\Delta\ NP_1\ \#ON\ NP_2\ \#OF\ NP_3\rangle\!\rangle;$

†3 以後，手順を示すものを，変換規則と呼ぶ．これは f: D over L→L* なる写像を意味する．与えられた標題文からその常用語，句長などにおける条件に基づいて，各句にいかなる role を付加すべきかを決定するものであり，これにより，与標題文の主題を表示するメタ言語を構成する．

(i) $NP_1=\Lambda_\wedge((W(NP_2)\in U_\wedge NP_2\Rightarrow\!\!\!\!\Rightarrow N)_\vee$
$(NP_2\Rightarrow N_\wedge NP_3\Rightarrow\!\!\!\!\Rightarrow N))$, $[\Lambda, P, M]^2$

(ii) $NP_1=\Lambda_\wedge(NP_2\Rightarrow N_\wedge NP_3\Rightarrow N)$, $[\Lambda, M']^2$

(iii) $NP_1=\Lambda_\wedge(\urcorner(ii)_\wedge\urcorner(ii))$, $[\Lambda, M, R]^1$

(iv) $(W(NP_2)\in U_\wedge NP_2\Rightarrow\!\!\!\!\Rightarrow N)_\vee$
$(NP_2\Rightarrow N, NP_3\Rightarrow\!\!\!\!\Rightarrow N)$, $[P, P, M]^2$

(v) $NP_2\Rightarrow N_\wedge NP_3\Rightarrow N$, $[P, M']^2$

(vi) $\urcorner(iv)_\wedge\urcorner(v)$, $[P, M, R]^1$

〔7〕 《#⊿ NP_1 #ON NP_2 #OF NP_3 #OF NP_4》;

(i) $NP_1=\Lambda_\wedge(W(NP_2)$
$\in U_\wedge W(NP_3)\in U)$, $[\Lambda, P, P, M]^2$

(ii) $NP_1=\Lambda_\wedge(NP_2\Rightarrow N_\wedge NP_3\Rightarrow N)$
, $[\Lambda, M', R]^1$

(iii) $NP_1=\Lambda_\wedge(\urcorner(i)_\wedge\urcorner(ii))$, $[\Lambda, P, M, R]^1$

(iv) $(W(NP_2)\in U_\wedge W(NP_3)\in U)$
, $[P, P, P, M]^2$

(v) $(NP_2\Rightarrow N_\wedge NP_3\Rightarrow N)$, $[P, M', R]^1$

(vi) $\urcorner(iv)_\wedge\urcorner(v)$, $[P, P, M, R]^1$

〔8〕 《#⊿ NP_1 #ON NP_2 #FOR NP_3》;

(i) $NP_1=\Lambda_\wedge(W(NP_2)\in U_\wedge NP_2\Rightarrow N)$
, $[\Lambda, P, M]^2$

(ii) $NP_2=\Lambda_\wedge(\urcorner(i))$, $[\Lambda, M, O]^1$

(iii) $W(NP_2)\in U_\vee NP_2\Rightarrow N$, $[P, P, M]^2$

(iv) $\urcorner$(iii) , $[P, M, O]^1$

〔9〕 《#⊿ NP_1 #FOR NP_2》;

(i) $W(NP_1)\in U_\wedge NP_1\Rightarrow N$, $[P, M]^2$

(ii) $\urcorner(i)$, $[M, O]^1$

〔10〕 《#⊿ NP_1 #FOR NP_2 #OF NP_3》;

(i) $W(NP_1)\in U_\wedge NP_1\Rightarrow N$, $[P, 2]$

(ii) $\urcorner(i)$, $[M, O']^1$

〔11〕 《#⊿ NP_1 #TO NP_2》;

(i) $W(NP_1)\in U$, $[P, M]^2$

(ii) $\urcorner(i)$, $[M, O]^1$

〔12〕 《#⊿ NP_1 #TO NP_2 #OF NP_3》;

(i) $W(NP_1)\in U$, , $[P, 2]$

(ii) $\urcorner(i)$, $[M, O]^1$

〔13〕 《#⊿ NP_1 #IN NP_2》; , $[M, C]^0$

〔14〕 《#⊿ NP_1 #WITH NP_2》; , $[M, D]^1$

〔15〕 《～#TO NP》; , $[O]^1$

〔16〕 《～#IN NP》; , $[C]^0$

〔17〕 《～#BY NP》; , $[C]^0$

〔18〕 《～#ON NP》; , $[C]^0$

〔19〕 《～#WITH NP》; , $[D]^1$

〔20〕 《～#FOR NP》; , $[O]^1$

X…WORD
O…PREPOSITION

Fig. 3 A Method for Scanning of a Sentence

5. 実　験

5.1 実験方式

計算機実験では，まず前置詞，常用語，空語，ED型およびING型の語のファイルを作製する．入力文は，**Fig. 3** に示すように2回捜査される．最初は，前置詞（図中O）を捜し，しかる後，文頭に向かって，1単語ずつ捜査して，常用語にはU，ED型語にはE，空語にはP，ING語にはI，一般の名詞にはN，形容詞にはA，ANDには，その機能によりQまたはDがマークされかつ前置詞位置，句の長さが求められる．この過程のフローチャートを **Fig. 4** に示し，また上記の入力文記号変換の一例を **Fig. 5** に示した．図中で上方の欄は，格納されている各ファイルの内容の一部であり，各例の下には，記号変換の結果，前置詞位置，句長が示されている．前置詞位置は，Fig. 3 に示すように，まず標題文中の前置詞から捜査するため，たとえば，Fig. 5 の i）では，"OF" が3語目に，"FOR" が5語目，"ON" が9語目にあり，最後の "." が12語目にあることが求められる．句長は，やはりi）を例にとれば，"OF" が3語目として求められた後，Fig. 3 に示すように文頭へと逆方向に各語の機能を決定していくが，その際，空語，ING型などの語を除いた語数を求める．たとえば，A COMPARISONでは，Aは空語であるから，句長は1である．つぎのMETHOD は 1，GENERATING NORMAL DEVIATES は GENERATING を除いて，2となる．とくに，iii）などでは，"ON" ではじまるため，第1句の句長は0である．これは，前述の $NP_1=\Lambda$ に相当する場合である．こうした条件が，変換規則の条件 α として用いられる．

検索への展開のため結果を2次情報として DRUM に書き込みをさせるが，OPTION で TYPE OUT させている．

使用計算機は FACOM230-10，使用言語は COBOL である．PROGRAM は，DATA DIVISION 600 エントリー PROCEDUR DIVISION 2000 ステップ程度である．

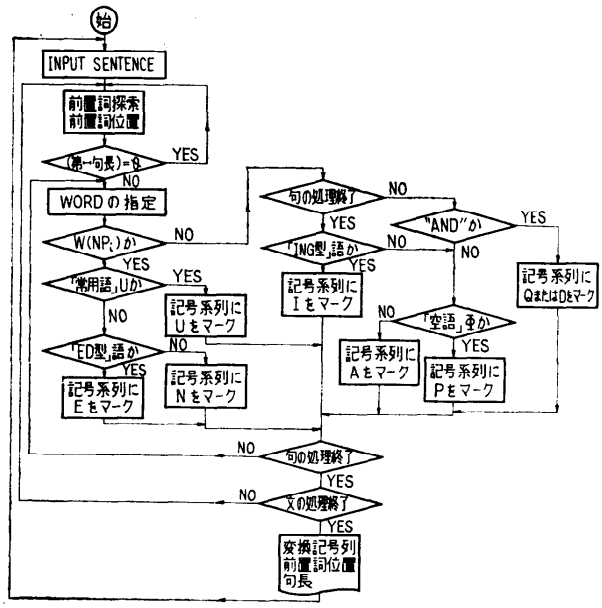

Fig. 4 Flow of Sybolic Translation of Input Sentens

```
   U: ANALYSIS    METHOD    COMPARISON .....
  ED: APPLICABLE TO    REALISED BY    RECOGNISED BY .....
 ING: COVERING    SWITCHING    PROGRAMING .....
   φ: A    AN    THE    THEIR    ITS    GENERAL    LARGE    VERY
PREP: OF    FOR    ON    IN    WITH    BY    TO    .
CODE: F    R    O    I    W    Y    T    Z
i) A COMPARISON OF METHOD FOR GENERATING NORMAL DEVIATES ON DIGITAL COMPUTER.
        P U F U RI A N O A N Z
        3 5 912
        1 1 2 2
ii) A GENERAL ANALYSIS OF VARIANCE SCHEME APPLICABLE TO A COMPUTER WITH A VERY
   LARGE MEMORY.
        P P U F A N E T P N W P P P N Z
        4 81116
        1 2 1 1
iii) ON ALGEBRAIC THEORY OF AUTOMATA.
        O A N F N Z
        1 4 6
        0 2 1
```

Fig. 5 Examples of Symbolic Translation of Input Sentence.

データーとしては，IEEE COMPUTER, INFORMATION & CONTROL, JOURNAL of A.C.M. など記載の文献標題を用いた．

5.2 結果の例

実験結果の一例を **Fig. 6** に示す．表現の方法は，前記の Meta-言語の定義に従がっている．それぞれの場合について，考察を加える．

（1） 変換規則をそのまま適用して結果の得られる構造の浅い例である．こうした例に関しては，結果の良好であることは明らかであろう．

（2） 比較的構造の浅い例であり系列としての前置詞機能と，基本的変換規則の組み合わせで結果の得られるものであるため，結果はやはり良好である．

（3） 構造が深くなるに従って，ロールを各句に附加できず，前置詞を含む合成句に対してロールを附加する必要が生じてくる．そのような例としては

（i） それぞれの句の表現が簡単すぎて，表示する事象がばくぜんとなり，ロールを指定することが意味を持たなくなる．

（ii） 前置詞の順列が設定した変換規則ではカバー

(1)

AN EXAMPLE OF A SELF-ORGANIZING SYSTEM.
M A SELF ORGANIZING SYSTEM
P AN EXAMPLE

AN INTERNAL SORTING METHOD FOR DIGITAL COMPUTERS.
M AN INTERNAL SORTING METHOD
O DIGITAL COMPUTERS

GENERAL RECURSIVE FUNCTIONS OF NATURAL NUMBER.
M GENERAL RECURSIVE FUNCTIONS
R NATURAL NUMBER

AN INTRODUCTION TO STATISTICAL COMMUNICATION THEORY.
M STATISTICAL COMMUNICARION THEORY
P INTRODUCTION

TWO DIMENSIONAL ARRAYS WITH DESIRABLE CORRELATION PROPERTIES.
M TWO DIMENSINAL ARRAYS
D DESIRABLE CORRELTION PROPERTIES

(2)

PROPERTIES OF A NEURON WITH MANY INPUTS
M A NEURON
P PROPERTIES
D MANY INPUTS

UNIFORM ASYMPTOTIC THEORY OF DIFFRACTION BY A PLANE SCREEN.
M UNIFORM ASYMTOTIC THEORY
R DIFFRACYION
C A PLANE SCREEN

STATISTICAL ESTIMATION OF THE INTRINSIC DIMENSIONALITY OF DATA COLLECTION.
M THE INTRINSIC DIMENSIONALITY
R DATA COLLECTION
P STATISTICAL ESTIMATION

A PROCEDUR FOR THE DIAGONALIZATION OF NORMAL MATRIX.
M THE DIAGONALIZATION
P A PROCEDURE
R NORMAL MATRIX

STRATEGIC APPROACHES TO THE STUDY OF BRAIN MODELS.
M BRAIN MODELS
P STRATEGIC APPROACHES
THE STUDY

A NOTE ON THE TWO ARMED BANDIT PROBLEM WITH FINITE MEMORY.
M THE TWO ARMED BANDIT PROBLEM
P A NOTE
D FINITE MEMORY

THREE LEVELS OF LINGUISTIC ANALYSIS IN MACHINE TRANSMISSION.
M LINGUISTIC ANALYSIS
P THREE LEVELS
C MACHINE TRANSMISSION

(3)

A COMPARISON OF METHODS FOR GENERATING NORMAL DEVIATES ON DIGITAL COMPUTERS.
M NORMAL DEVIATES
P A COMPARISON
METHODS
C DIGITAL COMPUTERS

ON THE P-RANK OF THE DESIGN MATRIX OF A DIFFERENTE SET.
M THE P RANK OF THE DESIGN MATRIN
R DIFFERENTE SET

A NOTE ON PRESERVATION OF LANGUAGES BY TRANSDUCERS.
M PRESERVATION OF LANGUAGES
P A NOTE
C TRANSDUCERS

ON AN APPRICATION OF DYNAMIC PROGRAMMING TO THE SYNTHESIS OF LOGICAL SYSTEM.
M DYNAMIC PROGRAMMING
P AN APPLICATION
O SYNTHESIS OF LOGICAL SYSTEM

ON THE FORMATION OF A CONVERGING SHOCK WAVE IN A GAS OF VARIABLE DENSITY.
M A CONVERING SHOCK WAVE
P THE FORMATION
C GAS OF VARIABLE DENSITY

(4)
SOLUTION OF ALGEBRAIC AND TRANSCENDENTAL EQUATION ON AN AUTOMATIC DIGITAL COMPUTER.
M ALGEBRAIC EQUATION
TRANSCENDENTAL EQUATION
P SOLUTION
C AN AUTOMATIC DIGITAL COMPUTER
NEW FOMULAS FOR COMPUTING INCOMPLETE INTEGRALS OF FIRST AND SECOND KIND
M INCOLPLETE INTEGRALS
P NEW FORMULAS
R FIRST KIND
SECOND KIND
SELECTION AND ORDERING OF FEATURE OBSERVATIONS IN A PATTERN RECOGNITION SYSTEMS.
M FEATURE OBSERVATIONS
P SELECTION
ORDERING
C PATTERN RECOGNITION SYSTEM
A GENERALIZATION OF ALGOL AND ITS FORMAL DEFINITIONS.
M ALGOL
P A GENERALIZATION
FORMAL DEFINITIONS

(5)
MODERN COMMUNICATION PRINCIPLES WITH APPLICATION TO DIGITAL SIGNALING.
M MODERN COMMUNICATION PRINCIPLES
D APPLICATION TO DIGITAL SIGNALING
SCENE ANALYSIS USING BY CONCEPT OF A MODEL.
M SCENE ANALYSIS USING
C CONCEPT OF A MODEL
AN ALGORITHM FOR THE DETERMINATION OF THE POLYNOMINAL OF BEST MINIMAXAPPROXIMATION TO A FUNCTION DEFINED ON FINITE POINT SET.
M ALGORITHM
O THE DETERMINATION OF THE POLYNOMINAL OF BEST MINIMAX APPROXIMATION
A FUNCTION DEFINED ON FINITE POINT SET

Fig. 6 An Example of Automatic Subjects Analysis

できない．現時点では，“IN”，“BY” などの条件規定の前置詞に続いて現われる “OF”，“FOR” などの前置詞の機能を設定していないので，“IN”，“BY” 以下のすべての系列を何と考えて，ロールを附加させざるを得ない．

などが考えられる．(i)は，表現自体の問題であるため，本考察とは独立の問題である．これに対して，(ii)の場合は，さらに資料の解析を行ない，変換規則の完備化を行なわなければならない．

(4) “AND” の出現する例である．例示したものは，考慮した “AND” の文形に一致しているものであり，結果は妥当なものである．しかしながら，前述「AND の処理」で解説したように，不十分な結果例も存在する．“AND” の用法に関して，より一般的な用法の解析の必要性がある．

(5) 不十分な結果の例である．例に示すように，特殊な表現に対しては結果が悪い．これは，本考察において，変換規則の設定が統辞論的方法に基づいていることが，原因である．少し具体的に解説すれば，

(i) 構造が深く，用意された変換規則でカバーできない．例示した最後のものなどは，その顕著なものである．こうした構造の深いものに対して，新たに変換規則を設定すればよいわけであるが，その場合には，現在のロールだけでは，各句の表現事象を表示できない場合も考えられよう．したがって，構造の深い標題の処理は，ロール指定，変換規則の設定の両者において，種々の問題が残ることになる．

(ii) 構造の深さとは独立に，推測していないような表現方法が標題中に出現する場合，結果が悪くなるのは当然である．例示の前2例はそうしたものである．こうした欠点は，新たに，with application to, by(the) use of などの熟語形の登録を行なえば，改善することができる．

以上で，結果の考察を終わる．

単語間の意味の関連，ロールと単語自身との関連などを推測しうる手順が求まれば，質の向上は顕著なものとなることが期待されよう．

6. 本考察を利用した検索システムの一例

6.1 考察すべきシステムの条件

本考察では，role-indicator を標題を構成する句に附加している．この理由は，Keyword 単独の意味よりも，phrase としての Keyword が持つ意味を重視するためである．Keyword 単独の場合には，論理演算の組み合わせによる単純 Matching 方式で十分であるが，本考察のような場合は，phrase としての Matching を考慮する必要があり，単純な方法では不十分であり，本考察を十分に活用する高度な方法から，容易な方法までのすべての方法をかねそなえていることが必要となる．以下で，質問形式，探索指令，および Matching の方法を概説する．

6.2 質問形式

（i） 質問文

質問文としては，a）標題文そのものを質問文とする．b）Meta-言語形式で，質問対象事項に，role-indicator を附加したリストを質問文とすることが考えられる．a）の場合は，主題分析と同様な処理が行なわれ，Meta-言語形式に統一されて，b）の場合に帰着される．したがって，b）の場合に関して，一般的考察を加えれば十分である．

（ii） 質問文の与え方

質問文を与える際，重要な点は同意内容を示す語の処理である．一般に，検索システムでは，こうした同義語・関連語を処理するために，“シソーラス”を用意するが，これは，単語の一般的な意味での同義性・関連性を列挙したものであり，phrase中での単語の同義性にはほとんど無力であり，かつ新語の出現に対しても無能力である．このような新語，phrase としての同義語は，user の知識にたよらなくてはいけない．また，そうすることが，呼び出し率の向上のために，欠かせぬ条件である．しかし，そのような同義な語による質問は，user の知識の深さに従属するため，同義質問文の数は，任意数設定しうるようにする必要がある．よって，質問文の与え方としては

1) 標題文そのものを質問文とする．
2) K 個（K≥1）の同義な Meta-言語形式の質問文．

が考えられる．

6.3 システムの形態

システムの形態としては，会話型がよいと考えられる．会話形式とは，user の質問を満足しうる解答の量が格納文献資料中にいくら存在するかを，システムが応答し，希望条件に合致しない場合，つぎにとるべき処理方式をシステムが指定し，user はその指定に従って質問形式を手直しして，システムとの会話を行ない，user の満足できる結果を得るまで，処理を続行する形式を意味する．システムとの会話形式の概要は，以下のとおりである．

〈i〉 標題文形式の質問文の場合

質問文である標題文は，主題分析により，Meta-言語形式に統一され，シソーラスを用いて，同義の質問文が作成され，探索作業をへて，その主題分析結果と，探索された文献とを出力する．user は，結果が満足できないものであれば，主題分析結果を参照として，つぎの Meta-言語形式での質問が可能である．

〈ii〉 Meta-言語形式での質問文の場合

user は，1つ以上の任意個の同義の質問文と，条件となる最低必要件数を設定する．システムは，こうして与えられた質問によって，探索対象を求め，つぎに示す Matching によって，合目的文献を求めるが，与えられた条件を満足しない場合，以下の指令を出力する．

Q1；同義な表現があれば，追加せよ．なければ，Q2 の指令を行なう．

Q2；表現の簡易化を行なう．表現の簡易化とは，それぞれの句の修飾語中で，比較的重要でない語を削除したり，また修飾部の短い語は，上位概念の語へ手直しすることを意味する．

Q3；中心事項 M を限定する条件をゆるめる．すなわち，重要性の低い role-indicator に対応する句を削除する．

Q4；role-indicator を無視して，Keyword の論理 Matching 方式とする．

6.4 探索と Matching の方法

自然言語を直接取り扱うため，語尾変形の処理などは当然必要であるが，単語に関する Matching に関しては種々の方法があるので，現時点では，単語の Matching の方法よりも，句としての Matching に関して解説する．

句の Matching に関しては，句の意味という概念の設定が必要となるが，ここでは，句の意味を，深層レベルでは取り扱わず表層レベルでの単語の組み合わせとする．具体的にいえば，句Qの意味は，各単語の表層的な形での機能と順序も含めた組み合わせから構成される．句Qと句Pとの同義とはそれぞれの句を構成する単語とその組み合わせが同様である場合と考えよう．たとえば，$Q=\omega_3\omega_2\omega_1$，$P=\omega_4'\omega_3'\omega_2'\omega_1'$ において，$\omega_i=\omega_i'(1\leq i\leq 3)$ である場合，また $\omega_1=\omega_1'$，ω_2

$=\omega_2'$, $\omega_3=\omega_4'$ である場合，さらに $\omega_1=\omega_2'$, $\omega_2=\omega_4'$, $\omega_3=\omega_4'$ のような場合なども，句Qと P句とは同義であると解釈する．もちろん，上例の同義性に，階層は存在する．同義性の強いものほど，Matching がよいと考える．以下に，形式化した同義性の定義を記述する．

〔定義 6-1〕

（i） 句Qの意味を m(Q) とかく．

（ii） 句Qと Pにおいて，句Pの意味が句Qの意味と同義であることを，$m(Q)\subseteq m(P)$ とかく．

（iii） 句Qと Pにおいて，句Pの意味と句Qの意味が一部同義であることを，$m(Q)\subset m(P)$ とかく．

〔定義 6-2〕

句Qが，$Q=\omega_m\cdot\omega_{m-1}\cdot\cdots\cdot\omega_1$，句Pが $P-\omega_n'\cdot\omega_{n-1}'\cdot\cdots\cdot\omega_1'(n\geq m)$ なる構成であるとき，句Pと Q句との同義性は，つぎのことを意味する．

（i） $m(Q)\subseteq m(P)\Leftrightarrow(\forall i),\ (\exists j\geq i),\ \omega_i=\omega_j'$

（ii） $m(Q)\subset m(P)\Leftrightarrow(\forall i),\ (\exists j),\ \omega_i=\omega_j'$

質問文で与えられた，ある role-indicator を附加された句の長さを $N(Q_x)$ とかく（句長は主題分析の定義によるものとし，添数Xは role-indicator を示す．）このとき，格納されている事項中の同じ role-indicator を附加された句の長さを，$N(D_{1x}), N(D_{2x}), \cdots, N(D_{Kx})$ とするとき

$$N(Q_x)\leq N(Di_x)\quad 1\leq i\leq K$$

なる条件を，質問文中で指定されたそれぞれの role-indicator に対して満足する文献が，対象資料として限定される．そうして，指定された対象資料に対して

$$m(Q_x)\subseteq m(D_{ix}),\ m(Q_x)\subset m(D_{ix})$$

を満足する文献を，output と考える．

より詳細な事項は，他の機会を得て，報告したいと考えている．

7. 結 言

標題文という，あいまい度の低い自然言語の部分集合を対象として，索引の自動的作成に関する基礎的考察を行なった．構造の浅い標題のもつ主題分析に関しては十分な結果を得たが，深い構造の標題に関しては不十分な点が少なくない．深い構造をもつ標題の主題の分析のためには，Syntax のみでなく，WORD の役割などの考察も必要となろう．現在の検索システムにおける自動化が，自然言語の統計的処理に基づいて行なわれている限り，質の向上は期待できない．本考察は，原理的な段階であり，すぐ実用可能なものではない．しかし，本考察に示した方向への研究を行なうことにより，人間の労力を要せぬ，かつ従来よりも質的に向上した主題分析，および検索システムの構成が，可能であると考える．

自然言語処理，検索システム構成などに熟知された諸氏のご批判・ご助言を期待する．

最後に，助言をいただいた水本雅晴氏，三上和敬氏および研究室諸氏に感謝する次第である．

参考文献

1) 情報処理学会：情報検索特集，情報処理，Vol. 7, No. 6 (1966)
2) 中井 浩：主題分析法，標準抄録文の書き方，およびその記号化への試論，JICST. Vol. 4, No. 8
3) 中村幸雄：主題の構造，情報処理 I，(C・11・1) 共立出版．(1969)
4) 高橋達郎：情報検索，東洋経済新報社．(1968)
5) W. S. Cooper: Fact Retrieval and Deductive Question-Answering Information Retrieval Systems, J. of ACM, Vol. 11, No. 2, (1964), pp. 117～137
6) L. S. Coles: An On-Line Question-Answering System with Natural Language and Pictorial Input, Stanford Research Institute. Technical Report. June, 1968
7) B. Raphael: SIR: Semantic Information Retrieval, Semantic Information Processing, pp. 33～135, THE MIT Press. (1968)
8) 安部，豊田，田中：主題分析の自動化に関する一考察，信学会オートマトン研究会資料，A 70-9 (1970-05).

（昭和45年6月30日受付）