情報学基礎 29－3
(1993. 5. 18)

抄録からの主題文の自動抽出

原田隆史* 細野公男* 野美山浩** 諸橋正幸**
* 慶應義塾大学文学部図書館・情報学科
** 日本ＩＢＭ東京基礎研究所

現在用いられているキーワードの自動抽出手法は，全ての文を対象としてキーワードを抽出するため，必ずしも文献の主題概念を表現しない語も抽出してしまうという問題点がある。
そこで，本研究では，主題概念を表現するキーワードをできるだけ効果的に抽出するために，抄録文のうち文献の主題を表現する文（主題文）を自動的に抽出するための手法を開発した。
文の構文的特徴や文中に出現する特別な表現を元にして主題文の抽出を行うことにより約81.1%の主題文を正確に抽出することができた。また，抽出した主題文中のキーワードを用いて検索実験を行った結果，検索もれをそれほど増加せずに検索ノイズを減少させることができた。

AUTOMARIC EXTRACTION OF THE SUBJECT BEARING SENTENCES

Takashi Harada* Kimio Hosono* Hiroshi Nomiyama** Masayuki Morohashi**
* School of Library and Information Science, Keio University, Mita, Minato-ku, Tokyo.
** Tokyo Research Laboratory, IBM Japan Ltd., Sanbancho, Chiyoda-ku, Tokyo.

In order to automatically extract good free keywords for content designation from abstracts, it must be effective and efficient to single out subject bearing sentences at first, and then extract those keywords from them.

This study, first of all, describes charcteristics of a method developed to automatically discriminate such sentences from those which show premises and conclusions, based on the particular expressions appeared in abstracts and the characteristics of syntactic structure in them.

Then this paper reports the result of experiment where the method was applied to the abstracts in the field of computer science.

Ⅰ．抄録を対象とするキーワードの自動抽出

A．キーワード自動抽出の一般的問題

現在のオンライン情報検索においては，通常の日本語文で表現された検索質問をそのままの形で用いて検索することはできない[1)2)]。要求する主題概念をキーワードに置き換え，このキーワードと文献中に含まれるキーワードとの照合によって検索が行われることになる。

キーワードを決定する方法は，付与索引方式と抽出索引方式の２通りに大別される[3)]。付与索引方式は，文献の主題分析を行い，その結果得られた主題概念を的確に表現するキーワードを文献中で使用されている語とは独立に付与する方式である。付与索引方式で与えられるキーワードはシソーラスなどであらかじめ用意されたものから選択する方式がとられることが多い。しかし，付与索引方式は一般に高度な知的判断が必要とされ，現在のところ人手で行う必要があることからキーワードの決定には多大の労力や時間・費用を必要とする。

一方，抽出索引方式は，文献中に出現した語句をそのままキーワードとして使用する方式である。この方式では，文献中のキーワードの持つ特徴を明らかにすればよく，主題を分析する過程が必要ないことから迅速にキーワードが決定できる。そのため，コンピュータを用いてキーワードを自動的に抽出する研究が盛んに行われている。コンピュータを用いた自動化の手法としては，語の出現頻度特性を利用する方法や用語辞書・不要語辞書を用いる方法，構文解析を用いる方法などがあげられる[4)5)]。

これらの方法のうち，構文解析を用いる方法は自然言語処理技術の発達にともなって急速に研究が進められている。日本語を対象とした構文解析による抽出索引法の研究としては，木本によるINDEXERシステムの開発[6)]や絹川らの研究[7)8)]，細野らの研究[9)]がある。これらの研究においては，構文解析を行うことによって主語と述語との対応，修飾語と被修飾語との対応関係などを明らかにし，その結果を用いてキーワードを抽出することを試みている。

しかし，これらの構文解析を用いた手法においても，各文単位での分析結果をもとにして，文中のすべての語を対象としたキーワードの抽出を行っている。このように，すべての語を対象としてキーワードの抽出を行った場合，以下の問題がある。

1) 得られたキーワードが，文献の主題として述べられている内容を表現していない可能性がある
2) キーワードの持っている重要性の差を考慮した検索を行うことができない。

中でも1)の問題は大きな問題である。たとえば，「これまで使われていたＡシステムは……が問題であった。そこで本論文では，新しくＢシステムを開発した。」という文章を対象にキーワードの抽出を行った場合，一般に，抽出索引方式では「Ａシステム」と「Ｂシステム」のどちらもキーワードとして抽出されることになる。しかし，文献の中で述べられている内容は「Ｂシステム」に関するものであって，「Ａシステム」について述べているのではない。もし，「Ａシステム」という語を用いて検索を行った場合，この文献は検索ノイズとなる。このように，全文を対象としてキーワードの抽出を行った場合，主題を表現していない文にキーワードが出現していることによる検索ノイズを避けることができない。

キーワードを決定する際の原則として索引作成マニュアル[10)]には，実際に実験・研究された事実のみをキーワードとし，他人の業績，将来のことについてはキーワードとはしないと規定されている。この原則に基づいてキーワードの選択作業を行うためには，実験・研究の結果，将来の展望などが記述された部分はキーワード抽出の対象とせず，文献中で主として述べられている内容のみを適切に表現する語をキーワードとして抽出することが望ましい。

B．主題文を抽出する対象としての抄録

従来，このようなキーワードを抽出しようとする研究の多くは，文献の内容を表現する抄録，標題をキーワード抽出の対象としている[4)]。これは，文献の全文を抽出の対象とした場合，以下の問題が存在するためである。

1) 内容が広範囲にわたるため処理や分析に労力がかかる。
2) 論文中における記述，表現の仕方が著者ごとにまちまちであり，文章構造にも統一性がない可能性

がある。

それに対し，抄録や標題は限られた長さで文献の内容全体を効率よく記述しているため，分析が本文を対象とする場合に比較して容易であり，キーワード抽出の対象として適していると思われる。また，抄録は作成基準に基づいて専門家の手で作成されるため，記述の仕方にも統一性があると考えられる。さらに，抄録や標題には本文中で述べられている研究そのものの記述だけではなく，研究の占める学問上の位置づけや，応用面の価値なども記述されているため，キーワードを自動的に抽出するための対象として適切であると考えられる。

日本科学技術情報センターの情報部作業マニュアルには原著的論文の抄録に盛り込むべき内容として，1)前提説明，2)目的・主題範囲，3)方法論，4)結果，5)考察・結論，6)注記があげられている[11]。

しかし，６つの項目の全てがキーワード抽出の対象として適切であるとはいいがたい。たとえば，「1)前提説明」は研究・開発・調査などの背景，先行研究などについて述べられる部分である。本稿では，この文を「前提文」と呼ぶことにするが，前提文に出現するキーワードは，その論文の内容に関するものではなく先行研究の内容に関するものも含まれる。したがって，この部分に出現する語を検索の対象とすることは検索ノイズの原因となる。

また，研究の方法，結果，考察を記述している「3)方法論」，「4)結果」，「4)考察・結論」の部分は，１つの文に方法論と結果がまとめて記述される例も見られるなど，はっきりと区別することが困難であることが多い。本稿では，これらの文をまとめて「結果文」と総称するが，その結果文中に出現するキーワードは，研究の目的を達成するために用いた手法に関するものや，将来の展望，今後の課題などに関するものが含まれ，当該論文の中心課題とは異なるものが多い。そのため，この部分に出現する語を検索対象にすることも検索ノイズの原因となる。

一方，「2)目的・主題範囲」は，その論文で何を扱い，何をしたのかについて述べている部分であり，きわめて重要である。情報検索を行う場合には，論文の扱う範囲，内容に関する部分に記述されたキーワードを対象に照合が行われることが望ましいと考えられるからである。また，「6)注記」は，論文の主目的外であっても価値のある知見や重要な情報が記述されている部分である。そして，検索もれを最少限にとどめるためには，これも検索の主たる対象と考えることが妥当である。したがって，これらの文を主題文と呼ぶことにすれば，キーワード抽出は主題文を対象にすることが望ましいといえる。

Ⅱ．主題文自動抽出の考え方と要件

A．過去における主題文自動抽出の試み

実際の抄録中に含まれる各文が上記の前提文，主題文，結果文のどの部分に属するものであるのかを，人間の手で比較的容易に判断することは可能である。しかし，そのためには多くの時間と費用が必要とされる。したがって，この判断を自動的に行うことができれば，主題文の自動抽出が可能になり，その結果，前述のようなキーワード自動抽出の高度化が可能となる。

主題文の自動抽出に関わる研究として，発表者らのものがある[12)13)]。

この研究は，まず，前提文，主題文，結果文の持つ表層的な特徴および，ある種類の文の次にどの種類の文が出現するかという文の接続関係の特徴を分析し，分析された特徴をもとに各文の種類を自動的に判断することを試みたものである。そして，これらの特徴をもとにすれば，抄録の第１文目に主題文がある場合は97%，それ以外の場合は40%の抄録について完全な判断ができるという結果が得られている[12)]。ただし，文の接続関係の特徴に基づいて文の判断を行った場合，文章の内容を十分に反映できないという問題があったため，引き続きその解決方法として，文の接続関係の代わりに文の構文的な特徴を元にして判断を行う手法を開発した[13)]。

本稿では，この表層的な特徴だけでなく，文の構文的な特徴もふまえて文の種類を機械的に決定する方法による結果を示すとともに，さらに，主題文のみを対象としたキーワード自動抽出の有効性について報告する。

B．主題文中に含まれるキーワード数の調査

主題文の自動抽出にあたっては，まず，抄録中から

抽出されたキーワードが主題文にどの程度出現するのかを調査する必要がある。これは，各抄録文から抽出されたキーワードが他の文にはほとんど出現せず，主題文中により多く出現していれば，キーワード自動抽出を主題文に限定することの意義が明らかになるためである。

そこで，JICST科学技術文献ファイル電気工学編に含まれる353抄録の1060文を対象に，「前提文」，「主題文」，「結果文」中に含まれるキーワードの数を手作業で調べた。

キーワードが主題文から直接，または主題文中に示された代名詞によって示されるものとして，どの程度抽出されるかを示したのが第1表である。

第1表に示すように，1060文中に出現した2866個のキーワードのうち2466個(86.0%)が主題文中に含まれる。また，前提文や結果文に出現するキーワードのうちの一部は主題文にも出現しており，主題文のみを対象とした場合に抽出できないキーワードは，2488個中137個(5.5%)にすぎなかった。

5.5%のキーワードが主題文以外の文のみに含まれていることから，キーワードの決定にあたって，必ずしも主題文のみから語を抽出すればよいと断定することはできない。しかし，主題文から抽出される語に対して，他の種類の文から抽出される語よりも高い重要性を与えるなどの手法が有効であることは明らかである。

第1表　各文の数とキーワード数

文の種類	文の数	キーワード数	主題文にないキーワード数
前提文	172文(16.2%)	230(8.0%)	85
主題文	718文(67.7%)	2466(86.0%)	-
結果文	161文(11.0%)	170(5.9%)	52
その他	9文(0.8%)	0(0.0%)	0
計	1060	2866	

Ⅲ．前提文，主題文，結果文の特徴

A．調査対象

抄録中の前提文，主題文，結果文を正しく判断するためには，実際の抄録中の各文の持つ特徴を明らかにする必要がある。これは，本研究においては，どのような特徴を持てばどの種類の文である確率がどのくらいであるかという規則を作成し，この規則に基づいて判断を行うからである。

分析の対象にする抄録としては，訓練された抄録者によって同一の基準のもとに書かれたものが大量に得られることが望ましい。また，主題文以外についても分析を行うため，原文献が取り扱っている主題を主として伝える指示抄録ではなく，取り扱う主題の他に，研究の背景や結果などについても記述した報知的抄録または半報知的抄録を対象とすることが適切である。

また，主題文のみを対象とした検索を行った場合と，抄録中の全文を対象とした検索を行う場合の結果を比較するため，特定のテーマについて記述された抄録の集合が必要である。そこで，本研究では，ＪＩＣＳＴ科学技術文献ファイル電気工学分野の抄録を用いた。

JICST科学技術文献ファイル電気工学編に含まれる353抄録1060文を対象として，前提文，主題文，結果文の特徴の分析を行い，それに基づいて文の種類を決定するための規則を作成した。この際，仮に1つのルールだけでは文の種類を決定したとしても，複数の規則を組み合わせることで，より有効な判断ができることがあるため，作成した規則の組み合わせの規則も作成した。

文の種類を決定するために用いた文の特徴は，以下の4点である。

1) 文と文とをつなぐ接続詞
2) 複文の中で用いられる接続表現
3) 主題文中に特有の表現
4) 文末表現

B．接続詞の分析

抄録では，まず前提文，主題文，最後に結果文という順序で記述される場合が多い[12]。一般に，内容的に大きく異なる文章を記述する場合には，接続詞を用いて文と文との関係をわかりやすくすることがあるため，前置詞と主題文，主題文と結果文との間には接続

詞が存在する可能性があると考えられる。

そこで，文頭に表れる接続詞を対象とし，その接続詞で接続される前後の文のカテゴリーがどのように変化するかを分析した。

ただし，抄録文中では，字数に制限があるため接続詞が省略されることが多く，本研究で対象とした抄録中に接続詞は1060文中に53個が出現するにとどまった。第２表にその結果を示す。

第２表　接続詞による文の種類の変化

接続詞	前後での文の変化	出現回数
また	前提文 -> 前提文	2（ 7.1%）
	主題文 -> 主題文	21（75.0%）
	結果文 -> 結果文	4（14.2%）
	結果文 -> 主題文	1（ 3.6%）
さらに	前提文 -> 前提文	1（ 8.3%）
	主題文 -> 主題文	9（75.0%）
	結果文 -> 結果文	2（16.7%）
しかし	前提文 -> 前提文	2（40.0%）
	主題文 -> 主題文	2（40.0%）
	結果文 -> 結果文	1（20.0%）
そして	主題文 -> 主題文	3（100%）
そこで	前提文 -> 主題文	5（100%）

第２表に示すように，「また」においては28個中27個，「さらに」，「そして」，「しかし」ではすべての場合で文のカテゴリーの変化はなかった。これは，「また」，「さらに」，「そして」が累加あるいう添加の働きを持つ接続詞であるためと考えられる。

また，「しかし」もそれまで述べた事柄と異なる意見や事実を付け加えるときに用いられる接続詞であり，意味的には反対の事柄が接続されていたとしても，文のカテゴリーには変化はないものと判断できる。

この結果から「また」，「さらに」，「そして」，「しかし」を文頭にふくむ文の場合には，その前の文のカテゴリーが認識できれば，当該文の種類も判断できると考えられる。

それに対し，「そこで」は，すべて前提文から主題文への変化をともなう場合に用いられていた。これは，「そこで」という接続詞は，前の事柄の結果ある行為がなされるという，状況と行為の因果関係を示すものであるためと考えられる。この結果から，「そこで」で接続される前の文は前提文，後の文は主題文である可能性が高いといえる。

C．複文の分析

一般に，抄録では字数に対する制限が存在するために複文が含まれることがある。このような複文においては，文の前半と文の後半とを結ぶ表現によって接続関係が判断可能である。この接続関係を元にして文の種類をとらえることができる。

第３表に，各種類の文に表れる意味ごとの接続表現を示す。今回分析対象とした抄録中に含まれる複文は382文であり，その中に含まれる接続表現は422個であった。

第３表　各文における接続表現の出現頻度

		前提文	主題文	結果文	合計
従属接続	条件	7	3	5	15
	原因	13	9	5	27
	目的	6	22	2	30
	方法	10	147	20	177
	様式	9	18	7	34
	例示	1	0	0	1
	時間	4	1	0	5
等位接続	並列	8	86	10	104
	逆接	7	3	4	14
	平接	4	8	3	15

第３表に見られるように，方法の意味を表す接続表現（例：「ことにより」）では177個のうち147個(83.1%)が，また，目的の意味を表す接続表現（例：「ため」）では30個のうち22個(73.3%)が主題文で使われるものであった。

逆に，条件の意味を表わす接続表現（例：「～とすると」）は，15個のうち3個(20.0%)だけしか主題文中には出現しなかった。

これらのことから，複文においては，方法，目的，条件の意味を表す接続表現をとるかどうかによって，主題文かどうかを特定できる可能性が高いと考えられる。

D．主題文に特有の表現

抄録では，字数の制限などのため一度出現した言葉を別の語で簡潔に表現する場合が多い。このような一度出現した語を指示する語のうち，主題文に特有の表現として以下の３種類の表現を見い出すことができた。

1）標題を示す語

「標記～」，「標題～」などがこれに該当する。本研究で対象とした抄録中でこのような表現は42文に出現した。このうち，主題文が39文，前提文が２文，結果文が１文であった。

2）「本研究」，「本文」，「本論文」などの表現

「本稿」，「本報」，「本文」，「本論」，「本誌」「本論文」などがこれに該当する。本研究で対象とした抄録で，このような表現は29文に出現した。これらの表現はすべて主題文中に出現した。

3）「ここでは」

本研究で対象とした抄録中，「ここでは」は20文に出現し，そのすべてが主題文であった。「ここでは」という表現は他と区別してその論文で取扱った内容について述べる時に用いられる表現である。そのため，論文の主題文中で多く用いられると考えられる。

E．文末表現

日本語において，文末の表現は1)文末文節直前の助詞または助詞相当表現，2)文末文節の語幹，3)文末文節の語尾の３つの構成要素から構成される。

たとえば，「～について説明した。」という文末の場合には，以下のように分けられる。

文末文節直前の助詞等　・・・・・ について
文末文節の語幹　・・・・・ 説明
文末文節の語尾　・・・・・ した

発表者らは，以前に各構成要素ごとに分析を行い，これらが文の種類に応じた特徴を持つことを指摘している[12]。しかし，その分析では以下の問題点があった。

1）文末文節を複数の文節が修飾していた場合に，動詞の直前に記述された表現しか分析対象としていない

2）動詞の分類を行わず各動詞ごとに分析を行ったため，分析対象となる動詞数が少ない

そこで，本研究では，構文解析を行って動詞の直前以外に記述された表現の分析を行った。また，国立国語研究所の分類語彙表[14]をもとにして，各文の述語を分類し，どのような述語表現がどの種類の文に多く用いられているかを調べた。

構文解析には，現在機械翻訳システムで使用されている日本語解析ワークベンチ（ Japanese Analysis Work Bench ）を用いた。このシステムでは，まず入力した抄録文の形態素解析を行い，その情報をもとに，構文解析を行って文節の係り受け関係を構文木で表現する。

分析の対象となる述語は，単文，複文にかかわらず，文末の述語，すなわち構文解析によって得た構文木の１段目に出てくるもののみとした。

「文末文節の直前の助詞または助詞相当表現」については，以前に指摘した[12]「～について」以外に，主題文に特有の表現として「～につき」「～に関して」「～かを」という表現を抽出することができた。すなわち，これらの表現が出現した349文のうち主題文が290文(83.1%)を占め，結果文は56文(16.0%)，前提文は３文(0.9%)であり，主題文の中で特に多く見られる表現であることが明らかとなった。ただし，これ以外の表現については，各文の種類による差異は見られなかった。

「文末文節の語幹」については，分類語彙表の分類に基づいて作成した動詞の分類カテゴリーごとに集計を行った結果，各分類カテゴリーに属する動詞の出現する文の種類に大きな偏りがあることが明らかとなっ

た。特に主題文で多く出現する傾向が見られた分類カテゴリーの例を第４表に，特に主題文以外で多く出現する傾向が見られた分類カテゴリーの例を第５表に示す。

第４表　主題文に特有の動詞カテゴリー

動詞のｶﾃｺﾞﾘｰ（代表例）	前提文	主題文（割合）	結果文	計
談話(述べ)	0	75(100%)	0	75
説明(解説，概説)	0	74(100%)	0	74
紹介・斡旋(紹介)	0	55(100%)	0	55
会議・議論（議論，論じ，提案）	0	32(100%)	0	32
叙述(技術，論述)	0	25(100%)	0	25
製造(開発，製造，作)	1	31(96.9%)	0	32
研究(研究，検討)	0	30(96.8%)	1	74
呈示・指定（示，指摘，示唆）	0	61(87.1%)	9	70

第５表　主題文以外に特有の動詞カテゴリー

動詞のｶﾃｺﾞﾘｰ（代表例）	前提文	主題文（割合）	結果文	計
成立・発生（出来，成，生じ）	15	17(40.5%)	10	42
使用(用い，利用)	7	13(32.5%)	20	40
存在(あ，存在)	20	3(9.7%)	8	31

第４表に示すように，「談話」というカテゴリーの「述べ」という動詞は最も出現頻度が高く，76文に出現するが，それらはすべて主題文に含まれるものであり，主題文を他の文と識別する際に非常に有効であると考えられる。同様に，「会議・議論」というカテゴリーの「議論」，「論じ」，「討議」，「提案」は合計32文に出現したが，すべて主題文中に出現するものであった。また，「叙述」というカテゴリー中の動詞は25文に出現するが，これもすべて主題文中に出現した。

したがって，文末文節の語幹によって特に前提文と主題文を見分けることは可能といえる。なお，述語が形容詞あるいは形容動詞である文は，いずれも主題文中に出現しないという結果が得られた。

「文末文節の語尾」（動詞の語幹に続く付属語）にも，文のカテゴリーごとに特徴が見られた。

たとえば，「開発」という語幹は主題文中では「開発する」や「開発した」という表現として使われるが，前提文中では「開発している」，「開発されつつある」という表現として使われる。同様に，「示」は主題文では「示す」または「示した」という表現で使用されるが，結果文では「示された」という表現で使用されることが多い。また，「この方法では役にたたない」など，否定の付属語をとる文は，いずれも主題文以外であった。

さらに，「～であろう」などのような推量を示す付属語や，「～できる」などの可能を示す付属語は，主として将来の展望を述べる際に使用されるものであり，結果文に多く出現している。また，「～である」などのような断定を示す付属語も主題文には少ない表現としてあげることができる。

Ⅳ．抽出した特徴に基づく主題文の自動抽出

A．文単位の分析

Ⅲで示した特徴から129個の規則を設定した。規則は，条件部と判断部からなり，実際の文の種類の割合に応じてどの程度文の種類の判断に有効かという数値が付与されている。

たとえば，「文末の動詞が"研究"というカテゴリーに属する動詞であるならば，その文は主題文である」という規則には，第４表に示す実際の分析結果に対応して，0.968という数値が付与されることになる。

また，同時に複数の規則が成立する場合，その成立する規則の組み合わせによっては，２つの規則に付与された数値の組み合わせからは予想できないほど文の種類決定に大きな影響を与える場合がある。そこで，129個の規則の条件部の組み合わせを用いて911個の組

み合わせ規則を作成し，これも用いた[13]。

実際の判断では，判断の対象となる文を解析した結果が条件部と合致した場合に，判断部の結果が採用される。

規則を作成するのに用いた抄録とは別の152抄録536文を対象として文の分類を行った結果を第６表に示す。

第６表　文の種類判断実験の結果

		システムの判断			
		前提文	主題文	結果文	計
人による分類	前提文	53	29	3	85
	主題文	60	275	4	339
	結果文	51	33	28	112
	計	164	337	35	536

第６表に示すように，人が主題文と判断した339文のうち275文(81.1%)がシステムによっても主題文だと判断された。また，システムが主題文だと判断した337文のうち人の判断と一致した文の割合は，81.6%であった。

この結果は，発表者らが文と文との接続に関する分析を元に行った実験[12]の結果（人手によって主題文と判断された文の71.6%をシステムが主題文と判断した）よりもかなり高い値である。

B．抄録単位の分析

上記のように判断した各文の種類を基に，抄録単位で適切なキーワードを抽出することが可能であるかどうかの分析を行った。

そのために，適切な検索結果が既知の検索質問に対して，抄録の全文を対象とした検索と主題文のみを対象とした検索とを行い，その検索ノイズおよび検索もれの変化を分析した。

検索対象としては，JICST科学技術文献ファイル電気工学編に含まれる抄録のうち，「データベース」あるいは「インタフェース」のいずれかの語を含む98抄録を用いた。

分析対象とする98抄録のうち，65抄録が「データベース」で検索した場合に適合する抄録，42抄録が「インタフェース」で検索した場合に適合する抄録であった。すなわち，29抄録については「データベース」および「インタフェース」のいずれの語で検索した場合においても適合することになる。

「データベース」で検索した場合に適合する抄録としては，ある特定の分野のデータベースを構築したという報告，そのデータベースの内容・構成・利用方法の説明，管理方式の検討，そのデータベースを利用するメリットなどについて述べたものがあげられる。また，「インタフェース」で検索した場合に適合する抄録としては，自然言語インタフェース（自然言語問い合わせシステムや自然言語質問システムを含む）の開発や概要の説明，自然言語インタフェース開発支援ツールの開発や概要の説明を行うものがあげられる。

「データベース」および「インタフェース」という語を用いて行った検索の結果を第７表に示す。

第７表　検索実験の結果

「データベース」で検索（適合件数 65件）				
	検索件数	適合件数	精度	再現率
全文を対象	87	64	70.1%	98.5%
主題文を対象	70	61	87.1%	93.8%
「インタフェース」で検索（適合件数 42件）				
	検索件数	適合件数	精度	再現率
全文を対象	54	42	77.8%	100%
主題文を対象	46	40	87.0%	95.2%

第７表に示すように，「データベース」という語で検索した場合には，抄録中の全文を対象とした場合に比較して主題文のみを対象とした場合に検索された抄録数は，85件から70件(80.5%)へと17件減っている。これに対し，検索もれは，１件から４件へと３件増加

しているにすぎない。また，「インタフェース」という語で検索した場合には，全文を対象とした場合に比較して主題文のみを対象とした場合に検索された抄録数は54件から46件(85.1%)へと８件減少している。これに対し，検索もれは２件増加するにすぎない。

検索もれとなった６件のうち，「データベース」での検索における３件および「インタフェース」での検索における１件は，主題文の判断が正しく行われなかったことによる検索もれであった。また，それ以外のものは当該文献が製品紹介的な内容を持っており，検索対象とする語が前提文のみに出現したものであった。

以上から，主題文のみを対象とした検索を行った場合，全文を対象とした検索を行う場合と比較して，それほど検索もれを増加させずに検索時における検索ノイズを減らすための１手法を示すことができたといえよう。

V．おわりに

本研究では，抄録を構成する文を前提文，主題文，結果文の３種類に分類し，文の表層的および構文的特徴に基づいてこれらの文の種類の自動識別を試みた。

特に情報検索における検索ノイズの低減のためには主題文中のキーワードを対象として検索を行うことが必要であると考え，主題文であるかどうかの判定をすることを中心に実験を行った。

文の表層的な特徴，構文的な特徴に基づいて129個の規則を作成し，これをもとにして抄録から主題文を機械的に抽出する実験を行った結果，主題文の81.1%を正しく抽出することができた。また，主題文のみを対象として検索実験を行った結果，検索対象を主題文に限定することで，検索もれをそれほど増加させずに検索ノイズを減少させることができることを明らかにできた。

主題文以外の文においてのみ出現するキーワードも存在することから，主題文のみを対象とした検索ですべての状況に対応できるわけではない。しかし，主題文を対象にすることによって，検索効率が大きく向上することを示すことができたと言えよう。

主題文を対象とした検索において検索ノイズとなった抄録の多くは，主題文に出現した語のうち，本来はキーワードと判断すべきでない語との照合が行われたものであった。このような検索ノイズを除去するためには，主題文に出現した語のうち，どの語がキーワードであるかを判断しなければならない。

今後は，主題文抽出のための規則の精緻化を行うと共に主題文から適切なキーワードを自動的に抽出する方法についても検討する必要があろう。この場合，代名詞や「標記の」，「標題の」などの表現が実際のキーワードとして何をさすのかを明らかにすることが必要であろう。

さらに，抄録からの主題文の抽出だけにとどまらず，全文からの主題文の抽出についても考えていきたい。

引用文献

1) 杉山健司ほか．自然言語理解に基づく情報検索システムIRIS．情報処理学会自然言語処理研究会報告，NL58-8，p.1-8（1986）．

2) 佐藤正光ほか．特許情報検索のための日本語質問文解析．情報処理学会論文誌，Vol.25，No.3，p.365-371（1984）．

3) 細野公男編．情報検索．東京，雄山閣，1991．259p.

4) 諸橋正幸．自動索引付け研究の動向．情報処理，Vol.25，No.9，p.918-925（1984）．

5) 図書館・情報学ハンドブック編集委員会．図書館情報学ハンドブック．東京，丸善，1988．p.574-579.

6) 木本晴夫．日本語新聞記事からのキーワードの自動抽出と重要度評価，電子情報通信学会論文誌．D-I，Vol.J74-D-1，No.8，p.556-565（1991）

7) 絹川博之ほか．日本語情報検索システムにおけるキーワード自動抽出．日立評論．Vol.64，No.5，p.74-78（1982）

8) 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.21，No.3，p.200-207（1980）

9) 細野公男ほか．日本語文章からのキーワード自動抽出．情報処理学会第35回（昭和62年後期）全国大会．5S-5，p.1277-1278（1987）

10) 日本索引家協会編. 索引作成マニュアル. 東京, 日外アソシエーツ, 1983, 237p.
11) 日本科学技術情報センター情報部. 4. 抄録作業. 情報部作業マニュアル. 東京, 日本科学技術情報センター, 1978, 4-01-1 - 4-36-2.
12) 原田隆史ほか. 抄録からの主題文の自動抽出. Library and Information Science, No. 29, p. 125-137 (1991)
13) 野美山浩ほか. 事例ベースを用いた発見的規則制御の最適化. 情報処理学会自然言語処理研究会報告, NL-89-8, p. 57-64(1992)
14) 国立国語研究所. 分類語彙表. 東京, 国立国語研究所, 1964, 362p.