情　報　学　基　礎　31－8
（１９９３．　９．　１４）

抄録からのキーワードの自動抽出

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，文中に出現する全ての語を対象として抽出を行うため，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，文の構文的特徴や文中に出現する特別な表現を元に，主題概念を表現する語のみをキーワードとして抽出することを試みた。文の構文的特徴および特徴的な表現をもとに192個の規則を作成することによって約82.1%のキーワードを正確に抽出することができた。

AUTOMATIC KEYWORD EXTRACTION FROM ABSTRACTS

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., Shimotsuruma, Yamato-shi, Kanagawa.

In order to automatically extract good free keywords for content designation from abstracts, it is necessary to analyze characteristics of subject bearing keywords in abstracts.

This paper, first of all, describes charcteristics of a method developed to automatically discriminate subject bearing keywords 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 reports the successful result of experiment where the method was applied to the abstracts in the field of computer science.

Ⅰ．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)]。このうち，抽出索引方式には，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)]。いずれの方法を採用した手法においても，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われている。しかし，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。

とくに，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……が問題であったので，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，従来の手法では「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。

そこで，本研究では，文献の主題内容を表すキーワードの持つ特徴を分析し，より適切なキーワードのみを自動的に抽出する手法の開発を試みた。

Ⅱ．キーワードの出現特徴の分析

Ａ．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，本文を対象とする場合に比較して分析が容易であり，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。

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。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抄録の1535文を対象に，抄録を文単位で構文解析し，抄録中のキーワードの出現特徴の抽出を試みた。なお，抄録文の解析においては，丸山らによる「日本語解析ワークベンチ(Japanese Analysis Workbench)」[5)]を使用した。

Ｂ．キーワードの決定と特徴の分析

情報検索システムへの応用という観点から考えると，抄録文すべてを分析することは適当ではない。すなわち，抄録中によくあらわれる表現であり，かつ解析によってキーワードの決定に貢献できる可能性がある表現のみを対象として，分析を行うことが望ましい。

JICST科学技術文献ファイルに収録された電気工学分野の抄録文約13000件を対象に，「設計」および「コンピュータ」という語の直後にどのような表現が出現するかを調査した結果によれば[6)]，これらの語の直後の表現として，比較的出現頻度が高く，またキーワードの決定が可能であると思われるものとしては以下の表現があげられる。

1) キーワードの直後に助詞または助詞相当表現がきているもの
2) キーワードの直後に格助詞「の」がきているもの
3) キーワードを前方要素とする複合語
4) キーワードの直後に並列の関係を示す「および」「並びに」「と」のような語が出現しているもの

また，キーワードを決定するために用いることのできる表現として上記以外に以下があげられる。

5) 時を表す語とキーワード
6) 否定形の文とキーワード
7) 形容詞，形容動詞，副詞とキーワード
8) 指示語とキーワード

そこで，本研究では1)2)および5)～8)に焦点をあてて，主題概念を表すキーワードのもつ特徴を明らかにすることを試みた。

具体的な手順は以下の通りである。

a) 抄録中の各文に含まれる語のうち，主題概念を表しているキーワードのみを人手で抽出する
b) 抄録中の各文を構文解析して，構文解析樹を作成する
c) 構文解析樹上である語と他の語がどのように接続されているかを確認し，その接続関係を元に主題概念を表すキーワードの特徴を分析する。

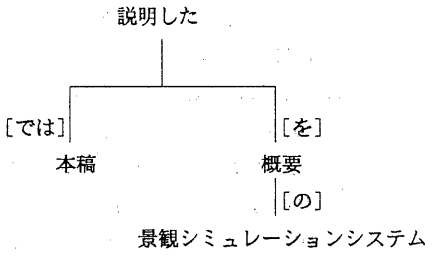
C．助詞または助詞相当表現とキーワード

日本語において単語の意味的役割をあらわすものとして，単語の後に続く助詞や助詞相当表現（以下は助詞等と記す）がある[7]。そこで，キーワードと接続される助詞等の持つ特徴を分析することにより，キーワードの特徴を分析した。この際，単に助詞等との関係だけではなく助詞等で結ばれる述語動詞についても注目することとし，助詞等および述語の組み合わせとキーワードとの関係について分析を行った。

具体的には，出現回数の多い助詞「を」，「が」，「は」および助詞相当表現「について」を抽出し，それぞれの助詞あるいは助詞相当句ごとにどのような動詞に掛かっているかを分析した。さらにそれらの助詞等と動詞の組み合わせにかかる語がキーワードであるかどうかを調べた。こうして，ある助詞あるいは助詞相当表現と動詞の組合わせがかかった場合，その語がキーワードとなるのかならないのかを分析した。

この分析の際，助詞等の直前に出現する語がキーワードであるかどうかという分析だけではなく，構文解析樹の上で助詞等に対する葉の部分に出現する語すべてについてキーワードであるかどうかの判断を行った。

たとえば「本稿では，景観シミュレーションシステムの概要を説明した」という文の場合，「～を説明した」と直接接続されている語は「概要」であり，キーワードである「景観シミュレーションシステム」とは直接の接続関係はない。しかし，この論文においては，「景観シミュレーションシステム」を「説明した」ことは明らかである。したがって，キーワードと「～を説明した」との間には接続関係がないと判断することはキーワードを正しく抽出できない原因となる。このように，キーワードが助詞等と直接は接続されていなくても，意味的には接続関係があると考えた方がよい場合が数多く見られた。そこで，ある助詞等が直接接続されている語を含む構文解析樹の枝の下までを対象としてキーワードとの接続関係があるかどうかを判断した。したがって，第１図の場合，「概要」という語だけでなく「景観シミュレーションシステム」という語も「～を説明した」と接続されていると考える。



第１図　構文解析樹と助詞等と接続される語の範囲

助詞等と動詞の組み合わせごとに，助詞等の前の語がキーワードであるかどうかについて調べた結果を第１表に示す。ここで，動詞については国立国語研究所の分類語彙表[8]に基づいてグループ化を行った。

第１表に見られるように，助詞相当表現「について」の場合には，ほぼすべての動詞についてその直前の語がキーワードであった。また，助詞「を」，「が」，「は」については動詞の分類コードによってキーワー

ドとなるかどうかを判断することができた。

ただし，同じ分類コードを持つ動詞であっても各動詞ごとに違うキーワードとの接続関係を持つ例も存在した。このような動詞の例を第１表中に＊をつけて示す。たとえば，「を」と分類コード(2.15)の動詞の場合，動詞が「改善する」ではキーワードと接続される割合が83.3%であるのに対し，「減少する」では0%と大きな違いが存在した。

第１表　助詞等＋動詞の組み合せとキーワード

○：キーワードと接続された数

×：キーワードと接続されなかった数

比率：キーワードと接続される割合

助詞等	(分類コード)　動詞	○	×	比率
について	(2.15)統轄する　等	4	0	100.0%
	(2.30)考察する　等	22	0	100.0%
	(2.31)記述する，解説する	120	1	99.2%
	説明する　等			
	(2.35)紹介する	10	1	90.9%
を	(2.15)進む，減少する　等	142	112	55.9%
	＊改善する	5	1	83.3%
	＊～化する	29	6	82.9%
	＊分ける，分離する	2	7	22.2%
	＊強化する	0	4	0%
	＊変化する	0	3	0%
	＊減る，減少する	0	6	0%
	(2.30)分かる，決める　等	263	113	69.9%
	＊試みる	7	0	100.0%
	＊分析する	7	0	100.0%
	＊検討する	25	1	96.2%
	＊測定する	9	2	81.8%
	(2.31)記述する　等	145	28	83.8%
	＊提案する	15	0	100.0%
	＊論じる	9	0	100.0%
	＊述べる，記述する	37	2	94.9%
	＊解説する	12	1	92.3%
	＊説明する	33	4	89.2%
	(2.34)行う，実施する　等	62	36	63.2%
	(2.35)紹介する　等	53	18	74.6%
	＊紹介する	45	6	88.2%
	(2.38)用いる，開発する等	188	68	73.3%

助詞等	(分類コード)　動詞	○	×	比率
が	(2.12)実現する，できる等	47	47	50.0%
	＊存在する	0	5	0%
	(2.15)進む，減少する　等	23	34	40.3%
	＊減る，減少する	0	7	0%
	(2.30)分かる，決める　等	13	28	31.7%
	(2.34)実施する　等	6	11	35.3%
	(2.37)与える，持つ　等	8	7	53.3%
	(2.38)使う，開発する　等	20	9	69.0%
	＊開発する	10	2	83.3%
	(3.12)可能な，必要な　等	10	32	23.8%
	(3.19)多い，少ない　等	3	17	15.0%
は	(2.11)基づく，関係する等	13	16	44.8%
	(2.12)依存する，できる等	17	17	50.0%
	(2.15)進む，減少する　等	32	33	49.2%
	(2.30)示す，図る　等	28	28	50.0%
	＊示す	1	7	12.5%
	＊図る	5	0	100.0%
	(2.31)説明する，述べる等	14	1	93.3%
	(2.34)行う	11	8	57.9%
	(2.35)紹介する　等	5	9	35.7%
	(2.37)持つ，有する　等	17	5	77.3%
	＊提供する	7	1	87.5%
	(2.38)用いる，開発する等	28	14	66.7%
	＊利用する	5	0	100.0%
	(3.12)可能な，必要な　等	25	20	55.6%
	(3.19)大きい，多い　等	8	17	32.0%
	(3.37)重要な，大切な　等	6	9	40.0%

D．語＋「の」＋名詞の組み合わせとキーワード

たとえば，「本論文ではINDEXERシステムの概要を説明した」という文では，「INDEXERシステム」という語がキーワードであるが，このキーワードの後ろの「～の概要」という表現がキーワードであることを強く示唆している。この「概要」という語のように，キーワードではないが，キーワードを探す手がかりとなる名詞が存在すると考えられる。このような例としては以下における下線部の表現があげられる。

・ここでは特性データ交換と関係するデータ・ネデリング活動の現状について討議する。
・ビジュアルシステムの特徴，宇宙関連シミュレーションなどについて記述

第２表　語＋の＋名詞の組み合わせ

名詞の分類	名　　詞	語がｷｰﾜｰﾄﾞである割合
概念	概念(14/14)，概要(18/20)	94.1%
特徴	特徴(12/12)，特長(4/6)	88.9%
動向	動向(18/20)，傾向(3/4)	87.5%
使用	利用(22/26)，使用(5/7)，応用(10/13)	80.4%
開発	構築(7/7)，試作(3/3)，改善(8/14)，開発(35/46)，作成(4/4)，更新(4/5)	77.2%
現状	現状(10/13)	76.9%
評価	効果(5/6)，効率(3/4)，性能(4/4)，評価(17/25)，問題点(10/12)，有用性(3/4)	76.4%
例示	事例(6/9)，例(14/18)	74.1%
手法	手法(6/9)，技法(5/5)，法(23/33)	72.3%
研究	研究(17/26)，分析(5/7)，解析(13/16)，試験(6/10)	69.5%

（ ）内は，分母にその語が出現した回数，分子にその語の前の語がキーワードであった回数を示す。

そこで，抄録中の「の」＋名詞という表現をすべて抽出し，「の」の前の部分にキーワードがくる可能性の高い名詞があるかどうかを調べた。その結果，第２表に示す名詞を抽出することができた。

E．時を表す語とキーワード

一般に，抄録中で過去の研究の問題点に触れている場合にはキーワードの出現する可能性は低いと考えられる。反対に，その論文であつかった研究において開発したシステムなどについて説明している部分は，現在のこととして述べられていることが多く，キーワード出現の可能性も高い。

そこで，「時」に関する語とキーワードの出現の関係について調べた。まず，「時」を表す語を抜き出し，次いで構文解析樹においてその語が接続されている動

第３表　時を表す語とキーワード

時制	時を表す語	ｷｰﾜｰﾄﾞを含む割合
現在	最近(11/14)，現在(10/10)，現代(2/2)，最新(6/7)，始めに(3/3)，初期(2/3)，～後(4/5)，その後(2/3)，～以降(1/1)，～以来(2/2)	86.0%
未来	将来(4/6)，未来(1/1)，今後(5/6)，次代(1/1)，新世代(1/1)，第２世代(1/1)	81.3%
過去	過去(3/6)，以前(0/2)，従来(2/19)，～年前(1/3)，具体的な年月日(5/20)，過去～年間(1/3)，当時(0/1)，～年間(6/14)，～年代(2/6)，～時代(2/5)	27.8%

（ ）内は，分母にその語が出現した回数，分子にその語の前の語がキーワードであった回数を示す。

詞から下のサブツリーにキーワードが含まれているかどうかを調べた。

第３表に，抜き出した「時」を表す語を現在，過去，未来に分けてそれぞれの語の出現した回数と，そのうちキーワードを含んでいた回数を示す。

第３表に見られるように，「最近」，「現在」のように現在を表す語と，「将来」，「今後」のように未来を表す語が出現すると，キーワードを含む可能性が高いことが明らかとなった。逆に，「以前」「従来」「～年～月～日」のように過去を表す語が出現すると，キーワードを含む可能性が低くなることが明らかとなった。

F．否定形の文とキーワード

たとえば，「本論文ではＡ方式については触れていない」といった場合，「Ａ方式」という語はその論文の主題を表しているとはいえず，キーワードとはならない。このように，文が否定されている場合，その文の中にはキーワードは含まれない可能性がある。そこで，否定形の文と，キーワード出現との関係に着目して分析を行った。すなわち，「～ない」という表現を抽出し，構文解析樹でこの表現と接続される語がキーワードであるかどうかを分析した。

その結果，否定の形をとる表現は全部で５１個あり，そのうち４１個(75.9%)がサブツリーにキーワードを全く含んでいなかった。

また，否定形の述語の下のサブツリーに存在する文節の数と，そのうちキーワードを含む文節の数をかぞえた。分析対象とした296文節のうち，キーワードを含む文節は２２(7.4%)であった。今回分析の対象とした抄録全体の文節の数は12168で，そのうちキーワードを含む文節は3869(31.7%)であった。全体でキーワードの割合が約30%であるのに体して，否定の方の部分ではキーワードの割合は10%にも満たない。このことからも，否定の形の文の中には，キーワードは出現しにくいと判断できる。したがって，否定の形での文であるかどうかという情報もキーワード出現の特徴として用いることができるだろう。

G．形容詞，形容動詞，副詞とキーワード

主題を表現するキーワードにおいては，その語を協調する目的から形容詞等が修飾する表現が用いられることがある。

そこで，形容詞，形容動詞，副詞とキーワードとの関係に着目して分析を行った。具体的には，形容詞，形容動詞，副詞をすべて抜き出し，それらが修飾する語がキーワードであるかどうかを調べた。その結果，比較的キーワードを修飾しやすいと判断できるものとして，形容詞「新しい(30回の出現のうち22回がキーワードを修飾)」，形容動詞「最適な(15回の出現のうち12回がキーワードを修飾)」を見いだすことができた。また，副詞「特に」については，「特に～した」という表現として用いられ，９回の出現すべてにキーワードが含まれていた。これら３つの語もキーワード出現の特徴として利用できると考えられる。

H．指示語とキーワード

字数の限られた抄録の中では，同じ表現の繰り返しを避けるために，「これ」，「それ」，「これら」といった指示語がよく用いられる。限られた字数の中で繰り返し使われる語は重要度が高いと考えられることから，指示語で示される語はキーワードである可能性が高いと思われる。

そこで，まず抄録中から指示語を抽出し，それらの指示語がキーワードを指しているのかどうかを調べた。その結果を第４表に示す。第４表に示すように，「この」，「その」，「それら」で示される語については，キーワードを指す可能性が高いといえることが明らかとなった。しかし，「これ」，「それ」，「これら」で示される語については，数も少なくキーワードであるかどうかを判断することはできなかった。

第４表　後＋の＋名詞の組み合わせ

その　(91/101: 90.1%)	これら(6/10 : 60.0%)
それら(7/8　: 87.5%)	それ　(4/7　: 57.1%)
この　(77/90 : 85.6%)	これ　(12/26 : 16.1%)

Ⅲ. キーワードの自動抽出

A. 抽出方法

Ⅱ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"改善する"という動詞であり，かつ，この動詞と「と」という助詞で接続される語であるならば，その語はキーワードである」という規則には，第２表に示す実際の分析結果に対応して，0.833という数値が付与されることになる。これらの規則の中にはキーワードを抽出できる確率が低い規則も含まれているが，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，本研究では野美山らによるシステムを用いて条件部を組み合わせ，911個の組み合わせ規則を作成してこれも用いた[9]。

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。

B. 抽出結果

規則を作成するのに用いた抄録とは別の152抄録717文を対象として，上記のシステムによってキーワードの自動抽出を行った。その結果を第６表に示す。

第６表　キーワード自動抽出の結果

システム / 人手	ｷｰﾜｰﾄﾞ	非ｷｰﾜｰﾄﾞ	計
ｷｰﾜｰﾄﾞ	1375	300	1675
非ｷｰﾜｰﾄﾞ	1199	2640	3839
計	2574	2940	5514

第６表に示すように，人間がキーワードと判断した1675語のうち1375語(82.1%)がシステムによってもキーワードだと判断された。また，システムがキーワードだと判断した2574語のうち人の判断と一致したキーワードの割合は，53.4%であった。

本システムがキーワードの意味解析をせずに構文解析のみを用いていることを考えればこの結果はかなり高い値であると考えられる。

Ⅳ. おわりに

本研究では，文の内容を適切に表すキーワードが文中のどのような構造上に出現するかを分析し，文の構造を利用したキーワード抽出を行った。

その結果，人間が主題内容を表すキーワードと判断した語のうち80%以上の語を正しくキーワードと判断することができた。このことは，語の構文的な特徴および文中に出現する特徴的な表現をもとにして主題内容を表すキーワードを抽出する手法の有効性を示すものと考えることができる。

しかし，システムが抽出したキーワードのうち，人間もキーワードだと判断した割合は半分程度にとどまった。このことは，不必要な語をキーワードとして抽出してしまうことを意味しており検討課題であろう。今後，以下の点を中心に研究を継続する予定である。

1) 規則の再検討と追加

キーワードの出現する条件あるいは出現しない条件に関する規則を追加する必要があるだろう。そのためには，まず今回の分析における観点を再検討し，より詳細な検討を行うことが必要であろう。

たとえば，今回は助詞相当表現として「について」のみを取上げたが，ほかの助詞相当表現についても分析してみる必要があるだろう。そのためには，より多くの抄録を分析する必要がある。また，同一抄録中における出現頻度などの情報も加えてみるなどの対策も考えられよう。

2) キーワード抽出の対象となる文を限定する

本研究では，抄録中のすべての文を対象としてキーワードの抽出を行ったが，今後は対象を主題を表す文に限定して検討するなどの方法も考える必要があろう。たとえば，以下の例では，後半の文が主題を表していると考えられる。すなわち，前半の文は論文で述べることの前提として一般的な話をしているもので，この論文の内容を表すキーワードは存在しないと考えられる。

「通信や制御システムが故障すると，オペレータは艇の情報や制御を失うが，故障に耐性があるコンピュータシステムは誤差や欠陥を克服して指定さぎょうを行う。故障対処法，ＲＯＢシステムの概要，故障耐性向上法などを述べる。」

しかし，現在のシステムでは，「通信」「制御システム」「コンピュータシステム」「指定作業」という語がキーワードであると間違って判断されることになる。

このノイズを取り除くためには，前半の文をキーワードの対象としないことが必要となる。著者らは抄録中の主題を表す文を抽出する研究を現在行っており[10]今後は，これらの研究をまとめていく予定である。

謝辞

本研究を行うにあたり御協力いただいた関根さゆり氏（日立製作所）に感謝します。

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) Maruyama，Hiroshi et.al．An Interactive Japanese Parser for machine translation．COLING'90，Vol.2，p.257-262（1990）
6) 原田隆史ほか．構文解析に基づくキーワードへのロール決定方法の高度化．第27回情報科学技術研究集会発表論文集．東京，日本科学技術情報センター，1990，p.37-42．
7) 長尾真編．言語の機械処理，講座現代の言語．東京，三省堂，1984．p.67-71．
8) 国立国語研究所．分類語彙表．東京，国立国語研究所，1964，362p．
9) 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)
10) 原田隆史ほか．抄録からの主題文の自動抽出．情報処理学会情報学基礎研究会報告，FI-29-3，p.17-26(1993)