情　報　学　基　礎 33－7
（１９９４．　５．　１８）

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，1)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こと，2)文の構造的特徴や文中に出現する特別な表現をもとに，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という２つの手法から適切なキーワードのみを抽出することを試みた。主題文の抽出およびキーワードの持つ特徴を表現する規則を作成することによって必要なキーワードのみを抽出できることが確認できた。

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 must be effective and efficient to single out subject bearing sentences, and to analyze characteristics of subject bearing keywords in abstracts.

This paper, first of all, describes characteristics of a method developed to automatically discriminate subject bearing sentences and keywords from those which show premises and conclusions, based on the particular expressions appeard 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)5)]。現在，実用に供されているシステムの多くは不要語辞書を用いる方式を採用している。また，構文解析を用いる方法も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。

しかし，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。

そこで，発表者らは適切なキーワードのみを自動的に抽出することを目的に，これまで文献の主題内容を表す文を抽出する研究および，キーワードの持つ特徴を分析する研究を行ってきた[6)7)]。本研究では，これらの成果をまとめ，自動抽出されるキーワードの精度を高めることを目的として実験を行った。

Ⅱ．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。

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文を対象に実験を行った。

Ⅲ．抄録中からの主題文の自動抽出

一般に，抄録中には「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」を示す内容が記述される[8)9)10)11)]。

しかし，これら６つの項目がすべて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，先行事例について述べた部分である。これを「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以外に先行研究の内容に関するものも含まれることになる。

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」「5)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。ここでは，これらをまとめて「結果文」と総称するが，この結果文中に出現するキーワードは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ことがある。

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，検索もれを最小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。これを「主題文」と呼ぶ。キーワードの抽出は，この主題文を対象にすることが望ましいと考えられる。

このような抄録中の前提文，主題文，結果文を正しく判断するために，過去において発表者らは実際の抄録中の各文の持つ特徴の分析を行い，各文について以下の４つの特徴を明らかにした[6]。

1) 文と文とをつなぐ接続詞

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結果，「また」「さらに」「そして」「しかし」を含む文の場合には，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ことを明らかにした。また，「そこで」で接続される場合，接続詞の前の文は前提文，後の文は主題文である可能性が高いことが明らかにした。

2) 複文の中で用いられる接続表現

複文においては，文の前半と公判とを結ぶ表現によって接続関係が判断可能であることを明らかにした。すなわち，複文が，方法，目的，条件の意味を表す接続表現をとる場合には，主題文である可能性が高いと考えられる。

3) 主題文中に特有の表現

主題文に特有の表現として，以下の表現があることを明らかにした。

・標題を示す語
・「本研究」「本文」「本論文」などの表現
・「ここでは」という表現

4) 文末表現

日本語においては，文末の表現は1)文末文節の助詞また助詞相当表現，2)文末文節の語幹，3)文末文節の語尾とう３つの構成要素から構成される。これらのうち，「文末文節の直前の助詞または助詞相当表現」については，「～について」「～につき」「～に関して」「～かを」という表現が出現した場合に主題文である可能性が高いことを明らかにすることができた。

「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い，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。たとえば，「談話」というカテゴリー動詞（たとえば「述べ」という動詞）が出現した場合，それは主題文である可能性が高い。また，「会議・議論」というカテゴリーの動詞も主題文に出現する可能性が高いことが明らかとなった。

さらに，「文末文節の語尾」にも文のカテゴリーごとに特徴が見られた。たとえば，「～であろう」などのような推量を示す付属語や，「～できる」などの可能性を示す付属語は結果文に出現する。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。

これらの特徴から規則を設定し，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することができた。

Ⅳ．キーワードの出現特徴の分析

キーワードの出現特徴を分析するために以下の手順で実験を行った。

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と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。

その結果，以下の６つの点からキーワードの特徴を明らかにすることができた。

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について

述語動詞および助詞（または助詞相当表現）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，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」については動詞の種類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。

2）キーワードの直後に各助詞「の」がきているもの

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，概念，特徴，動向，使用等のカテゴリーに属する名詞と結ばれる語がキーワードである可能性が高いことを明らかにした。

3）時を表す語とキーワード

「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結果，「最近」，「現在」，「最新」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。

4）否定形の文とキーワード

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た。

5）形容詞，形容動詞，副詞とキーワード

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」，形容動詞「最適な」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」という表現のばあいには，キーワードが含まれている可能性が高いことが明らかになった。

6）指示語とキーワード

「この」，「その」が，キーワードを指す可能性が高いことが明らかとなった。

これらの特徴から規則を設定し，これをもとにして抄録からキーワードを機械的に抽出する実験を行った。実験の評価は，システムが抽出したキーワードのうちどのくらいの割合が人手で抽出したものと一致したかという値と，人出で抽出されたキーワードのうちどのくらいの割合をシステムが抽出できたかという値から判断した。この研究では，人間がキーワードと判断した語のうち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語のうち人間の判断と一致したキーワードの割合は，53.4%であった。

V．主題文中のキーワードの抽出

Ⅲ章およびⅣ章で示した主題文の特徴およびキーワードの特徴を元にしてキーワードの自動抽出を試みた。主題文の自動抽出実験結果とキーワードの自動抽出実験を組み合わせる場合，その組み合わせ方として以下の２つが考えられる。

1）主題文中のキーワードのみを抽出する。

2）主題文である条件をキーワードである条件と同様のものとみなしてキーワードの自動抽出を行う。

A．主題文中のキーワードのみの抽出

JICST科学技術文献ファイル電気工学編に含まれる152抄録717文のうちシステムが主題文であると判断した408文のみを対象としてキーワードの自動抽出を行った。その結果を第１表に示す。

第１表　主題文からの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	756	208	964
非ｷｰﾜｰﾄﾞ	428	1527	1955
計	1184	1707	2919

第１表に示す通り，人間がキーワードと判断した964語のうち756語(78.4%)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1184語のうち人間の判断と一致したキーワードの割合は，63.9%であった。これは，キーワー

ドを正しく抽出できない場合も増加するが，誤って非キーワードであると判断することも少なくなることを意味している。

B．主題文である条件をキーワードである条件と同様のものとみなしてキーワードの自動抽出

Ⅲ章に示した研究の準備段階で，キーワードが主題文中に多く含まれていることが明らかとなった。そこで，主題文を抽出するために用いた規則をキーワードを特徴づける規則と同等のものとみなしてキーワードを抽出する実験を行った。その結果を第２表に示す。

第２表　主題文からのキーワード自動抽出の結果

人手＼システム	ｷｰﾜｰﾄﾞ	非ｷｰﾜｰﾄﾞ	計
ｷｰﾜｰﾄﾞ	1251	424	1675
非ｷｰﾜｰﾄﾞ	1084	2755	3839
計	2335	3179	5514

第２表に示す通り，人間がキーワードと判断した1675語のうち1251語(74.7%)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2335語のうち人間の判断と一致したキーワードの割合は，53.5%であった。これは，主題文であるという規則を追加した結果として正しく抽出されないキーワードが増加することを意味しており，主題文を抽出するための規則をキーワード抽出のための規則に追加して使用することはあまり効果的ではないことが明らかとなった。

Ⅵ．おわりに

私たちは，今までキーワードの持つ構文的な特徴および特徴的な表現に基づいてキーワードを自動抽出する手法と，主題文であるかどうかを自動的に判断する手法について独立に実験を重ねてきた。

第Ⅴ章に示すように，この両者を単純に組み合わせるだけでは必ずしも適切なキーワードのみを抽出することができるとはいえないが，今回の実験で「コンピュータ」「システム」という語のように文の構造や表現上の特徴だけではキーワードと判断されてしまう語が主題文中に出現するものという制限をかけることによってキーワードと判断されない例も見つけることができた。このような場合には検索ノイズの低減に役立つと考えられる。

このような例は他にも存在すると考えられる。どのような特徴を持つキーワードの候補が文の種類による制限を受けるのかについては今後の検討課題となろう。

また，本研究では，キーワードであるかどうかという観点から語の分類を行っているが，キーワードかどうかという判断だけではなく，キーワードである可能性を計算するということも有効であろう。

特に研究の初期の段階において情報検索を行った場合，検索モレよりも検索ノイズが問題になることがある。また，利用者が思いついたキーワードで検索を行った時に非常に多くの文献がヒットしてしまってどの文献から読み始めていいのか判断に困るということが起こりえる。

このような場合に，システムがより適合度が高いと考えられる文献から順に出力することができれば利用者の情報要求を満たすことができると考えられる。本システムでは，キーワードであるかどうかを判断するのに語がキーワードであるかどうかを数字で算出する方法を採用している。また，主題文であるかどうかの判断についても同様である。この算出された値を適合度とみなして適合度順出力をするという応用も考えることができよう。適合度順出力への応用については，早急にとりあげたいテーマであり，近日中に実験を行いたいと考える。

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. 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)
7) 原田隆史ほか. 抄録からのキーワードの自動抽出. 情報処理学会情報学基礎研究会報告. FI-31-8, p.55-62(1993)
8) 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. 36p.
9) 中村幸雄. 講座 論文と抄録の書き方5. 情報の科学と技術. Vol.39, No.9, p.353-360 (1989)
10) 溝口歌子. 抄録法. ドキュメンテーション研究. Vol.23, No.5, p.157-163 (1973)
11) Liddy, E.D. The Discource-level Structure of Empirical Abstracts: an Exploratory Study. Information Processing & Management, Vol.27, No.1, p.55-81 (1991)