Library and Information Science No. 29 1991

抄録からの主題文の自動抽出

Automatic Selection of the Subject Bearing Sentences from Abstracts

原田隆史, 丸山 宏, 佐藤元美,
Takashi Harada, Hiroshi Maruyama, Motomi Satoh,
細野公男, 諸橋正幸
Kimio Hosono, Masayuki Morohashi

Résumé

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

This paper, first of all, describes characteristics of a method developed to automatically discriminate such sentences from those which show premises and conclusions, based on the relative position of the sentences in texts and the paticular expressions appeared in them.

Then this paper reports that the result of the experiment where the method was applied to the abstracts in the field of computer science. In 286 out of 386 abstracts (69.4%), almost all sentences include in them were classified successfully. Furtheremore, some of the sentences were recognized conectly in 65 abstracts. As far as the first sentence is a subject bearing one, it was identified so with very high probability i.e. 97.1%.

I. はじめに

A. キーワード検索方法の改善

B. 抄録中の主題を表わす部分

原田隆史：慶應義塾大学文学部図書館・情報学科助手
Takashi Harada, Teaching Assistant, School of Library and Information Science, Keio University, Mita, Minato-ku. Tokyo.
丸山 宏：日本 IBM 東京基礎研究所
Hiroshi Maruyama, Researcher, Tokyo Research Laboratory, IBM Japan Ltd., Sanbancho, Chiyoda-ku, Tokyo.
佐藤元美：東京海上火災
Motomi Satoh, The Tokio Marine and Fire Insurance Co. Ltd., Marunouchi, Chiyoda-ku, Tokyo.
細野公男：慶應義塾大学文学部図書館・情報学科教授
Kimio Hosono, Professor, School of Library and Information Science, Keio University, Mita, Minato-ku, Tokyo.
諸橋正幸：日本 IBM 東京基礎研究所
Masayuki Morohashi, Researcher, Tokyo Rerearch Laboratory, IBM Japan Ltd., Sanbancho, Chiyoda-ku, Tokyo.
1992年2月28日受付

I. はじめに

A. キーワード検索方法の改善

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，キーワードを論理演算子で結び付けることによって検索式を作成し，検索が行われる。たとえば「コンピュータを用いたロボットの設計に関する文献を検索したい」という検索質問の場合には，キーワードとしてコンピュータ，設計およびロボットという語を用い，これらのキーワードを論理積で結び付けて「コンピュータ AND ロボット AND 設計」という検索式で表現される。これらのキーワードと文献の索引語とを照合し，「コンピュータ」「ロボット」および「設計」という3つの索引語を持つ文献が検索されることになる。

文献の索引語を決定する方法としては，付与索引方式と抽出索引方式の2通りがある[3)]。付与索引方式は，文献の主題分析を行って主題概念を的確に表現する索引語を決定し，これ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索引語は，シソーラスなどであらかじめ用意されたものから選択されることが多い。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，索引作業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索引語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コンピュータを用いた自動化の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，用語辞書・不要語辞書を用いる方法は，DIALOG などの商用オンライン情報検索システムの多くで実際に用いられている索引手法である[3)]。また，構文解析を用いる手法についても，自然言語処理技術の発達にともなって近年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)]などがある。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。

たとえば，INDEXER では，まず不要語を除去することによってキーワードの候補となる語を抽出し，その候補となる語が持つ論文上での表現の特徴や，頻度情報などに関する規則（連体修飾語はキーワードとしないなど）をもとに絞り込む方法が採用されている[6)]。絹川らは，外電記事文中から文節を切り出し，名詞文節と述語の関係，名詞文節内の名詞の意味分類，名詞文節中の格助詞の情報を用いてキーワードを抽出するとともに，そのキーワードにロール（主体，客体，時，場所，活動など）を付与することを試みている[7),8)]。また，細野らは，キーワードになりえない句の統語パターンを記述し，これを用いて文章中から不要な句を削除する方法でキーワードを抽出している[9)]。これらの構文解析を用いる手法は，いずれもまだ実験段階であるが，用語辞書や不要語辞書を用いる方法に比較してキーワードを抽出する精度を高めることを可能にしている。

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て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) キーワードの持つ重要性に差が生じる可能性がある。

中でも 1) は大きな問題である。

たとえば,「コンピュータを設計の対象として扱っている文献」という検索質問を「コンピュータ AND 設計」という検索式で表わし,抄録中の文と照合して検索することを考える。この検索式中のキーワードと,「設計用のコンピュータのしくみについて述べる。しかし,コンピュータを設計することについては触れていない」という抄録から抽出された索引語との照合を行う場合,この抄録の後半部分に出現する「コンピュータ」という語は設計の対象であるため,上記の検索式に対する適合文献と判断されることになる。しかし,抄録中の「しかし」以降の文は,この文献では扱っていないことを記述する文であり,文献の主題とは関係がない。このように,すべての文を対象として索引語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。

この問題を解決するためには,当該論文が伝えようとしている主題を適切に表現するキーワードを抽出する手法を考える必要がある。

索引作成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索引し,他人の業績,将来のことについては索引しないと規定されている。この原則に基づいて索引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索引の対象とせず,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。

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部分であることから,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。そこで,本研究においても,抄録を対象として分析を行うこととする。

B. 抄録中の主題を表わす部分

抄録作成機関である日本科学技術情報センター (JICST) 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以下があげられている[11]。

1) 前提説明
   研究・開発・調査などの経緯,背景,定義など
2) 目的・主題範囲
   研究,開発,調査などの目的,理由,取り扱っている主題の範囲
3) 方法論
   研究,開発,調査などに用いた理論,原理,条件,対象,材料,手段,方法,手順,精度など
4) 結果
   研究,開発,調査などから得られた知見,すなわち,実験的・理論的結果,得られたデータ,認定された関係,観察結果,得られた効果・性能など
5) 考察・結論
   結果の分析・検討・結果の比較・評価,問題提起,今後の課題,仮説,応用,示唆,勧告,推論,予測など
6) 注記
   研究・開発・調査の主目的外であるが価値のある知見や情報で重要と思われるもの

上記の 6 項目は,JICST の抄録だけに特有な内容というわけではない。中村は,報知的抄録の作り方として,原論文の論じている範囲と目的,研究実行の手段と方法,結果の 3 点を記述するとしている[12]。また,溝口は,抄録の記述の内容として,目的,方法,結果,結論をあげている。なお,溝口は結果と結論については,一緒に記述してもよいが,事実と推測とは区別すべきであることも指摘している[13]。この 2 人があげた内容は,JICST の情報部作業マニュアル中の 2) 目的,主題範囲,3) 方法論,4) 結果について詳しく書くようにという指針と重なっている。さらに,Liddy は,抄録作成の専

門家へのアンケート調査と，既存の抄録の分析を行い，英語論文の抄録の構造を明らかにしている[14]が，その構造は上記の 6 つの項目とほぼ一致している。中村，溝口，Liddy らの論文から，JICST の情報部作業マニュアルに記述された 6 項目が，通常，抄録で述べる内容をほぼ網羅するものといえよう。

ただし，6 つの項目のうち，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。本研究では，これらの文をまとめて結果文と総称する。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ため，そのキーワードを検索対象とすることは検索ノイズの原因となる。

抄録の中には，その研究の内容や主題ではなく，過去の研究の問題点や研究の背景について記述しているものが存在する。これは，上記の「1) 前提説明」を示す部分に該当する。本研究ではこの文を前提文と呼ぶことにする。前提文中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ため，そのキーワードを検索の対象とすることは検索ノイズの原因となる。

「2) 目的・主題範囲」については，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るとされる。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。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。したがって，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。これらの文を主題文と呼ぶことにする。

以上のように，本研究では抄録中の文を「前提文」「主題文」「結果文」の 3 つに分類して分析を行った。

主題文からキーワードを抽出することの有効性に関しては，梅田の調査がある[15]。この調査では，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 文について「前提文」「主題文」「結果文」中に含まれるキーワードの数を調べており，1060 文中に出現した 2866 個のキーワードのうち 2603 個 (90.8%) が上記の主題文中に含まれることを明らかにしている[15]。約 10% 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，索引語の決定にあたって，必ずしも主題文からのみ語を抽出すればよいということはないが，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付加するなどの手法が有効であることを示しているといえよう。

実際の抄録中に含まれる各文がどのような役割を持っており，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は，人手では比較的容易に判断することが可能である。しかし，そのためには多くの時間と費用が必要とされる。この判断を自動的に行うことができれば，検索ノイズを低減することが可能になると考えられる。現在まで，コンピュータでこのような文の判断を行うための研究はあまり行われておらず，自動的に文の種類を判断するための規則も明らかとはなっていない。そこで，本研究では実際の抄録を人手によって，前提文，主題文，結果文に分類し，これら各文の持つ表層的な特徴を分析し，この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手法の開発を試みた。

II. 前提文，主題文，結果文の特徴の分析

A. 分析対象

本論文で分析の対象にする抄録は，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ものであることが望ましい。主題文以外についても分析を行うため，主として原文献が取り扱っている主題を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，本研究においては，前提文，主題文，結果文の持つ一般的な特徴を抽出することを目的とするが，一般的な特徴を把握するためには多くの抄録を処理する必要がある。そこで，本研究では，分析対象として，JICST 科学技術文献ファイル電気工学分野に含まれる 1988 年 4 月分の抄録を採用した。

JICST 抄録作成マニュアルの「抄録に盛り込む内容」に，「前提説明」の部分は簡単に触れる程度か，省略しても良いと記述されているように，あまり短い抄録には前提文がほとんど存在しないことが予想される[11]。本研究では，前提文，主題文，結果文というそれぞれの文が持つ特徴を抽出することを試みることから，前提文や結果文が存在しない抄録ばかりを分析対象とすることは適切ではない。そこで，200 字以上の長さを持つ 345 抄録

1630 文を対象として前提文，主題文，結果文の持つ特徴の分析を行った。なお，分析対象中の前提文の総数は 236 文，主題文の総数は 404 文，結果文の総数は 990 文であった。

B. 前提文，主題文，結果文の特徴

抄録中の前提文，主題文，結果文を正しく認識するためには，各文の表層的な特徴だけでなく，文の意味を把握する必要がある。一般に，文の意味を自動的に判断することは困難であるが，日本語においては，その助詞や助動詞相当表現といった表層的な特徴をもとに，ある程度意味解析を行えることが指摘されている[16)-19)]。

本研究では，文の表層的な特徴として以下の 3 つの点に注目し，分析を行った。

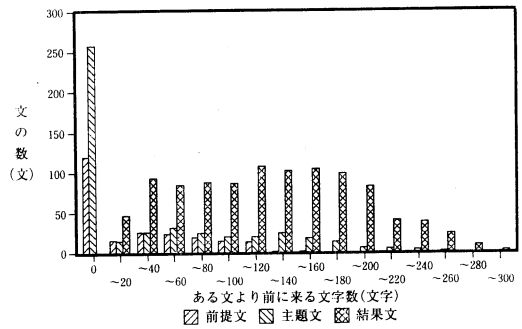
1) 抄録中での文の位置
2) 各文中の特徴的な表現
3) 文末表現

1. 抄録中での文の位置

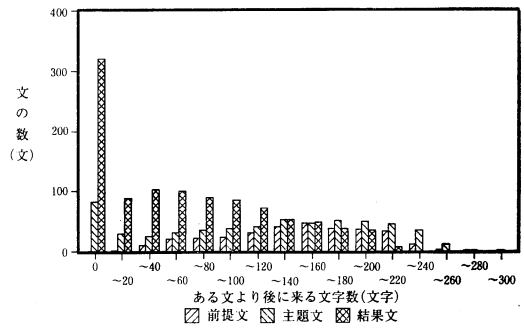
抄録中では，前提文，主題文，結果文の順に記述されることが多い。また，前提文がなく主題文が抄録の第 1 文目から書かれることもある。これらの判別に際しては各文の抄録中での位置の情報が有効であると考えられる。そこで，ある文より前にある文字の数およびある文より後にある文字の数を数えた。ある文より前の文字数についてのグラフを第 1 図，ある文より後の文字数についてのグラフを第 2 図に示す。第 1 図に見られるように，ある文より前の文字数については，主題文と結果文については主題文が抄録の先頭に出現する数が多いという点以外には大きな特徴は見られなかった。しかし，前提文については文の前に 140 文字以上の文字がある場合はほとんど存在しないという特徴が見られた。また，ある文より後の文字数については，それが 120 文字以下の場合には結果文である可能性が高いことが明らかとなった。

2. 主題文の特徴ある表現

主題文の中でよく使われる表現を見ていくと，「本研究では～，本論文では～，ここでは～」などの「本～」という表現，および「ここでは～」という表現が比較的多く出現した。それらの表現が各種類の文のなかでどのくらい使われているかを第 1 表に示す。第 1 表にみられるように，「本～」という表現は主題文では 20 文，結果文では 6 文で使われているが前提文では全く使用されていなかった。また，「ここでは～」という表現は，主題文では 9 文，結果文では 3 文出現したが前提文では全く出現しなかった。したがって，これらの表現は，前提文と主題文の識別に役立つと考えられる。



第 1 図　前提文，主題文，結果文の文頭からの位置



第 2 図　前提文，主題文，結果文の文末からの位置

第 1 表　「本～」「ここでは～」という表現の出現傾向

	前提文	主題文	結果文
「本～」	0	20	6
「ここでは～」	0	9	3
全　体	236	404	990

また，抄録作成テキストに，標題中の長い語句は，抄録中で繰り返し使用せず，「標題の」とか「標題化合物」などの語で重複を避ける[20)]とあるように，「標記の」や「標題の」「題記の」は，抄録の主題をまとめた標題の中の語を文の中で使うときに代用するものであり，これらの語は主題文中で使われることが多い。すなわち，これらの表現を含む文は主題文である確率が高いと考えられる。

3. 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) 文末文節の語尾という 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，以下のように構成要素に分けられる。

文末文節直前の助詞または助詞相当表現……について
文末文節の語幹　　……説明
文末文節の語尾　　……した

1) 文末文節直前の助詞または助詞相当表現

日本語において，単語の意味的役割を表すものとして，助詞または助詞相当表現がある[18]。吉田は自然文による構造的意味関係を担う表現は，助詞，助動詞，補助用言および，接続詞を慣用句的な複合表現にまで拡大解釈したものであるとし，それらの表現の中には文節間の関係を指示する表現があることを明らかにしている[18),19)]。

たとえば，「飯を喰う」「人について語る」における「を」「について」のような格助詞的表現，「彼しか知らない」「彼にまで疑われる」における「しか～ない」「にまで」のような副助詞的表現，「点とか線」「大きさや重さ」における「とか」「や」のような並列関係を示す表現，などがその例である。

抄録中において見られる助詞または助詞相当表現としてどのようなものがあるかについては，原田らが明らかにしている[16]。本研究では，原田らが抽出した助詞または助詞相当表現について，それが前提文，主題文，結果文のどの文に多くみられるかについて分析を行った。

その結果，助詞または助詞相当表現のうち多くの表現については，前提文，主題文，結果文のいずれかに多く出現するという傾向は見られなかった。ただし，一部の表現については出現する文の種類に大きな偏りが見られた。たとえば，「～について」については，出現した 198 文のうち主題文が 141 文 (71.2%) を占め，結果文は 56 文 (28.3%)，前提文は 1 文 (0.05%) であった。すなわち，「～について」という表現は主題文の中で特に多く見られる表現であり，主題文であるかどうかを判定するために使用することができると考えられる。

2) 文末文節の語幹

文末文節の語幹は，一般に文章の述語に相当する部分である。抄録では文の簡略化のため主語を省略することはあるが述語を省略することはない。したがって，文の特徴を明らかにする上で文末文節の語幹に出現する動詞は大きな手がかりとなることが考えられる。

そこで，前提文，主題文，結果文のそれぞれについて各文で出現する語幹を分析した。その結果，各種類の文で出現する傾向に大きな違いがあることが明らかとなった。このうち，特に主題文で多く出現する傾向がみられた語幹の例を第 2 表に示す。第 2 表に見られるように，「述べ（る）」「示（す）」「紹介（する）」などの表現が主題文に多くみられ，前提文にはほとんど使われていないことが明らかとなった。この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が可能となろう。これらの動詞は，叙述，説明，呈示，報告などを示すものと考えられる。

3) 文末文節の語尾

文末文節の語尾としては時制を示すもの，能動態，受動態の区別を示すもの，肯定，否定の区別を示すものなどが存在する。たとえば，「～する」という語は能動態，肯定形であり現在形である場合に用いられるが，過去を示す場合には「～した」となり，受動態の場合には「～される」，否定形の場合には「～しない」という表現がとられる。

このような，文末文節の語尾に関する各文の種類ごとの特徴を分析した。その結果を第 3 表に示す。第 3 表に見られるように，主題文においては「～する」や「～した」などが多く，「～ている」「～である」などは少ないという傾向が見られた。また主題文では名詞（体言止め）や動詞の原形で終わっているものが多く，形容詞で終わっているものは少なかった。

前提文においては，「～する」「～した」という表現は比較的少なく，「～ている」「～である」などの表現や「ない」のような否定を示す表現が多く出現した。また形容詞で終わっているものも多かった。結果文には「～ない」「～なかった」「～できる」などの表現や形容詞止めが多く出現する傾向が見られた。

これは，主題文が当該論文の主な内容を示す文であるところから，研究の意義を肯定的に強く表現する表現が使われるのに対し，前提文や結果文では先行研究や背景，さらに将来の展望を示すことが多いため，否定形や受身形の表現が使われるものと考えられる。

4. 前提文，主題文，結果文を特徴づける表現

以上をまとめると以下のようになる。

1) ある文より前の文字数が 140 文字以上である場合には前提文ではない可能性が高い
2) ある文より後の文字数が 120 文字以下である場合には結果文である可能性が高い
3) 「ここでは～」「本～」「表記の～」「標題の～」「題記

第2表 文末の文節の語幹の出現傾向

	前提文	主題文	結果文
述べ	1	95	20
紹介	0	26	6
説明	0	26	24
考察	0	18	8
提案	3	45	12
示	3	19	26
行	0	33	19
検討	0	27	16
調べ	0	27	18
記述	2	17	10
開発	2	9	7
解析	0	7	3
作製	0	7	3
研究	1	13	2

第3表 前提，主題，結果文の文末文節の語尾

文節の語尾	前提文	主題文	結果文
サ変名詞＋した	11	304	393
サ変名詞＋する	35	93	235
～られた・された	8	6	25
～られる・される	12	2	51
～できた	0	1	19
～できる	12	3	63
～させた	1	5	5
～させる	1	0	10
～ない	20	4	37
～なかった	1	1	7
～ている	60	10	79
～ていた	0	0	4
～である	70	17	132
～であった	1	0	35
動詞＋た	10	127	208
動詞の終止形	35	102	219
形容詞（ないを除く）	12	0	17
名詞（体言止め）	3	75	158

の～」「この論文では～」という表記が含まれる場合には主題文である可能性が高い

4) 文末文節の前に「～について」がある場合には主題文である可能性が高い

5) 文末文節の語幹が「述べ」「紹介」「説明」「考察」「提案」「示す」「行」「検討」「調べ」「記述」「開発」「解析」「作成」「研究」である場合には主題文である可能性が高い

6) 文末の語尾が「～した」である場合や体言止めである場合には主題文である可能性が高い

7) 文末文節の語尾が「～ない」「～ている」「～させる」「～される」「～できる」「～ない」「～なかった」「～である」「～であった」である場合または形容詞止めである場合には主題文でない可能性が高い

III. 機械による文の種類の自動判断

A. 機械による文の種類の判断方法

上記のように，抄録中の前提文，主題文，結果文にはそれぞれの表層的な特徴が見られることが明らかとなった。そこで，この抽出された表現をもとに文のパターン化を行い，その組合せパターンごとに前提文，主題文，結果文である確率を算出して，パターンごとに抄録中の各文の種類を判断することを試みた。文のパターン化は各文が上記の 7 つの特徴に対応する，以下の条件のそれぞれに該当するかどうかに基づき，その組み合わせで行った。7 つの条件に対して「ある」か「ない」の 2 つの場合が考えられるため，組み合わせの総数は 2 の 7 乗である 128 通りのパターンが考えられるが，分析対象とした 345 抄録 1630 文中に実際に出現したパターンは 19 通りであった。

1) ある文より前の文字数が 140 文字以上であるかどうか

2) ある文より後の文字数が 120 文字以下であるかどうか

3) 「ここでは～」「本～」「表記の～」「標題の～」「題記の～」「この論文では～」という表記が含まれるかどうか

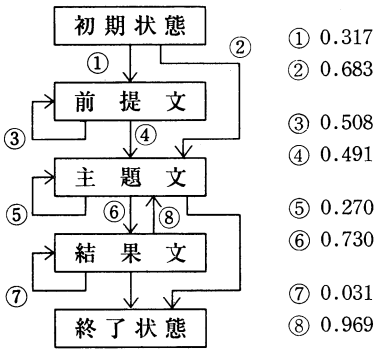
4) 文末文節の前に「～について」があるかどうか

5) 文末文節の語幹が「述べ」「紹介」「説明」「考察」「提案」「示す」「行」「検討」「調べ」「記述」「開発」「解析」「作成」「研究」であるかどうか

6) 文末の語尾が「～した」であるか，または体言止めであるかどうか

7) 文末文節の語尾が「～ない」「～ている」「～させる」「～される」「～できる」「～ない」「～なかった」「～である」「～であった」であるか，または形容詞止めであるかどうか

しかし，このように各文の表層的な特徴だけに注目して，前提文，主題文，結果文を決定した場合，結果を示



第3図　順序機械モデルと基礎遷移確率

してから研究の動機・目的を示すというような，抄録の構造としては不自然な判断がなされることもありえる。例えば，ある抄録中の第一文に対応するパターンからは結果文である確率が最も高く，第二文に対応するパターンからは前提文である確率が最も高い場合には，結果文が抄録の第一文目，前提文が第二文目ということになるが，調査した 353 抄録中にこのような例は存在しなかった。

そこでこのような問題を取り除くために，文の種類の判断を，その文の表層的なパターンにだけではなく，前の文の種類も考慮して行うものとした。すなわち，ある文の種類は当該文の表層的なパターンに基づく確率と，当該文の前の文の種類に基づく確率との積が最大となるものと判断した。この前の文の種類に基づく確率を以下では基礎遷移確率と呼ぶことにする。文の流れのモデルを第 3 図に示す。

第 3 図に示すように，第 1 文の判断にあたっては初期状態からの遷移と考える。この場合，前提文または主題文と判断されることはあっても結果文と判断されることはない。もし，第 1 文が前提文であると判断された場合，第 2 文の判断は前提文からの遷移と考えることになる。したがって，第 2 文は前提文または主題文と判断されることはあっても結果文と判断されることはない。同様に第 2 文の種類に基づいて第 3 文の判断がなされることになる。また，最終文については終了状態への遷移が可能であることが条件となる。第3図に示すように，終了状態へは主題文または結果文のみからの遷移が可能であるため，最終文が前提文と判断されることはない。

例えば，抄録の第一文が前提文か主題文かは初期状態からどちらの文へ遷移するかによって決まるが，それぞれの文への基礎遷移確率は 0.317 と 0.683 である。もし，第一文の表層的な特徴のパターンに対応する前提文の確率が 0.46，主題文の確率が 0.06，結果文の確率が 0.48 であるとすると，これらと基礎遷移確率との積をとり数値の大きい方へと遷移することになる。すなわち，前提文が 0.14582，主題文が 0.04098，結果文は 0 となるため，第 1 文は前提文と判断される。表層的な特徴だけから見れば結果文と判断される可能性が最も高いが，結果文への基礎遷移確率は 0 であるため結果文と判断されることはない。第 2 文についても前提文からの遷移となるので前提文か主題文のいずれかとなる。

B. 機械による文の種類の判断結果

1. 抄録単位での結果

上記の 345 抄録 1630 文で明らかになった基礎遷移確率およびパターンごとの各文の出現確率をもとに，JICST 科学技術文献ファイル電気工学分野の 1988 年 5 月分に含まれる 386 抄録を対象として，抄録中の前提文，主題文，結果文の機械による分類を行った。その結果を第 4 表に示す。

第 4 表に見られるように，分析対象とした 386 抄録のうち，268 抄録 (69.4%) について前提・主題・結果文を完全に判断することができた。また，主題文の一部を正しく予測できたものは 65 抄録 (16.8%) であった。主題文についてシステムの判断が，まったく予測と異なったものは 53 抄録 (13.7%) であった。

第4表　抄録中の前提文・主題文・結果文の機械による抽出結果

機械による識別と人手による判断が完全に一致	268 抄録
判断の一部分が一致	
主題文ではない文も主題文として抽出した	12 抄録
主題文の一部を抽出できなかった	53 抄録
主題文の一部が抽出できず別の文を抽出した	1 抄録
全く誤った主題文を抽出した	52 抄録

第5表 抄録の構成と判断結果

抄録の構成	抄録数	正判断抄録数*1	一部正判断抄録数*2	誤判断抄録数*3
主題文数が1文で前提文がない	240	233(97.1%)	6(2.5%)	1(0.4%)
主題文数が1文で前提文がある	75	30(40.0%)	5(6.7%)	40(53.3%)
主題文数が複数で前提文がない	49	3(6.2%)	46(93.9%)	0(0%)
主題文数が複数で前提文がある	22	2(9.1%)	9(40.9%)	11(50.0%)

*1：完全に正しく判断できた抄録の数
*2：一部の主題文について正しく判断できた抄録の数
*3：判断結果が全く誤っていた抄録の数

すべての文を正しく判断することができた 268 抄録のもつ構造パターンを第 5 表に示す。第 5 表に示すように，263 抄録については文中の一つの文のみが主題文で，主題文ではじまりその後はすべて結果文という構造，もしくは，主題文の前は前提文で後は結果文という構造を，残りの 5 抄録は，複数の主題文が存在する構造をもつものであった。これらの内容について，以下に主題文の数と位置ごとにまとめる。

a) 主題文が一つである抄録

i) 第 1 文が主題文である場合

第 1 文のみが主題文である 240 抄録中の 97.1% にあたる 233 抄録について抄録中の全ての文の種類を正しく判断することができた。また，一部の文を正しく判断できなかった 7 抄録のうち 6 抄録については，第 1 文を含む複数の文が主題文であるという判断をすることができた。判断結果が全く誤っていた抄録は，通常の抄録の構造とは異なり第 2 文および第 3 文に研究の背景が述べられている抄録であった。このような，前提文が主題文の後に位置する構造を持つ抄録は数が少ない。したがって，第 1 文のみが主題である場合には，本システムによって，ほぼ正しく主題文を判断することができるといえよう。

ii) 第 1 文以外が主題文である場合

2 番目以降の一つの文のみが主題文である 75 抄録中の 30 抄録 (40.0%) について，抄録中の全ての文の種類を正しく判断することができた。また，一部の文を正しく判断できなかった 45 抄録のうち 5 抄録については，主題文を含む複数の文が主題文であるという判断をすることができた。

主題文を正しく判断することができなかった抄録においては，前提文中の文に対して主題文であるという判断を，また主題文に対して結果文であるという判断を行っており，主題文を前提文と誤って判断するケースはみられなかった。

	文末の文節の表現	当該文の前の文字数	人手による判断	機械による判断
第1文	～系である。	0	前提文	前提文
第2文	～皆無といってよい。	30	前提文	前提文
第3文	～強くもない。	72	前提文	前提文
第4文	～提案する。	95	主題文	主題文
第5文	～考えられる。	120	結果文	結果文

第4図 連続した前提文を正しく判断できた抄録の例

	文末の文節の表現	当該文の前の文字数	人手による判断	機械による判断
第1文	～ものである。	0	前提文	前提文
第2文	～を踏む。	45	前提文	主題文
第3文	～している。	83	前提文	結果文
第4文	～考えられる。	105	前提文	結果文
第5文	～不可欠である。	120	前提文	結果文
第6文	～述べる。	150	主題文	結果文
第7文	～できる。	168	結果文	結果文

第5図 連続した前提文を正しく判断できなかった抄録の例

文末の文節の表現	当該文の前の文字数	人手による判断	機械による判断
第1文　～考察する。	0	主題文	主題文
第2文　～提案する。	52	主題文	主題文
第3文　～述べる。	92	主題文	主題文
第4文　～される。	122	結果文	結果文
第5文　～できる。	160	結果文	結果文

第6図　連続した主題文を正しく判断できた抄録の例

文末の文節の表現	当該文の前の文字数	人手による判断	機械による判断
第1文　～定義した。	0	主題文	主題文
第2文　～実行した。	75	主題文	結果文
第3文　～を包含する。	160	主題文	結果文
第4文　～される。	200	結果文	結果文
第5文　～できる。	236	結果文	結果文

第7図　連続した主題文を正しく判断できなかった抄録の例

そこで，前提文を正しく判断できた抄録とできなかった抄録の比較を行った。前提文と主題文の判断において有効と考えられる要因としては，当該文の前の文字数および，文末の文節の表現が考えられる。たとえば，前提文を正確に判断している抄録における前提文の文末表現および，当該文の前の文字数としては第4図に示すものがあげられる。

これらの表現は，主題文でないことを強く示唆するものである。このように，三つの文すべてが主題文でないことを示唆するものである場合には，前提文を正しく判断していくことが可能となる。

それに対し，第5図に示す例の場合には，前提文を正しく判断することができなかった。

第5図で，第1文および，第5文の文末は，主題文ではあまり見られない表現ではあるが，第2文目の文末は前提文よりも，主題文で，よくみられる表現である。また，第3文目の文末表現は，前提文だけではなく結果文においてもよく見られる表現である。このような場合には，システムは第2文を前提文ではなく主題文であると判断することになる。

しかし，限られた文字数の抄録の中で前提文を5文続けて述べた場合，当該論文の内容を表現する文章が制限されてしまうという問題があると思われる。また，前提文中には研究の開発などの経緯，背景，研究動機などが記述される。そこでは，研究の概要を説明するために必要な説明文や，過去の研究成果が記述されることがある。文脈間の意味的な解析を行わないでこのような記述と主題文とを判断するのは困難である。

特に著者抄録においては，過去の研究成果と現在の研究成果の区別があいまいな抄録や，字数の制約などから体言止めなどを多用する抄録もしばしばみられる。本研究のように文の構文情報をもとにした情報検索技術の高度化を試みるためには，このような抄録作成の標準化をおし進める必要があるといえよう。

b)　主題文が複数ある抄録

本研究で対象とした抄録のうち，複数の主題文を持つものは71抄録であった。このうち，5抄録(7.1%)について，すべての文の種類を正しく判断することができた。また，すべての文を正しく判断することはできなかった66抄録のうち，55抄録については主題文の一部が一致し，11抄録については全く誤った判断がなされた。

全く誤った判断がなされた抄録のうち，前提文から始まっているにもかかわらず，主題文が第一文目にきていると判断された抄録が8抄録と大きな割合を占める。これは，a) ii) に示したように，前提文の表現の特徴を明らかにすることが困難なためであると考えられる。

連続した主題文を正しく判断できる抄録の例を第6図に，連続した主題文を正しく判断できなかった抄録の例を第7図に示す。

第6図の第1文～第3文および，第7図の第1文～第3文は，共に主題文によく見られる表現である。しかし，第6図の第1文～第3文が主として主題文のみによく見られる表現であるのに対して，第7図の第1文～第3文の文末に見られる動詞や「～した」という表現は，主題文と結果文の両方に多く出現する表現であった。主題文が連続して存在する抄録の数は少ないことから，このような場合には基礎遷移確率を取り入れることによって主題文が結果文と誤って判断されることにな

第6表 文ごとの判断結果

機械による判断結果	人手による判断結果			判断結果の正解率
	前提文	主題文	結果文	
前提文	83	3	2	94.3%
主題文	55	336	14	83.0%
結果文	37	130	1041	86.2%
正しく判断できた割合	47.4%	71.6%	98.5%	

る。

この問題を解決するには，基礎遷移確率にたより過ぎるのではなく，さらに主題文の分析を行い，抄録の最後に出てくる主題文や結果文と結果文の間にはさまれて出てくる主題文の特徴ある表現を見つけて条件に加える必要があるだろう。

それに対し，第6図に見られる「～考察する」「～提案する」「～述べる」はいずれも主題文であることを強く示唆している。ただし，このように主題文であることを強く示唆する表現を持つ文が連続する例は少ない。そこで，主題文であることを強く示唆する表現を持つ文が連続しない場合には主題文の一部が結果文と判断されてしまうことになる。

その他，抄録中に主題文が一度現れ，その後に結果文がきて再び主題文が現れる場合が21抄録で見られた。これらの部分は，情報部作業マニュアルに示す「6) 注記」に該当することが多いと考えられる。すなわち，研究・開発・調査の主目的外であるが，価値のある知見や情報で重要と思われるものである[11]。主たる論文の主題ではなくとも，検索もれを最小限にとどめるためにはこれらの部分についても主題文と判断できることが望ましい。しかし，この後半の主題文を正しく判断することは非常に困難である。これは，「結果文の次が主題文」という基礎遷移確率が0.031と非常に小さいため，もし結果文の次の主題文が，主題文である確率の高いパターンを満たしていたとしても，基礎遷移確率との積はかなり小さくなってしまうためである。本研究でも，結果文の次の文が主題文であると正しく判断できた例は存在しなかった。

このような場合に，正しく主題文を判断するためには，接続詞に関する分析も必要であろう。たとえば，本研究で分析対象とした抄録には，「次いで」という接続詞で文がつながれる場合が2件，「また」「ここでは」で接続される場合が各2件，「さらに」，「最後に」，「次に」の場合が各1件ずつ見られた。また，文末に出てきた特徴のある表現としては「～にも言及」「～についても記述」「～方法も示す」があったが，これらの中にはいずれも副助詞である「も」が出現している。本研究で対象とした抄録中では，主題文の後に結果文が出現し，その後再び主題文が出現するという例が少なかったため，これらの表現が文の種類を判断するのに有効であるかどうかについて，はっきりした傾向をつかむことができなかった。今後の検討課題であろう。

2. 文単位の結果

次に，抄録単位の分析結果において前提文，主題文，結果文の予測が一致しなかった抄録について，抄録中のどの文がどのように誤って判断されたのかを明らかにするため，文単位での判断結果について分析を行った。その結果を第6表に示す。

第6表に見られるように，文単位の判断結果としては主題文を結果文と誤って判断した場合が多かった。また，前提文を主題文もしくは結果文と誤って判断した場合も大きな割合を占めた。これは，前提文を特徴づける表現が位置の情報以外にはほとんどなく，主題文である特徴を持たない文が抄録の最初から続く場合に前提文であるという消極的な判断を行っているためと考えられる。したがって，前提文が複数文連続して出現する場合には，そのすべてを前提文と判断することは困難である。

IV. 終わりに

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，文の表層的な特徴を基にし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語との照合を行うことが必要であると考え，主題文であるかどうかを判定することを中心に実験を行った。

文の表層的な特徴として7つをあげ，これをもとにして抄録からその主題を表す部分を文ごとに機械的に抽出する作業を行った結果，69.4% の抄録について正確に抽出することができた。また，主題文の一部を正しく判断できたものも含めると 86.2% の抄録について主題文の判断をすることが可能であった。これは，本研究の手法が意味解析や構文解析を行わない表層的な特徴のみを用いていることを考慮すると比較的高い値であり，今後，構文解析の手法などを取り入れるための基礎データとして有効な結果を示すことができたと考えている。

本研究で主題文を正確に判断できなかった抄録の多くは，表層的な特徴を持たない文が前提文として連続して出現するものであった。このように，文の表層的な特徴を持たない文については基礎遷移確率が大きく影響を与えることになり，基礎遷移確率から予想される文の種類と異なる場合には判断を誤る結果となる。

この問題を解決するには，各文の持つ特徴をより詳しく分析し，単独の文の持つ情報から正しく文の種類を決定するルールの高度化が必要とされる。日本語処理技術の発達にともなって文章の高度な構文解析が可能となってきており，この技術の導入が望まれる。特に，抄録においては，文章をつなげて複文構造とする例が見られるが，この複文の前半と後半で文の種類が変わる場合がある。たとえば，「標準 IC のレイアウトは大域的経路選定と詳細経路選定に分けられるが，本文では新しい大域的経路選定を提案し，従来の分岐ツリーアルゴリズムよりも，全体のチャネル密度を 3% 少なくした。」という文の場合を考える。この文の前半は，「標準 IC のレイアウト」についての一般的な常識を述べた部分で，「前提」に分類される。しかし，「本文では」以降からは，この論文で扱った「主題」の範囲を述べている。本研究では主題文を正しく判断することを中心に考えたため，このような前提と主題の両方の役割を持つ文を主題文と判断して実験を行ったが，複文を分解する規則を明らかにできれば判断力はより高まると考えられる。

また，本研究では文末文節の助詞または助詞相当表現と文末文節の語幹（動詞）を独立に分析したが，実際には助詞または助詞相当表現と動詞との組み合わせで判断する方が適切なことがある。たとえば，「示（す）」という語は「～は」という助詞に続いて出現する場合に主題文でない例が多くみられる。このように，文末文節の語幹を独立して分析するのではなく助詞または助詞相当表現との組み合わせで考えようとする場合，どの助詞とどの動詞とが結びついているのかを明らかにした上で分析を行う必要がある。このような対応関係は構文解析の手法を導入することで実現可能であると考えられる。

さらに，代名詞や「標記の」「標題の」「題記の」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も必要であろう。今後は，構文解析の手法も取り入れてより正確な判断を行うことをめざしていきたいと考える。

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，NL 58-8，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，1991，259 p.
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自動抽出と重要度評価．電子情報通信学会論文誌．D-I，Vol. J 74-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，1983，237 p.
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，1978．36 p.
12) 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)
13) 溝口歌子．抄録法．ドクメンテーション研究．Vol. 23，No. 5，p. 157-163 (1973)
14) Liddy, E. D. The Discourse-level Structure of Empirical Abstracts: an Exploratory Study. Information Processing & Management, Vol. 27, No. 1, p. 55-81 (1991)
15) 梅田栄廣．抄録からの主題文の自動抽出．慶應義塾大学．1992，67 p．卒業論文.
16) 原田隆史他．キーワード間関係の規定による情報検索ノイズの低減．第 26 回情報科学技術研究集会発表論文集，p. 139-146 (1989)

17) 原田隆史他．構文解析にもとづくキーワード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，p. 37-44 (1990)
18) 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.
19) 吉田将．日本語の規格化に関する基礎的研究．昭和58年度科学研究費補助金一般研究 (B) 研究成果報告書．1984.
30) 日本科学技術情報センター情報部．JICST 抄録作成テキスト．東京，日本科学技術情報センター，1987，64 p.