計算言語学 25－2
（1981. 2. 6）

シナリオを用いて構造化されたキーワードをアブストラクトから抽出する一手法

堀 浩一　斉藤 忠夫　猪瀬 博
（東京大学 工学部）

A concept of structured key words is proposed based on a linguistic examination of abstract of paper to be used for information retrieval systems. Structured key words represent the deep structure of an abstract of a technical paper. A method for extracting structured key words from abstracts is given. Structured key words are extracted top-down way according to a scenario which describes the direction of understanding sentences. A result of an experiment is described to show the feasibility of the method.

§1 はじめに

我々の研究室では、現在、ビデオディスクに1次情報としての原文書画像を、磁気ディスクに2次情報を蓄え、N1ネットワークを介し、2次情報により、必要な文献をオンライン検索し、1次情報のコピーをファクスから得る原文書データベースシステム（仮称）を研究開発中である。データベースが大規模化した場合には特に、キーワードの有効な抽出を自動的に行なう事が望ましい。こうした目的で、文献検索システムの将来の姿を探るために、蓄えるべき2次情報とその抽出法についての基礎的な考察と小規模な実験を行なったので御報告する。

文献検索システムは、図1のようにとらえる事ができる。従来、抽出する2次情報は、著者名、タイトル、雑誌名、キーワード等であり、キーワードの自動抽出は、統計学的手法（statistical method）によるものが主であった[1]。

検索の精度や、再現度を向上させるために、キーワードに重みをつけたり、ロールを付与したりする事も考えられている。

しかし、欲する文献は、どのような文献か、あるいは、ある文献が何について述べているのかを直接表現できるのは自然言語のみである事を考えると、2次情報としては、単純なキーワードでなく、言語学的手法を用いて文章の意味を理解した上で必要な情報を抽出する事が望ましい。

本報告では、文献抄録の深層構造を反映する構造化キーワードの考えを提案し、構造化されたキーワードを抽出する手法を示し、その実現性を評価するための小規模な実験の結果を述べる。

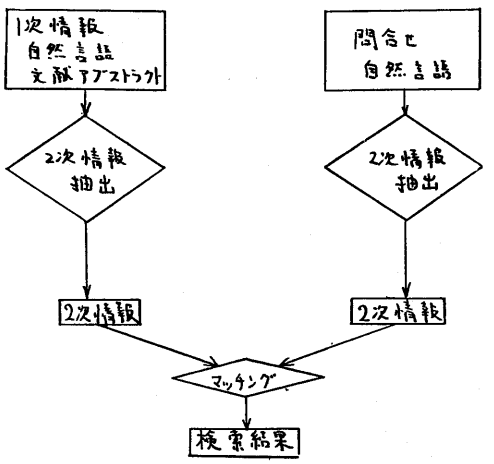

図1　文献検索システム概念図

§2 構造化されたキーワード

文献をその内容で検索しようとするためには、どの程度複雑で、どれくらいの量の情報を抽出し蓄えておけば有効であろうか。

従来の研究例としては、図2のような日本語記事のロール付インデクス[2]、図3のような特許請求文のspecification table、図4のような言語処理の文献のspecification table[3]、モンタギュー文法を用いたlogicに基くもの[4]などが発表されている。

これらの従来の研究では、その使用目的に応じて、キーワードの構造が、人間によりトップダウンに定められているように思われる。

ロール(①主体 ②客体 ③時 ④場所 ⑤活動 ⑥主題) + インデクス

図2 ロール付インデクス(絹川ら[2])

COMPOSITION: (OBJ-D:は, COMP-D:{は})
FUNCTION: (PRED-F:__, AG-D:は, C_1)
QUALITY: (PRED-Q:__, OBJ-D:は, C_2)
CONNECTION: (PRED-C:__, OBJ-D:は, PARIC-D:__, CHANNEL-D:__, M:__)
where C_1 ≡ PAT:__, INSTR:__, SO:__, G:__, M:__, COND:__
C_2 ≡ COMPAR:__, DEG__, CIR:__.

図3 特許請求文specification table (西田ら[3])

COMPOSITION: (OBJ-S:は, COMP-S:{は})
FUNCTION: (PRED-F:__, AG-S:は, C_3)
INSTRUMENT: (PRED-F:__, AG-S:__, C_3, INSTR-S:__, USE-S:__, USE-S:は)
USE: (PRED-F:__, AG-S:__, C_3, INSTR-S:は, USE-S:__)
where C_3 ≡ PAT:__, SO:__, G:__, M:__

図4 自然言語の分野の文献のspecification table (西田ら[3])

言語学的手法により、文献アブストラクトの深層構造として得られるキーワードを構造化されたキーワードと呼ぶ事にする。

本研究は使用目的に応じて、構造の複雑さを定めて、ふさわしいキーワードを抽出する事ができる汎用のシステムを作る事を目的としている。

実用化システムとして用いるためには、構造化キーワードの表現法、構造化キーワードの抽出法、及び辞書の構成法は、できる限り単純明解でなければならない。

構造化キーワードの表現方法としては、フレーム、セマンティックネットワーク、述語論理等が考えられる。ここでは、使用目的が、検索のみで、静的である事を利用して、単純な図5に示すようなフレームによる表現法を用いる事とする。1つのフレームは1つの概念を表わし、フレーム名といくつかのスロットから成る。スロットはフレームまたは最小構成要素の単語(キーワード)である。このフレームは、概念を階層的に表現する枠組であり、推論などの知識操作を目的としたMinskyらのフレームの概念とは異なる。図5は論文というフレームの例を示している。

構造化キーワードのフレームは、次のような手続きを経て定めるべきであると考える。

第1段階:文の表層構造に近い一般的フレームにアブストラクトと問い合わせ文をあてはめる事により分析を行ない、共通の、より深層に近い特殊なフレームを(人手により)見出す。

第2段階:この特殊なフレームを中心に文献検索システムを構成し、フレー

ムにあてはまらない文献や問合せが出現した時は、必要に応じて、フレームの修正を行なう。

第3段階：上記2段階を経て、一般に文献アブストラクトとして表現される文章と、検索システムによる問合わせに使用される文の深層構造を知り、定形化できた後では、そのフレームにそって、アブストラクトを一般の論文の著者に書いてもらうようにする。

第3段階については、文章の流通範囲が限られている応用（例えば、一研究室内の研究のドキュメンテーション）に対しては、自然言語を工学的に扱うという意味で有意義かもしれない。学会のような大組織でも、労力を投入すれば、論文誌のアブストラクトのフレームと、さらに表層の文体まで規定する事も将来期待できよう。

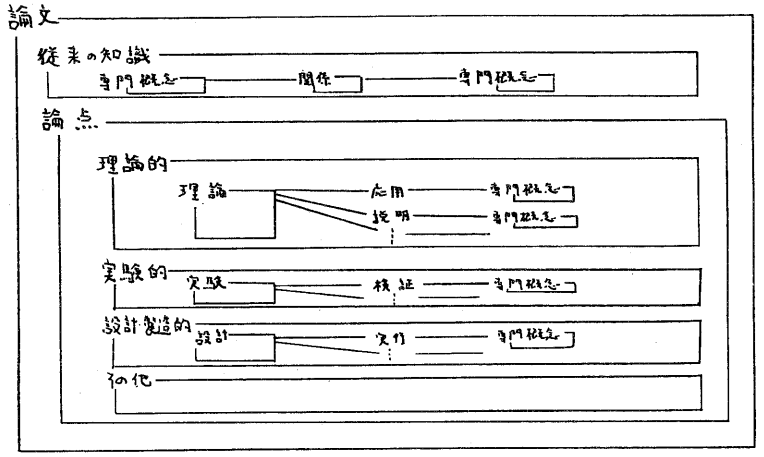

図5 フレームの例

§3 シナリオ

まず、第2章で述べた構造化キーワードを、文献アブストラクトから抽出する方法を考える。

このためには、複数の文からなる文章から、その意味的つながりを理解して、深層構造を得る事が必要であり、これは談話理解の問題にたちいる事になる。一般に談話理解は、自然言語処理の問題の中でも難しい問題とされているが[5]、対象を、意味のあいまいさの少ない文献アブストラクトに限定すれば、困難は軽減される。さらに、もし、第2章の最後に第3段階として述べたように論文アブストラクトのフレームと表層の文体を規定してしまう事ができれば、問題は、より簡単になる。

自然言語処理の手法は、図6に示すbottom-upに逐次処理していく方法と、図7に示すSchankらの主張するように表層から直接深層構造に変換する方法[6]の2つに大別できよう。本研究では、取り扱う文章を限定している事と、目的が検索に必要な情報を抽出するだけであり構文解析が完璧に行なわれる必要はないので、各段階にあいまいさ処理の伴なう図6の方式より、図7の方式が適当であると考える。

形態素分析→構文解析→意味解釈→モデルとの対応

図6 逐次処理方式

図7 direct parsing

自分の興味（フレーム）にそって文を読んでゆき、単語の意味から直接深層構造を構成していくというSchankらの方法は、一般に、人間の文章理解の方式に近いと考えられるが形式的取扱いが難かしく、また知らない世界の文章を受け付けない等の欠点を有するとされている。本研究では以下述べるように、ATN（augmented transition network）を拡張して用いる事により、図7の方式を実現し、これらの欠点をカバーする事にした。構文解析の規則を記述・実行するパーザに対応するものとして、単語レベルから直接フレームにあてはめていく規則を記述・実行する機構をここではシナリオと呼ぶ事にする。

ATNはパーザとして広く用いられているが、本来0型の記述能力を有しているから、シナリオとしても用いる事ができるはずである。例えば、テロリズムという図8のようなフレームに対するシナリオの概略は図9のように記述できる。

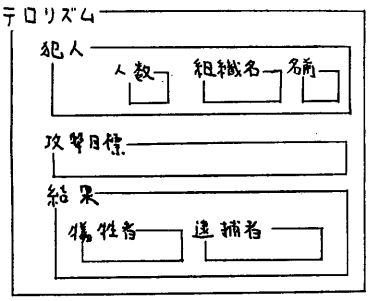

図8 テロリズムを表わすフレーム

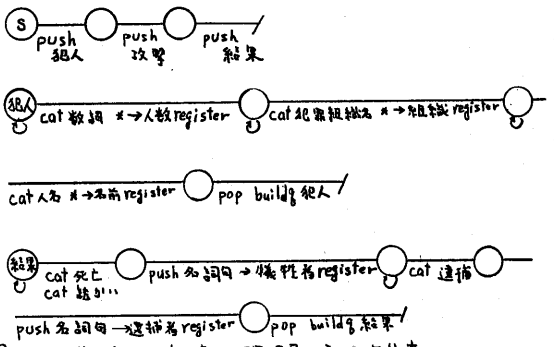

図9 テロリズムの記事を理解するシナリオ

本研究で使用するパーザとしてWoodsのATN[7]とほぼ同じATNのインタプリタを作製した。ただし単語のカテゴリとして、品詞と意味を区別してどちらも指定できるようにし、予期しない意味の単語に出会った時に、インタラプトをかけて、別のシナリオを呼び出して、理解の方向を分岐できるようにした。ATNを用いた理由は、1）記述形式を明解にできる事　2）シナリオもATNも本質的にトップダウンで都合が良い事　3）構文解析のみを記述できるので、未知の世界の文章も一応受け入れられるようにできる事　等である。今回の応用では、特に別に意味処理のプログラムなどを用意する必要はなく、ATNのみで充分であった。

単語の意味の記述法もいろいろ考えられようが、実用化システムでは単純であるほど望ましいと考え、理論、実験等の意味の分類のみを与える事にした。実際の分類については次章に示す。

さて、実際には、構文解析、意味処理、フレームへの構成を、同時に平坦に記述する事はそれほど容易ではない。例えば、"A method of file clustering is given in this paper. It can be applied to large files." も、"A method of file clustering applicable to large files is given." も等しく図10に示すような構造化キーワードを与えるべきであろう。そこで、シナリオをレキシコン・ドリブン・ルーチン、グラマ・ドリブン・ルーチン、

フレームドリブンルーチンと呼ぶ同時に走る三つの部分に分けて記述する事にする。レキシコンドリブンルーチンは、意味を知った単語に出会った時グラマドリブンルーチンに対し割込をかけてローカルな意味処理を行なうルーチンである。例えば、応用という意味の単語"applied"に出会うと、何への応用かを探し、処理結果の例えば(applied (to) large files)をレジスタに入れ、その処理が行なわれた事をふすフラグを立てる。グラマドリブンルーチンは、おおざっぱな構文規則に従って文を流し読みしていくためのルーチンである。また構文的役割によって意味が異なってくる単語に対処するため、レキシコンドリブンルーチンの割込に対するマスクを制御する役割も実行する。マスク制御とは、例えば、"describe"は論点を記述する単語に分類してあるので、"This paper describes a theory of ～."という文を読むと、"describe"という単語に出会った時割込がかかって、この文は著者が論点を記した文であるというフラグが立つが、"The model which describes the principle is ～."という文の"describe"ではフラグが立たないようにする機能である。フレームドリブンルーチンはフレームにそってトップダウンに文章理解の道筋を定め、レキシコンドリブンルーチンの処理結果を再配置してフレームにあてはめていくためのルーチンである。例えば、論点として、"関係モデルの理論"をレキシコンドリブンルーチンから受けた時には、フレームドリブンルーチンは、その理論に関する、"基礎(based on ～)","応用(applied to)","説明(explains ～)"等の記述をレキシコンドリブンルーチンから得る事を期待し、処理する。指示代名詞については、正確な処理をせず、期待する意味の記述だけ拾っていくという単純な方法としている。例えば、前述の例の"A method of file clustering is given in this paper. It can be applied to large files."では、論文の論点が、"file clustering の手法"であると理解し、次に期待していたappliedという記述があらわれるので、"その file clustering の手法が large files に応用できる"と理解する。

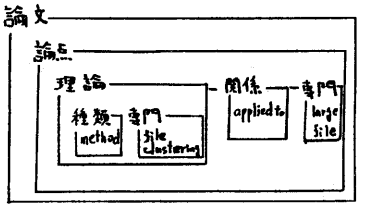

図10 構造化キーワードの例

レキシコンドリブンルーチンとグラマドリブンルーチンとフレームドリブンルーチンを並列に実行する方法として、現在、Unix オペレーティングシステムのパイプの機能を用いてグラマドリブンルーチンとフレームドリブンルーチンをパイプでつないで並列実行し、レキシコンドリブンルーチンはグラマドリブンルーチンに割込をかけて実行するという方法をとっている。

フレームや単語の意味の分類が詳細に定まらない段階ではグラマドリブンルーチンの比重が重く、対象とする文章を絞り、フレームや単語の意味を詳細に定めるほどグラマドリブンルーチンの比重は軽くなる。

§4 実験

以上述べた考えに基いて、小規模な実験を行なった。対象はACM transaction on Database Systems の1978年12月(vol.4)から1979年12月(Vol.4)までの1年分全31文献のアブストラクトである。語彙は約1000単語、使用した計算機はDEC LSI11/03(Unix オペレーティングシステム)、及びアイ電子AICOM C5である。ATNインタプリタは、Pascalで記述した。

4－1　第1段階

まず、分析の段階として、文献アブストラクトには、「どういう事が書かれているか」を調べる。一般的フレームとして、図11に示すフレームを用意した。

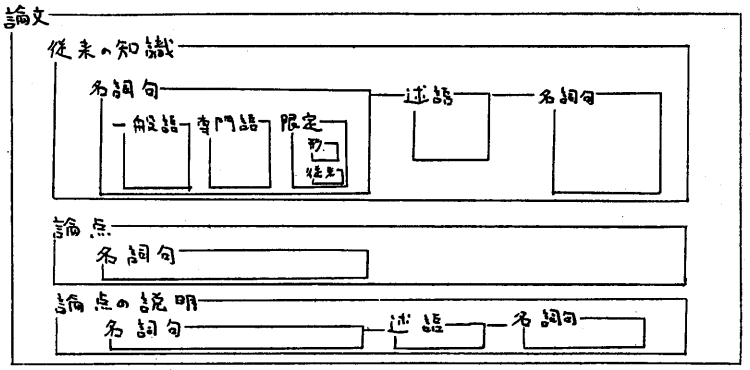

図11 論文のフレーム(I)

単語の意味を、論点記述、専門語、一般有意語、無意語の4つにだけ分類する。論点記述とは、describe, show, give, propose, introduce 等の、論点を示すのに用いられる単語である。専門語とは、専門家でなければ意味のわからない単語である。無意語とは、冠詞、前置詞等のそれ自身は意味をもたない単語で、これら以外を一般有意語とした。

この段階では、シナリオはほとんどグラマ・ドリブン・ルーチンのみである。

図12に入力したアブストラクトと、その出力例を示す。

入力

A classified, or clustered file is one where related, or similar records are grouped into classes, or clusters of items in such a way that all items within a cluster are jointly retrievable. Clustered files are easily adapted to broad and narrow search strategies, and simple file updating methods are available. An inexpensive file clustering method applicable to large files is given together with appropriate file search methods. An abstract model is then introduced to predict the retrieval effectiveness of various search methods in a clustered file environment. Experimental evidence is included to test the versatility of the model and to demonstrate the role of various parameters in the cluster search process.

出力

```
(JURAI=( : FILE + CLASSIFIED CLUSTERED -> IS <- ONE : + ( : RECORDS + RELATED SIMILAR -> GROUPED <- ITEMS WAY : ) (JURAI=( : FILES
 CLUSTERED -> ADAPTED <- STRATEGIES METHODS : SEACH FILE UPDATING + BROAD NARROW SIMPLE ) (JURAI=( : + -> ARE <- : + AVAILABLE ) (I
 _RONTEN= METHOD : FILE CLUSTERING FILES + INEXPENSIVE APPLICABLE LARGE (TADASHI= METHODS : FILE SEARCH + APPROPRIATE ) ) (IP_RONTEN
 : MODEL + ABSTRACT (TADASHI= PREDICT EFFECTIVENESMETHODS ENVIRONMENT : RETRIEVAL SEARCH FILE + VARIOUS CLUSTERED ) ) (IP_RONTEN= E
 IDENCE : + EXPERIMENTAL(TADASHI= TEST VERSATILITY DEMONSTRATE ROLE : MODEL PARAMETERS CLUSTER SEARCH PROCESS + VARIOUS ) )
```

図12　入出力例(I)

4－2　第2段階

第1段階の結果から、文献検索システムの2次情報として有用と思われるものを(人手により)拾い上げ、その形を分類し、図11のフレームより特殊な、図13に示すフレームを作成した。単語の意味を図14に示すように分類した。図13のフレームにあてはめるためのシナリオの一部を図15に示す。図12と同じアブストラクトに対する出力例を図16に示す。

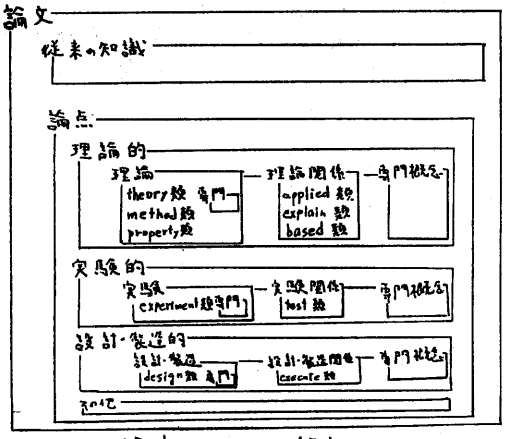

図13 論文のフレーム(II)

分類		単語の例
理論	theory類	theory model concept notion
	method類	method technique strategy
	capability類	capability characteristic property
理論関係	applied類	applied used extended
	explain類	explain predict represent
	based on類	based extention use
実験		experiment
実験関係		test explore
設計・製造		design implementation
設計・製造関係		execute improve
論点記述		describe give show discuss
専門		file storage query
その他有意		large small good
無意		a the in on

図14 単語の意味の分類

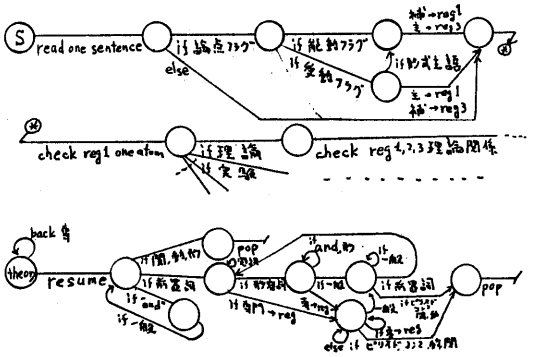

図15 シナリオの一部

```
ronten=
  riron
    method-rui
      FILE CLUSTERING
    applied to-rui
      LARGE FILES
  riron
    theory-rui
    explain-rui
      RETRIEVAL SEARCH CLUSTERED FILE
  jikken
    jikken
    test
```

図16 出力例(II)

この結果、31文献中27文献(87%)については、一応図13のフレームにあてはまる、何らかの情報が抽出できた。うまくいかない文章の例を図17に示す。図17の例にもみられるように、"capability of relational model"という代りに、"whether relational model is good or not"というように1つの概念を、よりやさしい、いくつかの単語を用いて言いかえてある場合には、本研究の方法ではうまくいかない。これに対処するためには、より強力な知識ベースと推論機構が必要であろう。

処理時間は1sentenceあたり40秒ほどである。

Record structures are generally efficient, familiar, and easy to use for most current data processing applications. But they are not complete in their ability to represent information, nor are they fully self-describing.

図17 うまく処理されない文章の例

4-3 第3段階

図13のフレームに基いて論文アブストラクトを生成する生成シナリオを作成した。これを用いる場合にはユーザはシステムの示す文体と単語群から適当なものを選択して、文章を作成するようにする。図18にその出力例を示す。

METHOD OF NATURAL LANGUAGE UNDERSTANDING IS GIVEN. THIS METHOD IS BASED ON CONCEPT OF FRAME. THIS METHOD IS APPLIED TO METHOD OF AUTOMATIC INDEXING. CONCEPT OF TOP-DOWN UNDERSTANDING IS GIVEN. THIS CONCEPT OVERCOMES PROBLEM OF AMBIGUITY OF NATURAL LANGUAGE.

図18 生成シナリオ出力例

§5 おわりに

構造化されたキーワードの考えを提案し、その抽出法を与え、小規模な実験を行なった。

今回与えたシナリオをそのまま大規模な文献集合に適用すれば、おそらくフレームにあてはまらない文章が続出して実用には供せないと考えられる。しかし、大規模なシステムにおいても、現在よりきめ細かな検索ができるようになる事は大いに望まれる所である。1つの解決法としては、第2章で第3段階として述べたように、アブストラクトや本文の構造と文体を規定する事も考えられよう。自由に書かれた文章と理解できるようにするためには、今回の手法に、何らかの形で知識の蓄積と推論の機構を導入する事を検討しなければならない。

小規模な個人用の文献データベースに対しては、本研究の手法は有効であると考える。例えば、毎月、"database"という key word を含む文献のみ、大規模なデータベースから検索してきて蓄えておき、本研究のシステムで整理し、研究に用いるといった使用法が考えられよう。

科学技術文献の文献検索システム以外への応用は今後の検討課題である。

参考文献

[1] 中井,"機械補助索引(MAI)について", 情報管理, vol.19, no.4, Jul.1976

[2] 綱川本村,"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, May, 1980

[3] Nishida,F.,Takamatsu,S.,"Structured-Information Extraction and Retrieval from short Texts",German-Japanese Workshop on I&D,Sept.27-29,1980

[4] Nishida,T.,Doshita,S.,"THE FRAMEWORK OF KNOWLEDGE REPRESENTATION AND ITS RETRIEVAL IN LGS-THE LITERATURE GUIDE SYSTEM",Proc. of 6th. IJCAI,1979

[5] 田中,"談話理解の構造", 情報処理, vol.20, no.10, Oct., 1979

[6] Schank,R.C.,et al.,"PARSING DIRECTLY INTO KNOWLEDGE STRUCTURES",Proc. of 6th. IJCAI,1979

[7] Woods,W.A.,"Transition Network Grammars for Natural Language Analysis ", Communications of the ACM,vol.13,no.10,Oct.,1970