



係り受け関係に基づく文献の検索*

高松 忍** 藤田 米春** 西田 富士夫**

Abstract

This paper presents a method transforming a Japanese noun phrase or clause expressing a subject into a canonical function form which consists of a function symbol and several terms with their corresponding case indicators.

The canonical function form is constructed so that various Japanese expressions of a subject can be transformed into a unique and concise function form as much as possible.

Ambiguities of complicated dependency relations between verbs and nouns can be removed using more precise categories in the related special field.

Furthermore, an efficient retrieval method is presented which retrieves almost all documents implied contextually, as well as conceptually, by a request.

1. ま え が き

文献検索において、文献の主題をどのように表現するかの問題は呼出し率や適合率に関連して重要な問題である。従来提案されているものの中には、文献のキーワードに role-indicator や weight を付加したり、キーワード間をリンクしたりする方法がある。しかし類似の内容をもつものが多い特許文献などの場合、精密で漏れが少ない検索を行うには構文と意味の両面からの研究が必要であると思われる。

本稿では、文献の標題文や特許請求範囲文などにおいて同じ意味内容を表わす名詞節や名詞句は一つの標準形式の関数表現に変換し、これをその文献の主題表現とし呼出し率の向上と検索の能率化をはかっている。また、節や句における動詞と名詞間の係り受けのあいまいさは、対象に関するカテゴリの組を用いてその解消をはかっている。さらに、検索に関連して、限定・修飾条件や全体と部分などの格関係に基づく係り受け構造の包含関係から検索要求に含まれる下位の文献をすべて検索する一つの能率的な方法を述べている。

2. 名詞節・名詞句の関数表現への変換

この章では、標題文や特許請求範囲文などの名詞節や名詞句から、係り受け関係を明確に表わした関数表現に変換する手法について述べる。

2.1 名詞節・名詞句の関数表現

文献の標題文や特許請求範囲文などは、

‘アルカリ化により処理したフェライト薄膜’ (1) のような名詞節や名詞句で表わされることが多い。(1) によって表わされる対象は、製造方法が、アルカリ化による処理’に限定され、成分が‘フェライト’に限定された‘薄膜’に関する事項の集合と考えられ、(1)を

‘薄膜 (COMPO:=フェライト, OBJ¹:=処理した (時制:=過去, MEANS:=アルカリ化成一により))’ (2)

なる関数表現で表わすことができる。

一般に、名詞節や名詞句によって表わされる対象は、いくつかの物や属性や事象などによって修飾・限定された物や属性や事象などの一つの集合と考えられ、次式 (3) のような関数表現で表わすことができる。

$$f(K_1:=t_1-j_1, K_2:=t_2-j_2, \dots, K_n:=t_n-j_n) \quad (3)$$

ここに、 f は名詞節や句の最右端の名詞語 (‘受け’ とよぶ) よりなる関数記号を表わし、 t_i ($i=1, 2, \dots, n$) は f を修飾・限定する項 (‘係り’ とよぶ) を表わ

* A Document Retrieval Based on Dependency Relations by Shinobu TAKAMATSU, Yoneharu FUJITA, and Fujio NISHIDA (Department of Electrical Engineering, Faculty of Engineering, University of Osaka Prefecture).

** 大阪府立大学工学部電気工学教室

す。また、 K_i ($i=1, 2, \dots, n$) は f と t_i 間の係り受け関係を明示する格ラベルであり、 t_i の後置記号 j_i は ‘の’ などの所有の格助詞の語または空語である。

項 t_i は、名詞語、式 (3) のような関数表現または次式 (4) のような述語詞 (動詞、形容 (動) 詞) 節を表わす関数表現:

$$\hat{p}(L_0 := s_0, L_1 := s_1 - j_1', L_2 := s_2 - j_2', \dots, L_m := s_m - j_m') \quad (4)$$

である。ここに、 p は ‘受け’ の述語詞 (動詞、形容 (動) 詞) の語、 s_i ($i=1, 2, \dots, m$) は p を修飾・限定する ‘係り’ の項で、名詞や副詞の語または式 (3) や式 (4) のような関数表現であり、その後置記号 j_i' は格助詞や接続助詞の語または空語である。

L_0 は述語詞の語の時制、態、使役や可能を指定する格をまとめて表わし、 s_0 は L_0 に入る助動詞の語を表わす。 L_0 を特殊格とよび、 L_0 以外の格 L_i ($i=1, 2, \dots, m$) を普通格とよぶ。普通格の内、主格 (SBJ)、目的格 (OBJ)、補語格 (COMPL) およびそれらの三つの格に入る項のカテゴリは述語詞の語ごとに単語碎書に与えられる。普通格のその他の主な格およびそれらの格に入る項のカテゴリを **Table 1** に示すが、それらの格は、この論文では、各述語詞の語に共通であるとする。ここに、普通格のカテゴリは、後述の構文分析における格ラベルの決定や下位のカテゴリの名詞語を含む文献の検索などのために、**Table 2** に示すように階層的に分類しておく。なお、述語詞の語のカテゴリは、述語詞の語の語幹などをとって名詞化した語 (事象名詞とよぶ) のカテゴリとする。

つぎに、式 (3) の名詞語 f のもつ格ラベル K_i について述べる。

式 (3) において、 f が事象名詞の語のとき、格ラベル K_i ($i=1, 2, \dots, n$) は、 f に対応する述語詞の語に対する項 t_i の格ラベルとする。

f が事象名詞以外の名詞 (非事象名詞とよぶ) で、項 t_i が述語詞または事象名詞の語であるときには、 f が t_i に係る格 L_i の逆を示す格ラベル L_i^{-1} を t_i の格ラベルとし、 $f(L_i^{-1} := t_i)$ とする。例えば、述語詞節、薄膜を処理した’の関数表現 ‘処理した (OBJ=薄膜-を)’ に対し、名詞節 ‘処理した 薄膜’ の関数表現を ‘薄膜 (OBJ=処理した)’ で表わす。

その他の場合、すなわち、 f が非事象名詞の語で、項 t_i が非事象名詞の語であるとき、 f と t_i の間で ‘所有する’ などの動詞の語が省略され、‘の’ などの所有の格助詞の語におきかえられたものと考えられる。

Table 1 Normal cases common to all predicates.

格 (ラベル)	道具材料格 (INST)	条件格 (COND)	場所格 (LOC)	時間格 (TIM)
カテゴリ	物・現象体	属性	場所	時間
後置記号	により, で, から	で, において	に, において	に, で

格 (ラベル)	模態格 (MOD)	手段格 (MEANS)	目標格 (PURP)	提示格 (EXH)
カテゴリ	形容	事象	事象	物・事象
後置記号	—	により, で, て	に, ために	において

Table 2 Classification of categories.

カテゴリ名	例		
事物	非事象	材料……………誘電体, 半導体, フェライト, …	
		物 部品……………ダイオード, 薄膜, PET, …	
		装置……………交換器, 分波器, 導波路, …	
		現象体……………波, 光, パルス, …	
		属性……………屈折率, 圧電性, 圧力, 温度, …	
	数量……………100°C, 10 ⁻⁸ トル, …		
	場所……………水中, 地上, …		
	時間……………実時間, 即時, …		
	事象	状態	形容……………最適, 一定, 高, 多, …
		所有……………所有, …	
含有……………含有, …			
動作		機能……………検波, 整流, 増幅, …	
		現象……………反射, 透過, 発生, …	
操作	使用……………使用, 利用, 応用, …		
形成……………形成, 処理, 作製, …			
その他の操作……………蒸着, 測定, 設計, …			

Table 3 Case labels for pairs of nouns belonging to non-events.

受けの語 f のカテゴリ	格 (ラベル) K_i	係りの語 (句) t_i のカテゴリ	助詞の語 j_i	例
物・現象体	属性格 (ATR)	属性		線状アンテナ
部品	成分格 (COMPO)	材料	の	フェライト薄膜
装置	成分格 (COMPO)	部品		薄膜トランジスタ
部品	提示格 (EXH)	装置	の	半導体装置の基板
材料	提示格 (EXH)	部品	における	薄膜の導電材料
属性	属性値格 (VAL)	数量		610°C の温度
属性	対象格 (GSD*)	物・現象体	の	薄膜の抵抗率

* GSD: der Gegenstand

この場合、 f と t_i の意味関係は ‘物とその成分’、‘物とその属性’ や ‘属性とその属性値’ などにはほぼ限られ、このような意味関係を表わす格ラベル K_i は、**Table 3** に示すように、 f と t_i の属するカテゴリの組ならびに助詞の語 j_i から定められる。

2.2 変換

名詞句や名詞節は、非終端記号 <NPC> (Noun Phrases or Clauses) から生成され、その主な生成規則を **Table 4** (次頁参照) に示す。ここに、角括弧 [] は括弧内の記号が括弧とともに省略されることがあることを示す。

<NPC> は、連言、選言の接続詞 <Cc>, <Cd> で接

Table 4 Production rules of noun phrases and clauses.

分類	番号	生成規則
(a) 並列句	a-1	$\langle NPC \rangle \rightarrow \langle NPC_1 \rangle \langle C_c \rangle \langle NPC \rangle$
	a-2	$ \langle NPC_1 \rangle \langle C_d \rangle \langle NPC \rangle$
	a-3	$ \langle NPC_1 \rangle$
	a-4	$\langle NPC_1 \rangle \rightarrow \langle NP \rangle \langle NC \rangle$
(b) 連体修飾関係	b-1	$\langle NP \rangle \rightarrow \langle NPC \rangle [\langle JK_N \rangle] \langle NP \rangle$
	b-2	$ \langle N \rangle$
	b-3	$\langle NC \rangle \rightarrow \langle PC(i=2) \rangle [\langle C_c \rangle] \langle NC \rangle$
	b-4	$ \langle PC(i=2) \rangle \langle C_d \rangle \langle NC \rangle$
	b-5	$ \langle NC_1 \rangle$
	b-6	$\langle NC_1 \rangle \rightarrow \langle PC(i=3) \rangle \langle NP \rangle$
(c) 連用修飾関係	c-1	$\langle PC(i) \rangle \rightarrow \langle PC_1(i=2) \rangle [\langle JC_1 \rangle] \langle PC(i) \rangle$
	c-2	$ \langle PC_1(i=4) \rangle \langle JC_2 \rangle \langle PC(i) \rangle$
	c-3	$ \langle PC_1(i) \rangle$
	c-4	$\langle PC_1(i) \rangle \rightarrow \langle NPC \rangle \langle JK_F \rangle \langle PC_1(i) \rangle$
	c-5	$ \langle AD \rangle \langle PC_1(i) \rangle$
	c-6	$ \langle P(i) \rangle$

続されたいくつかの名詞句 $\langle NP \rangle$ や名詞節 $\langle NC \rangle$ からなる (a-1~4). $\langle NP \rangle$ は、いくつかの係りの名詞句や名詞節 $\langle NPC \rangle$, 'の' などの所有の格助詞 $\langle JK_N \rangle$ と、受けの名詞 $\langle N \rangle$ からなり (b-1~2), $\langle NC \rangle$ は、係りの述語詞節 $\langle PC(i=3) \rangle$ と、受けの名詞句 $\langle NP \rangle$ などからなる (b-3~6). ここに、 i の値 0~5 は、終止、未然、連用、連体、假定、命令の活用形を表わす. $\langle PC(i) \rangle$ は、いくつかの係りの名詞句や名詞節 $\langle NPC \rangle$, 格助詞 $\langle JK_F \rangle$ と、受けの述語詞とそれに続く数個の助動詞からなる述語詞句 $\langle P(i) \rangle$ などからなる (c-1~6). ここに、 $\langle JC_1 \rangle$, $\langle JC_2 \rangle$ と $\langle AD \rangle$ はそれぞれ、連用形、假定形の接続助詞と副詞を示す.

Table 4 に示すように、係り受け関係を示す生成規則はすべて、

$$\langle R \rangle \rightarrow \langle D \rangle [\langle J \rangle] \langle G \rangle \quad (5)$$

なる形で表わされる. ここに、 $\langle G \rangle$ は受けの語 g を最右端の語として含む語の列を、 $\langle D \rangle$ は受けの語 g へ係る語の列を、 $\langle J \rangle$ は格助詞などの助詞の語 j を生成する非終端記号である.

構文分析は、主に順位文法と従来の意味分析の手法を用いて行う⁷⁾.

Table 4 から順位表を作ると、過半数の順位関係は一義的に定まるが、その他の二重順位関係に対しては、従来のように、ある一つの名詞のカテゴリが述語詞の語の一つの格のカテゴリに含まれるかどうかを比較・照合することが必要である.

このような方法により、(5)の生成規則を適用してその右辺の記号列を $\langle R \rangle$ におきかえるとき、 $\langle D \rangle$ に対応して構成された関数表現 d を用いて、 $\langle R \rangle$ に対

応する関数表現:

$$g(\dots, K_i = d - j, \dots)$$

を構成する. ここに、格ラベル K は、前節で述べたように、 g または d が事象名詞 (節や句) や述語詞 (節や句) のときには、対応する述語詞の語の格構造と Table 1, 2 を参照して定め、 g と d がともに非事象名詞 (句や節) のときには、 g と d の属するカテゴリの組ならびに助詞の語 j から Table 3 を参照して定める.

例1 名詞句 '光導波系の固有モードの解析' において、'光導波系'、'固有モード' と '解析' はそれぞれ、'物'、'属性' と '事象' のカテゴリに属する. これより、'解析する' の格構造: (SBJ=人・装置-が、OBJ=属性-を) と Table 3 を参照して、Table 4 の (b-1) などの生成規則を繰返し適用すれば、

'解析(OBJ=固有モード (GSD=光導波系-の) -の)'

なる関数表現が構成される.

以上述べた構文ならびに意味分析により、多くの場合、関数表現への変換を確定的に行うことができるが、中には従来の意味分析では係り受けのあいまいさが解消できない場合がある. 例えば、

$$\dots t_{1n} j_{1n} \dots t_{1i} j_{1i} p_{tr_1} j_{r_1} t_{r_2} \dots \quad (6)$$

なる語の列において、 p と t_{r_i} ($i=1, 2$) との間の係り受けのあいまいさは解消しない. すなわち、いま ' $t_{1n} j_{1n} \dots t_{1i} j_{1i} p$ ' が $\langle PC(i=3) \rangle$ に還元され、 t_{r_1} と t_{r_2} がともに述語詞の語 p の格 K_r のカテゴリに属するとき、 $\langle PC(i=3) \rangle$ が t_{r_1} , t_{r_2} のいずれに係るか不明である.

このような場合、述語詞の語の二つ以上の格がとるさらに詳しい下位のカテゴリの組または専門的なカテゴリの組を利用すると、あいまいさが解消される場合が多い. いま、Table 5 (次頁参照) のように、述語詞の語 p の複数個の格 K_1, \dots, K_r に

$$p(\dots, K_i = t_{i_1}, \dots, K_r = t_{r_1}, \dots) \\ \rightarrow t_{i_1} \in C_{i_k} \wedge \dots \wedge t_{r_1} \in C_{r_k}$$

を満たすいくつかの下位のカテゴリの組 (C_{i_k}, \dots, C_{r_k}), $k=1, 2, \dots$ が存在するものとする. このとき、(6)において、' t_{i_1}, \dots, t_{i_n} ' の一つが C_{i_k} に属し、 t_{r_1} と t_{r_2} のいずれか一つの項 t_r だけが C_{r_k} に属するときには、 $\langle PC(i=3) \rangle$ は t_r に係るものと決めることができる.

例2 '六方結晶構造を有する高仕事関数の合金' において、'六方結晶構造' と '仕事関数' なる語は、

Table 5 A part of the word-dictionary.

見出し語	品 詞	上位カテゴリ
六方結晶構造	<N>	属 性
仕事関数	<N>	属 性
合金	<N>	材 料
材料	<N>	物

見出し語	品 詞	カテゴリの組
有する もつ 含む	<P(i=0)>	{SBJ-OBJ 部品-材料 装置-部品 物-属性 物-機能・現象
構成する	<P(i=0)>	{SBJ-OBJ 部品-装置 材料-部品

Table 5 からわかるように、‘有する’という述語詞の語の下位カテゴリの組に属さず、‘六方結晶構造’と‘合金’なる語が下位カテゴリの組：(OBJ=属性, SBJ=物)に属する。

これより、‘六方結晶構造を有する’の <PC(i=3)> は、‘高仕事関数’の <NP> でなく、‘高仕事関数の合金’の <NP> に係るものと決めることができる。

3. 関数表現の標準化

係り受け構造をもつ標題文などを対象とする文献検索で重要なことは、同じ意味的な関係をもつ見出しは、同じ表現形式に統一し、かつ所要記憶容量や検索の速さの関係からその表現形式を簡潔にすることである。簡潔な表現形式としては、自然語表現に名詞語だけからなる複合名詞句表現がある。これと並行して、前章で述べた関数表現の標準化をつぎのような手順で行う。

始めに、文形の変換と述語詞の語の名詞化ならびに除去により、標題文中の述語詞節や句の関数表現を、複合名詞句表現に対応する簡潔な関数表現に変換する。つぎに、同じ意味内容をもち、形が異なる関数表現があれば、同義変換により標準の表現に変換する。

[1] 文形の変換

対象間の基本的な意味関係は同じで表現形式が異なる文形に、受動形と能動形、使役形と非使役形がある。受動形や使役形の関数表現は、つぎの変換規則(7)~(9)により能動形で非使役形の関数表現に統一するものとする。ここに、記号‘=’は同義関係を示し、特殊格のラベル *V* と *C* はそれぞれ、態格と使役格を示す。また、ACTORは能動形において主格となる受動形における格ラベルであり、CACTOR (causa-

tive actor) は使役形において主格となる非使役形における格ラベルで、その項が‘物’のカテゴリに属するなら道具材料格 INST、‘事象’のカテゴリに属するなら手段格 MEANS である。

(a) 受動-能動変換

$p(V=受動, SBJ=t_1-が, ACTOR=t_2-により, \dots) = p(V=能動, SBJ=t_2-が, OBJ=t_1-を, \dots)$ (7)

(b) 使役-非使役変換

(i) *p* が自動詞のとき

$p(C=使役, SBJ=t_1-が, OBJ=t_2-を, \dots) = p(C=非使役, SBJ=t_2-が, CACTOR=t_1-により, \dots)$ (8)

(ii) *p* が他動詞のとき

$p(C=使役, SBJ=t_1-が, IOBJ=t_2-に, DOBJ=t_3-を, \dots) = p(C=非使役, SBJ=t_2-が, OBJ=t_3-を, CACTOR=t_1-により, \dots)$ (9)

例3 補獲させる (C=使役, SBJ=デバイス-が, IOBJ=補獲中心-に, DOBJ=電子-を) = 補獲する (C=非使役, SBJ=補獲中心-が, OBJ=電子-を, INST=デバイス-により)

なお、この他、動詞の連体形における現在形と過去形、例えば‘~を含んだ膜’と‘~を含む膜’や、可能の助動詞を含む形と含まない形、例えば‘~を形成しうる金属’と‘~を形成する金属’などのように、意味関係がほとんど同じで表現が異なる文形の組がある。ここでは簡単のため小さい意味の差異は無視し、可能の助動詞を含まない現在形に統一するものとする。

このように変換した後では、述語詞の語の特殊格の項は一意的に指定されるので、関数表現の特殊格およびその項を省略することができる。

[2] 述語詞の語の名詞化と除去

[1] で変換された関数表現の中に含まれるすべての述語詞の語を、その語幹などをとって名詞化した事象名詞の語でおきかえ、各項のすべての後置記号を除去する。

例4 2.1の関数表現(2)は、

‘薄膜 (COMPO=フェライト, OBJ⁻¹=処理 (MEANS=アルカリ化成))’なる関数表現に簡略化される。

2.1で述べたように、Table 1, 3 に示す格ラベルが表わす意味関係として、‘物’に対して成分 (CO-

MPO) や属性 (ATR), '事象' に対して道具材料 (IN-ST) や手段 (MEANS) などがある。このような意味関係を示す格ラベル K_i を含む表現:

$$f(K_i := t_i) \quad (10)$$

に対応する名詞句に対し, '所有', '含有', '使用', '応用', '形成' や '処理' などの事象名詞の語 p' を用いた表現:

$$f(L_i^{-1} := p'(L_i := t_i)) \quad (11)$$

に対応する等価な名詞句が存在する。

式 (10) は, 式 (11) に比べ事象名詞の語 p' を用いないより簡潔な表現であり, 文献検索においてより能率的である。これより, 式 (11) の事象名詞の語 p' が予め定めた特定の事象名詞に属する場合には, 式 (11) の表現を式 (10) のような簡略化した表現に変換し, 表現を統一する。ここに, 格ラベル K_i は, **Table 6** に示すように, 受けの語 f のカテゴリ, 係りの語 (句) t_i のカテゴリならびに事象名詞の語 p' に対する語 (句) t_i の格ラベル L_i から定められる。

f に対する t_i の格ラベル K_i が **Table 6** に設けられていない場合, すなわち式 (11) の名詞句表現に対応する式 (10) の等価な名詞句表現が存在しない場合には, 格ラベル K_i として式 (11) における t_i の格ラベル L_i を用いるものとする。

例5 例4の関数表現においては, '処理' が除去可能な事象名詞の語で, '薄膜', 'アルカリ化成' はそれぞれ, '物', '事象' のカテゴリに属し, さらに '処理' に対して 'アルカリ化成' の格ラベルは MEANS である。これより, **Table 6** を参照して, 例4の関数表現は,

'薄膜 (COMPO := フェライト, OBJ⁻¹ := アルカリ化成)' に簡略化される。

[3] 同義変換

前節まで, 関数表現を複合名詞句表現に対応する形にまで簡略化してきたが, 簡略化した関数表現において, 修飾関係が逆転しても, もとの表現と同一の意味内容を指定すると考えられる表現が存在する。このような修飾関係には,

- (a) 物 f のもつ動作 f' とその動作の様態 f'' の関係
- (b) 物 f のもつ属性 f' とその属性の形容 f'' の関係
- (c) 物 f のもつ属性 f' とその属性の操作 f'' の関係

などがあり, 「 f'' の f' をもつ f 」は「 f' が f'' の

Table 6 Removal of verbal nouns.

事象名詞 p'	格ラベル L_i	受けの語 f のカテゴリ	係りの語 (句) t_i のカテゴリ	格ラベル K_i
所有	OBJ	物・現象体	属性	ATR
	SBJ	属性	物・現象体	GSD
所含	OBJ	部品	材料	COMPO
	SBJ	装	部品	COMPO
形成・処理	MEANS	物・現象体	事象	OBJ ⁻¹
	OBJ	事象	物・現象体	INST
使用	PURP	物・現象体	事象	MEANS
		事象	事象	INST ⁻¹ PURP

Table 7 Sets of cases for equivalence-relations.

	カテゴリの組			格ラベルの組			
	f	f'	f''	K_1	K_1'	K_2	K_2'
(a)	物	動作	形容	SBJ ⁻¹	MOD	EXH ⁻¹	SBJ
(b)	物	属性	形容	ATR	SBJ ⁻¹	EXH ⁻¹	SBJ
(c)	物	属性	操作	ATR	OBJ ⁻¹	EXH ⁻¹	OBJ

f 」と同義であると考えられ, 次式のような関係が成立つ。

$$f(K_2 := f''(s, K_2' := f'(t), r)) = f(K_1 := f'(K_1' := f''(s), t), r) \quad (12)$$

ここに, K_1, K_1', K_2, K_2' は, **Table 7** に示すような格ラベルの組である。また, s, t, r は任意のいくつかの項とそれらの格ラベルをまとめて示す記号 (空の場合も含む) である。

ここでは, 式 (12) の右辺の形のを標準形とし, 式 (12) の左辺の表現を右辺の標準形に変換する。

例6 (a) 回路 (EXH⁻¹ := 容易 (SBJ := 検出 (OBJ := 故障))) = 回路 (SBJ⁻¹ := 検出 (OBJ := 故障, MOD := 容易)) すなわち, '故障検出容易な回路' は '故障を容易に検出する回路' と同義とする。

(b) 物質 (EXH⁻¹ := 高 (SBJ := 導電率)) = 物質 (ATR := 導電率 (SBJ⁻¹ := 高)) すなわち, '導電率が高い物質' は '高導電率の物質' と同義とする。

(c) トランジスタ (EXH⁻¹ := 最適化 (OBJ := 線形特性)) = トランジスタ (ATR := 線形特性 (OBJ⁻¹ := 最適化)) すなわち, '線形特性が最適化されたトランジスタ' は '最適化された線形特性をもつトランジスタ' と同義とする。

2.1 の関数表現 (3) から, 以上述べた [1], [2], [3] の変換を施してえられる表現を標準形とよぶ。

4. 検索

質問の形式は、

(a) 文献標題文のような名詞節または名詞句の自然語表現

(b) 3. で述べた標準形の関数表現

の2種類が考えられるが、(a)の形は前章で述べた方法により(b)の標準形に変換されるので(b)の形について考察する。なお、本章では検索の点から、標準形の中に含まれる名詞語をキータームとよぶ。また、あるキータームに対し、その下位のカテゴリに属するキータームを下位のキータームとよぶ。

検索用ファイルには、公知のように逐次ファイル(serial file)と逆ファイル(inverted file)があるが、逐次ファイルの場合も同様であるので、ここでは逆ファイルを用いる場合について述べる。

逆ファイルの中のキータームは、下位のキータームを含む文献の検索を能率的に行うため、Table 2 のようなカテゴリを最下位のキータームまで階層的に分類しておく。そして、キータームに続いて Fig. 1 のように、これらのキータームを標準形の関数表現の中を含む文献の番号とそのキータームの関数表現における位置記号を付加しておく。ここに、キータームの位置記号をつぎのように表わす。

(i) 標準形の先頭キータームの位置記号を空とする。

(ii) 位置記号 α をもつキータームの格 K に含まれる先頭キータームの位置記号を $\alpha \cdot K$ とする。

例えば、'設計(OBJ=抵抗器(COMPO=FET))'におけるキータームの内、'抵抗器'と'FET'の位置記号はそれぞれ、'OBJ'と'OBJ.COMPO'である。

Fig. 1 に半導体部門におけるファイルの一例を示す。図のファイルには、つぎのような標題文の文献番号を位置記号とともに記録している。

① '磁気記憶用フェライト薄膜'

薄膜 (COMPO=フェライト, INST⁻¹=磁気記憶)

② 'パーマロイ薄膜のトルク曲線'

トルク曲線 (GSD=薄膜 (COMPO=パーマロイ))

③ 'スパッタリングによる Bi₄Ti₃O₁₂ 薄膜の作製'

作製 (OBJ=薄膜 (COMPO=Bi₄Ti₃O₁₂,

OBJ⁻¹=スパッタリング))

検索は以下に示す様式(1),(2)の順に行われるものとする。

様式(1) 質問の関数表現に含まれるキータームま

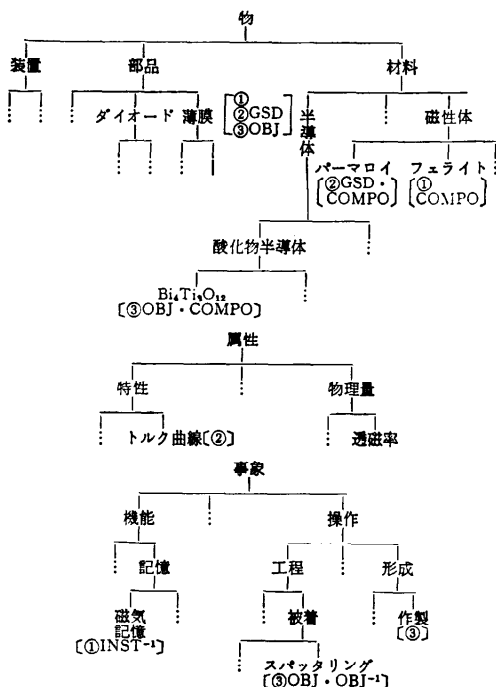


Fig. 1 A part of the inverted file of titles.

たはその下位のキータームを含む文献の番号とそのキータームの位置記号を、逆ファイルから従来のキータームマッチングにより検索する。

様式(2) 様式(1)により検索された文献の中から、つぎに述べる方法により、質問の関数表現に包含される文献を検索する。

様式(2)による検索に用いられる包含関係には、カテゴリの上位・下位関係に基づくものと、全体と部分の修飾関係に基づくものがある。なお、ここでいう包含関係とは、例えば、薄膜 \subseteq 部品、フェライト \subseteq 磁性体や薄膜 (COMPO=フェライト, OBJ⁻¹=アルカリ化成) \subseteq 部品 (COMPO=磁性体) のように、関数記号や関数表現で指定される事物の集合の包含関係とし、通常の集合の包含を示す関係記号 ' \subseteq ' で表わす。

以下に、それぞれの包含関係の規則とそれらに基づく検索手順について述べる。

(i) カテゴリの上位・下位関係に基づくもの
2個の関数表現

$$t \triangleq f(K_1=t_1, K_2=t_2, \dots, K_m=t_m)$$

$$t' \triangleq f'(K_1'=t_1', K_2'=t_2', \dots, K_n'=t_n')$$

において、

$$f \subseteq f' \wedge \bigvee_{1 \leq j \leq n} \exists (K_i = K_j' \wedge t_i \subseteq t_j') \quad (13)$$

が成立つなら、 $t_i \subseteq t_j'$ が成立つものとする。

標準形における格ラベルに関する情報は、前述のように、ファイルでは位置記号で表わされるので、質問に包含される文献を検索する手順は (13) によりつぎのようになる。

手順1 キータームとして、質問の関数表現に現われるキーターム t_j' ($j=1, 2, \dots, n$) またはその下位のキータームを含み、位置記号として、 t_j' の位置記号 α_j' ($j=1, 2, \dots, n$) を含む関数表現の文献を検索する。

すなわち、逆ファイルから様式 (1) により検索される t_j' またはその下位のキータームの内、その位置記号が α_j' と一致する文献の番号の集合を S_j とするとき、

$$\bigcap_{j=1}^n S_j$$

が検索される文献番号の集合である。

例えば、上述の手順1により、Fig. 1の逆ファイルから質問 '作製 (OBJ=薄膜 (OBJ⁻¹=被着))' すなわち、'被着による薄膜の作製' に対して、番号 ③ の文献が検索される。

(ii) 全体と部分の修飾関係に基づくもの

全体と部分の修飾関係には、物とその中に含まれる物との関係、物とその属性との関係、目標事象とそれに含まれる手段事象との関係などがある。主題表現のこのような修飾関係においては、'部分 f を含む全体 f' の研究、調査' は '全体 f' に含まれる部分 f に関する研究、調査' を包含するものと考えられ、 f' または f のあとに研究、調査などの語が省略されているものと考えると、次式のような包含関係が成立する。

$$f(t, K_1=f'(s)) \subseteq f'(s, K_2=f(t)) \quad (14)$$

ここに、 t や s は任意のいくつかの項とその格ラベルをまとめて示す記号 (空の場合も含む) であり、 K_1, K_2 は Table 8 に示すような全体と部分の修飾関係を表わす格ラベルの対である。

さらに、(14) の右辺において、 $K_2=f(t)$ による限定を除いても (13) の規則により、

$$f(t, K_1=f'(s)) \subseteq f'(s) \quad (15)$$

の包含関係が成立つ。

いま、2個の関数記号 f, f' に対し、(14) または (15) が成立つものとする。(14)、(15)の規則からわかるように、左辺の f の引数に現われる全体 $f'(s)$ の各キータームは、右辺の対応するキータームに対し位

Table 8 Case-pairs of a part and the whole.

K_1	K_2
EXH (提示格)	COMPO (成分格)
GSD (対象格)	ATR (属性格)
PURP (目標格)	MEANS (手段格)

置記号 K_1 を余分に含む。逆に、右辺の f' の引数に現われる部分 $f(t)$ の各キータームは、左辺の対応するキータームに対し位置記号 K_2 を余分に含む。

これより、(14)、(15)に基づき質問に包含される文献を検索する手順はつぎのようになる。

手順2 質問の関数表現に現われるキーターム t_j' とその位置記号 α_j' に対し、様式 (1) により検索される t_j' またはその下位のキータームの内、その位置記号が

α_j' に Table 8 の左列の格ラベル K_1 を加えた記号列、または、 α_j' から Table 8 の右列の格ラベル K_2 を除いた記号列

からなる文献の番号を検索する。このような文献番号の集合を S_j とし、後は手順1と同様に行う。

例えば、質問 '分波器 (COMPO=共振器)' すなわち、'共振器で構成された分波器' に対し、上述の手順2により、'共振器 (ATR=リング状, EXH=分波器)' すなわち、'分波器におけるリング状共振器' なる標題文の文献が検索される。

5. シミュレーション結果

標題検索の小規模な計算機シミュレーションを行った。使用計算機はミニコン FACOM U-200、使用言語はアセンブリ語であり、構文分析と標準化のプログラムに約 18KB、検索のプログラムに約 3.4KB の容量を要した。以下に入出力結果の例を示す。

(22) 'パーマロイ薄膜のインピーダンス特性'

(52) 'パーマロイ薄膜の耐食特性'

なる標題文はそれぞれ、

((22) (インピーダンストクセイ (GSD (ハクマク (COMPO パーマロイ))))))

((52) (タイシヨクトクセイ (GSD (ハクマク (COMPO パーマロイ))))))

の標準形に変換され、これらの文献は質問:

'磁性体薄膜の特性'

((R3) (トクセイ (GSD (ハクマク (COMPO ジセイタイ))))))

に対して検索された。また、

(68) 'MOS トランジスタにおけるコンダクタンスパラメータの精密測定'

(51) 'MOS 電界効果型トランジスタの基板濃度測定' なる標題文はそれぞれ、

((68) (ソクテイ (OBJ (コンダクタンスパラメータ (GSD モストランジスタ))) (MOD セイミツ)))

((51) (ソクテイ (OBJ (ノウド (GSD (キバン (EXH モスデンカイコウカガタトランジスタ))))))

の標準形に変換され、これらの文献は質問：

'MOS トランジスタにおける物理量の測定'

((R 4) (ソクテイ (OBJ (ブツリリョウ (GSD モストランジスタ))))))

に対して検索された。文献 (22), (52), (68) は、カテゴリの上位・下位関係に基づく包含関係 (13) により、文献 (51) は、(13) と全体と部分の修飾関係に基づく包含関係 (15) により検索されたものである。

6. む す び

以上、本文で述べたように、節や句に現われる係り受け関係を確定するために、動詞を実質的なリンク要素として細分化したカテゴリの組を導入することにより、これらの関係をかなりの程度確定することができた。

また、検索漏れを防ぐために、二、三の係り受け表現の同義関係、包含関係を導入したが、標題検索では本稿に述べたものでかなりのものをおおうものと考え

る。格関係を含む複雑な構造の検索を行うわりには、その検索速度や二次ファイルの所要記憶容量は従来の逆ファイルのキータームマッチングとあまり変わらず、検索効率の面でも十分実用的であると思われる。

本稿では主に文献標題文に対して述べたが、特許請求範囲文に対しても、基本原理は標題文の場合とほぼ同様であり、本稿の手法を適用しうると考えられる。

参 考 文 献

- 1) 安部, 豊田, 田中: 文献標題を利用した索引作製の自動化に関する考察, 情報処理, Vol. 11, No. 12, pp. 699~710 (1970).
- 2) 長尾, 辻井, 田中: 意味および文脈情報を用いた日本語文の解析, 情報処理, Vol. 17, No. 1, pp. 10~28 (1976).
- 3) Y. Wilks: An Intelligent Analyzer and Understander of English, Commun. ACM, Vol. 18, No. 5, pp. 264~274 (1975).
- 4) W. M. Turski: On a Model of Information Retrieval System Based on Thesaurus, Inform. Stor. Retr., Vol. 7, pp. 89~94 (1971).
- 5) 中村: 情報処理 I, 共立出版 (1969).
- 6) F. W. Lancaster, 松村訳: 情報検索の言語, 日本ドクメンテーション協会 (1976).
- 7) 西田, 高松: 制限された日本文の論理式への変換と情報の抽出, 信学会論文誌 D, Vol. J 59-D, No. 8, pp. 515~522 (1976).

(昭和 52 年 9 月 26 日受付)

(昭和 53 年 4 月 17 日再受付)