Vol. 25 No. 2 情報処理学会論文誌 Mar. 1984

技術抄録文からの関係情報の自動抽出†

高松 忍†† 日下浩次†† 西田富士夫††

本論文は，特許請求範囲文などの技術抄録文（日本語）から，抽出項目を指定したフレームを用いて構造情報を抽出し，関係形式などにデータベース化する手法を与えている．従来，キータームの自動インデクシングの研究が行われているが，ここでは一歩進めて，キーターム間の関係情報を自動的に抽出する手法について考える．抽出すべき情報は，ある主題タームについて述べたいくつかのサブフレームから成る仕様フォーマットにより指定する．仕様フォーマットは技術分野ごとに設定し，その各サブフレームは格構造の形で記述する．入力文の構文解析は，動詞の格構造パターンや各分野の専門知識を用いて，ボトム・アップ法で行い，格構造形式の内部表現に変換する．続いて，内部表現を仕様フォーマットの形に標準化し，標準化した内部表現から仕様フォーマットを用いてターム間の関係情報を抽出する．このようにしてえられた抽出情報は，関係形式などのファイルに蓄積される．

1. まえがき

近年，自然言語処理技術やデータベース技術の進展とともに，科学情報などの知識データベースシステムの研究が盛んに行われている．本来，知識とは，ターム間の関係や事象間の関係を表す構造的なものであり，このような情報を技術テキストから自動的に抽出・収集することは，知識データベース構成の基本問題である．従来，キータームの自動インデクシングの研究が行われているが，ここではさらに一歩進めて，キーターム間の指定された関係情報を自動的に抽出する手法について考える．

本論文では，特許請求範囲文などの技術抄録文を対象とし，これから，抽出項目を指定したフレームを用いて構造情報を抽出し，関係表の形式などにデータベース化する手法を与えている．

抽出すべき情報は，いくつかのサブフレームから成る仕様フォーマットにより指定する．仕様フォーマットは，各技術分野ごとに設定し，その各サブフレームは格構造の形で記述する．

初めに，動詞構文パターンとその格構造パターンや各分野の専門知識を用いて技術抄録文を解析し，格構造形式の内部表現に変換する．続いて，内部表現を仕様フォーマットのサブフレームに類似した形に標準化し，標準化した内部表現から仕様フォーマットを用いて構造情報を抽出する．このようにして得られた抽出情報は，関係検索を行うため，関係データベースの形などに変換して蓄積される．

2. 抽出項目の仕様

抽出すべき構造情報は，ある主題項目について述べた何個かのサブフレームで表し，その枠組を仕様フォーマットとよぶ．式(1)にその形式を示す．

$$\{L^j;(K_1^j\text{-}C_1^j:\underline{\ },\cdots,K_i^j\text{-}C_i^j:\underline{t},\cdots,K_{n_j}^j\text{-}C_{n_j}^j:\underline{\ })\mid j=1,2,\cdots,m\} \quad (1)$$

式(1)において，各サブフレームの左端のラベル $L^j(j=1,2,\cdots,m)$ は，そのサブフレームの名前を示し，その右の表現は，いくつかの格ラベルと意味カテゴリ名の対から成る格構造を示す．ここに，$K_k^j(k=1,2,\cdots,n_j)$ は格ラベルを示し，C_k^j はその格に入るタームの属すべき意味カテゴリ名を示す．また，コロン ':' の右の下線は，その格のスロットを示し，とくに，二重下線は，そのサブフレームにおける主ターム（主題）t が入る格のスロットを示す．

表1(a)，(b)は，半導体分野に関する仕様フォーマットを示す．表1(a)は，半導体装置一般に関する仕様フォーマットを示し，機能 (FUNCTION)，性質 (PROPERTY)，構成関係 (COMPOSITION)，位置関係 (LOCATION) およびプロセス (PROCESS) の五つのサブフレームから成る．主装置の構成要素に関する情報は，構成要素を主タームとしてこの表を繰り返し適用することにより抽出される．

表1(b)は，同じく半導体装置に関して，製造プロセスを主題として情報抽出する仕様フォーマットを示し，プロセス (PROCESS)，原理・手段 (PRINCIPLE) およびサブプロセス (SUB-PROCESS) の三つのサブ

† Automatic Extraction of Relational Information from Technical Abstracts by SHINOBU TAKAMATSU, HIROJI KUSAKA 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

表 1(a) 半導体装置に関する仕様フォーマット
Table 1(a) A specification format for semiconductor devices.

FUNCTION; (PRED-Physical ACTion: —, AGent-PRODUCTS: *t*, OBJect-PHYSOBJ ∪ QUANTity: —, CONDition: —)
PROPERTY; (PRED-ATTRibute: —, OBJ-QUANT ∪ PROPERTY: —, LOCation-PRODUCTS: *t*, DEGRee-VALue: —, PARTICipant-PRODUCTS: —)
COMPOSITION; (OBJ-PRODUCTS: *t*, COMPonent-PRODUCTS: —)
LOCATION; (OBJ-PRODUCTS: *t*, LOC-PRODUCTS: —)
PROCESS; (PRED-PACT: —, GOal-PRODUCTS: *t*, OBJ-PHYSOBJ: —, INSTRument-PHYSOBJ: —, LOC-PHYSOBJ ∪ PHYSLOC: —, COND: —, MANNer: —, MEANS-PACT: —)

表 1(b) 半導体装置の製造プロセスに関する仕様フォーマット
Table 1(b) A specification format for the manufacturing processes of semiconductor devices.

PROCESS; (PRED-PACT: *t*, OBJ-PHYSOBJ: —, GO-PRODUCTS: —, INSTR-PHYSOBJ: —, LOC-PHYSOBJ ∪ PHYSLOC: —, COND: —, MANN: —)
PRINCIPLE; (PRED-PACT: *t*, MEANS-PRINCIPLE ∪ PACT: —)
SUB-PROCESS; (OBJ-PACT: *t*, COMP-PACT: —)

フレームから成る．サブプロセスの情報を指定したレベルまで抽出するために，これらを主タームとして表1(b)が繰り返し適用される．

3. 構文解析

3.1 格構造の構成

技術抄録文の構文解析は，動詞構文パターンの上に構成した格構造パターンを用いて行われる．ここに，動詞構文パターンは，文献 5) の格助詞パターンの分類にほぼ従っている．筆者らは，この動詞構文パターンの上に格ラベルと意味カテゴリの対を付与し格構造パターンを構成した．ここに，意味カテゴリの体系および格ラベルは文献 7) と同じものを用いる．

表2にパターン VP 9{主格＋対格＋与格} に属する動詞語の格構造パターンを示す．第1行目は，VP 9 の動詞構文パターンにおけるおもな構成要素の統語上の役割名であり，第2行目以下は，動詞語の意味カテゴリごとに構成した格構造において，格ラベルとその格に入る語の意味カテゴリ名の対の組を示す．ただし，VP 9 パターンでは，主格は動詞語の意味カテゴリにかかわらず AGent-HUMan をとり，技術文ではほとんどの場合省略されるので，この部分は表から除いた．

表 2 動詞パターン VP 9 の格構造
Table 2 Case structures of verb pattern VP 9.

VP 9	対 格 (を)	与 格 (に)
PRED-Physical TRANSfer (2151～2154, 2156)*	OBJ-PHYSOBJ (12, 14～15)	(に向けて) GOal-PHYSOBJ ∪ PHYSLOC (117)
PRED-POSSessive TRANS (237)	OBJ-OBJECT (12～15)	(に対して) RECIPient-PHYSOBJ
PRED-Mental TRANS (231)	OBJ-MENTOBJ (130～132)	(に対して) RECIP-HUM (12)
PRED-CONNECTion (235, 2155)	OBJ-PHYSOBJ	PARTIC-PHYSOBJ
PRED-PRODuction (2121～2123, 2386)	OBJ-PHYSOBJ	(において) LOC-PHYSOBJ ∪ PHYSLOC
PRED-USE (23852)	OBJ-THINGS (1～3)	(のために) PURPOSE-ACTion (215～219, 23, 25)
PRED-ATTRibute TRANS (2157～2158)	OBJ-PHYSOBJ ∪ QUANTity (114, 1192～1197)	GO-VAL (1191)

* ( )内の数字は，国立国語研の分類語彙表に基づいて付けた意味カテゴリの分類コードを示す．意味カテゴリ間の階層関係の判定にはこの分類コードを用いる．すなわち，コード c_1 の長さがコード c_2 の長さより短く，c_1 が c_2 の上位桁に含まれるなら，c_1 の意味カテゴリは c_2 の意味カテゴリの上位カテゴリである．

動詞構文パターンの構成要素の統語上の役割を示すものとしては，表2の与格の欄に示すように，格助詞の他に，'に対して'，'において' などの複合助辞がある．これらの後置詞は，格助詞よりもさらに用法が限定されており，深層格を決める上に有効な手がかりを与える．

構文解析において，動詞語に対する係りの語の格ラベルは，係りの語の統語上の役割とその意味カテゴリならびに動詞語の構文パターンと意味カテゴリから，上記のような格構造パターンの表を用いて整合するものを選定する．各語の統語上ならびに意味上のカテゴリや動詞語のとる構文パターン名などは，単語辞書に記載しておく．

例1：次の各文は，文法的には同じ与格の後置詞'に'を含んでいるが，表2の格構造パターンと動詞語の意味カテゴリとから，異なる意味上の格ラベルがつけられる．

(1) 半導体基板表面に　酸化膜を
LOC-PHYSLOC (1175)　OBJ-PRODUCTS (14)
形成する
PRED-PROD (2122)

(2) ゲート構造の設計に
PURPOSE-THINKACT (2308)
2次元解析を　用いる
OBJ-THINKACT (2307)　PRED-USE (23852)

(3) 炉の温度を　500℃ に
OBJ-QUANT (1193)　GO-VAL (11911)
高める
PRED-ATTRTRANS (21582)

(4) コレクタ電極を　電源に
OBJ-PRODUCTS (14)　PARTIC-PRODUCTS (14)
接続する
PRED-CONNECT (21554)

動詞の構文パターン部に含まれない副詞句や副詞節の動詞語に対する格関係，ならびに，名詞句における係りの語の受けの名詞語に対する格関係は，文献6)とほぼ同様な手法により定められる．

構文解析は，以上に述べた格構造パターンなどを用いて行い，文および名詞句の内部表現をそれぞれ，式(2)，式(3)のように構成する．

$$(K_0\text{-}C_0 : t_0,\ K_1\text{-}C_1 : t_1, \cdots, K_n\text{-}C_n : t_n) \quad (2)$$

$$t_0'(K_0'\text{-}C_0' : *,\ K_1'\text{-}C_1' : t_1', \cdots, K_m'\text{-}C_m' : t_m') \quad (3)$$

ここに，t_0は文における受けの述語を，t_0'は名詞句における受けの名詞語を示す．$t_j(j=1,2,\cdots,n)$，$t_j'(j=1,2,\cdots,m)$はそれぞれ，t_0, t_0'を修飾する係りの句や節の内部表現を示し，これらはまた式(2)または式(3)の形をとる．$K_j\text{-}C_j$ $(j=0,1,\cdots,n)$，$K_j'\text{-}C_j'$ $(j=0,1,\cdots,m)$はそれぞれ，t_j, t_j'の格ラベルとその意味カテゴリ名の対を示す．また，記号'*'は名詞句の格構造における受けの名詞語t_0'の位置を示す．

構文解析は，拡張LINGOL[4]の手法に基づいたボトム・アップ法で行う．解析途中において，整合しない格構造が生じればその格構造を棄却し，整合する格構造だけを並列に構成していく．

3.2 専門的知識の利用

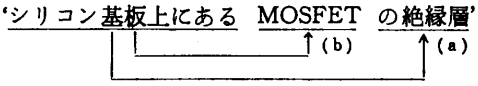
特許請求範囲文などは，複数個の節や句で修飾された多くの句や節を含み，このため，文法的制限や前節で述べた一般的な格構造パターンによる制限だけでは，係り受け関係にあいまいさが生じることが多い．この場合には，入力文における2項関係が専門分野における一般的関係の事例などになっており，その分野の一般的知識から説明できるかどうかで解決できることが多い．

専門的知識は，対象の性質や対象間の関係を表す構造的なものであり，これを仕様フォーマットのサブフレームに類似した関係表の形（以下，知識表とよぶ）で記述する．ここで，知識表の属性名としてサブフレームの格ラベルを用いる．これらのサブフレームは，物とその構成要素，構成要素間の位置関係，物とその性質や機能，プロセスとその生成物やサブプロセスなどの基本的関係を表す．**表3**に，半導体装置(MOSFET)のフレームに関する知識表を示す．知識表は，専門分野ごとに，上位カテゴリの専門用語を用いて一般的に記述し，単語辞書中の専門用語には，その語またはその上位語を含む知識表の格ラベルへのポインタを記載しておく．

知識表のサブフレームは，表3のCOMPOSITION表やLOCATION表に示すように，'(～が～を) 含む'，'(～が～に) 在る'などのリンク述語を省いた句表現に近い表現や，PROCESS表に示すように，'(～することにより～を) 形成する'，'(～するのに～を) 用いる'などの一般的述語を省いた簡略表現で表している．

一方，入力文から構成された格構造では，上記のようなリンク述語や一般的述語を含んだいろいろな表現がある．したがって，知識表にアクセスするためには，入力文のいろいろな格構造を知識表のサブフレームの形に変換しなければならない．ここでは，**表4**に示すようなアクセス表により，入力文の格構造を知識表のサブフレームに対応づけ，入力文の正しい係り受け関係を同定する手順について述べる．

次のような名詞句を考えよう．



この名詞句を前節のように格構造パターンを用いて

表 3 半導体装置 (MOSFET) に関する知識表
Table 3 Knowledge tables of semiconductor devices (MOSFETs)

COMPOSITION

OBJ	COMP
MOSFET (S1132)*	{半導体基板(S12), 絶縁層(S131), ソース電極(S142), ゲート電極(S143), ドレイン電極(S144)}
半導体基板 (S12)	{ソース(S15), ドレイン(S16), ゲート(S17), チャンネル(S18)}

LOCATION

OBJ	LOC
絶縁層(S131)	-on 半導体基板 (S12)
ソース電極(S142)	-on ソース (S15)
ドレイン電極(S144)	-on ドレイン (S16)
ゲート電極(S143)	-on ゲート(S17) ∪ チャンネル (S18)
ゲート∪チャンネル (S17) (S18)	-between〈ソース(S15), ドレイン (S16)〉

PROCESS

プロセス順序	PRED	OBJ	GO	LOC
(1)	熱酸化 (S2411)	半導体基板 (S12)	絶縁層 (S131)	-on 半導体基板 (S12)
(2)	エッチング (S23)	絶縁層 (S131)	拡散窓 (S191)	-in 絶縁層 (S131)
(3)	拡散 (S211)	不純物 (S32)	{ソース(S15), ドレイン(S16)}	-in 半導体基板 (S12)
			ゲート (S17)	-between 〈ソース(S15), ドレイン(S16)〉
(4)	蒸着 (S2211)	金属 (S341)	ソース電極 (S142)	-on ソース (S15)
			ドレイン電極 (S144)	-on ドレイン (S16)
			ゲート電極 (S143)	-on ゲート∪チャンネル (S17) (S18)

* ( )内の数字は，日本特許情報センターのターム一覧表(半導体部門) に基づいて付けた階層分類のコードを示す．

一般的に解析すると，(a)，(b)の2通りの係り受け関係に対応して，次のような格構造が構成される．

(a′) (PRED-EXIS：ある，OBJ：絶縁層, (S131)
LOC：シリコン基板) (S121)

(b′) (PRED-EXIS：ある，OBJ：MOSFET, (S1132)
LOC：シリコン基板) (S121)

表 4 入力文の格構造から知識表へのアクセス表
Table 4 An access table from the case structures of input sentences to the knowledge tables.

入力文の格構造			知識表の格ラベル	知識表名
(a)	PRED-INCLusion(含む,…)		—	COMPOSITION
	OBJ		OBJ	
	PARTIC		COMP	
(b)	PRED-EXIStence (在る,…)		—	LOCATION
	OBJ		OBJ	
	LOC		LOC	
(c1)	PRED-PRODuction (形成する,…)		—	PROCESS
	OBJ		GO	
	LOC		LOC	
	MEANS	PRED	PRED	
		OBJ	OBJ	
(c2)	PRED		PRED	PROCESS
	OBJ		OBJ	
	PURPOSE	PRED-PROD (形成する,…)	—	
		OBJ	GO	
		LOC	LOC	

つぎに，この例について，(a′)，(b′) のいずれの格構造が知識表の一般的関係の事例になっているかを調べてみよう．

初めに，入力文の格構造における述語の意味カテゴリ (述語がなければ他の必須格の格ラベル) をキーとして表4のアクセス表をひき，入力文の格構造に対応する知識表にアクセスし，入力文の格ラベルの組に対応する知識表の格ラベルの組を検索する．続いて，入力文の格構造の格に含まれる専門用語の組に対し，その語またはその上位語よりなる組が知識表の中にあるかどうか，すなわち，知識表の一般的関係の事例になっているかどうかを調べる．

たとえば，(a′)，(b′) の格構造は，表4(b)から表3の LOCATION 表の格ラベルの組：(OBJ；LOC) に対応し，(a′) がこの表の第2行目の組：

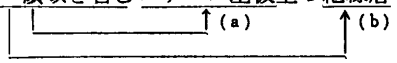
(絶縁層 (S131)；半導体基板 (S12))

の事例になっている．これより，係り受け関係が(a)のように定まる．

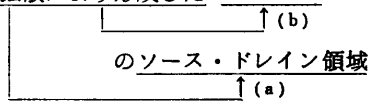
専門知識を用いても係り受け関係のあいまいさが解消しない場合には，システムと人間との対話によって解消する方式をとる．

例2：

(1) チャンネル領域を含むシリコン基板上の絶縁層



(2) 不純物の拡散により形成した MOSFET



(1) 名詞句(1)から，(a)，(b)の2通りの係り受け関係に対応して，次のような格構造が構成される．

(a′) (PRED-INCL：含む，OBJ：シリコン基板，(S121)
PARTIC：チャンネル領域)
(S18)

(b′) (PRED-INCL：含む，OBJ：絶縁層，(S131)
PARTIC：チャンネル領域)
(S18)

(a′)，(b′)の格構造は，表4(a)から，表3の COMPOSITION 表の格ラベルの組：(OBJ；COMP) に対応し，(a′)がこの表の第3行目の組の事例になっている．これより，係り受け関係が(a)のように定まる．

(2) 同様に，名詞句(2)から，

(a′) (PRED-PROD：形成する，
OBJ：ソース・ドレイン領域，
(S15, S16)
MEANS：(PRED：拡散，
(S211)
OBJ：不純物))
(S32)

(b′) (PRED-PROD：形成する，
OBJ：MOSFET，
(S1132)
MEANS：(PRED：拡散，
(S211)
OBJ：不純物))
(S32)

なる格構造が構成される．(a′)，(b′)の格構造は，表4(c1)から，表3の PROCESS 表に対応し，(a′)がこの表の第4行目の組の事例であり，これより，係り受け関係が(a)のように定まる．

4. 標準化

構文解析によって得られる内部表現のなかには，意味関係は同じであるが異なった形の内部表現がある．この章では，仕様フォーマットのサブフレームを標準形にとり，入力文から得られる内部表現の格構造をこれらの標準形に変換する手法について述べる．

仕様フォーマットのサブフレームは，'含む' などのリンク述語や '引き起こす' などの補文をとるリンク述語を含まない．このため，これらのリンク述語を含む内部表現からリンク述語を除き，OBJ 格などの必須格のラベルをリンク述語が表すより具体的な格ラベルで置きかえ，サブフレームの形に標準化する．また，サブフレームは，'成分' や '原因' のような格ラベルを表す名詞語をタームとして含まない．このため，これらの名詞語を含む場合には，これを同格表現に変換し，さらに，これらの語を同格関係にある具体的なタームで置きかえ，サブフレームの形に標準化する．

上記の内部表現の標準化は，以下に述べるような順序で段階的・組織的に行う．ただし，シソーラスによる類義語の標準化の問題にはふれない．

4.1 リンク述語の除去

'含む'，'成る'，'引き起こす'，'用いる' などのリンク述語を含む文は他にいろいろな等価な表現がある．ここでは，まず，リンク述語を除去して同格表現に変換する方法について述べる．

一般に，'t_1 は t_2 を t_0 する' という文の内部表現：

(PRED-C_0：t_0, K_1-C_1：t_1, K_2-C_2：t_2) (4)

に対し，'t_1 は t_2 を t_0 するものである' という同格表現：

(OBJ-C_1：t_1, APPOSition-C_1：
もの (PRED-C_0：t_0, K_1-C_1：＊, K_2-C_2：t_2)) (5)

が存在する．

式(5)において，t_0 が '(物 t_1 は物 t_2 を) 含む' などのリンク述語のときには，式(5)の APPOS 格の表現は，リンク述語を省略して，

s(COMPOSITE-PHYSOBJ：＊,
OBJ-PHYSOBJ：t_2) (6)

と簡単化され，s は 'もの' と表すことができる．

また，式(5)の t_2 が

(PRED-C_0'：t_0', K_1'-C_1'：t_1', …)

なる述語表現で，t_0 が '(t_1 は t_2 を) 引き起こす' などの補文構造 (埋め込み述語構造) をとるリンク述語のときには，式(5)の APPOS 格の表現は，リンク述語を省略して，

s(CAUSE-THINGS：＊,
PRED-C_0'：t_0', K_1'-C_1'：t_1', …) (7)

と簡単化される．

上記のようなリンク述語を含まない同格表現への変

表 5 リンク述語の除去
Table 5 Removal of linking predicates.

	C_0	t_0 の例	K_1-C_1	K_2-C_2	K_1*	K_2*
(I)	INCLusion	含む	OBJ-PHYSOBJ	PARTIC-PHYSOBJ	COMPOSITE	OBJ
	POSSESSION	有する	OBJ-PHYSOBJ	PARTIC-ATTRibute ∪ QUANT	OBJ	CHARacteristics
			OBJ-PHYSOBJ	PARTIC-PHYSOBJ	COMPOSITE	OBJ
	EXIStence	在る	OBJ-PHYSOBJ	LOC-PHYSOBJ ∪ PHYSLOC	OBJ	LOC
(II)	CAUSE	引き起こす	AG-THINGS	OBJ-ACTion	CAUSE	
	USE	用いる	OBJ-PHYSOBJ	PURPOSE-ACT	INSTR	
			OBJ-ACT	PURPOSE-ACT	MEANS	

換は，次式に示すように規則化することができる．

(PRED-C_0 : t_0, K_1-C_1 : t_1, K_2-C_2 : t_2)
→(OBJ-C_1 : t_1, APPOS-C_1 :
s(K_1^*-C_1 : ＊, K_2^*-C_2 : t_2)) (8)

(PRED-C_0 : t_0, K_1-C_1 : t_1, K_2-C_2 :
(PRED-C_0' : t_0', K_1'-C_1' : t_1', …))
→(OBJ-C_1 : t_1, APPOS-C_1 :
s(K_1^*-C_1 : ＊, PRED-C_0' : t_0',
K_1'-C_1' : t_1', …)) (9)

ここに，式(8)および式(9)はそれぞれ，**表5**の(I)および(II)の欄に示すようなリンク述語 t_0 に対して適用され，K_1^*, K_2^* は，表に示すように，リンク述語の意味カテゴリ C_0，格ラベル K_1, K_2 と意味カテゴリ C_1, C_2 から定められる．

例3：(1a)‘半導性ガラス特性を有する酸化物’
(2a)‘チャンネル領域を含むシリコン基体’
(3a)‘ゲート構造の設計に2次元解析を用いる’

(1a)～(3a)の内部表現はそれぞれ，(1b)～(3b)に示すように，リンク述語を含まない同格表現の形に変換される．

(1b) 酸化物 (PRED-POSSESSION : 有する,
OBJ-PHYSOBJ : ＊,
PARTIC-ATTR : 半導性ガラス特性)
⇨酸化物 (OBJ-PHYSOBJ : ＊,
APPOS-PHYSOBJ :
s(OBJ-PHYSOBJ : ＊,
CHAR-ATTR : 半導性ガラス特性))

(2b) シリコン基体 (PRED-INCLusion : 含む,
OBJ-PHYSOBJ : ＊,
PARTIC-PHYSOBJ : チャンネル領域)
⇨シリコン基体 (OBJ-PHYSOBJ : ＊,
APPOS-PHYSOBJ :
s(COMPOSITE-PHYSOBJ : ＊,
OBJ-PHYSOBJ : チャンネル領域))

(3b) (PRED-USE : 用いる,
OBJ-THINKACT : 2次元解析,
PURPOSE-ACT : (PRED : 設計,
OBJ : ゲート構造))
⇨(OBJ-THINKACT : 2次元解析,
APPOS-THINKACT :
s(MEANS-THINKACT : ＊,
PRED : 設計, OBJ : ゲート構造))

4.2 格ラベルを表わす名詞語の除去

‘t_1 は t_2 の成分である’などのように，格ラベルを表す‘成分’などの名詞語を含む同格表現は，‘t_2 の成分’を‘t_2 の成分であるようなもの’と考えて，前述のように，

(OBJ-PHYSOBJ : t_1, APPOS-PHYSOBJ :
s(COMPonent-PHYSOBJ : ＊,
OBJ-PHYSOBJ : t_2)) (10)

と表すことができる．

式(10)の同格表現において，t_1 と s は同格の関係にあるので，s すなわち＊記号を等価な t_1 で置きかえることができる．これより，式(10)の同格表現は，

(COMP-PHYSOBJ : t_1, OBJ-PHYSOBJ : t_2) (11)

と簡単化される．

他の‘原因’,‘手段’や‘用途’などの名詞語を含む同格表現についても同様であり，これらの名詞語の表す格ラベルを K_1^* とすると，上記の変換は，次式のように規則化することができる．

(OBJ-C_1 : t_1, APPOS-C_1 :

$s(K_1^*\text{-}C_1:*,\ K_2\text{-}C_2:t_2))$
$\rightarrow(K_1^*\text{-}C_1:t_1,\ K_2\text{-}C_2:t_2)$ (12)

ただし，$K_2\text{-}C_2:t_2$ は，格ラベル，意味カテゴリ名およびその格のタームの組の列を示す．

また，前節の式(8)と式(9)の右辺の同格表現も式(12)の規則により，それぞれ，

$(K_1^*\text{-}C_1:t_1,\ K_2^*\text{-}C_2:t_2)$ (13)

$(K_1^*\text{-}C_1:t_1,\ \text{PRED-}C_0':t_0',\ K_1'\text{-}C_1':t_1',\cdots)$ (14)

と簡単化される．

例4：例3(1b)～(3b)の右辺の同格表現はそれぞれ，(1c)～(3c)の表現に変換される．

(1c) 酸化物 (OBJ-PHYSOBJ：*，
CHAR-ATTR：半導性ガラス特性)

(2c) シリコン基体 (COMPOSITE-PHYSOBJ：
*，OBJ-PHYSOBJ：チャンネル領域)

(3c) (MEANS-THINKACT：2次元解析，
PRED：設計，OBJ：ゲート構造)

4.3 格構造の変換

前節では，リンク述語や格ラベルを表す名詞語を除去して，仕様フォーマットの形に変換する手順について述べた．しかし，仕様フォーマットの形をしていても，主題や受けの語の選び方によって，なお表現上形が異なる格構造の組が存在する．これらのなかには，

(a) '対象 t_1 と t_2 の合成物 t_3' と
'対象 t_3 の要素 t_1 と t_2'

(b) '動作 t_2 するために動作 t_1 する' と
'動作 t_1 することにより動作 t_2 する'

(c) {'対象 t_3 は属性 t_1 の物理量 t_2 をもつ'
'対象 t_3 の物理量 t_2 は属性 t_1 である'} と
'対象 t_3 は物理量 t_2 が属性 t_1 である'

に対応する格構造の組がある．(a)～(c)の格構造の組は，次式に示すように，相互に変換可能である．

(OBJ-PHYSOBJ：$\{t_1, t_2\}$,
COMPOSITE-PHYSOBJ：t_3)
$\rightleftarrows$(COMPonent-PHYSOBJ：$\{t_1, t_2\}$,
OBJ-PHYSOBJ：t_3) (15a)

(PRED-ACT：t_1, $K_1\text{-}C_1:t_1$,
PURPOSE-ACT：
(PRED-ACT：t_2, $K_2\text{-}C_2:t_2$))
$\rightleftarrows$(MEANS-ACT：(PRED-ACT：
t_1, $K_1\text{-}C_1:t_1$),
PRED-ACT：t_2, $K_2\text{-}C_2:t_2$) (15b)

{(OBJ-PHYSOBJ：t_3, CHAR-QUANT：
t_2(PRED-ATTR：t_1, OBJ-QUANT：*)),
(PRED-ATTR：t_1, OBJ-QUANT：
t_2(OBJ-PHYSOBJ：t_3,
CHAR-QUANT：*))}
$\rightleftarrows$(PRED-ATTR：t_1, OBJ-QUANT：t_2,
LOC-PHYSOBJ：t_3) (15c)

例5：(1) 例4(2c)の表現は，式(15a)により，つぎの表現と相互に変換可能である．

シリコン基体 (OBJ-PHYSOBJ：*,
COMP-PHYSOBJ：チャンネル領域)

(2) 名詞句 '低い抵抗率をもつ物質' の内部表現を前節の方法により変換した表現は，

物質 (OBJ-PHYSOBJ：*，CHAR-QUANT：
抵抗率 (PRED-ATTR：低い，
OBJ-QUANT：*))

である．この表現は，式(15c)により，名詞句 '抵抗率が低い物質' の内部表現：

物質 (PRED-ATTR：低い，
OBJ-QUANT：抵抗率，
LOC-PHYSOBJ：*)

と相互に変換可能である．

上記の変換を行うかどうかは，与えられた仕様フォーマットに依存する．たとえば，表1(a)の場合では，式(15a)～(15c)の右辺の表現が仕様フォーマットのサブフレームの形であり，したがって，この場合には，式(15a)～(15c)の左辺の表現を右辺の表現に変換する．

5. 項目の抽出とデータベース化

標準化された内部表現から，仕様フォーマットに指定される項目を抽出する．

初めに，内部表現からレベル1の主タームをとる．レベル1の主タームは，特許請求範囲文の場合，名詞句（節）の最後にくる受けの語であり，一方，技術論文のサマリ文の場合は，'提案する' や '述べる' などを述語とする文の OBJ 格の語である．ただし，これらの語が，'研究' や '方法' などの論文のスタイルを表す一般的な語であれば，これらの語の前にくる具体的な語をとるものとする．

つぎに，レベル1の主タームの語を含む内部表現をとり，仕様フォーマットの述語の意味カテゴリ（述語がなければ他の必須格の格ラベル）と主タームの格ラベルをキーとして，その内部表現に対応するサブフ

レームを同定する．サブフレームにマッチすれば，その内部表現から，同定したサブフレームに指定される項目を抽出する．

続いて，抽出されたレベル1の主タームの項目から，2章で述べたように，構成要素やサブプロセスなどの語をレベル2の主タームとし，これらの主タームに関し，上述と同様な抽出を行う．以下，指定したレベルまで同様な抽出を繰り返す．

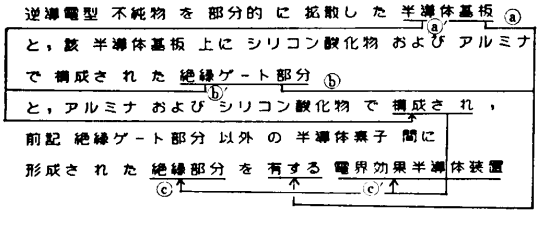
抽出された項目は，関係形式のファイルなどに蓄積される．関係形式の場合は，関係表を仕様フォーマットの各サブフレームごとに構成する．すなわち，サブフレームの名前を関係名とし，サブフレームの中の格ラベルを属性名とする．ただし，INSTRument などの自由格の部分は，別表として分離し，自由格の値がもとのどの組に属するかを示すため，その組番号をその表に加える．

6. 実験結果

技術抄録文の構文解析，標準化および項目抽出の計算機実験を行った．使用計算機は ACOS-700，使用言語は LISP であり，処理に必要な容量は作業領域を含めて約 300 k バイトである．構文解析から項目抽出までの処理時間は，インタプリタモードで技術抄録文1件（約150字）当り，約30秒であった．ただし，単語の辞書引き時間を含まない．構文解析および標準化に用いた規則は，活用などの変化を表すためインデックス付きとし，インデックス付きの規則がそれぞれ，約30個，約15個である．また，用いた格ラベルと意味カテゴリの個数はそれぞれ，約 25，約 35（内，事象のカテゴリ約 20）である．対象とした技術抄録文は特許請求範囲文など約 80 件（内，計算機処理したもの 10 数件，その機械辞書の語数約 600）である．

以下に，入出力結果の例を示す．

（a） 特許請求範囲文(標題：電界効果半導体装置)



文(a)では，係り受け関係に関して，ⓐまたはⓐ′，ⓑまたはⓑ′，ⓒまたはⓒ′のあいまいさがある．しかし，3.2 節で述べた方法により，専門知識を用いるとⓐ，ⓑ，ⓒの係り受け関係に定まり，つぎの内部表現が得られる．

（b） 内部表現

```
(電界効果半導体装置
 ((PRED : (有する))
①(PARTIC :  ---
   (*PARA*
    (絶縁部分
     ((PRED : (形成))
      (GO : ( * ))
      (LOC :
       (半導体素子
        ((OBJ : ( * ))
         (LOC : (以外 ((LOC : ( * )) (OBJ : (絶縁ゲート部分
((OBJ : ( * )) (DET (DEM))))))))))))
     ((PRED : (構成)) (OBJ : ( * )) (INSTR : (
      ③    *PARA* (アルミナ) (シリコン酸化物)))))
    (絶縁ゲート部分
     ((PRED : (構成))
     ②(OBJ : ( * )) ---
      (INSTR : (*PARA* (シリコン酸化物) (アルミナ)))
      (LOC : (半導体基板 ((OBJ : ( * )) (DET (DEM)))))))
    (半導体基板
     ((PRED : (拡散))
      (OBJ : (不純物 ((OBJ : ( * )) (CHAR : (逆導電型)))))
      (MANN : (部分的))
      (GO : ( * ))))))
 (OBJ : ( * ))))
```

上の内部表現では，簡単のため，述語の TENSE，VOICE および語の意味カテゴリー名を省略した．

内部表現(b)において，下線部①，②，③の表現はそれぞれ，以下のように標準化される．

（c） 標準形

```
① (電界効果半導体装置
    ((COMP :
      (*PARA*
         (絶縁部分  ・・・      )
         (絶縁ゲート部分 ・・・  )
         (半導体基板  ・・・     )))
     (OBJ : ( * ))))
②  (絶縁ゲート部分
     ((OBJ : ( * ))
      (COMP : (*PARA* (シリコン酸化物) (アルミナ)))
      (LOC : (半導体基板 ((OBJ : ( * )) (DET (DEM)))))))
③  (絶縁部分
     ((OBJ : ( * )) (COMP : (*PARA* (アルミナ)
        (シリコン酸化物)))))
```

標準化された内部表現から，表1(a)の仕様フォーマットを用いて，以下のような項目が抽出される．

（d） 抽出項目

（1） レベル1の抽出項目

```
***COMPOSITION***
 ((OBJ : (電界効果半導体装置)) (COMP : (*PARA*
    (絶縁部分) (絶縁ゲート部分) (半導体基板))))
```

（2） レベル2の抽出項目

```
***COMPOSITION***
 ((OBJ : (絶縁ゲート部分)) (COMP : (*PARA*
    (シリコン酸化物) (アルミナ))))
 ((OBJ : (絶縁部分)) (COMP : (*PARA* (アルミナ)
    (シリコン酸化物))))
***LOCATION***
((OBJ : (絶縁ゲート部分)) (LOC : (半導体基板)))
***PROCESS***
 ((GO : (半導体基板)) (PRED : (拡散)) (OBJ : (不純物))
    (MANN : (部分的)))
 ((GO : (絶縁部分)) (PRED : (形成)) (LOC : (半導体素子)))
```

7. むすび

本文では，おもに半導体分野の技術抄録文を例にとり，構文解析，標準化ならびに情報抽出の手法について述べた．この他にも，電子回路分野などの特許請求範囲文やサマリ文を対象として同様な仕様フォーマットを作成し，上記の手法について検討ならびに実験を行った．計算機実験から，本方法により比較的能率的に関係情報が抽出されることが確かめられた．とくに，特許請求範囲文における長い連体修飾の係り受け関係のあいまいさは，たんなる意味カテゴリのマッチングだけでは解消されない場合が多く，これらの解消には，多くの場合人間の介入を必要とする．本方法では，専門分野における一般常識的な知識を利用することにより，これらのあいまいさをかなりの程度（7割程度）減少させることができた．このようなあいまいさは，特許請求範囲文10件で約20～30カ所あるが，知識の利用により約8カ所に減少する．したがって，1件当り約0.8カ所程度に減少する．また，解析により得られた内部表現の格構造はほとんどの場合正しいものが得られ，標準化により，仕様フォーマットに指定した項目をほとんどすべて抽出できることが確認された．

参考文献

1) Rumelhart, D. E.: Notes on a Schema for Stories, in Bobrow, D. G. and Collins, A. (eds.), *Representation and Understanding*, pp. 211-236, Academic Press, New York (1975).
2) Hobbs, J. R.: Coherence and Interpretation in English Texts, Proc. 5th IJCAI, pp. 110-116 (1977).
3) 絹川，木村：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，Vol. 21, No. 3, pp. 200-207 (1980).
4) 田中，佐藤，元吉：自然言語処理のためのプログラミング・システム—拡張 LINGOL について—，信学論，Vol. J 60-D, No. 12, pp. 1061-1068 (1977).
5) 石綿：日本語動詞格支配の類型，総合研究(A)第2年次研究報告「知識工学の基礎とその応用に関する研究」，pp. 135-148 (1981).
6) 高松，藤田，西田：係り受け関係に基づく文献の検索，情報処理，Vol. 19, No. 12, pp. 1150-1157 (1978).
7) Nishida, F. and Takamatsu, S.: Structured-Information Extraction from Patent-Claim Sentences, *Inf. Process. Manage.*, Vol. 18, No. 1, pp. 1-13 (1982).
8) Nishida, F. and Takamatsu, S.: Japanese-English Translation through Internal Expressions, Proc. 9th COLING, pp. 271-276 (1982).
9) Nishida, F. and Takamatsu, S.: Semi-Automatic Indexing of Structured Information of Texts, 185th ACS National Meeting, Symposium on Natural Language Processing (March 1983).

(昭和58年4月27日受付)
(昭和58年9月13日採録)