見出し情報を用いたテキスト解析と情報抽出†

高松 忍†† 西田富士夫††

本論文は，技術関係などの説明的テキストを対象とし，見出し情報などを併用してテキスト解析を行い，主要な情報を抽出する手法を与えている．説明的テキストは，一般に，各記述ブロックごとに見出しを階層的にもち，各ブロックは，見出しに対応するフレームに従ってトップダウン的に記述される．このようなテキストの性質に着目し，標題や節の見出しからテキストのフレームを検索し，そのフレーム構造に関するトップダウン的な情報と文間の接続関係に関するボトムアップ的な情報を併用してグローバルなテキスト解析を行い，えられたテキスト解析結果から主要な情報を抽出したり，また，これからいろいろなレベルの要約文を生成したりする手法を与えている．

1. まえがき

最近，テキストの理解や要約やテキストからの情報抽出の研究が活発に行われている[1)~8),12)]．

われわれもまた，技術抄録などの比較的短いテキストから抽出項目を指定したフレームを用いて，構文解析してえた内部表現を標準化し，これから指定した内容の構造情報を抽出する手法の研究を行ってきた[9)~11)]．技術抄録のような比較的短いテキストでは，多くの場合，主要な項目だけがほぼ標準的なフレームに従って記述されている．このため，従来の方法のようにローカルなフレームとのマッチングによりフレームに合う主要な項目の値だけを抽出することができる．一方，比較的長い技術論文的なテキストでは，主要な項目と各種の詳細な記述が必要に応じて広範囲に分散して記述され，標準的なフレーム構造を設定することが難しい．したがって，このようなテキストに対し，従来の方法では，主要な情報を抽出することがかなり困難であり，グローバルなテキストの解析を必要とする．

この報告では，技術関係のテキストなどを対象とし，見出し情報などを併用してテキスト解析を行い，見出しに対応する項目の値を主要なものから順に能率的に抽出する手法を与えている．テキストの多くは標題や節などの見出しを階層的にもち，これに従ってトップダウン的に記述される．このようなテキストの性質に着目し，標題や見出しからフレームを検索し，その構造に関するトップダウン的な情報と文や文節間の関係に関するボトムアップ的な情報を併用してテキスト解析を行い，パラグラフや節などのブロックの主な内容を能率的に同定する一つの方法を述べている．そして，主要な情報を関係データベースへ抽出したり，また，要約文を生成する方法を提案し考察する．

2. 文間の接続関係

テキスト解析は文の主述語の意味的カテゴリや文間の接続情報のほかに，テキストの見出しに含まれる情報を用いて，ボトムアップ的かつトップダウン的に行う．文間の接続関係についてはこれまでいろいろの分類がなされているが[1)~3)]，ここではこれらを参考にして次のように分類する．

（1） 前書き的 (introductory)，補足的 (complementary)，言いかえ的 (rephrase)，例示的 (example)，付加的 (additional)

（2） 逐次的 (serial)，継起的 (temporal-succession)，因果的 (causal)，推論的 (reasoning)，並列的 (parallel)

（1）は文間に共通な主題タームなどを通して説明的なテキストを構成するのに必要な基本的な接続関係であり，この情報を用いてテキストにおけるトピックスやフォーカスや補足事項などに関して各文の役割を同定したり，情報の抽出に際し文に主要情報に関するレベルを与えるのに用いる．

（2）は，「…する前に…する」，「…した後に…する」，「…したので…する」，「…となるので…になる」，「…するとともに…する」のように，複文における従来の動詞の格構造に基づいたターム間の関係を文間の関係に拡張したものである．

ただし，上記の接続関係には包含関係があり，p なら q を $p \leqq q$ で表すとき，

† Text Analysis and Information Extraction Using Heading Information by SHINOBU TAKAMATSU and FUJIO NISHIDA (Department of Electrical Engineering, College of Engineering, University of Osaka Prefecture).

†† 大阪府立大学工学部電気工学科

言いかえ的≦補足的，例示的≦補足的，
因果的≦継起的，継起的≦逐次的，
推論的≦逐次的
が成立する．

2.1 前書き的関係

例えば式(1)のような構造をもち，前文が導入部で後文が本体部であるような関係である．

$$S_1: (PRED: p_1, K_{11}: t_1, K_{12}: t_2, K_{1r}: t_{1r})$$
$$S_2: (PRED: p_2, K_{21}: t_2, K_{2r}: t_{2r}) \quad (1)$$

ただし，p_1 は '述べる' などの Mental ACT 的な述語とする．$K: t$ は格名とタームの対を表し，二重下線のタームは主題を表す．

例1 S_1: この節では LFG について概説する．
S_2: LFG は文脈自由文法の一種であり，つぎの構造をもつ．

2.2 補足的関係

前文が本体部で後文が補足部の場合である．'ただし' や 'ここに' などの接続詞や式(1)のような文形などから同定する．なお，例示的関係や言いかえ的関係は補足的関係に属するものとする．

2.3 付加的関係

式(2)のように，後文が前文と同じ主題の異なる属性について記述している場合である．

$$S_1: (PRED: p_1, K_{11}: t_1, K_{1r}: t_{1r})$$
$$S_2: (PRED: p_2, K_{21}: t_1, K_{2r}: t_{2r}) \quad (2)$$

例2 S_1: P型半導体領域は拡散法によってN型半導体領域に形成される．
S_2: このP型半導体領域はキャリア注入に対しエミッタ領域として機能する．

2.4 逐次的関係

時間的または論理的な前後関係が文間に存在する関係で，継起関係，因果関係，前向き推論関係を含む．プロセス述語や関係述語などのような前後の文の述語のカテゴリや接続詞などにより同定する．

例3 S_1: 基板を酸化性雰囲気中で加熱して二酸化シリコン膜Aを形成した．
S_2: この膜の上にスパッタリング法により二酸化シリコン膜Bを積層した．

2.5 並列的関係

時間的または推論過程の記述において二つの文が同じ段階にあると考えられる関係で，場合わけによる記述や対比などの関係がこれに属する．

つぎにテキストの主要な部分を同定するため，文にレベルを設定する．S_1 を S_2 の前にあり，かつ，上述の文間の関係の一つをもつ最も近くにある文とする．このとき

(1) S_2 が S_1 に対し並列的，逐次的または付加的関係をもてば，一般に S_2 は S_1 と同じレベルにあるとする．ただし，S_1, S_2 のいずれか一方の主題タームが節の見出しの主題と一致するときこの主題タームを含む文のレベルを1レベルだけ高いものとする．

(2) S_2 が S_1 に対し補足的な関係をもてば，S_2 は S_1 より1レベルだけレベルが低いものとする．

(3) S_1 が S_2 に対し前書き的である場合，S_2 と S_1 とを同じレベルにあるとする．

なお，複文に含まれる副詞節や連体修飾節の従文は，その主文に必要な補足的情報であると考えられ，従文のレベルを主文のレベルと同じに設定する．一方，並列的関係や付加的関係でつながる重文の各文の間には上述と同様な文間の接続関係を設定し，また，上述の(1)のように，見出しの主題タームを含むか否かに従って重文の各文にレベルを設定する．

3. テキストの構造とフレーム

3.1 テキストの階層構造

テキストは，ある世界の物事をある観点から眺めたり，ある目的や意図をもって考察した書き手の一次元的な記述であり，その世界のいくつかの標準的なフレームに基づいて記述されているとみなすことができる．このような書き手の観点や意図を明らかにするために，テキストにはトップダウン的に標題や見出しやアブストラクトなどの短い記述の主題部を前置している．

図1はテキストの階層的構造を示す．図においてATTR$_i$ ($i=1,2,\cdots,n_1$) はテキストの主な属性名であり，ATTR$_{ij\cdots}$ ($j=1,2,\cdots,n_2$) はその子孫の属性名である．葉のノードの SENT$_{ij\cdots}$ というラベルはテキストのある文を表し，中間ノードの FRM$_{ij\cdots}$ はその子孫のノードにより構成されるあるパラグラフを表す．

このようにテキストの記述は主題の表す属性項目に関する記述を通じて詳細化され補充され，それらの属性項目や問題点の焦点事項として，いくつかの節やパラグラフや文が生成され，その世界に関連したより詳しい記述が付加される．

以上はテキストのトップダウン的な見方であるが，

ボトムアップ的に見れば，テキストは互いにコヒーレントな多くの文の集まりである．文の集まりはブロックごとに，主題を表す見出し部または導入部と述部を表す本体の部分とにまとめられる．ブロックが集まった全体も，また同様の構造をもっている．テキスト解析の主な仕事の一つは，その主題あるいは属性名に対応する述部あるいは属性値を同定することと考えられる．

技術報告などの説明的なテキスト(expository texts)の構造について考えてみよう．テキストの標題はテキストの主要な主題を示し，各節の見出しは標題に関するある属性項目を表し，それぞれの節に対してローカルな主題を表している．各節はこれらのローカルな主題のもとに，テキストの主題に対し付加的(additional)な属性項目の詳細を記述している．各節はいくつかのパラグラフからなり，各パラグラフは節の見出しに関するあるサブフレームについて記述している．

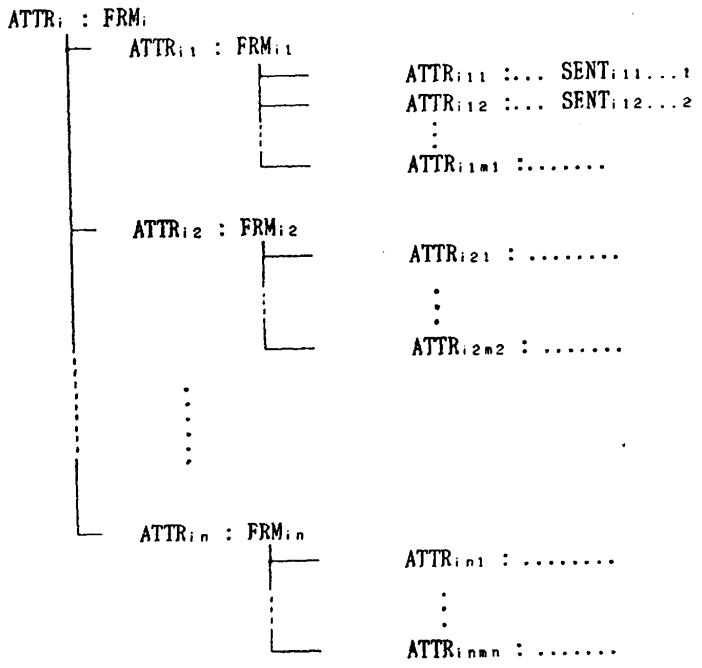

図1 テキスト構造
Fig. 1 A text structure.

3.2 フレームと例

説明的なテキストは，通常，まえがきやあとがきをもっているが，それらの節の構成はテキストの専門分野にあまり関係なく大同小異である．周知のようにまえがきの節は，そのテキストの標題の表す研究や技術開発などのこれまでの経過などこのテキスト記述の背景，目的，構成などのサブフレームからなる．また，あとがきの節は結果の要約や評価や検討すべき事項や今後の計画などのサブフレームからなる．テキスト本体の各節はその見出しが表す物事や属性項目に対応していくつかのフレームからなる．

各フレームは

$$frame\text{-}name(K_1-C_1 : t_1, \cdots, K_n-C_n : t_n) \quad (3)$$

なる見出し部といくつかの子のフレームからなる本体からなる．フレームの見出し部においてフレーム名に続き，引き数部の t_i $(i=1,\cdots,n)$ はフレームにおける構成要素を表し，K_i は t_i のフレームにおける格名(役割名)，C_i はこの格を埋めるべき構成要素 t_i のカテゴリ名を表す．フレーム内の主題を表す構成要素には二重下線を付ける．本体は見出し部に対し，属性値情報や焦点情報を与える部分であり，フレームを構成する子のフレームを直接，並置して表すか，構文規則の形で表す．また，子のフレーム間の接続関係を引用符‘ ’を用いて表す．

表1に自然言語処理などのソフトウェアシステムに関する技術テキストのフレームを示す．表1の(1)はソフトウェアシステムそのもののフレームを表し，トップレベルの焦点情報としてシステムの構成，機能や性質などの属性項目を，それらのフレームを用いて適宜，補足事項を伴いながら付加的，並列的に記述している．表1の(2)，(3)は，システムの FUNCTION や COMPOSITION のような属性項目のフレームを表し，同種のフレームを補足事項を伴いながら逐次的または並列的に繰り返し用いて，トップレベルの焦点情報を記述する記述形式を構文規則の形で与えている．また，システムが対象とするデータの構成や性質についても表1のフレームと同様なCOMPOSITIONや PROPERTY のフレームによって記述する．

以上のフレーム構造は前節で述べた図1のテキスト構造に対応する．すなわち，フレーム名と引数部がそれぞれ，図1の属性名 $ATTR_i$ とフレーム FRM_i $(i=1,2,\cdots,n_1)$ であり，その子フレームのフレーム名と引数部がそれぞれ下位の属性名 $ATTR_{ij}$ とフレーム FRM_{ij} $(j=1,2,\cdots,n_2)$ である．逆に，フレーム FRM_i，$FRM_{i\cdots k}$ などは $FRM_{i\cdots km}$ の祖先ノードの

フレームである．各フレームには主題が含まれ，これらの主題は図1のテキスト構造に対応する階層構造をもつ．

具象化し詳細化した各フレームを表層言語に展開すれば，フレームの大きさにより節やパラグラフや文間の関係でつながれたコヒーレントないくつかの文からなる一つのブロックになる．

逆に，テキストの各文は，その文の主題，述語のカテゴリや格構造を参照して対応するフレームを同定することにより，ある世界の具象化したフレームに還元することができる．一般的にフレームは上述のような物理的な世界のフレームと精神的なフレームに分類される．後者のフレームは書き手の考えやコメントや定義，仮定，推論などの記述に関するフレームである．いくつかの文を還元してえられる具象化したフレームは，パラグラフ間などのブロック間の関係により，先行するブロックの子フレームや兄弟フレームとして結合されて上位のフレームの子フレームとなり，テキストの内部表現を構成する．

後節で述べるように，テキストの構造を同定するため，表1のようなフレームと2節で述べた文間の関係などを用いてテキスト解析を行う．

表1 ソフトウェアシステムに関するテキストのフレーム
Table 1 Frames of texts on software systems.

```
(1)
DESCRIPTION (OBJect-SOFTWARE・SYSTEM: t)
    COMPOSITION (OBJ-SOFTWARE・SYSTEM: t,
                 COMPonent-SOFTWARE・SYSTEM: _)
    FUNCTION (PRED-MENTal・ACTion: _,
              AGent-SOFTWARE・SYSTEM: t, OBJ-DATA: _,
              GOal-DATA: _, INSTRument-DATA: _)
    PROPERTY (PRED-ATTRibute: _,
              OBJ-SOFTWARE・SYSTEM: t,
              DEGRee: _, COMPARison: _)
    IMPLEMENT (OBJ-SOFTWARE・SYSTEM: t, INSTR-
               PROGRAM・LANGUAGE: _, LOC-MACHINE: _)
    USE-FUNCTION (PRED-MENT・ACT: _,
                  AG-SOFTWARE・SYSTEM: _, OBJ: _,
                  GO: _, INSTR-SOFTWARE・SYSTEM: t)
(2)
FUNCTION (PRED-MENT・ACT: f, AG-SOFTWARE・SYSTEM: _
          OBJ-DATA: _, GO-DATA: _,
          INSTR-DATA: _, MEANS: FUNCTION*)
    COMPOSITION (OBJ-MENT・ACT: f,
                 COMP-MENT・ACT: {f1, f2, ···, fn})
    FUNCTION* ::=
        FUNCTION (PRED-MENT・ACT: fi, AG-SOFTWARE・
                  SYSTEM: _, OBJ: _, GO: _, INSTR: _)
      | FUNCTION* 'temporal-succession'
        FUNCTION (PRED-MENT・ACT: fj, AG-SOFTWARE・
                  SYSTEM: _, OBJ: _, GO: _, INSTR: _)
            ただし, fi, fj∈{f1, ···, fn}
(3)
COMPOSITION (OBJ-SOFTWARE・SYSTEM: t,
             COMP-SOFTWARE・SYSTEM: {t1, ··· tn})
    DESCRIPTION* ::=
        DESCRIPTION (OBJ-SOFTWARE・SYSTEM: ti)
      | DESCRIPTION* 'parallel'
        DESCRIPTION (OBJ-SOFTWARE・SYSTEM: tj)
            ただし, ti, tj∈{t1, ···, tn}
```

4. テキスト解析

4.1 前処理とフレームの検索

我々が試作したテキスト解析システム TAS (Text Analyzing System) はまずテキストの各文を解析しその内部表現を構成する．次に，アナフォラ解析により文に含まれる代名詞と省略語を同定する．従来のアナフォラ解析では，照応語や省略語の先行詞の候補として直前の文の主題ターム，直前の文の他の格のターム，前々文の主題ターム，…の順に順位を設け，これらの中から現在の文の格フレームの意味制約条件を満たすものを選定している[13),14)]．ところが，従来，省略語や代名詞の先行詞は直前の文の語かそうでなければ，省略語を含む文の祖先ノードの主題ターム，例えばパラグラフや節などの見出し語であることが多い．このため，本論文では，前節で述べたテキストの階層構造を用い，直前の文に現れるタームのほかに，直前の文の祖先ノードのフレームに含まれる主題タームの集合（現在の文，すなわち，省略語を含む文の祖先ノードの主題タームを含む）を参照し，これを先行詞の候補とする．したがって，従来の方法では祖先ノードのフレームの兄弟フレームに含まれる主題タームも候補として選ばれるが，本方法では兄弟フレームの主題タームは除外される．選ばれた候補の中から，後節で述べるテキスト解析で現在の文の直前の文に対する関係や省略語のカテゴリ条件などにより先行詞のタームを決定する．

続いて，文の内部表現を標準化する．標準化の主なものはタームに generic な動詞語や名詞語や格名を含むとき，これらを含まない表現に変換したり，文のフレームに対象と部分，対象と同格などの格を対として含むとき，これらをあらかじめ定めた標準的な表現に

変換したりする操作などである[9)~11)].

以上の標準化処理の後，TAS は3節で述べた表1のようなテキストのフレーム構造に関するトップダウン的な情報と，2節で述べた文間の接続関係などに関するボトムアップ的な情報を併用してテキスト解析を行う．

テキストのフレーム構造を検索するために，節の見出しやテキストの標題などを参照する．これらの見出しは名詞句やそれらの並列句の表現をとることが多い．t を中心名詞，$t_1, \cdots, t_m$ を連体修飾句に含まれる語とするとき，名詞句の内部表現は

$$t(K_0: *, K_1: t_1, \cdots, K_m: t_m) \qquad (4)$$

で表される．このようなテキストの見出しに対してこれをカバーする見出しをもつ次のようなフレームを順次，検索する．

(a) 中心名詞 t が 'もの' のカテゴリに属するとき，DESCRIPTION フレームの OBJ 格の要素のカテゴリ C_0 から $C(t) \subseteqq C_0$ をみたすフレーム：

DESCRIPTION (OBJ−C_0:＿) (5a)

を検索する．

ここで，$C(t)$ は語 t の直上のカテゴリを表し，包含関係 '$\subseteqq$' はカテゴリの上位・下位関係を意味する．

(b) 中心名詞 t が '事象' のカテゴリに属するとき，t から FUNCTION や COMPOSITION のフレーム名を求め，そのフレームの要素のカテゴリ $C_0, C_1, \cdots, C_m$ から

$C(t) \subseteqq C_0$, $C(t_1) \subseteqq C_1, \cdots, C(t_m) \subseteqq C_m$

をみたすフレーム：

frame-name(K_0−C_0:＿, K_1−C_1:_, $\cdots, K_m$−C_m:_) (5b)

を検索する．

なお，カテゴリ条件をみたすフレームが複数個ある場合には，最も下位のカテゴリの条件をもつフレームを検索する．

式(5)のフレームはテキストの主要部分を構成する第一レベルのフレームを与える．さらに，式(4)の $t_1, \cdots, t_m$ の各内部表現から，上と同様な手順により，主要部分に関連する次のレベルのフレームを検索する．このように検索したフレームを主要部から順に階層的に構成して見出しをカバーするテキストのフレームをつくる．

並列句の場合はそれぞれの名詞句に対応するフレームを検索すればよい．

例4

"日本語対話理解システムの構成"

なる見出しの内部表現は

構成 (PRED: *, OBJ: 日本語対話理解システム) ①

である．前述の(b)により，中心名詞 '構成' からフレーム名 COMPOSITION を求め，また，

C(日本語対話理解システム)

$\subseteqq$SOFTWARE・SYSTEM

の包含関係により，①から第一レベルのフレームとして

COMPOSITION (OBJ−SOFTWARE・SYSTEM: 日本語対話理解システム, COMP−SOFTWARE・SYSTEM: $\{t_1, \cdots, t_n\}$) ②

なる見出しをもつ表1(3)の COMPOSITION フレームを検索する．同様に，次のレベルのフレームとして

DESCRIPTION (OBJ−SOFTWARE・SYSTEM: t_i)

$t_i \in \{t_1, \cdots, t_n\}$ ③

なる見出しをもつ表1(1)の DESCRIPTION フレームを検索し，これらのフレームをレベル順に構成してこの見出しをもつテキストのフレームをつくる．

4.2 解析の概要

テキスト解析手順の流れ図の概要を図2に示す．テキストの各ブロックの解析結果を記録するため，スタックを設け，これに各ブロックのフレーム名，レベル数や入力文へのポインタなどを記録する．始めに，節の見出しやテキストの標題を参照してテキストのフレームを検索し，これらよりテキストのフレーム構造 F をつくる．また，テキストの文をさす入力ポインタ i を1にセットする．そして，次のようなテキスト解析手順に従い，図1のようなテキスト構造を出力する．

[テキスト解析手順]

ステップ 1: テキストの最初の文 S_1 のフレームをテキストのフレーム F を参照して同定する．同定できればステップ3へ，そうでなければステップ4へ．

ステップ 2: 現在の文 S_i とこれに先行する文との間の文間の関係を同定する．ここで，文 S_i に含まれる代名詞と省略語を同定するとともに文 S_i の内部表現を標準化する．その結果

(2.1)：文間の関係が付加的，逐次的または並列的であれば，テキストのフレームFを参照して先行するブロックの兄弟フレームとなる文 S_i のフレームを同定する．同定できればステップ3へ，そうでなければステップ (2.2) へ．

(2.2)：文間の関係が補足的であれば，先行するブロックの子フレームとなる文 S_i のフレームを同定する．同定できればステップ3へ，そうでなければステップ (2.3) へ．

(2.3)：文間の関係がない場合，テキストのフレームFのトップダウン的な情報と文 S_i の主題や述語のカテゴリなどのボトムアップ的な情報を併用して，文 S_i の親フレームでかつ先行するブロックの子フレームであるフレームを同定する．同定できればステップ3へ，そうでなければステップ4へ．

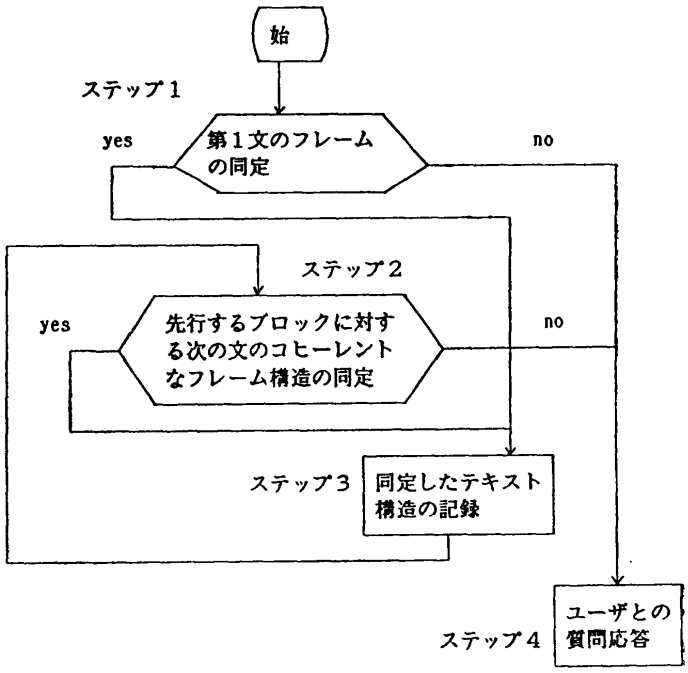

図 2 テキスト解析手順の流れ図
Fig. 2 The flow chart of the text analysis procedure.

ステップ 3：文S_iのレベル数，フレーム名，親フレーム名 (なければ空)，先行するブロックに対する文間の関係名 ($i=1$，すなわち，先頭文の場合は空) と文 S_i へのポインタをスタックに押し込む．後に続く文があれば，$i \leftarrow i+1$ としてステップ2へ，なければ，テキスト解析を終わる．

ステップ 4：停止してユーザの助けを求める．

解析手順のステップ2において，文 S_i が先行するブロックとある文間の関係をもつが，文 S_i に該当するフレームがなく，親フレームの兄弟フレームなどになっている場合には，文 S_i の主題，述語のカテゴリと文間の関係から文 S_i のフレーム名を与えるものとし，また，文 S_i に特記事項のマークを付ける．

また，文 S_i に該当するフレームがあるが，文の主題や文間の関係などから文 S_i のレベルが本来のフレームの標準的なレベルより高くなる場合には，文 S_i のレベルを上げてこれに特記事項のマークを付ける．

図 3，**図 4** に，自然言語処理に関する技術論文のテキストとその解析例を示す．

図3の例では，始めに節のヘディングから例4のように表1(3)の COMPOSITION フレームが検索される．まずステップ1により第①文が COMPOSITION のフレームにマッチするのでステップ3により図4①のようにテキストの部分構造を構成する．次に，第②文の第①文に対する接続関係が補足的 (complementary) 関係であるので，ステップ (2.2) により構成要素の表1(1)の DESCRIPTION フレームにシフトする．そして，その子フレームである FUNCTION フレームに第②文がマッチするので，ステップ3により図4②のようにテキストの部分構造を構成する．第③文の解析では，その文のフレーム構造，接続詞と前文までのテキスト構造から，ステップ2，ステップ3により，図4③の[ ]内に示すように省略語のタームを復元し，第③文を第②文に付加的 (additional) 関係で接続する FUNCTION フレームとして構成する．以下同様な手順を繰り返すと図4のテキスト構造がえられる．

図3のような技術テキストを対象として構文解析，標準化，テキスト解析，後節で述べる情報抽出と要約文生成の計算機実験を行った．システムは LISP 言語で記述し，計算機は ACOS-850 を使用した．構文解析から情報抽出までの平均処理時間は，インタプリタモードで約100語のテキストに対し約40秒であった．対象としたテキストは，半導体デバイスとハードウェア回路に関する特許抄録テキスト約 30 件，特許文献

の説明テキスト約 10 件と，ソフトウェアシステムなどに関する技術論文テキスト約 10 件である．比較的均一で標準的なフレーム構造をもつ特許関係のテキストについては約80-90%の部分が，技術論文的テキストについては約60-70%の部分が用意したフレーム構造に適合した．

技術論文的テキストにおいて，テキストの内容が見出しに対応する標準のフレームと一致しない子フレームや兄弟フレームを含むことが多い．例えば，“半導体装置の構成”を見出しとするテキストにおいて“その半導体装置の製造プロセスや特性”などの兄弟フレームが付加的に記述される場合などである．このような場合には，前述のように，文の主題，述語のカテゴリや文間の関係などのボトムアップ的な情報により標準フレームのどの位置のフレームかを同定し，これを特記事項として抽出する．

また，フレーム構造の同定には，2節で述べた文間の関係などの言語的知識や表1のような分野の一般的なフレームの知識のほかに専門用語シソーラスの知識を必要とする場合がある．例えば，2つの連続する文

S_1：“……翻訳システムについて述べる．”

S_2：“パーザは……”

のフレーム構造の同定には“翻訳システムの構成要素”についての知識を必要とする．

5. 情報抽出とテキスト生成

5.1 情報抽出

テキスト解析結果からの情報抽出の概略を述べる．テキスト解析によってえられたテキストの内部表現から，同定したフレームをガイドとして，レベル順にテキストの属する分野のフレーム情報を抽出し，関係形式のデータベースに蓄積する[9)~11)]．図3の例文のテキスト解析結果（図4）から構成した関係データベースを図5に示す．

標題： 対話領域の独立性を指向した日本語対話理解システム
2. システムの構成
①日本語対話理解システム MODUS は構文意味解析部と対話処理部から構成されている．②対話処理部は発話入力，発話生成，知識管理に関する制御を行なう．③また，入力文が与えられたときに，構文意味解析部を起動する．④構文意味解析部は日本語文を意味表現に変換する．
3. 構文意味解析部
⑤構文意味解析部は，拡張 CFG パーザ，単語辞書文法オブジェクト群から構成されている．
3.1 拡張 CFG パーザ
⑥パーザは，形態素解析部と構文解析部から構成されている．⑦これらは両方とも Lisp プログラムである．⑧形態素解析部は，まず文の先頭から辞書を引いて切り出し可能な単語を長い順に並べ，それぞれについて文の残りを切り出す．．．．．．⑨構文解析部はトップダウン予測付きのボトムアップアルゴリズムによって解析を行なう．．．．．．

図 3 入力テキスト
Fig. 3 An input text.

```
COMPOSITION(OBJ:日本語対話理解システムＭＯＤＵＳ,
            COMP:{構文意味解析部，対話処理部})                    ①
  DESCRIPTION(OBJ:対話処理部)
    FUNCTION(PRED:制御，AG:対話処理部,
             OBJ:{発話入力，発話生成，知識管理})                  ②
    FUNCTION(PRED:起動，[AG:対話処理部],
             OBJ:構文意味解析部,                                  ③
             COND:(PRED:与える，OBJ:入力文))
  DESCRIPTION(OBJ:構文意味解析部)
    FUNCTION(PRED:変換，AG:構文意味解析部,
             OBJ:日本語文，GO:意味表現)                           ④
DESCRIPTION(OBJ:構文意味解析部)
  COMPOSITION(OBJ:構文意味解析部,
              COMP:{拡張ＣＦＧパーザ，単語辞書文法
                    オブジェクト群})                              ⑤
    DESCRIPTION(OBJ:拡張ＣＦＧパーザ)
      COMPOSITION(OBJ:拡張ＣＦＧパーザ,
                  COMP:{形態素解析部，構文解析部})                ⑥
                  ..........
        DESCRIPTION(OBJ:形態素解析部)
          FUNCTION(PRED:引く，AG:形態素解析部,
                   OBJ:辞書)
                  ..........                                      ⑧
        DESCRIPTION(OBJ:構文解析部)
          FUNCTION(PRED:解析，AG:構文解析部,
                   MEANS:トップダウン予測付き
                         ボトムアップアルゴリズム)                ⑨
                  ..........
```

図 4 テキスト解析結果
Fig. 4 The result of the text analysis.

データベースの関係表は各サブフレームごとに構成する．すなわち，サブフレーム名を関係名とし，サブフレーム中の格名を属性名とする．また，タームや文のテキスト固有の重要性を保存するため，属性名としてテキストにおける文のレベル番号や文における主題タームなどを付け加える．

5.2 要約文の生成

4節のように構成されたテキスト解析結果からテキ

ストの要約文を生成する手順の概略を述べる．手順の流れ図を**図6**に示す．要約文生成の自然な方法は，指定された長さ以内でテキストの主要な部分を生成することである．本方法でも要約文の長さが指定されるものとし，その範囲内でテキストの中間表現から主要な情報をレベル順に取り出し，これより要約文を生成する．始めに，文のレベル番号 k を1にセットし，以下のような手順で生成する．

［要約文生成手順］

ステップ 1：テキスト解析結果から文や重文の各文について第 k レベルまでの中間表現を取り出す．取り出した中間表現から生成される文の長さが指定された長さの上限を超えないなら，$k \leftarrow k+1$ としてステップ1へ，

上限を超えるなら，第 $k+1$ レベル以下の中間表現を削除し，ステップ2へ．

ステップ 2：取り出した中間表現の第1レベルから順に，主題タームと文間の接続関係から単文，複文，重文の文型，文の主語と接続詞を選定し，ステップ3へ．ここで，埋め込み構造が二重以上にならないように，(a)のような文型選択規則を用いて文型を選定する．

ステップ 3：生成される文の長さが指定された長さを超える最下位レベルの部分については，これを超えないように(b)のような縮約化規則を用いて単文や複文の従属句の形に縮約し，ステップ4へ．ただし，指定された長さ以内に縮約できない場合は中間表現から最下位レベルの部分を削除する．

ステップ 4：中間表現を表層文に展開する．

(a) 文型選択規則

(1) 式(2)の付加的関係で接続する二つの文 S_1, S_2 において，主題タームを共有し，同じカテゴリの述語をもち，かつ S_1, S_2 がともに単文である場合には，文 S_2 の主題タームを省略し，S_1 と S_2 を一つの重文に整理する．

(2) 式(1)の補足的関係で接続する二つの文 S_1, S_2 において，S_1 と S_2 がともに単文である場合には，S_2 を連体修飾節の形で S_1 に埋め込み，S_1 と S_2 を一つの複文に整理する．

(b) 縮約化規則

(1) 補足的関係でつながる主要文 S_1 と補足文 S_2

COMPOSITION

OBJ	COMP
日本語対話理解システム	構文意味解析部，対話処理部
構文意味解析部	拡張ＣＦＧパーザ，単語辞書 文法オブジェクト群
拡張ＣＦＧパーザ	形態素解析部，構文解析部
……	……

FUNCTION

PRED	AG	OBJ	GO	MEANS
制御	対話処理部	発話入力，発話生成，知識管理		
起動	対話処理部	構文意味解析部		
変換	構文意味解析部	日本語文	意味表現	
引く	形態素解析部	辞書		
….	……	……	…..	
解析	構文解析部			トップダウン予測付きボトムアップアルゴリズム
…..	……	……	….	…….

図5 関係データベース
Fig. 5 A relational data base.

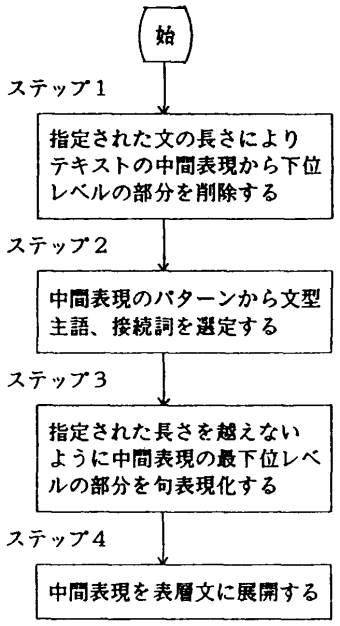

図6 要約文生成手順の流れ図
Fig. 6 The flow chart of the summary generation procedure.

において，S_2 からその主題タームを先行詞とする連体修飾節を作り，これを名詞句化して S_1 に埋め込む．

(2) 主要文 S と，その原因や手段を表し逐次的関係でつながる文 $S_1, \cdots, S_n$ において，各 S_i $(i=1, \cdots, n)$ を非必須格の部分を省いて動作名詞句化し，これらの並列句を主要文 S の原因格や手段格の副詞句として埋め込む．

(3) 因果的や継起的などの逐次的関係でつながる文 $S_1, S_2, \cdots, S_n$ において，中間のプロセス $S_2, \cdots, S_{n-1}$ を省略し，文 S_1 の OBJ 格や SOurce 格の部分を文 S_n に埋め込んだ形に縮約する．

図4のテキスト解析結果から，生成したレベルごとの要約の例を**図7**に示す．

レベル 1: 日本語対話理解システムは対話処理部と構文意味解析部からなる．
レベル 2: 日本語対話理解システムは対話処理部と構文意味解析部からなる．対話処理部は発話入力，発話生成，知識管理を制御し，構文意味解析部を起動する．構文意味解析部は拡張 CFG パーザ，単語辞書文法オブジェクト群からなり，日本語文を意味表現に変換する．

図7 生成テキスト
Fig. 7 Generated summaries.

6. むすび

4節で述べたように，計算機実験から本手法は特許文献や技術カタログのような均一で標準的な構造をもつテキストに対し特に有用であることが確かめられた．

技術論文的なテキストでは標準フレームと一致しない部分が多いが，本手法では，文の主題，述語のカテゴリや文間の関係などのボトムアップ的な情報を用いてこの部分のフレームを標準フレームと関連づけて同定し，これに特記事項のマークを付けて情報抽出などの処理を行っている．

テキスト解析によりえられた主要な情報は，5.1 節で述べたように，本来階層性をもたない均一な関係データベースの形に蓄積する．これにより，テキストの情報をいろいろな観点からフレキシブルに検索することができる．

テキスト解析結果からの要約文の生成などのテキストの生成については，手順の概略を簡単に述べた．テキスト構造に基づく文型の選定などについては，規則の精密化，体系化などさらに検討を進める必要がある．

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) Schank, R. C.: The Structure of Episodes in Memory, in Bobrow, D. G. and Collins, A. (eds.), *Representation and Understanding*, pp. 237-272, Academic Press, New York (1975).
3) Hobbs, J. R.: Coherence and Interpretation in English Texts, *Proc. 5th IJCAI*, pp. 110-116 (1977).
4) Gordon, L., Munro, A., Rigney, J. W. and Lutz, K. A.: Summaries and Recalls for Three Types of Texts, University of Southern California, Behavioral Technology Laboratories, Tech. Rep. No. 85 (1978).
5) McKeown, K. R.: Discourse Strategies for Generating Natural-Language Text, *Artif. Intell.*, Vol. 27, pp. 1-41 (1985).
6) Hahn, U. and Reimer, U.: TOPIC Essentials, *11th COLING*, pp. 497-503 (1986).
7) Tucker, A., Nirenburg, S. and Raskin, V.: Discourse and Cohesion in Expository Text, *11th COLING*, pp. 181-183 (1986).
8) Isahara, H. and Ishizaki, S.: Context Analysis System for Japanese Text, *11th COLING*, pp. 244-246 (1986).
9) Nishida, F. and Takamatsu, S.: Structured-Information Extraction from Patent-Claim Sentences, *Information Processing & Management*, Vol. 18, No. 1, pp. 1-13 (1982).
10) Nishida, F., Takamatsu, S. and Fujita, Y.: Semiautomatic Indexing of Structured Information of Text, *J. Chem. Inf. Comput. Sci.*, Vol. 24, No. 1, pp. 15-20 (1984).
11) 高松，日下，西田：技術抄録文からの関係情報の自動抽出，情報処理学会論文誌，Vol. 25, No. 2, pp. 216-224 (1984).
12) Nishida, F., Takamatsu, S., Tani, T. and Kusaka, H.: Text Analysis and Knowledge Extraction, *11th COLING*, pp. 241-243 (1986).
13) 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—文脈を考慮した処理，情報処理，Vol. 17, No. 1, pp. 19-28 (1976).
14) 吉本：談話処理における日本語ゼロ代名詞の扱いについて，情報処理学会自然言語処理研究会，NL 56-4 (1986).

(昭和63年1月11日受付)
(昭和63年6月24日採録)

高松　忍（正会員）

昭和23年生．昭和46年大阪府立大学工学部電気工学科卒業．昭和48年同大学院工学研究科修士課程修了．工学博士．現在，大阪府立大学工学部電気工学科講師．自然言語処理，知識情報処理の研究に従事．電子情報通信学会，人工知能学会，日本認知科学会各会員．

西田富士夫（正会員）

大正15年生．昭和25年京都大学工学部電気工学科卒業．現在，大阪府立大学工学部電気工学科教授．工学博士．自然言語処理，プログラムの仕様と詳細化，問題解決システムなどの研究に従事．著書「言語情報処理」など．電子情報通信学会，電気学会，計測制御学会，IEEE 各会員．