5J-7 要約支援システムCOGITO —「テキスト・パーサ」によるテキスト解析—

小松　英二　　安原　宏
(沖電気工業株式会社　総合システム研究所)

1．はじめに

要約支援システムCOGITOは、自然言語で記述された文書からテーブル形式の要約を生成することを目標としているシステムである。要約の際には、各文を文書全体の意図にそって解析することが必要であり、文脈解析が必要である。文脈解析としては、文書構造の解析がある。これには、従来、焦点による解析（Grosz[1]）や文と文の関係による解析（Hobbs[2]）等があったが、トップ・ダウンとボトム・アップの一方にかたよっており、また、注目している点も一面的である。本システムでは、文書構造の決定にあたって、これら2つの方法を融合した、より木目細かな解析を用いている。また、文書構造解析に際しては照応処理、文書からの情報抽出も同時に行なう。本システムでは、これらの処理部を文脈処理部とし、テキスト・パーサと呼ぶことにする。本稿では、テキスト・パーサの検討結果について述べる。

2．テキスト・パーサの概要

図1は、本システムの処理フローである。テキスト・パーサは、解析部で生成される、文の意味表現（北　他[3]）を入力とし、照応処理及び文書中の情報の抽出を行なう。図2はテキスト・パーサの出力例である。抽出情報は、a）対象情報と、b）文書構造情報の2つに区別されている。対象情報は、文書中の事物についての情報である。各文の内容は、焦点フレームと呼ばれる焦点を中心とした知識表現に変換した後、対象情報でノードとして用いる。また、文書構造情報は、文書の形式的構造や論理的展開を表わす情報である。各情報の詳細は、以下に述べる。

3．テキスト・パーサの処理方式

テキスト・パーサの処理方式をルーチン別に述べる。以下で、焦点の定義は、Sidner[4]に基づいており、文書の各時点でテーマとなっている事物や概念を表わす。

1）．焦点メカニズム・照応メカニズム

a）．焦点は2つ以上の文から決るものであり、まず、文型から焦点を予想し、期待焦点（Sidner[4]）を決定する。図3は、期待焦点決定のためのルールである。

本研究は新世代コンピュータプロジェクトの一環としてＩＣＯＴからの委託で行なわれたものである。

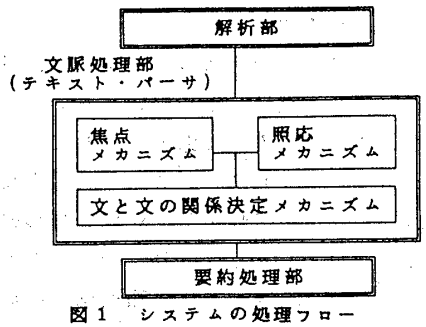

図1　システムの処理フロー

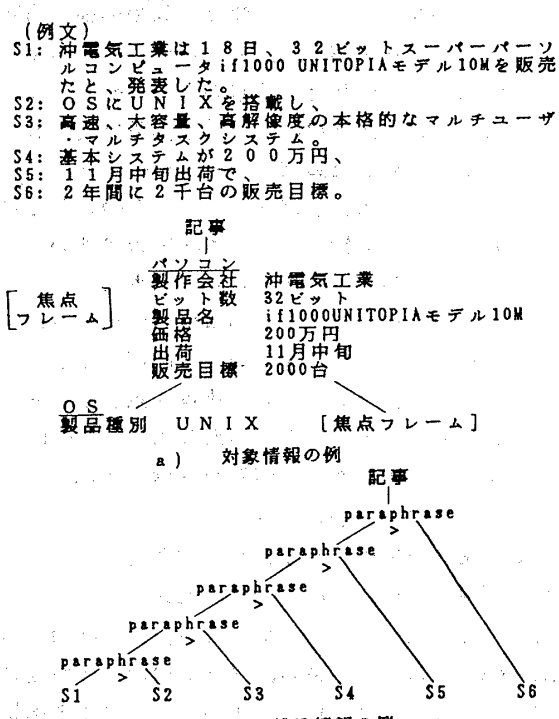

図2　テキスト・パーサの出力例

期待焦点決定ルール
1．「ダ」文では、主語と補語がis_a関係にあれば、主語が期待焦点。
2．1以外の「ダ」文では、補語が期待焦点。
3．特殊な動詞は、動詞が期待焦点。
4．3以外の他動詞文では、目的語が期待焦点。
5．3以外の自動詞文では、主語が期待焦点。
6．形容詞、形容動詞文では、主語が期待焦点。
補助ルール
1．上記のルールで決定した焦点が、属性の値の場合は、属性を推論し、期待焦点にする。
2．引用文は、埋め込み文の焦点を期待焦点にする。

図3　期待焦点決定のためのルール

Summarization Support System COGITO,-Text Analysis by Text Parser-

Eiji Komatsu,Hiroshi Yasuhara
OKI Electric Industry Co.,Ltd

b）．次に照応メカニズムによる照応処理、および、焦点の決定を行なう。照応は、期待焦点と意味素性を情報として、決定する。焦点は、原則的には期待焦点であるが、照応のパタンにより、それ以外の単語になることがある。本システムでは、Sidner[4] のアルゴリズムに基づき、期待焦点と後ろの文の照応の種類により焦点を決定する。

c）．文書が進むにつれて、焦点は移動する。本システムでは、焦点は、予め登録されている単語だけを用いる。図４は、焦点知識と呼ばれている互いに関連の深い項目のネットワークである。焦点は、焦点知識の項目のどれかに移動する。対象分野の構造知識は世界知識（安原[5] ）と呼ばれ、ループを許す木構造で表わされている。現在、焦点知識は、この知識を簡略化したものを用いている。焦点移動の候補が複数個あるときは、現在の焦点に近く、かつ、下位の項目に、優先的に移る。図５は、焦点知識のPrologによる表現例である。焦点知識は、総て二項関係で表されている。図６は、前述した焦点フレームのPrologによる表現例である。各文は、焦点フレームに変換されるが、前述したように焦点知識は、焦点移動の管理のためであるとともに、対象分野の世界知識にもなっている。このため、現在の焦点の項目に隣接した項目を用いて焦点フレームを生成し、焦点同士を結び付ける。図６は、前述した焦点フレームは、焦点知識をそのまま取り出した世界知識と、文書から取り出した具体的な数値や名称をマッチングした情報から構成されている。

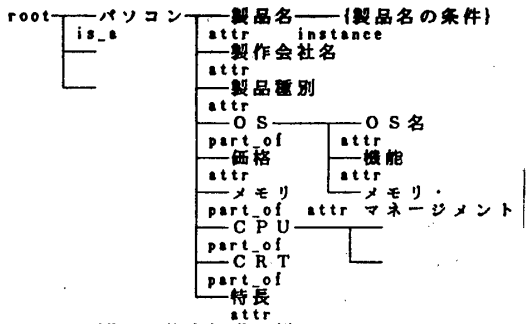

図４　焦点知識の例

```
is_a（ｒｏｏｔ，パソコン）．
attr（パソコン，製品名）．
instance（製品名，Ｘ）：－
        {（Ｘが製品名であるための条件）}．
```

図５　焦点知識のPrologによる表現例

```
is_a（ｒｏｏｔ，パソコン）．
attr（パソコン，製品名）．
        ・・・焦点知識から取り出された知識
instance（製品名，if1000UNITOPIAモデル10M）．
        ・・・文書から取り出した情報
```

図６　焦点フレームのPrologによる表現例

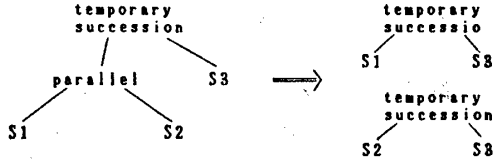

図７　文書構造情報からの情報抽出

３．文と文の関係決定メカニズム

対象情報を生成した後、文書構造情報を生成する。文と文の関係は、temprorary succession 、violated expectation、causal、paraphrase、example 、parallel、contrast、gap の８種類がある。文と文の関係の定義及び決定方法はHobbs[2]に基づいているが、焦点による解析結果を用ちいる点が、本システムの特徴である。木構造のリーフには、文番号が入る。また、文と文の関係には、各関係毎に、前後どちらの文が重要かの指標がついており、重要な方の文と新しい文の間で、関係を決定する。文書構造情報は、対象情報では表せない情報を抽出するためにもちいる。図７は、文章構造情報からの情報の抽出例である。特に、temporary succession、violated expectation、causalは、木構造の一部を、因果関係の情報として抽出する。

４．おわりに

テキスト・パーサは現在詳細検討中であり、部分的なインプリメントしか行なっていない。本システムでは、現在処理対象の文書を、情報処理関連の新聞記事に限定している。今後の課題としては、

１）．多様な文書への適用可能性

２）．パラグラフ等の文書構造の導入

３）．焦点知識の改良

を検討している。

参考文献：

[１] B.J.Grosz,Discourse Structure,
Stanford univ.Techinical Note 369(1985)

[２] J.R.Hobbs,Coherence and interpretation in
English texts,IJCAI85,pp110-116(1985)

[３] 北　他，要約支援システムCOGITO－言語解析部－，
本大会予稿

[４] C.L.Sidner,Focusing in the comprehension of
Definite anaphora,Artificial Intelligence,
The MIT press,pp267-330(1979)

[５] 安原，文書要約システムにおける世界知識の構造，
第３２回全国大会，pp110-116