自然言語処理 58 － 7
（1986. 11. 22）

要約支援システム COGITO

北 研二, 小松 英二, 安原 宏

沖電気工業（株） 総合システム研究所

要約支援システムCOGITOは、新聞記事を入力して、意味解析、要約処理をして、記事の要約として、表（テーブル）形式のリストを出力とするシステムである。

COGITOの特長の一つとして、コンピュータ上での文の意味表現形式を論理型表現の枠組で扱っており、知識表現、推論型との整合性を図りやすいということがあげられる。また、要約処理の際に世界知識を用いており、世界知識を用いるための基本的な手法として「焦点」および「文と文の関係」という2つを採用している。

本原では、COGITOでの自然言語の意味表現形式および要約処理方式について述べている。

Summarization Support System, COGITO

Kenji Kita, Eiji Komatsu, Hiroshi Yasuhara

Systems Laboratory, OKI Electric Industry Co., Ltd.
Shibaura 4-Chome, Minato-ku, Tokyo 108, Japan

COGITO is the text summarization support system which carries out the semantic analysis and the summarization processing of the input article and outputs the list as a summary of the article in the form of the table.

In COGITO, world knowledge is used for summarization processing, and as the basic method for utilizing world knowledge, "Focus" and "Relation of sentences" are adopted.

This paper describes the meaning expression formulae of natural language and the summarization processing system.

１．はじめに

現在、我々は世界知識およびそれに基づく推論等を論理型表現のパラダイムで統一的に実現する方式を研究中であり、これらの技術を用いた文書要約支援システムCOGITOを開発している。

既存の要約システムは、キーワードの頻度に基づいた統計的手法と、言語理論を用いた解析手法に大別できる。さらに、後者には、物語文法に基づくシステム(Rumelhart[1])、文書の因果連鎖を生成するシステム(Schank[2])、スクリプトを用いた方法(DeJong[3])、文と文の関係を木構造で表わす方法(Hobbs[4].etc)等がある。現段階では、キーワード方式がより実用的であるが、キーワードによる方式には限界があると考え、後者の解析的手法の検討を行なった。方式的には、「焦点」(Grosz[5])と「文と文の関係」(Hobbs[4])を基本的な方式として、さらに、大規模な世界知識を適用することを目標としており、キーワード以上の精度をもつシステムを目指している。本稿では、COGITOでの自然言語の意味表現形式、要約処理の概要等について述べる。

２．システム概要

COGITOは、日本語の意味解析を行なう解析モジュールと、その結果を用いて文のパラグラフを抽出し、要約テーブルを作成する要約モジュールから成る。

解析モジュールは、入力されたテキストを内部的な意味表現形式に変換する。意味表現形式は、基本的に格文法を基礎としており、格成分に相当するものをCOGITOのインプリメント言語であるPrologのユニット・クローズの集合に展開している。

要約モジュールは、解析モジュールで得られた意味表現形式を用いて、テキスト構造の解析を行なう。

また、要約処理の際に参照される知識ベースがあり、これは長期知識と短期知識とから構成される。長期知識は、世界知識、単語知識から成り、短期知識は、解析結果、要約結果等の動的知識であり、入力されたテキストが表現する知識の集合である。

システムへの入力は、当面、情報処理関連の新聞記事に限っており、出力としては要約テーブルと呼ぶ製品の属性と属性値から成るリスト形式のものが得られる。

システムは、現在のところ、C-Prologで記述され、VAX11/785上に開発しているが、将来的には、ICOTで開発された逐次型推論マシンPSIに移植する予定である。

３．意味表現形式

最近、知識をPrologのホーン節の集合で表現するという試みが小山、田中[6]で行なわれている。そこでは、“sem”述語と呼ばれる基本述語を用いて知識を表現しているが、以下で説明する“meaning”述語も同様の試みである。ただし、テンス、モーダリティ、接続関係等も“meaning”述語の枠組みの中で統一的に扱っているという点が異なっている。また、小山、田中[6]の場合は、概念辞書と知識表現を同一レベルで扱っているが、膨大な語彙を持つ必要のあるシステムでは、辞書と知識表現を切り離したほうがよいと考え、辞書記述は意味表現形式と完全に分離した。

本研究は第５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。

COGITOでは、文の意味は、Prologのユニット・クローズの集合で表現され、このために述語“meaning”が用意してある。述語“meaning”の一般形式は、

meaning(Concept#Id, Attr, Val, S_order).

である。ただし、“#”はオペレータとして宣言されている。また、述語“meaning”の各引数は、次のような意味を持つ。

Concept : 概念名, Id : 識別子,
Attr : 属性名, Val : 属性値,
S_order : 文インデックス.

文インデックスは、各文がテキスト中の何番目の文かとか、何番目の段落に表われたかというような文の出現順序を表わすのに用いる。

“meaning”述語の形式は、文内での概念と概念の間の関係に従い、幾つかのパターンに分けられる。

1). 格関係

格関係を表わす場合、第1引数に述語概念を、第2引数の属性名には深層格名、第3引数の属性値としてその深層格となる概念が入る。

[例]. meaning(Pred#1, actor, Actor#2, XXX).
meaning(Pred#1, obj, Obj#3, XXX).

2). 連体修飾関係

2つの概念'A'と'B'が連体修飾の関係にあるとき、第1引数と第3引数にはそれぞれ'A'、'B'が第2引数には連体修飾関係を示す名前が入る。

[例]. meaning('Chile'#1, loc, 'South America'#2, XXX).

3). テンス、モーダリティ

テンス、モーダリティも述語に対する属性として扱う。この場合、第2引数には'tense'あるいは'modality'が第3引数にその具体的な名前が入る。

[例]. meaning(XXX, tense, past, XXX).
meaning(XXX, modality, inference, XXX).

4). 接続関係

接続助詞などによって単文と単文が結びつけられる場合は、第1引数の概念名のところには、“causal_rel”（因果関係）等の接続の関係を示す名前が入り、属性名のところには各単文がその接続関係の中で果す役割を示す名前が入る。

[例]. 因果関係という関係で単文Aと単文Bが結びつけられ、Aが原因をBが結果を表わすとすると、

meaning(causal_rel#1, cause, A#XXX, XXX).
meaning(causal_rel#1, result, B#XXX, XXX).

のような意味表現が得られる。

このように、COGITOでの意味表現形式は、文の意味を細かなユニットに分解するところに特徴がある。こうすることにより、要約モジュールでの概念の取捨選択が容易になると期待される。また、意味表現をPrologのユニット・クローズに設定したことにより、Prologのユニフィケーション機構を用いて、必要な項目を取り出すことが楽にできる。

4．言語解析モジュール

COGITOは、文法記述言語にDCG形式を採用し、解析にはBUPを用いている。

DCGによる文法記述では、各文法カテゴリを述語と対応させ、その述語にカテゴリの持つ文法的な情報を引数として持たせることができる。COGITOでは、このことを利用し、各カテゴリにそのカテゴリを認識する過程で得られた中心的な概念およびその概念の修飾子リストを持たせている。そして、最終的なカテゴリである“sentence”を認識したときに、修飾子リストを分解し“meaning”述語に変換し、Prologデータベースにassertするようにしている。

図１に、入力文およびその意味表現を示す。

入力文：　沖電気工業は１８日３２ビットスーパーパーソナルコンピュータ
「ｉｆ１０００　ＵＮＩＴＯＰＩＡモデル１０Ｍ」を販売した、
と発表した。

意味表現：
```
meaning(発表#7, tense, past, []).
meaning(発表#7, agent, 沖電気#1, []).
meaning(発表#7, obj, 販売#5, []).
meaning(販売#5, tense, past, []).
meaning(販売#5, time, １８日, []).
meaning(販売#5, obj, ｉｆ１０００#3, []).
meaning(ｉｆ１０００#3, mod, [３２, ビット, スーパー,
        パーソナル, コンピュータ]#2, []).
```

図１．入力文とその意味表現

5．テキスト・パーサ

5-1. テキスト・パーサの概要

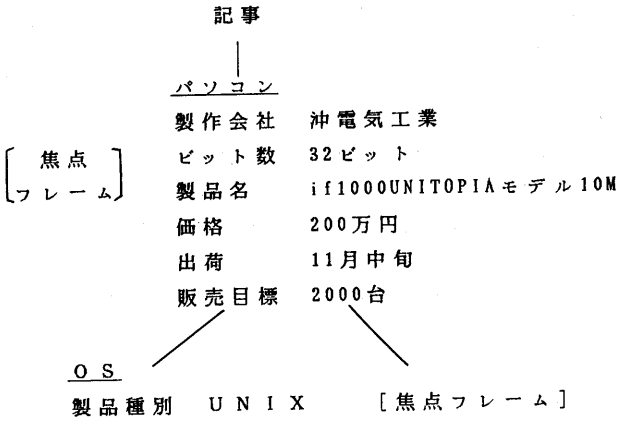
要約の際には、テキスト中の各文をテキスト全体の意図にそって解析することが必要であり、文脈解析が必要とされる。文脈解析としては、テキスト構造の解析がある。これには、従来、焦点による解析(Grosz[6])や文と文の関係による解析(Hobbs[4])等があったが、トップ・ダウンとボトム・アップの一方にかたよっており、また、注目している点も一面的である。本システムでは、テキスト構造の決定にあたって、これら２つの方法を融合した、よりきめ細かな解析を行なう。また、テキスト構造解析に際しては照応処理、テキストからの情報抽出も同時に行なう。COGITOでは、これらの文脈処理を行なう処理部をテキスト・パーサと呼んでいる。

テキスト・パーサは、解析モジュールで生成される文の意味表現である“meaning”述語を参照し、照応処理およびテキスト中の情報の抽出を行う。テキストから抽出される情報は、対象情報とテキスト構造情報の２つがある。対象情報は、テキスト中の事物についての情報であり、テキスト構造情報は、テキストの形式的構造や論理的展開を表す情報である。なお、COGITOでの焦点の定義はSidner[7]に基づいており、テキストの各時点でテーマとなっている事物や概念を表わす。

図２に対象情報とテキスト構造情報の例を示す。

入力テキスト：　S1：沖電気工業は１８日、３２ビットスーパーパーソナル
コンピュータif1000 UNITOPIAモデル10Mを販売したと、
発表した。
S2：ＯＳにＵＮＩＸを搭載し、
S3：高速、大容量、高解像度の本格的なマルチユーザー・
マルチタスクシステム。
S4：基本システムが２００万円、
S5：１１月中旬出荷で、
S6：２年間に２千台の販売目標。

対象情報：



テキスト構造情報：

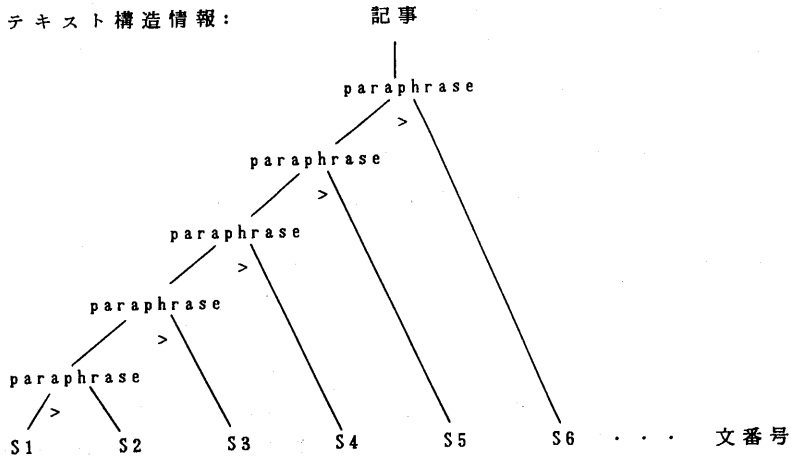


図２．対象情報およびテキスト構造情報

5-2. テキスト・パーサの処理方式

テキスト・パーサの中核を成す焦点メカニズム、照応メカニズム、文と文の関係決定メカニズムのそれぞれの処理方式を以下で説明する。

1). 焦点メカニズム・照応メカニズム

焦点は連続する2文（S1、S2とする）から決められる。まず、S1の焦点を予想し、期待焦点(Sidner[7])を決定する。この際には、期待焦点決定ルールを用いる。

次に照応メカニズムにより、S2の照応処理と、S1の焦点決定を行なう。照応は、それまでの焦点、S1の期待焦点、S2の格要素に対し、意味素性のチェックを行ないながら決定する。焦点は、原則的には期待焦点であるが、次の文の内容によっては、それ以外の単語になることがある。本システムでは、Sidner[7]のアルゴリズムに基づき、S1の期待焦点に対し、S2の照応の個数・指示物のパタンを考慮して、焦点を決定している。

文書が進むにつれて、焦点は移動する。焦点は焦点知識と呼ばれる互いに関連の深い項目のネットワーク上を遷移する。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。

2). 文と文の関係決定メカニズム

対象情報を生成した後、テキスト構造情報を生成する。文と文の関係の定義および決定方法はHobbs[4]に基づいているが、焦点による解析結果を用いる点が、COGITOの特徴である。また、文と文の関係には、各関係ごとに、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との間で、関係を決定する。図3に文と文の関係と重要性の指標の定義および決定方法を示す。木構造のリーフには、文番号が入る。テキスト構造情報は、対象情報では表わせない論理的な展開の情報を抽出するために用いる。

1．論理的な関係

1).	cause(CA)	S1とS2に因果関係がある。[<]
2).	temporary succession (TS)	S1の結果がS2の初期状態になっている。[=]
3).	violated expectation (VE)	S2.がS1から自然に推論される事柄に矛盾する。[>]

決定方法：　接続詞があるときは、接続詞により決定する。接続詞がないときは、それぞれの文の動詞の組み合わせにより決定する。

2．文型による関係

4).	parallel(PL)	S1とS2が並列関係にある。[=]
5).	contrast(CO)	S1とS2が対比される。[=]

決定方法：　2つの文の文型および対応する格要素の意味素性の比較により決定する。文型およびび対応する格要素の意味素性が一致するときは、parallel関係とし、ただ一つの意味素性が正反対になっているときは、contrast関係とする

3．焦点による関係

6). paraphrase(PP)	S2がS1の詳細化になっている。[>]
7). example(EX)	S2がS1の例になっている。[>]
8). implicit focus(IF)	S2がS1の暗示的な関連事項について述べている。[-]

決定方法：　2つの文の焦点同士の関係が、is_aならばexample、part_ofまたはattrならばparaphrase、implicitならばimplicit focusとする。

図3．文と文の関係と重要性の指標および決定方法

6．要約処理

まず、テキスト・パーサにより得られた対象情報とテキスト構造情報を評価する。焦点知識の各ノードは予め重みづけされており、指定した要約のレベルによって、対象情報の取捨選択を行なう。テキスト構造情報に以下の評価を行なう。

・パラグラフの最初の文ならば重みを＋1する。
・木構造の各関係ごとに、相対的な重みの大きい部分木を支配している文の重みを＋1する。

次に、要約のレベルに従い、そのレベルに近くなるように、各文の重みの大小により文を選び出し表示する。この際、文と文の接続が不自然にならないように一部の単語を省く。また、対象情報を要約テーブルと呼ぶ表形式で表示する。要約テーブルの例を図4に示す。

販売者			沖電気工業		
開発者			販売者と同じ		
販売物			32ビットスーパーパソコン		
ハード	本体	CPU	プロセッサ	M68010	
			クロック	10MHz	
		メインメモリ(RAM)	容量	標準	1Mバイト
				最大	8Mバイト

図4．要約テーブル

6．おわりに

本稿では、解析モジュールの意味表現形式および要約モジュールの「焦点」と「文と文の関係」の決定方式・抽出情報の内部フォーマット・要約の出力フォーマットについて述べた。現在、システムのプロトタイプを作成し、新聞記事の解析実験を行なっており、基本的な動作を確認している。

今後の課題として、多様な文書への適用可能性および世界知識の大規模化による精度の向上等がある。また、現在は処理対象の文書を情報処理関係の新聞記事に限定しているが、焦点知識等の改良を行ない、多様な文書への適用可能性について検討を進める予定である。

謝辞

本研究に当り、当研究所オフィスシステム研究部の椎野努部長並びにＩＣＯＴ第２研究室の横井俊夫室長から御指導を受けたことを感謝します。

参考文献

[1]. Rumelhart, D : Notes on a Schema for Stories, in Bobrow & Collins, eds., Representation and Understanding, Academic Press, 1975.

[2]. Schank, R. : Understanding Paragraphs, Tech. Report 6, Instituto per gli studi semanticie cognitive, Castagnola, Switzerland, 1974.

[3]. DeJong, G. : Prediction and Substatiation: A New Approach to Natural Language Processing, in Cognitive Science 3.

[4]. Hobbs, J.R. : Coherence and Interpretation in English Texts, in IJCAI85.

[5]. Grosz, B.J. : Discourse Structure, Stanford Univ. Tech. Note 369.

[6]. 小山、田中 : Definite Clause Knowledge Representation, in LPC '85.

[7]. Sidner, C.L. : Focusing in the Comprehension of Definite Anaphora, Artificial Intelligence, The MIT Press.