# 日本語テキストからの情報抽出システムにおける照応・省略解析

土井 伸一　　安藤 真一
NEC 情報メディア研究所

## 1 はじめに

現在我々の周囲には、新聞記事や技術論文を始めとする様々なテキストが氾濫している。しかし読者が本当に必要としているのは、それらのテキストから得られる情報のごく一部に過ぎない。最近では、事前に登録した特定のキーワードに関連する新聞記事だけを抽出して利用者に提供する「クリッピングサービス」も開始されているが、記事を読んで選択し、要約する作業のほとんどを人手で行っているのが現状である。従って、大量のテキストから有用な情報を効率的に検索・抽出するための支援技術が求められている。

我々はこのようなテキスト利用技術として、日本語の新聞・雑誌記事から指定の情報だけを抽出し、要約文を生成する情報抽出システム **VENIEX** (VENus for Information EXtraction) を試作した [1, 2]。本システムは、ユーザが指定した必要情報の項目リストと関連キーワードに基づいて、一定形式の情報を抽出する。

一般に新聞記事等では、抽出すべき情報はテキスト中に断片的に存在しており、テキスト全体の情報を抽出するには文ごとの情報を合成する必要がある [3]。この情報合成のキーとなるのが、文中の他の要素に言及する表現形式である照応表現である [4]。ここでは、代名詞や指示表現に加え、省略表現、同一名詞による指示等も含めて考える。照応表現に対する先行詞を決定することで、複数の文に出現している情報を関連付けることができる。

テキスト中の照応表現に関してはこれまでにも様々な研究が行われているが、照応の解析や談話構造の同定には多くの要因が関係し、現状の技術では十分な処理は難しい。しかし我々のシステムでは、キーワード間の依存関係の認定が主目的であり、キーワードに関係する照応表現のみを解析対象とする。これにより先行詞の探索範囲や省略の存在認定に関する知識を絞り込むことができ、実用的な文脈処理機構を実現できた。

本システムでは、始めに構文解析部が、入力テキストの各文の構文構造を解析しながらキーワード間の関係を認定し、各文から情報を抽出する。続いて文脈処理部が、キーワードに関する照応・省略表現、同一名称による指示等の解析規則に基づいて文ごとの情報を合成し、テキスト全体に対する情報を抽出する。本稿では、始めに情報抽出システムの概要を述べ、続いて、この文脈処理に基づく情報計算機構の詳細と評価を述べる。

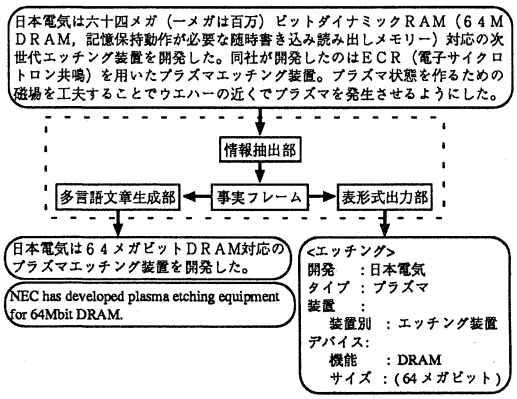


図 1: **VENIEX** 構成・入出力図

## 2 情報抽出システム VENIEX

試作した情報抽出システム **VENIEX** では、始めにユーザが、自分が読みたい情報の項目リストを指定するフレームと、同フレーム中の情報項目の内容を指定するためのキーワード集を準備する。システムはこれらを利用して電子テキストから必要な情報だけを自動抽出し、定められたフレームに当てはめた形式のデータと、日本語と英語の要約文で提示する。これまでに、半導体製造工程関連技術情報 =ME （マイクロエレクトロニクス）機能情報を抽出するシステムを実現している。以下、システム構成と情報抽出法、文章生成法について説明する。

### 2.1 システム構成

図 1に、試作した情報抽出システムの構成と入出力例を示す。本システムは、新聞記事に記述された 5W1H1D （誰が (who) 何を (what) いつ (when) どこで (where) なぜ (why) どのように (how) どうした (do_what)) からなる事実情報を抽出対象としており、これをフレーム構造を用いて表現している (以下、事実フレームと呼ぶ)。

我々が試作しているシステムは、この事実フレームとして、「ある企業が、半導体製造工程のある技術・装置をどうしたのか」という情報をテキストから抽出する。ここでポイントとなるのは、“企業”、“技術・装置”、及び両者の関係を示す用言 (開発、製造、販売、購入 etc.) からなる 3 項関係である。これに、企業の所在地、装置

が対応しているデバイス等の情報を加える。

本システムは、利用者が指定した事実フレームに当てはまる情報を入力テキストから抽出する情報抽出部と、抽出情報を表現する自然言語文章を生成する文章生成部からなる。情報抽出部は、抽出すべき情報に関連する語彙知識と、構文構造や照応表現といった言語的知識を用いて事実フレームを生成する [2]。また文章生成部は、事実フレームを中間言語に変換した後、複数言語の文章を生成する。この中間言語からの文生成には、機械翻訳システム PIVOT の文生成機構 [5] を利用している。

## 2.2 情報抽出

一般に新聞記事などのテキストでは、利用者が求める情報は一文内および記事内に分散して存在する。このため情報抽出では、記事内に存在する断片的な情報を収集し、正確に組み合わせる必要がある。本システムでは、

・抽出すべき情報に関連する特定の語彙（以下、キーワードと呼ぶ）の辞書を用意する
・各キーワードの意味構造として抽出すべきフレームの部分構造を与えておく
・一般的な構文・意味解析、文脈処理技術を用いて、キーワード間の関係を認定する
・認定した関係に従って、意味構造を合成する

手法で、記事全体の情報を抽出している。

### 2.2.1 キーワードの認定

本システムでは ME 機能情報用にキーワード辞書を用意し、形態素解析と同時にキーワードの認定を行っている。キーワードに対して与えた、ME 機能情報フレームの部分構造の例を図 2に示す。キーワード辞書には現在、企業名、半導体製造技術関連語、地名など、約 33,000 の語彙を登録している。

### 2.2.2 構文構造を利用した文内情報の抽出

ここでは通常の構文解析を行ってキーワード間の関係を認定し、係り受け関係の認定された 2 つのキーワードについて意味構造フレームを合成する。図 3に例を示す。

これによってキーワードの共起関係を利用するだけでは正確な抽出が困難な場合にも対処している。例えば、「A 社はスパッタリング用の材料を開発した」という文を考える。このとき開発されたのは材料であるにもかかわらず、単純な共起関係に基づく手法では「スパッタリングの開発」という情報を抽出してしまう可能性がある。しかし「スパッタリング」と「開発」の間には直接の係り受け関係はないので、我々の手法ではこの組合せを抽出対象から外すことができる。

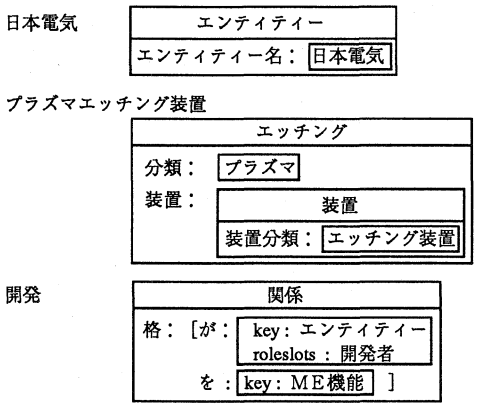


図 2: キーワードに与えた意味構造の例

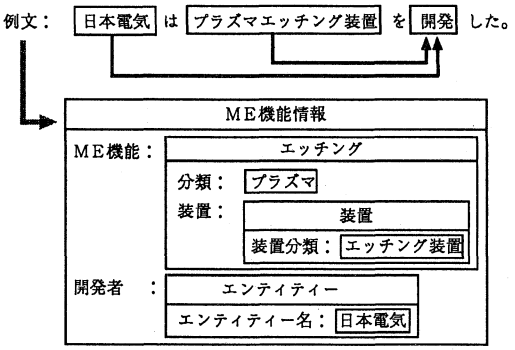


図 3: 構文構造を利用した情報抽出の例

### 2.2.3 企業名の推定

固有名詞である企業名を、あらかじめすべてキーワード辞書に登録しておくことは不可能である。そこで本システムでは、企業名がキーワードとして認定できなかった場合にも情報抽出を行うために、未知語の企業名としての機能推定を行っている。

例えば新聞記事では、企業名を表す語彙に対して「社」などの接辞を伴ったり、「日電アネルバ (府中市)」のように括弧で地名が続くパターンが頻繁に用いられる。そこでこれらの局所的なパターンを利用し推定を行う。

さらに構文構造を利用した企業名の推定も行っている。例えば「XXX がエッチング装置を開発した」という文で XXX が未知語の場合を考える。ここで図 2に示した「開発」の意味構造は、「が格」に企業 (エンティティー) が入ることを表している。そこで本システムは、この文で XXX を企業名であると推定する。

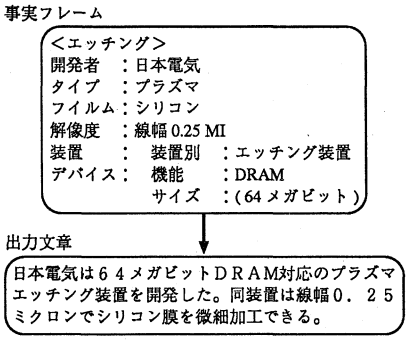


図 4: 文章生成の例

## 2.3 文章生成

文章生成部では、まず事実フレームを生成文一文に対応する単位に分割する。続いてフレーム内の関係構造を深層格関係へ変換して、PIVOT 中間言語に写像し、複数言語の文章を生成する。一般に、記事から抽出した事実フレームの構造が複雑で多数のスロットが存在する場合、そのすべてをまとめて１文で表現すると、構造の複雑な読みにくい文になってしまう。本システムではこれを避けるために、情報を適宜、何文かに分割して出力する。図 4に生成する文章の例 (日本語のみ) を示す。

# 3 情報抽出システムにおける文脈処理

我々は新聞記事から抽出すべき半導体製造技術情報として、企業名と、技術・装置名、及び両者の関係を表す用言 (開発、販売、購入 etc.) からなる 3 項関係を中心にしたフレームを定義した。従って、テキスト中に出現する照応表現の中で、企業名と技術・装置名に関するものだけを解析対象とすればよい。以下、照応表現に対する先行詞の認定、及び認定後の情報計算法を説明する。

なお、本システムは文法を換えるだけで文内の解析と文間の解析に同一の枠組で対処している。まず文節間文法によって一文ごとに解析を行って各文から情報を抽出する。続いて、キーワードに関する照応表現の解析規則を記述した文間文法に基づいて文ごとの情報を合成し、テキスト全体に対する情報を抽出する。

## 3.1 指示表現による照応 (同一指示)

新聞記事においては、指示表現として「同社」「同装置」等の接頭語や「この機種」等の連体詞を伴った照応表現が多く出現する (代名詞はあまり出現しない)。指示表現に関する文脈処理は、出現した企業名および技術・装置名を先行詞候補として保持しておき、指示表現が出現した際に候補中から適切なものを先行詞として選択することで実現する。今回のシステムでは新聞記事の分析に基づき、直前に出現した企業、技術・装置を先行詞とした。ただし企業に関しては、文章の情報構造を反映させるため前文中の「が格」に対応するものを優先する。

・ＣＩＴアルカテル社は、キヤノン販売と合弁で半導体製造装置販売会社「アルカンテック」を設立すると発表した。同社は従来エッチング装置を販売してきた。

ここでの「同社」は、「アルカンテック」「キヤノン販売」ではなく、前文の「が格」である「CIT アルカテル社」を指していると解析する。

・アプライド・マテリアルズ・ジャパン（ＡＭＪ）は二十日、イオンエッチング装置を開発した。この装置は四メガビットダイナミックＲＡＭまで処理できる。

ここでは、「この装置」が直前のエッチング装置を指していることを認定することで、対応するデバイスが 4M ビットの DRAM であるという情報を前文で示された 3 項関係に付加し、情報を合成することができる。

## 3.2 指示表現による照応 (非同一指示)

・日本企業では最大手のニコンが１６Ｍ量産対応のステッパーを発売した。キヤノンも同様の機種を発売しており、現在半導体メーカーが性能評価をしている。

キヤノンが発売したのも「１６Ｍ量産対応のステッパー」であることを認定するには、「この装置」の場合と同様に、「同様の機種」という照応表現が直前の装置であるステッパーを指していることを解析すればよい。しかし、「ニコンが発売したステッパー」と「キヤノンが発売したステッパー」は別のものである。従って属性だけをコピーし、別の装置として扱うことで対処する。

## 3.3 構文構造による照応

・住友金属工業は次世代ＬＳＩ用エッチング装置の商品化に取り組んでいる。早期の受注に向けて本格的な営業活動に入る方針だ。
住金が商品化するのはプラズマエッチング装置。

上記のような「～する (した) のは～」という形の強調構文では、その前半部に記述される内容はテキスト中で既知のものでなければならない。従ってこの場合も、指示表現等の場合と同様に、前方に存在する対応要素を認定して情報計算を行う。これによりこの例では、全体として「住友金属工業がープラズマエッチング装置をー商品化する」という関係を抽出できる。

### 3.4 省略の認定と補完

省略の場合には、指示表現の場合と異なって明示的な指標が存在しないので、まず省略がおきていることの認定が重要である。我々のシステムでは、関係を表すキーワードの辞書に、どの格要素にどのキーワード(企業、技術・装置)が来るかを記述している(図2参照)。従って、この格スロットが埋まったかどうかを見ることで、容易にキーワードの省略を認定できる。認定後の省略要素の補完は、照応表現の先行詞決定と同一の手順で行う。

・ラム・リサーチは、ドライエッチング装置で東南アジア地域の四五%のシェアを持つ。日本では住友金属工業と提携してエッチング装置を製造・販売している。

この例では、関係を表すキーワード「製造・販売する」の「が格」が省略されているので、直前の文の「が格」であるラム・リサーチによって補完する。

### 3.5 名称の同一性の判断

「同」や「この」等の指示表現がが出現していなくても、企業名の場合には名称の一致で照応関係を認定することができる。しかし略称や語尾の社の有無等の異表記が存在するので、名称の同一性の判断は簡単ではない。

我々のシステムでは、

1) 企業名辞書に与えた、別称・通称・略称辞書
   ex.) 日本電気←→日電←→NEC
   ex.) 住友金属工業←→住友金属←→住金
2) テキスト中で明示された別称も取り込む
   ex.) アプライド・マテリアルズ・ジャパン(AMJ)
3) 語尾の“社”を除いた語頭からの部分一致
   ex.) ＣＩＴアルカテル←→ＣＩＴ社
   ex.) ラムリサーチ←→ラ社

によって同一性を判断する。3) は特に、未知語を企業名として推定した場合に有効である。

### 3.6 文脈処理に基づく情報合成の例

図5に、指示表現と構文構造による照応表現の処理によってME機能情報を統合する例を示す。まず第2文の「同社」が第1文の「日本電気」に対する照応であると認定する。また第2文が強調構文であることから、「プラズマエッチング装置」と第1文の「エッチング装置」が対応することを認定し、両者の持つME機能フレームをマージする。これによって、「日本電気が64メガビットDRAM対応のプラズマエッチング装置を開発」という情報を得ることができる。

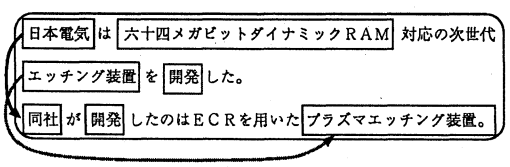


図 5: 文脈処理の例

## 4 評価

上述した文脈処理機能を組み込むことにより、新聞記事中の各文に分散して存在している情報を抽出・合成し、フレーム構造と要約文として出力するシステム **VENIEX** を実現した。作成した文脈処理機構は、分析した100記事について、約7割の記事に対処可能であった。

また、文脈処理機構を含む情報抽出部全体について、人手で作ったフレームとシステムの出力を比較して評価を行った。新聞記事100記事に対するブラインドテストでは情報適合率と情報再現率の平均で約50%という結果が得られている。同じ評価値を人間の作業に対して適用すると60-80%になるという報告があり[6]、キーワード辞書と解析規則の充実により人間の水準に近付けることができると考えている。

今後は文脈処理や未知語機能推定などの機能を、各機能ごとに詳細に評価し改良していく予定である。例えば半導体製造技術情報の抽出に関しては、企業の先行詞として、直前ではなく段落頭の「が格」を優先すべき場合も考えられる。これを含め、特に先行詞の認定や同一性の判断に関する条件を精緻化していく。

## 参考文献


[1] Muraki,K., Doi,S. and Ando,S. “NEC:Description of the VENIEX System as used for MUC-5”, *Proc. of MUC-5*, pp.147-160, 1993

[2] 安藤, 土井“新聞記事からの情報抽出システム –指定情報の抽出と多言語文章による提示–”, 人工知能学会第8回全国大会, 24-3, 1994

[3] Hobbs,J. “The Generic Information Extraction System”, *Proc. of MUC-5*, pp.87-92, 1993

[4] Wakao,T. “Reference Resolution Using Semantic Patterns in Japanese Newspaper Articles”, *Proc. of Coling 94*, pp.1133-1135, 1994

[5] Okumura,A., Muraki,K., Akamine,S. “Multilingual Sentence Generation from the PIVOT interlingua” , *Proc. of MT SUMMIT III*, pp.67-71, 1991

[6] Sundheim,B.M. “Overview of the Fourth Message Understanding Evaluation and Conference”,*Proc. of MUC-4*, pp.3-21, 1992