キーワードと構文構造に基づく情報抽出システムにおける文脈処理

2M−2

土井 伸一　安藤 真一　村木 一至

NEC 情報メディア研究所

1. はじめに

近年、新聞記事等の電子化テキストが大量かつ容易に入手できるようになってきており、情報検索・テキスト圧縮等、自然言語処理技術を活用した電子化テキスト利用技術が求められている。

筆者らはこのようなテキスト利用技術として、予め定義された一定の形式の情報をテキストから抽出するシステム **VENIEX** を作成した[1]。新聞記事等のテキストでは、抽出すべき情報は一般にテキスト中の各文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に抽出した情報を合成する必要がある。本稿ではこの情報合成に用いる文脈処理の手法について報告する。

2. テキストからの情報抽出と文脈処理

我々のシステム **VENIEX** は、抽出すべき情報を表現する語彙知識を格納したキーワード辞書を持つ。システムはこの辞書を参照して入力テキストの各文の構文構造を解析することでキーワード間の依存関係を同定し、情報を抽出する。しかしながら各キーワードはテキスト中に分散して存在するため、一文から得られる情報は断片的なものに過ぎず、テキスト全体から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。

ここで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。ここでは、代名詞や指示表現だけではなく、省略表現(ゼロ照応)、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(あるいは省略に対する補完要素)を決定することにより、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。

テキストに出現する照応表現に関してはこれまでに、新聞に出現する照応表現の分析[2]や、文章の談話構造との関連付け[3]、同一名詞による指示の解析[4, 5]等、様々な研究が行われている。しかし、照応の解析や談話構造の同定には多くの要因が関係するため、現状の技術ではすべての照応表現の十分な解析はできない。特に、日本語には名詞句の定/不定を表現する文法形式が存在しないため、同一の名詞による指示が個体としても同一のものを指しているか否かの判断は困難である。

しかし我々のシステムは、キーワードの存在とキーワード間の依存関係の同定が目的なので、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、文脈処理機構を実現できた。以下、新聞記事からの半導体製造技術情報の抽出を例にとり、具体例に基づいて **VENIEX** による解析法を示す。

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得られる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、対応するデバイス、企業の所在地等の情報を加えたフレームを定義した。従って企業名と技術・装置名に関する照応表現だけが解析対象となる。照応表現に対する先行詞の同定、及び同定後の情報計算は、新聞記事約100記事を分析して、以下の手順で行うこととした。

3.1. 指示表現による照応(同一指示)

新聞記事では、「同社」「同装置」等の接頭語や「この機種」等の連体詞を伴った指示表現が多く出現する。指示表現に関する文脈処理は、出現した企業名と技術・装置名を先行詞候補として保持し、指示表現が出現した際に適切なものを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。

Context Analysis in Information Extraction System based on Keywords and Text Structure
Shinichi DOI, Shinichi Ando, Kazunori MURAKI
NEC Information Technology Research Laboratories

・アプライド・マテリアルズ・ジャパン（ＡＭＪ）は二十三日、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すことを同定することで、対応するデバイスが4M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。

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。

ここで、「キヤノンが−１６Ｍ量産対応のステッパーを−発売した」という3項関係を同定するには、「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。しかしこれは「ニコンが−ステッパーを−発売した」という3項関係とは別のものである。従って属性だけをコピーして、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、その主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対応する要素を同定して情報計算を行う。これにより上記の例では、全体として「住友金属工業が−プラズマエッチング装置を−商品化する」という3項関係を抽出できる。

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必要である。我々のシステムでは、関係を表す用言の辞書に、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ので、この格スロットが埋まったかどうかを見ることで、容易に省略の存在を認定できる。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販売している。

この例では、関係キーワード「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

指示を表す接頭語や連体詞などが出現していなくても、特に企業名の場合には、名称の一致だけで照応関係を同定することができる。しかし企業名には、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) 企業名の辞書情報に別称をできるだけ与える
ex.) 日本電気←→日電←→NEC
2) テキスト中で明示された別称を取り込む
ex.) アプライド・マテリアルズ・ジャパン (AMJ)
3) 先頭からの部分一致 (語尾の“社”は除く)
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。

4. おわりに

上述した文脈処理機能を組み込むことにより、テキスト (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、一定の形式で出力するシステム **VENIEX** を実現した。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。

今後はこれらの機能をより多くの記事によって評価する予定である。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、先行詞の同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] 安藤, 土井, 村木“キーワードと構文構造に基づくテキストからの情報抽出システム”, 情報処理学会第47回全国大会, 2M-3, 1993
[2] 柴田, 田中, 福本“新聞社説記事における照応現象”, 情報処理学会第40回全国大会, 5F-4, 1990
[3] 内藤, 島津“談話意味構造の極小拡大を求めることによる照応処理について”, 情報処理学会自然言語処理研究会, 72-4, 1989
[4] 桃内“同一名詞による照応表現について”, 情報処理学会自然言語処理研究会, 68-7, 1988
[5] 野垣内, 飯田“キーボード会話における名詞句の同一性の理解”, 情報処理学会自然言語処理研究会, 72-1, 1989