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

2M－3

安藤 真一　土井 伸一　村木 一至
NEC　情報メディア研究所

1. はじめに

近年、大量の電子化テキストが利用できるようになり、これらから有用な情報を得るための支援技術として要約や情報抽出が研究されている。要約技術には頻出するキーワード、文末表現、接続表現などを利用する手法がある [1, 2, 3]。通常要約すべき主題はユーザの視点によって異なり、ユーザがこれを選べることが必要である。しかし、現在の要約技術では固定した視点からの情報しか提示することができない。

一方、情報抽出技術は取り出すべき情報を予め詳細に定義し、これを正確に取り出す技術である。ここでは文章内容に対する視点を明確に定めているため、限られた空間内で情報を扱うことができる。このため情報抽出技術は一般的な意味解析やモデル化を必要とする要約技術に比べて実現性が高い。さらに、抽出すべき情報の種類を複数用意することにより、複数の明確な視点を提供できる要約技術として考えることができる。

以上の観点から、我々は要約の基礎技術として、ある1つの視点からの情報を新聞記事から抽出する情報抽出システム **VENIEX** を試作した。本システムは抽出すべき分野に依存した語彙(以後、キーワードと呼ぶ)の知識を構文構造に応じて組み合わせることによって情報を抽出している。

2. キーワードと構文構造に基づく情報抽出

要約の視点として「事実」を考える。「事実」を「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)」という5W1H1Dからなる情報と定義すると、これは7つのスロットを持つフレームとして表現することができる。さらに各スロットに埋めるべき語彙の分野を限定し、フレーム内におけるスロット間の関係を固定することにより、視点はある特定分野の明確なものとなる。この視点に従ってテキストから情報抽出することを考える

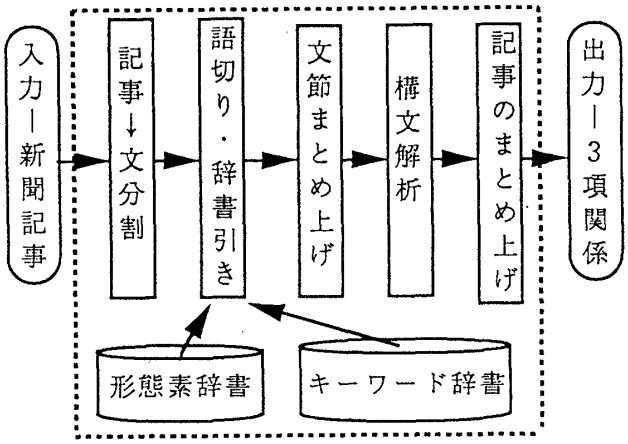

図1　システムの構成

と、抽出すべき情報はスロットを埋めるキーワード語彙と複数のキーワードからなるフレーム構造である。ただし、フレーム構造内で各スロットは特定の関係を持っているため、出現したキーワード間の関係がスロット間関係と一致することがフレーム構造を抽出する条件となる。

本システムはキーワード間関係を解析する手法として構文構造を利用する。ここでテキストのキーワード集合に対しフレーム構造が抽出できると認定された場合、すなわちキーワード間関係が抽出すべきフレームのスロット間関係に一致する場合、各キーワードをそれぞれ対応するスロットに埋めてフレームを生成し、これを抽出結果として出力する。

以下では固定した視点の一例として半導体製造技術を取り上げる。ここでは、whatを「レイヤリング」「エッチング」などの半導体製造技術、do_whatを「開発」「販売」などの4概念と限定することにより、視点を明確化した。ここで抽出すべきフレームのスロット間関係は「誰が半導体製造技術をどうした」という「who － what － do_what」の3項関係とした。また、例えばwhoに埋まる企業の存在する場所といった各スロット毎の詳細情報も定義し、これも抽出対象とした。

3. システムの機能と構成

図1に本システムの構成を示す。

Information Extraction System based on Keywords and Text Structure
Shinichi ANDO, Shinichi DOI, Kazunori MURAKI
NEC Information Technology Research Laboratories

本システムは3項関係の抽出において企業名、半導体製造技術名、関係語 (do_what スロットに埋められる語) や地名などのキーワードを利用している。これらキーワードは詳細情報を含めた構造でキーワード辞書に格納されている。本システムでは構文解析によって得られた構文構造に従い、キーワード間の関係を計算する。ここでは言語構造のレベルによって計算可能なキーワード間関係が変化するため、文節、文、記事の3つのレベルに分けて説明する。

文節レベルの処理では文節の認定と同時に、キーワードの現れた文節をマークする。ただし、1文節中に複数のキーワードが現れた場合にはそれらのキーワードの組み合わせやパターンをキーとして、キーワード情報の合成や一方のキーワードの削除などを行っている。例えば、「日電アネルバ（本社東京）は」という文節には企業名、地名の2種のキーワードが出現しているが、このとき、地名キーワード情報を企業キーワード情報の中の存在場所に書き込み、合成を行う。

さらに1文全体の構文構造に従って、キーワードを含む文節間の関係を認定し、3項関係の抽出を行う。ここでは主に do_what を示す用言の持つ格パターンやキーワード同士の修飾関係を利用する。例えば、

日本企業では最大手のニコンが今年16M
量産対応のステッパーを発売した。

という文では関係語キーワード「発売」のが格に企業名キーワード「ニコン」が、を格に技術名キーワード「ステッパー」が入る。このとき、3つのキーワード間に関係が成立し、「ニコン − ステッパー − 販売」の3項関係を抽出する。

最後に各文に分散しているキーワード情報、キーワード間関係をマージして、記事全体の情報を抽出する。ここでは文脈構造を推定し、照応や省略の対応関係を利用する [4]。

4. キーワード推定

テキスト内にキーワードが存在しない場合、本システムは情報抽出することができない。しかし特に企業名などが固有名詞であることを考えると、これら全てを登録しきることは困難である。このため、本システムはキーワード推定を行っている。ここでは文節まとめ上げ部と構文解析部でそれぞれ異なったパターンを利用する。

文節まとめ上げ部ではキーワードの並び方、文字種、特定の語より構成されるパターンを利用することによって、キーワード推定を行っている。主なパターンを以下にあげる。

1 企業名推定
- ・文字列 +「社」
- ・文字列 + 特定の表現を含む括弧句
  (例) 日電アネルバ（本社東京）
- ・地名 + 企業名
  (例) 日本IBM

2 装置の商品名推定
- ・技術名 +「文字列」
  (例) エッチング装置「ILD-4031」

構文解析部では構文構造を利用し、企業名推定を行っている。例えば、企業名の位置に未定義語を埋めることにより3項関係が生成できた場合、この未定義語を企業名として認定する。

5. おわりに

本稿では構文解析によってキーワード間の関係を認定しながら、事実を表すフレームからなる情報を抽出する情報抽出システムについて述べた。本システムで新聞記事約900記事に対して半導体製造技術情報の抽出実験を行なったところ、再現率 (正解のうち抽出できた情報の割合 [5])41.9%、適合率 (抽出した情報のうち正解の割合 [5])58.9%という結果を得た。再現率の低さはキーワードとする語彙、特に関係を表す語彙が100語強と少なかったことに起因していると考えられる。また適合率の低さは文毎に抽出した情報のマージが十分でないため、同じ結果が複数現れた結果であると考えられる。以後、詳細な評価を行い、上記2つの問題点を含めた課題を考慮して、より本格的なシステムを構築する予定である。

参考文献

[1] 渡辺, 辻井, 長尾 “文の表層上の手掛かりを用いたテキスト構造解析”, 情報処理学会第32回全国大会予稿集, Vol.2, pp.1633-1634 (1986.10).

[2] 江原 “抄録化のためのトリガ語の分析”, 情報処理学会第42回全国大会予稿集, Vol.3, pp.180-181 (1992.3).

[3] 豊浦, 有田 “テキストの内容を表すワードマップ作成の試み”, 情報処理学会研究報告, Vol.92, No.87, pp.17-23 (1992.11).

[4] 土井, 安藤, 村木 “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”, 情報処理学会第47回全国大会予稿集 (1993.10)

[5] “MUC-4 EVALUATION METRICS”, Proceedings Fourth Message Understanding Conference(1992)