新聞記事からの情報抽出と多言語インデックス付与

4Q－9

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

新聞記事には政治、経済、技術など有効な情報が多く含まれており、国内外問わず新聞記事の提供サービスへのニーズは高い。しかし単に大量の記事を提供しても、そこにはユーザにとって不要な情報が大量に含まれるため、ユーザは利用可能な情報を選別することが困難になる。

このため新聞記事の提供サービスでは情報選別が不可欠であり、また幅広いユーザに対応する多言語化が必要であると考える。この2つの組合せによってユーザの知りたい情報に関する部分の翻訳結果を提供することで、ユーザはその結果を直接利用したり、元記事へのインデックスとして利用することができる。そこで我々は、ユーザが予め知りたいと指定した情報を新聞記事から抽出し、多言語で文章化するシステムを試作した。特に本稿では抽出対象としてマイクロエレクトロニクス(ME)機能情報を取り上げ、「誰がどのような半導体製造技術をどうしたのか」という情報を抽出し、多言語で文章化する手法について述べる。

2. 全体構成

本システムはユーザに指定された情報として「事実」を取り上げ、その表現に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)）から成る事実フレームを用いた。さらに例えば半導体製造技術の分類といった詳細情報も扱い、これをスロット毎にフレーム構造で表現した。以下ではこのフレームを下位構造フレームと呼ぶ。ユーザは、下位構造フレームを含む事実フレーム全体の構造と、各スロットを埋める語彙の分野を指定することで抽出すべき情報を指定することができる。

図1にシステムの構成を示す。

本システムは大きく分けて、ユーザに指定された事実フレームを抽出する情報抽出部と、抽出された事実フレームから自然言語文章を生成する文章生成部からなる。事実フレームの詳細な構造は予めユーザの指定によって固定されているため、独立に各モジュールを開発、評価し、置換することができる。また特定分野の決まったフレーム構造を扱えばよいため、各モジュールでの処理も比較的簡素な処理で行うことができる。

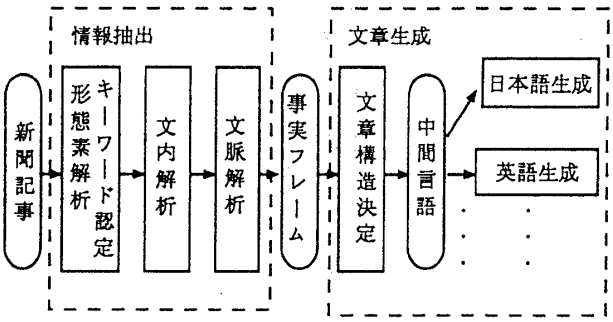

図 1: システムの構成

本システムでは情報抽出に、指定された内容に関する複数種の語彙(以下キーワードと呼ぶ)の組合せと構文構造に従って事実フレームを生成する情報抽出機構 [1, 2] を利用した。また文章生成では事実フレームから文章構造を決定し、これを機械翻訳システムPIVOTの中間言語へ写像する機構に加え、多言語を生成する文生成機構 [3] を利用した。以下にそれぞれについて述べる。

3. 情報抽出

ある決まった分野の情報を抽出する場合、分野に依存した特定語彙が手がかりになる。すなわち、関連語彙の存在がME機能情報の存在を示していると考えられる。本システムは対象分野の関連語彙を中心に解析するため、キーワード辞書を設けている。関連語彙は一語でその分野における意味が決まるため、事実フレームの一部が埋まった部分フレーム構造を意味構造として持つ。

さらに情報抽出では構文構造を利用している。システムはキーワードを優先して構文解析を行い、係り受け関係の認定された2つのキーワードについてその部分フレームの合成を行っている。これにより単純なキーワードの共起関係だけでは正確な抽出の困難な場合にも対処している。例えば

Information Extraction from Newspaper Article and Multi-lingual Indexing
Shinichi ANDO, Shinichi DOI, Kazunori MURAKI
NEC Information Technology Research Laboratories

「住金は日電アネルバが開発したスパッタリング装置を販売している」という文の場合、単純な共起関係の利用では「日電アネルバによる開発」、「住金による販売」だけを正確に抽出することは難しい。また「日電アネルバはスパッタリングの材料を開発した」という文では、開発したのは材料であるにも関わらず「スパッタリングの開発」を抽出してしまう。本システムは構文構造により、合成すべきキーワードの組を選択している。

また複数の文に跨る事実フレームを記事全体で統合する文脈処理機構やキーワード辞書に存在しない企業名を推定するキーワード推定機構も有している。

情報抽出部分については人手で作ったフレームと出力を比較することによって評価することができる。新聞記事100記事について行った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%であった。同じ評価値を人間が行った同じ作業に対して計算すると60-80%になるという報告があり[4]、キーワードの充実によりほぼ人間に近い結果を出すことができると考えている。

4. 文章生成

通常、文章生成では出力すべき話題の選択とこれに応じた文の組み立てが問題となる。しかし本システムでは情報抽出部が話題の選択を行っている。その話題とはユーザが予め指定した内容、すなわち事実フレームである。また事実フレームは基本的に5W1H1Dの構造を持ち、詳細情報フレームも事実フレームとの関係が予め曖昧性なく定義されている。このため、ユーザの指定した内容に関する知識を用いて、文に変換することが比較的容易にできる。例えばME機能情報の場合、入力された事実フレームが「who - what - do-what」だけであれば、各スロットの値を「誰が何をどうした」のような文形に写像すればよい。

ただし、各スロットは下位構造フレームを持つ場合があり、下位構造フレームのスロット値は、上位スロット値を修飾する語彙であると考えることができる。このため1文内に下位構造まで含めると構造の複雑な、長い文となってしまう。本システムでは提示する文を分かりやすいものとすることを目的として、文章構造を決定している。ここでは1文の長さがある一定以上になる場合に下位構造の情報を分割し、別の文として生成した。また分割に際して各文に納める情報の選択が必要となるが、本システムでは下位構造フレームのスロットと上位スロットとの関係を用いた。具体的には下位のスロット値が同一文に入ることにより増加する文節数と、上位下位のスロットの結束性に基づき優先度を設け、これを利用した。

入力:
日本電気は六十四メガ（一メガは百万）ビットダイナミックＲＡＭ（６４ＤＲＡＭ、記憶保持動作が必要な随時書き込み読み出しメモリー）対応の次世代エッチング装置を開発した。同社が開発したのはＥＣＲ（電子サイクロトロン共鳴）を用いたプラズマエッチング装置。プラズマ状態を作るための磁場を工夫することでプラズマをウエハーの近くで発生させるようにした。

⇓

出力:
日本電気は６４メガビットＤＲＡＭ用のプラズマエッチング装置を開発した。
NEC has developed plazma etching equipment for 64Mbit DRAM.
NEC ha developeado plazma_etching equipo para 64Mbit DRAM.

図2: 入出力例

本システムは文章構造を決定した後、1文毎にPIVOTの中間言語へ写像し、多言語生成モジュールを用いて多言語文章を生成する。

図2に入出力例を示す。

5. おわりに

本稿では新聞記事からユーザの指定した情報を抽出し、これを多言語文章で出力するシステムについて報告した。これによって新聞記事からユーザにとって有用な情報を抽出し、多言語で提供することができる。しかしそのためには、ユーザが新たに抽出すべき情報を指定したとき、情報抽出部と文章生成部がこれに適応する分野移行性が問題になる。今後、解析と生成の機能向上と共に、この機構を設計していく予定である。

参考文献

[1] 安藤、土井、村木“キーワードと構文情報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回全国大会予稿集, Vol.3, pp.83-84 (1993.10).

[2] 土井、安藤、村木“キーワードと構文情報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, Vol.3, pp.81-82 (1993.10).

[3] A.Okumura, K.Muraki and S.Akamine “Multilingual Sentence Generation from the PIVOT interlingua” , *Proc. of MT SUMMIT III*, pp.67-71 (1991.7).

[4] Beth M. Sundheim, “Overview of th Fourth Message Understanding Evaluation and Conference”, *Proceedings of Fourth Message Understanding Conference*, pp3-21 (1992.6)