自然言語処理 124－13
（1998．3．13）

トップダウンなパターン解析に基づく情報抽出

西野 文人[*1] 落谷 亮[*1] 木田 敦子[*2] 乾 裕子[*2] 桑畑 和佳子[*3] 橋本 三奈子[*3]
{nisino,ochi}@flab.fujitsu.co.jp,
{akida,hinui}@ibs.or.jp,
{waka,hasimoto}@aisys.se.fujitsu.co.jp
[*1] 富士通研究所 [*2] 計量計画研究所 [*3] 富士通

新聞における企業活動に関するの記事など事象が明確に記述される文書では、情報抽出における抽出精度の向上や処理の複雑さを低減するために、事象構造の制約により名称抽出などの部分構造抽出を行なうのが効果的と考えられる。このようなに考えに基づき、トップダウン処理のパターン解析により事象構造を決定し、事象構造を制約として名称認識を行ない，さらには実世界との対応付けを行なうシステムを作成した．このシステムを用いて新聞の組織合併情報，新製品情報からの情報抽出の実験したところ，組織名適合率 80～90%，合併事象の抽出率 55～75% を得た．

Information Extraction using Top-down Pattern Analysis

Fumihito NISHINO[*1], Ryo OCHITANI[*1]
Atsuko KIDA[*2], Hiroko INUI[*2]
Wakako KUWAHATA[*3] and Minako HASHIMOTO[*3]

[*1] Fujitsu Laboratories Ltd.
[*2] The Institute of Behavioral Sciences
[*3] Fujitsu Ltd.

The information on the event structure improves the quality of information extraction and reduces the complexity of the process for the documents that express the events clearly, such as newspaper articles on corporate activities. We developed an extraction system that generates event structures by the topdown pattern analysis and extracts named entities based on the restriction given by the event structures. After the pattern analysis, the system relates the extracted entities with the real world entities. An experiment of extraction from news articles on corporate mergers and new products shows 80-90% precision for the name of organizations and 55-75% precision for the merger events.

1 はじめに

大量の文書の中から情報ニーズに合致した文書を見つけ出すのが情報検索であり，その文書を課題解決に適するように加工するのが情報抽出であると言われている [1]．そしてこの情報抽出の研究は米国では一連の MUC (Message Understanding Conference) の中で促進されてきた [2]．日本語を対象とした情報抽出も，MUC の多言語版として MET (Multilingual Entity Task)[3] の中での日本語名称認識 (Named Entity) タスクのコンテスト [4] をはじめとして様々な研究が行なわれており，その手法としては，既存の形態素解析をチューニングして利用するもの [5]，形態素解析結果を再加工するもの [6]，形態素解析を行なわずに字面のパターンマッチング処理で行うもの [7, 8, 9] などがあり，またそのパターンマッチングの高速化の研究なども行なわれている [10, 11]．そして MET コンテストによれば日本語の名称はかなり高い精度で認識できることが報告されている [4]．

名称認識タスクは文章中の固有名や時間表現，数値表現などを認識するものであるが，実際に文書を検索したり文書中の内容を整理して提示することで課題解決の支援を行うときには，単に名称を認識するだけでなく，記事全体から事象構造を抽出すること，すなわち MUC の中の Scenario Template タスクの実現も必要になってくる．すなわち，文章中に出現する単語に対して，例えば「A 社が X を発売する．」という文では「A 社」が <組織名> であり，「X」が <製品名> であり，「発売する」が <行為> であるというように，文章中の各部分に対して属性を付与していくというだけでなく，事象（この例では <新製品の発売>）を定義し，その事象における各属性（ <販売会社>, <新製品情報> など）に対して属性値（「A 社」，「X」など）を埋め込み，さらにその名称や日時などを実世界で特定することが重要になってくる．

そこで我々は新聞記事のような半定型文書に対し，トップダウンに事象構造を決定し、事象構造を制約として名称認識を行ない，さらには実世界との対応付けを行なうシステムを作成した．本稿では我々のシステムを紹介し，企業合併情報や新製品情報などの事象を抽出対象とした情報抽出実験の結果について示す[1]．

2 全体方針

我々の目標は事象構造を把握することであるが，実際の新聞記事では実際の事象の必要要素がどのように記述されているであろうか．通常，新聞記事では 1 文めに事象の全体の要約が比較的定型的な形で記述され，2 文め以降でこれを補足する情報が記述されているのではないかと考えられる．しかし，1 文めの分析だけで事象構造を把握するのに充分であろうか．そこで企業の合併事象を例にとって，日経新聞における 1 文めに出現する情報を調査したところ，独立企業同士の合併では 1 文め中に合併する企業名が出現するのに対して，親会社があって，その子会社同士が合併したり，吸収合併のような場合には必ずしも合併する企業名が 1 文め中には出現しないことや，合併後の新組織の情報については多くの場合 2 文め以降に記述されていることがわかった（この調査に対する詳細な報告は [12] で述べる）．新製品情報や研究開発情報などでも同様に 1 文めに要約が述べられ，2 文め以降で必要情報が補足されている．そこで，我々は，1 文めを詳しく分析することで事象構造の枠組を定め（この調査に対する詳細な報告は [13] で述べる），1 文めだけでは埋め込まれずに残った属性の情報を 1 文めの分析で得られた手がかりを利用して 2 文め以降から求めて埋めるという処理の方針を立てた．

3 エンジン設計方針

システム設計ではまず形態素解析を利用するか利用しないかが大きく方針のわかれるところである．形態素解析は未知語，特にひらがな名の名称に弱い．大量の固有名辞書を用意してもなお未知語の問題はつきまとうし，逆に大量の固有名を辞書に登録すると形態素解析に色々な副作用が生じる（例えば単に「ん」のような会社名が実在する）．新聞のような一般人を相手とした文章では特別な知識を有さない一般の人が正しく固有名を認識できるような表記上の工夫がなされて

[1]本研究開発は IPA 創造的ソフトウェア育成事業の支援によるものである

いるはずである．そこで我々は大規模な辞書に頼った形態素解析を行なうのではなく，様々な表記上の手がかりに基づいた字面のパターンマッチング処理を基本とすることにした．システムの構成を図 1に示し，以下ではこのシステムの特徴とその考え方について述べる．

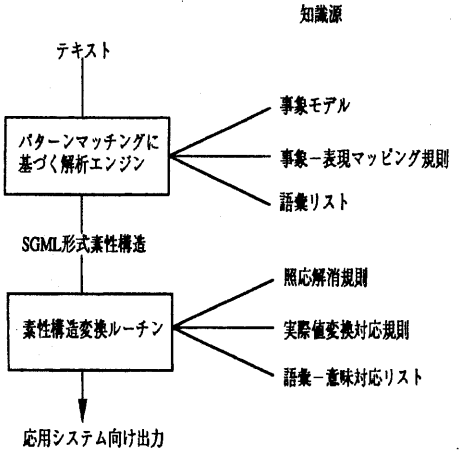

図 1: システムの構成

3.1 トップダウンアプローチ

多くのシステムでは，記事を解析するのに，日付，固有名などをまず認識し，それから順に大きな構造を組み上げていくようなボトムアップのカスケード型の処理を採用している [14, 15]．これに対して我々はパターンマッチングに基づく解析フェーズと素性構造変換の 2 フェーズとし，解析フェーズでは記事全体を入力として受けとり（文の切りだしは行なわない），事象モデル，事象－表現マッピング規則，語彙リストを知識源として利用して，まず事象判定を行い，その判定にそって必要な項目を抽出し，必要に応じて抽出された項目の細部をさらに抽出していくというアプローチをとった．図 2に事象－表現マッピング規則の例を示す．ここで module ～ end で囲まれている部分が一つの事象（あるいは実体）と表現とのマッピング規則であり，1 つの事象に対する表現のバリエーションを記述しているものである（なお# で始まる行はコメントである）．抽象化された概念は抽出項目として「<」と「>」とで囲んで示している．抽出項目は意味的な役割を示す名前のほかに，実体の型が宣言されており，その型に応じて細部情報が解析される．例えば，図 3の記事の 1 文めの解析では <販売主体組織情報> としてまず「薩摩酒造（鹿児島県枕崎市、本坊喜一郎社長）」の部分が抽出され，その後 <販売主体組織情報> の型である組織型の事象－表現マッピング規則にしたがって <組織名> や <組織所在地>，<組織社長名> が抽出される．このようにトップダウンに解析を進めることで，文脈に応じたパターンマッチングを可能としている．

3.2 表現と事象のマッピング

抽出したい事象の枠組を表現する事象枠組規則と，名称や述語の語句の構成のバリエーションを表現する語句構成規則の組合せにより，表現と事象のマッピングを見通しの良いものとしている．すなわち，事象枠組規則によって，抽出すべき名称の文脈（外部証拠）を処理し，また，語句構成規則によって，抽出すべき名称の語や句の構成（内部証拠）のバリエーションを内部証拠の文字列パターン型として処理することで，規則記述の見通しの向上を図っている．

3.3 語彙リスト

同一のことを表現するのに色々なバリエーションがあるが，語彙レベルのバリエーションまでを含めて事象－表現マッピング規則にすべて列挙するのでは規則が繁雑になり，規則の管理が容易でなくなってしまう．そこで，語彙リストの中で，ある語彙集合に属性を定義し，その属性を用いてパターンマッチングを行うようにしてある（規則中で& 名前; が属性名を指定しているものである）．

3.4 属性付与

あるパターンにマッチした時，本文中にはない情報を付与したくなることがある．そこで，パターンマッチが成功した時に，属性を付与することができるようにした．これにより，例えば時制が過去なのか未来なのかといったようなことを属性として与えている．

```
%
module 販売情報・用言型0

## α ハ/ガ β ヲ ##

<デ格要素>& デ;<販売主体組織情報>& ハ/ガ;<販売日付><販売製品情報>& ヲ;
<販売主体組織情報>& ハ/ガ;<デ格要素>& デ;<販売日付><販売製品情報>& ヲ;
<販売主体組織情報>& ハ/ガ;<販売日付><デ格要素>& デ;<販売製品情報>& ヲ;
<販売主体組織情報>& ハ/ガ;<販売日付><販売製品情報>& ヲ;<デ格要素>& デ;
<販売主体組織情報>& ハ/ガ;<販売日付><販売製品情報>& ヲ;

# 以下略

end

module 組織情報

## 運営対象 ##
<運営対象> を & 運営; する <組織情報>
<運営対象>& 運営; の <組織情報>
<運営対象> を営む <組織情報>

# 以下略

end
```

図 2: 情報抽出規則例

3.5 素性構造変換

パターンマッチングの解析結果は基本的には文を分割して分割した各部分の役割を明確にしたものにすぎない．情報検索や情報整理をするためには数値・日付の解析や照応表現の解析，必要項目の取り出しによる整形などが必要になってくる．そのためにはパターンマッチングの解析結果をこれらの処理が行いやすい形式にしておくのがよい．そこで我々はパターンマッチング解析の結果を SGML 表現による素性構造とした．すなわち，各構造は 1 つの構造名（SGML タグ名）と，0 個以上の素性構造ないし属性名－属性値のペア（SGML の属性ないしタグ内の構造）と，必要に応じて参照用のラベル（SGML 属性で表現される）を持たせる構造とした．図 4に記事分析結果の SGML 構造を示す．

3.6 実世界値との対応

情報を整理するという観点では，表現から実体へのマッピングが一つの重要な役割をはたす．検索もれを防ぐために表記のゆれ処理や同義語展開処理が行なわれることは多いが，逆にノイズを減らす処理も必要である．例えば，「日本電気」という名前の組織には少なくとも「日本電気株式会社」と「株式会社日本電気」という別の組織が存在する[2]．しかし新聞には実体を特定するために正式名称での記述や業種，本社地，社長名等の記述もなされているのでこれらの情報から実体を特定することが必要である．

また，数値，日付は単にその文字列の範囲を認定するだけでなく，情報検索や情報整理ができるように絶対値に変換する．例えば，「一日に発売する」が 1998 年 3 月 20 日の記事ならば，「 <date

[2]帝国データバンクによる

```
<article>
<number>152841 </number>
<date>19940309 </date>
<medium> 本紙地経面 </medium>
<page> 九州B </page>
<id>940309-2101 </id>
<headline> 薩摩酒造、きょう発売、樽で貯蔵熟成の麦しょうちゅう。 </headline>
<body> <p>　薩摩酒造（鹿児島県枕崎市、本坊喜一郎社長）は九日、カシの樽（たる）で貯蔵熟成した麦しょうちゅう「琥珀（こはく）の夢」＝写真＝を鹿児島、熊本、宮崎の三県で発売する。三月中には九州全県に広げ、年内には全国で販売する。 </p>
<p>　琥珀の夢は原料に二条大麦を使用、貯蔵熟成したため、琥珀色で香りが高く、オンザロックやストレートで味わいの深さを楽しめる。アルコール分は二五度。参考小売価格は消費税込みで一・八リットル瓶入り一本千六百円。 </p>
</body>
</article>
```

図 3: 新製品情報記事例

value='1998-4-1'> 一日 </date> に発売する」のような属性値を付与し，「一日に発売した」という記事ならば，<date value='1998-3-1'> という属性値を付与する．「来月」，「上旬」のような語から実際の数値に変換するための情報は語彙－意味対応リストとして保持している．

また，新聞記事では，「同」とか「両」がよくつかわれており，例えば，「xx 社（社長 yy）と zz 社（同）」のような「同」が社長が yy であることを示していることを認識する必要がでてくる．参照には属性名－属性値を指すもののほかに，属性値だけを指すもの，属性名を指すもの，文字列を指すものなどがあり，推論を必要とするものもあるが，現在は表現上，括弧で括られたものだけに対応している．

4 実験・評価

情報抽出の規則の作成には 1990 年から 1994 年までの日本経済新聞の記事を利用し，実験・評価としては 1996 年の日本経済新聞の記事および 1991 年，1992 年の毎日新聞の記事を利用した．なお実験対象としては組織合併に関する話題と新製品に関する話題を選び，あらかじめ本文の文字列検索で絞り込んだ記事集合を利用した．

事象判定では，どの情報までを事象抽出の対象とするかは議論の残るところであるが，おおよそ，合併事象で 72%，新製品情報で 91% の再現率を得ている（対象は日本経済新聞）．うまく事象判定できていない原因としては，複文（例：「合併，～としてスタートした」）や，複合表現の一部の述語（例：「合併する構想を発表した」），他の語を使っての表現（例：「～を一つにする」）があった（詳細は [13] で述べる）．

次に事象判定された文に対して解析を行ない，合併元組織名，合併後新組織名，製品販売元組織名が正しく抽出できたかを MUC-5 にならって部分正解は 0.5 で乗じたカウント [16] で適合率，再現率を求めた（表 1）．さらに，合併事象に関しては，記事中に存在する必須要素がすべて埋められたかどうかとして，合併事象を 1) 独立組織どうしの合併，2) ある組織の子会社などの関連組織どうしの合併，3) ある組織の子会社などの関連組織と独立組織の合併，の 3 つの型に分けて抽出率を求めた（表 2）．なお，ここでは必須要素としては合併するすべての組織と新組織の名前とした．

```
<記事内容 文末表現=発表述語なし>
    <記事内容・用言型0 field=製品販売情報 アスペクト=未来>
        <販売情報・用言型0>
            <販売主体組織情報>
                <組織名> 薩摩酒造 </組織名>
                <販売組織補足情報 type=組織補足情報 >
                    <要素1 要素数=2 type=要素 >
                        <組織所在地> 鹿児島県枕崎市 </組織所在地>
                    </要素1>
                    <要素2 要素数=2 type=要素 >
                        <氏名> 本坊喜一郎 </氏名>
                        <役職名> 社長 </役職名>
                    </要素2>
                </販売組織補足情報>
            </販売主体組織情報>
            <販売日付 type=date value="1994-03-09">
                <日> 九日 </日>
            </販売日付>
            <販売製品情報>
                <製品名連体修飾句> カシの樽（たる）で貯蔵熟成した </製品名連体修飾句>
                <製品情報>
                    <種別> 麦しょうちゅう </種別>
                    <製品名> 琥珀（こはく）の夢 </製品名>
                    <写真> =写真= </写真>
                </製品情報>
            </販売製品情報>
            <デ格要素>
                <販売地域> 鹿児島、熊本、宮崎の三県 </販売地域>
            </デ格要素>
        </販売情報・用言型0>
    </記事内容・用言型0>
</記事内容>
<価格 type=price unit=円 value="from 1600 to 1600">
    <金銭> 千六百 </金銭>
</価格>
```

図 4: 記事分析結果

井出ら [10] は抽出すべき情報とその周辺の文字列との関係から学習データに基づいて機械的に作成したテンプレート（事象－表現マッピング規則）を用いて，1994 年の日本経済新聞の製品紹介記事で実験をしている．そこにおける製品販売元組織名の抽出率は適合率，再現率とも 75 ～ 80% であった．評価セッ

	日本経済新聞		毎日新聞	
	適合率	再現率	適合率	再現率
合併元組織名	72%	70%	91%	88%
合併後新組織名	100%	63%	100%	35%
製品販売元組織名	96%	67%	82%	51%
トータル	80%	69%	89%	70%

表 1: 名称抽出

	日本経済新聞 抽出率	毎日新聞 抽出率
1) 独立合併型	55%	68%
2) 関連組織合併型	50%	75%
3) 関連組織−独立組織合併型	70%	89%
トータル	57%	74%

表 2: 事象抽出

トは異なるが，難易度などは我々の評価セットとほぼ同等と考えてもよいだろう．これと比べると我々の課題は再現率の向上，すなわち表現のバリエーションへの対応であることがわかる．特に，合併後新組織は 2 文め以降に出現することも多く，それへの対応が課題である．

5 おわりに

我々はテキストを構造化することで情報検索や情報整理に役立てようとしている．一つの応用として，情報抽出によって構造化した文書に対して情報検索を行なえるようにし，検索結果を整理して提示するシステムを作成した．図 5は実際に焼酎製品の情報を検索した結果の表示例である．

実用的なシステムを目指すには今後情報抽出の再現率の向上とともに様々な事象に対応していくことが大きな課題となっている．今回，人手によって規則の開発を進めてきたが，この経験を活かして規則作成支援，あるいは規則の自動学習について検討していきたいと考えている．

参考文献

[1] Cowie, J. and Lehnert, W.: Information Extraction, *Communications of the ACM*, Vol. 39, No. 1, pp. 80–91 (1996).

[2] Grishman, R. and Sundheim, B.: Design of the MUC-6 Evaluation, in *Proc. Sixth Message Understanding Conference(MUC-6)*, pp. 1–11 (1995).

[3] Merchant, R., Okurowski, M. E. and Chinchor, N.: The Multilingual Entity Task(MET) Overview, in *Proc. Tipster Text Program(Phase II)*, pp. 445–447 (1996).

[4] Maiorano, S. and Wilson, T.: Multilingual Entity Task(MET): Japanese Results, in *Proc. Tipster Text Program(Phase II)*, pp. 449–451 (1996).

[5] Takemoto, Y., Wakao, T., Yamada, H., Gaizauskas, R. and Wilks, Y.: NEC Corporation and University of Sheffield: Description of NEC/Sheffield System Used For MET Japanese, in *Proc. Tipster Text Program(Phase II)*, pp. 475–476 (1996).

[6] 江里口善生，木谷強：パターンマッチング手法による名称特定処理の有効性の検討，情処研報，NL115-10, pp. 67–73 (1996).

注釈情報一覧表（製品販売）

全記事数 ：13 異なる記事数 ：13 表示記事数 ：1−13

組織名	製品種	製品名	価格	発売日	
高千穂酒造	高濃度の高級焼酎（しょうちゅう）		2,500	1996/11/12	長期貯蔵焼酎２製品、
ニッカウヰスキー	韓国産しょうちゅう			1994/04/05	ニッカ、韓国産しょう
白露酒造	高級焼酎	季の詩（ときのうた）	2,000	1992/03/28	白露酒造、トウモロコ
舞酒造	びん入りの純米焼酎（しょうちゅう）	舞心	650	1991/08/01	花の舞酒造、びん入り
新得酒造公社	乙類焼酎	サホロ・ラトニュー	630	1990/01/10	新得酒造公社、下旬に、
舞酒造	米しょうちゅう	舞心（まいごころ）	1,200	1990/08/06	酒かす、しょうちゅう、
白花酒造	バーボン風味の焼酎（しょうちゅう）	バーボニック焼酎Ｂ＆Ｓ		1990/10/01	白花酒造が発売、バー
雲海酒造	香りの高い本格焼酎（しょうちゅう）	香り仕込	300	1991/03/20	雲海酒造、香り高い冷
薩摩酒造	麦しょうちゅう	琥珀（こはく）の夢	1,600	1994/03/09	薩摩酒造、きょう発売、
薩摩酒造	特製化粧箱入りの焼酎（しょうちゅう）	皇太子御成婚・慶祝ラベル	1,500	1993/06/01	皇太子さまご成婚記念
壱岐焼酎協同組合	開かれる「長崎『旅』博覧会」の記念ボトル入り麦しょうちゅう	壱岐っ子	2,800	1990/08/09	「旅」博記念ボトル入
白露酒造	芋焼酎（しょうちゅう）	酎党和平（ちゅうとうわへい）	730	1991/02/26	湾岸戦争の早期終結を
宝酒造	本格純米しょうちゅう	よかいち	1,250	1993/03/05	宝酒造、しょうちゅう

全記事数 ：13 異なる記事数 ：13 表示記事数 ：1−13

図 5: 構造化文書の検索結果

[7] 木谷強：固有名詞の特定機能を有する形態素解析処理，情処研報，NL90-10, pp. 73–80 (1992).

[8] 松尾比呂志，木本晴夫：抽出パターンの階層的照合に基づく日本語テキストからの内容抽出法，情報処理学会論文誌，Vol. 36, No. 8, pp. 1838–1844 (1995).

[9] 井出裕二，藤吉誠，永井秀利，中村貞吾，野村浩郷：テンプレートを用いた新聞記事からの製品情報抽出システム，情処研報，NL115-12, pp. 83–90 (1996).

[10] 井出裕二，藤吉誠，永井秀利，中村貞吾，野村浩郷：構造化テンプレートを用いた新聞記事からの製品情報抽出，情処研報，NL118-2, pp. 7–14 (1997).

[11] 江里口善生，木谷強：富田一般化 LR パーザを用いた情報抽出，情報処理学会論文誌，Vol. 38, No. 1, pp. 44–54 (1997).

[12] 木田敦子，乾裕子，桑畑和佳子，橋本三奈子，落谷亮，西野文人：情報抽出のための新聞記事テキスト分析，言語処理学会第 4 回年次大会 (1998).

[13] 桑畑和佳子，橋本三奈子，木田敦子，落谷亮，西野文人：新聞記事を対象とした企業動向に関する事象構造の抽出，言語処理学会第 4 回年次大会 (1998).

[14] Grishman, R.: The NYU System for MUC-6 or Where's th Syntax?, in *Proc. Sixth Message Understanding Conference(MUC-6)*, pp. 167–175 (1995).

[15] Hobbs, J. R., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M. and Tyson, M.: FASTUS: A Cascaded Finite-State Transducer for Extracting Informatio from Natural-Language Text, in Roche, E. and Schabes, Y. eds., *Finite-State Language Processing*, pp. 383–406, The MIT Press (1997).

[16] Chinchor, N. and Sundheim, B.: MUC-5 Evaluation Metrics, in *Proc. Fifth Message Understanding Conference(MUC-5)*, pp. 69–78 (1993).