自然言語処理 140－9
（2000．11．21）

複数製品の紹介記事からの製品情報抽出

赤松 順子† 高尾 宜之‡ 永井 秀利† 中村 貞吾† 野村 浩郷†

†九州工業大学 情報工学部 知能情報工学科
‡ソニー(株)
† E-mail: {akamatsu,nagai,teigo,nomura}
@dumbo.ai.kyutech.ac.jp
‡ E-mail: ytakao@sm.sony.co.jp

我々は従来の研究において，1記事中に単一の製品が紹介されている記事を対象とし，抽出項目とその周辺の文字列を記述したテンプレートを用いた字面処理の手法の有効性を確認してきた．しかし，新聞記事には1記事中に複数の製品を紹介している記事も存在し，それらの記事に対して，従来の手法を適用しても正しく抽出を行うことができなかった．

本論文では，単一の製品を抽出するテンプレートを複数の製品が紹介されている記事に対応させるため，テンプレートの拡張を行った．しかし，複数の製品が記述されている記事には構造が複雑な文章が存在し，そのような文からは字面の情報のみを用いて正しい抽出を行うことが困難であるため，構文解析を行い，抽出項目間の係り受け関係を利用した手法についても合わせて提案する．そして，テンプレートの手法との比較を行うことによって複雑な構造の記事からの抽出率を向上させることができることを示す．

Information Extraction from Newspaper Articles of Multiple Products

Junko Akamatsu †, Yoshiyuki Takao ‡,
Hidetoshi Nagai †, Teigo Nakamura † and Hirosato Nomura †

† Department of Artificial Intelligence, Kyushu Institute of Technology
‡ Sony Ltd.
† E-mail: {akamatsu,nagai,teigo,nomura}
@dumbo.ai.kyutech.ac.jp
‡ E-mail: ytakao@sm.sony.co.jp

We have reseached a textual analisyis method for information extraction from newspaper articles described only one product with templates which describe the relationship between information to be extracted and its surrounding strings and have confirmed the effectiveness of our method by experiments.However, some newspaper articles describe information about multiple products and our previous system failed to extract correct information from such articles.

In this paper,we extended our system to be able to deal with the articles about multiple products. But there are some sentences described with complicated structure in them, so it was difficult to extract from such sentence with template. As its antidote, we experimented the way to use dependency information then compared the results with template and show the efficiency with dependency information about extracting correct information.

1 はじめに

大量の電子化されたデータを用いることが可能となった現在，氾濫している情報を管理する技術が要求されている．計算機による情報抽出システムは計算機可読な文書情報を自動的に処理し，必要な情報のみを取り出してくることを可能にする．

情報抽出を行うにあたっては，構文解析や意味解析を行い，その解析結果を用いる方法が考えられるが，定型性があり，かつ抽出対象が明確である文書に対しては，これらの解析を行わずに，字面処理によって高速に情報の抽出を行うことができる．このような処理を行う抽出手法としては，抽出する項目とその周辺の文字列を記述したテンプレートと入力文章とのマッチングによるものがある．

これまで新聞記事の新製品紹介記事のうち，1記事に1つの製品を紹介している記事を対象に製品情報の抽出を行ってきた [5]．しかし，記事の中には複数の製品が同時に紹介されているものも存在し，このような記事から正しい抽出を行うことができなかった．また，複数の製品情報を含む場合，抽出された情報についてそれぞれの製品毎の情報の対応付けを行わなければならない．

本論文では1記事中に複数の製品を記述している記事を対象とした．このような記事は抽出すべき情報が並列構造で表現されることが多く，単数の製品を紹介している記事と比べて文章の構造が複雑・多様化している．そこで，これまでに用いていたテンプレートを拡張した表層処理による手法と，構文解析の結果を利用した手法を提案し，それを用いた実験により各々の手法の比較を行う．

2 複数製品紹介記事の分析

複数製品紹介記事の記述パターンを分析するため，このような記事に対し抽出すべき表現にタグ付けを行った．

[1] では日経新聞1994年版の453記事から複数製品紹介記事の分析が行っているが，今回，新たに毎日新聞1991年～1995年5月までの複数製品紹介記事895記事の分析を行った．その結果として加わった記述パターンは頻度が少ないものであった．そこで各抽出項目の出現数と表記パターンによる対応付けの詳細については [1] を参照することとする．

ここではテンプレート，構文解析，両手法の処理で直接用いる階層関係の定義についてのみを2.2節に記す．

2.1 抽出項目の設定

複数製品紹介記事の抽出を試みる内容である「抽出項目」とそれを表すタグを表1に示す．

従来の1記事中に1つの製品が紹介されている記事からの情報抽出では「製品の細分類」を除く5種類を抽出項目としていた．しかし，複数の製品紹介を含む記事には「製品名」が同じでもバージョンや販売方法の違い等で複数の製品を記述している記事があるため，「製品の細分類」という項目を新たに設けた．

分析対象 895 記事に対して，抽出項目を示すタグと対応関係を示すリストを付与したデータを用いて記事の分析を行った．タグは $\langle item\# \rangle$ と $\langle /item\# \rangle$ の間に囲まれた文字列が $item$ という情報を示し，#は抽出項目別の通し番号である．リストは，1つの製品情報 i を i=$\langle c, k, n, s, p, d \rangle$(表1参照) の6つ組と定義し，1記事中の製品情報 I は個々の製品情報の集合とする．

$\langle c1 \rangle$服部セイコー$\langle /c1 \rangle$（０３・３５６３・２１１１）は，ウエアの上から装着可能な$\langle k1 \rangle$時計$\langle /k1 \rangle$「$\langle n1 \rangle$スパイアー・ジーコンパクト型ミラー付きウオッチ$\langle /n1 \rangle$」「$\langle n2 \rangle$同提げ時計$\langle /n2 \rangle$」＝写真を$\langle d1 \rangle$２８日$\langle /d1 \rangle$に発売する。コンパクトタイプの腕時計はふたの裏側に鏡がついている。(中略) 価格は$\langle p1 \rangle$６０００円$\langle /p1 \rangle$。
I = $\langle (c1, k1, n1, \phi, p1, d1), (c1, k1, n2, \phi, p1, d1) \rangle$

タグ	役割
c	販売元
k	製品種別
n	製品名
s	製品の細分類
p	価格
d	発売日

表 1: 抽出項目のタグ

2.2 抽出項目の階層関係

タグ付けされた記事から，1記事中での抽出情報の出現頻度をもとに複数製品記事を記述パターンを分析した．「販売元」「発売日」はほとんどの場合，1記事中に1箇所ずつ記述され，全ての製品に対応する．そこで，「製品種別」「製品名」「細分類」「価格」(k, n, s, p) の対応関

係を調べるため，出現頻度毎に分類した．各項目属性値を2つ以上存在する(“複”)，1つ存在する(“単”)，存在しない(“ϕ”)の3つの属性値をつけて調査した．頻度が高い組合せを表2に示す．

(k, n, s, p)	頻度(記事数)
(単, 単, 複, 複)	287記事
(単, 複, ϕ, 複)	131記事
(単, 複, ϕ, 単)	79記事
(複, 複, ϕ, 複)	60記事

表2：1記事中の出現パターンの分類

表2の各項目の対応関係を調べると，「製品種別」の抽出情報が1つで「製品名」の抽出情報が複数の場合，必ず「製品名」の抽出情報の全てが「製品種別」の抽出情報と対応した．この他にもこのような抽出項目どうしの対応関係が見られた．そこで，抽出項目どうしの対応関係を基にして図1のような階層を設定した．この図は上位項目が単数で下位項目が複数の場合，上位項目にすべての下位項目が対応することを表している．

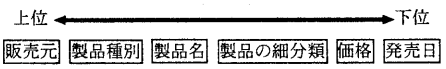

図 1: 抽出項目の階層関係

3 テンプレートを用いた情報抽出

2章で行った分析を基に，複数の製品情報を抽出するテンプレートを作成し，パターンマッチングによる抽出法を述べる．

3.1 複数項目テンプレート

情報抽出処理のためのテンプレートの定義を行う．テンプレート作成に用いる用語を示す．

- **●抽出情報:** 抽出項目に対応する情報を表す文字列
- **●パターン:** パターンマッチングの対象となる文字列長1文字以上の文字列
- **●固定パターン:** 抽出対象に頻出する特徴的な文字列(“発売する”，“販売する”など)
- **●ワイルドカード:** パターンマッチング上，文字列長0以上の任意の文字列とマッチしうるシンボル

続いて，複数製品紹介記事を抽出するためのテンプレートを作成する方法を述べる．まず，テンプレート作成用データに2.1節で定義したタグ付けを行い，項目間の関係を表すリストを作成する．

- **●タグ付けした文章:** 発売するのは奥行き1.64メートルの「<s1> KA 1</s1>」(<p1> 百三十二万円 </p1>)と同1.78メートルの「<s2> KA 2</s2>」(<p2> 百四十二万円 </p2>)。
- **●項目間の関係を表すリスト:** s1p1， s2p2

次に，タグ付けされた文章から固定パターン，正解データの前後1文字を残しテンプレートを作成する．

- **●固定パターン:** 発売するのは
- **●テンプレート:**
  発売するのは*「<s1>」(<p1>)*「<s2>」(<p2>) *: (s1p1， s2p2)

3.2 テンプレートの重み付け

抽出処理ではテンプレートの集合を用いてパターンマッチングを行うため，1つの文章に複数のテンプレートがマッチする可能性がある．そのなかで最も有効なテンプレートを選択する基準としてテンプレートに重みを与えておき，その情報を利用する．

各テンプレートをテンプレート作成用データにマッチングさせ，

$$\text{重み} = \frac{\text{情報抽出が成功した文の数}}{\text{マッチした文の数}} \times \text{抽出する抽出情報の個数}$$

の式で重みを付与した．

3.3 抽出項目間の対応付け

一つの項目について複数の情報が抽出された場合，それぞれの情報が他の項目のどの情報と対応しているかという対応関係をつける必要がある．以下のように1文中での対応付け，1記事中での対応付けを行う．

○1文中の対応付け

1文中での対応付けはテンプレートを用いて行う．作成したテンプレートには，複数の項目がある場合は，その項目間の対応関係を表すリストを付加してある．テンプレートとのパターンマッチに成功した場合は，そのリストに基づいて1文での複数項目間の対応付けを行う．

○1 記事中の対応付け

テンプレートにより抽出された抽出情報と対応関係のうち，複数の文に分かれて存在する情報を1つの製品の情報として対応付ける．このとき2.2節で示したに項目間に上位と下位の関係を利用する．

3.4 テンプレートを用いた抽出システム

テンプレートを用いた情報抽出実験システムについて述べる．システムの実装にあたってはRuby[4]を使用した．本システムの概要図を図2に示す．

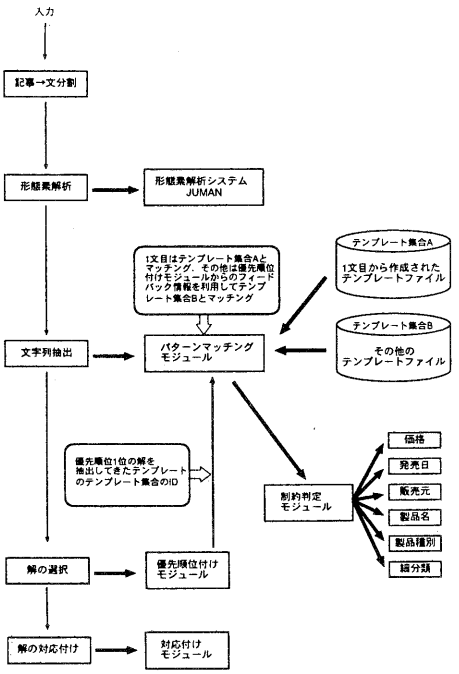

図2: テンプレートを用いた情報抽出システム

各モジュールの処理

（形態素解析モジュール）

記事の中の1文を入力として受けとり，形態素解析システムJUMAN[2]を用いて各単語の分割情報および品詞情報を得る．これらの情報は制約判定モジュールで利用される．

（パターンマッチングモジュール）

記事の中の1文を入力として受けとり，3.2節で付与した重みが高いテンプレートから順にマッチングを行い，抽出した情報を出力とする．

「販売元」「製品種別」「製品名」「細分類」のマッチングは最短一致で行い，ワイルドカードのマッチングは最長一致で行っている．

（制約判定モジュール）

パターンマッチングの結果抽出された文字列を渡され，制約判定の合否をパターンマッチングモジュールに返す．抽出項目に対する制約には，抽出項目に依存する制約と抽出項目に依存しない一般的な制約があり，制約判定の結果が1つでも失敗すれば，制約判定モジュールは不合格である．

＜＜抽出項目に対する制約＞＞

1. 抽出項目に依存する制約
   - **製品種別および販売元制約判定**
     単語の区切り情報と品詞情報を用いて制約判定を行う．また，「価格」や「発売日」の表記パターンにマッチしたら失敗とする．
   - **価格および発売日制約判定**
     「価格」は「数字＋通貨単位」，「発売日」は「数字＋“年(月または日)”」などの特徴的な表記パターンをテンプレート中に記述してあり，その表記パターンにマッチした文字列が制約判定モジュールに渡されるので，ここでは制約判定は用いない．
   - **製品名および細分類制約判定**
     「製品名」や「細分類」は販売する側が自由に付けているために，制約は用いない．ただし抽出項目に依存しない制約はかける．
2. 抽出項目に依存しない制約
   単独では意味のある句として成り立たないものを排除するための制約として，以下に示すものを用いた．
   - 括弧の対応がついてなければ失敗
   - 読点から始まるならば失敗
   - 禁則開始文字（「ぁ」や「ゃ」など）で始まるならば失敗

（解の優先順位付けモジュール）

パターンマッチングモジュールから渡される解候補の中で，ひとつの製品の抽出項目に対して複数の解が存在する場合，マッチしたテンプレートの重みを用いて優先順位をつけ，最も優先順位が高いものをその記事の抽出情報として選択し出力する．

（抽出情報の対応付けモジュール）

解の優先順位付けモジュールでその記事の抽出情報が選択された後，3.3 節で示した対応関係のルールを用いて対応付けを行う．

4 構文解析を用いた情報抽出

3 章では字面情報を用いたテンプレートの手法を述べてきたが，複数の製品を紹介する記事の中には，字面の情報のみでは正しい抽出を行うのが困難とみられる複雑な文章が存在する．このような文章からの抽出を行うために構文解析を用いた手法について述べる．

4.1 固定パターンの定義

抽出したい情報は，製品を発売する，または発売に関連することを意味する文節に係ることが多い．そのような文節を固定パターンと定義し，係り受け情報を用いて情報抽出を行う．

分析用データ 895 記事をタグ付けしたものと，日本語構文解析器 KNP[3] による出力をもとに抽出項目周辺の係り受け関係を調べた結果，記事本文 3934 文中に抽出項目が含まれる文節は 5188 箇所存在した．抽出項目は，1 文中の任意の位置に抽出項目が複数存在する場合 (表 3-1) と，特定の表現の直前のに存在する場合 (表 3-2) があった．(表 3-1,3-2 中の数値はその表現に係る抽出項目の個数を表す)

“発売”	1928	“チェンジ”	44
“追加”	83	“開発”	33
“販売”	78	“できる”	32
“発表”	57	“改良”	31
“売り”	52	“搭載”	28
“設定”	45	“開始”	18

表 3-1：抽出項目 (任意の位置) の係り先になりやすい表現

“〜円”	746	“〜機種” 他	432
“」”	630	抽出項目	128

表 3-2：抽出項目 (直前) の係り先になりやすい表現

4.2 固定パターンに係る抽出項目の分析

4.1 節で決定した固定パターンに係る文節が，係り受けの際にどの格形式で係るかによって抽出項目であるかを判定する．

これらの固定パターンに係る抽出項目の格形式のうち頻度が高いものを記事の 1 文目の組合せを表 4-1，2 文目以降の組合せを表 4-2 に示す．なお，固定パターンは出現位置や組合せによって係り受けが変わるため，格形式が類似する組合せを A〜E の 5 種類にに分類した．

	A(653)	B(122)		C(28)	
		“発売”等に係る	“追加”等に係る	“発売”等に係る	“発表”等に係る
販売元 C	未格	未格			未格
発売日 D	無格 隣接 カラ格	無格 隣接 カラ格		無格 隣接 カラ格 二格	
製品種別 K	ヲ格	ヲ格*1	二格 ヲ格 ト格 同格連体	ヲ格	
製品名 N	ヲ格	ヲ格*1	二格 ヲ格 ト格 連体	ヲ格	
細分類 S		ヲ格*1	二格 ヲ格 ト格 連体		
	ト格Nヲ格N 連体Nヲ格N 連体Sヲ格S		二格Nヲ格S 連体Nト(ニ)格N 連体S(N)ヲ格S(N) ニ格Nト格Sヲ格S (ト) (ニ)		

*1：この場合“追加”等に係る項目はない

表 4-1：固定パターン出現形態別抽出項目の格形式 (1 文目)

	A(97)	D(119)	E(397)
販売元 C	未格 ガ格	ガ格	
発売日 D	隣接 カラ格 二格		
製品種別 K	未格 ヲ格 同格連体	未格 二格 デ格	未格 ガ格 隣接
製品名 N	未格 ヲ格 連体	未格 連体 二格 ト格	ガ格 未格 隣接
細分類 S	未格 ヲ格 連体	未格 連体 二格 ト格	ガ格 デ格 隣接 未格 無格
		ト格S(N)デ格S(N) (未)	

表 4-2：固定パターン出現形態別抽出項目の格形式 (2 文目以降)

- A(1 文目，2 文目以降):“発売” を示す表現のみの文
- B(1 文目のみ):“発売” を示す表現に “追加” 等 (*1) が係る文
- C(1 文目のみ):“発売” を示す表現が “発表” 等 (*2) に係る文
- D(2 文目以降):“追加” 等のみの文
- E(2 文目以降):価格の表記が含まれる文

*1 追加，開発，改良，設定，チェンジ
*2 発表，開始，始める

4.3 係り受けルール

4.2 節の固定パターンと格形式を利用して，解候補を抽出する．その解候補に以下のルールを適用する．

(1) 解候補が以下の条件に当てはまる場合，削除する．

– 同一抽出項目の解候補AとB(文字列長:A＞B)においてBがAの部分文字列である場合
– 「販売元」「製品種別」「製品名」「細分類」の解候補が価格，発売日，固定パターンの表記パターンにマッチする場合

(2) 次の条件に当てはまる解候補は削除し，その直前の文節が“同格連体，ノ格，隣接”のいずれかの格形式でに係る場合はその文節をを解とする．

– 解候補が“～機種(品目，モデル)，新製品(商品，モデル)”に準ずる表現を含む文節の場合．

(3)「製品名」の直前の文節が“同格連体”の場合はその文節を「製品種別」の解候補に追加する．

4.4 構文解析を用いた抽出システム

構文解析を用いた情報抽出実験システムについて述べる．

システムの実装にあたっては3.4節と同じくRubyを使用した．本システムの概要図を図3に示す．

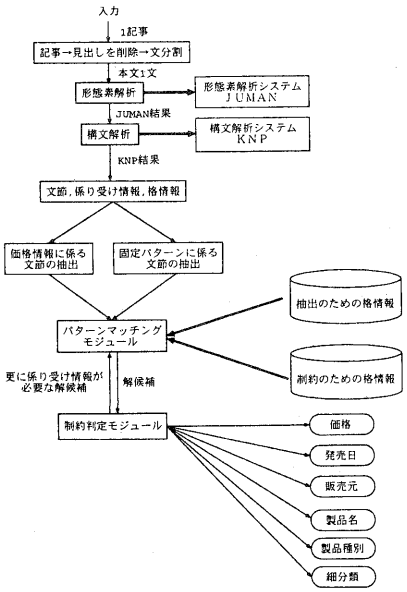

図3：構文解析を用いた情報抽出システム

各モジュールの処理

3.4節および4.1節～4.3節と重複する説明は省略する．

(構文解析モジュール)

JUMANの結果を入力として，KNPを用いて構文解析を行う．出力形式はリスト形式を用いる．

(文節情報作成モジュール)

KNP結果から，パターンマッチングモジュールに必要な，文節，係り受け情報(各文節の係り先)，格情報(ガ格，ヲ格，同格連体など)を抜粋し，文節情報とする．

(パターンマッチングモジュール)

文節情報作成モジュールの結果をそれぞれの固定パターンに係る文節組について格情報ファイルとマッチングを行ない，抽出に成功したすべての情報を制約判定モジュールに渡す．

また制約判定モジュールの結果，更に係り受け解析が必要な場合，制約のための格情報ファイルとマッチングを行ない解を決定する．

(制約判定モジュール)

パターンマッチングモジュールから渡された解候補について4.3節に示したルールを適用し，更に係り受け解析結果が必要であると判断した場合はパターンマッチングモジュールに再度渡す．

5 実験と評価

3.4節のテンプレートを用いた抽出システム，4.4節の構文解析を用いた抽出システムを使用して，新聞記事の複数製品記事からの情報抽出実験を行う．

実験では，毎日新聞1991年から1995年までの複数製品紹介記事1007記事を使用した．1991年1月から1995年6月までの895記事をテンプレート作成用データ，および構文解析における分析用データとし，1995年7月から1995年12月までの112記事を評価用データとした．

まず，2つのシステムを用いた実験の抽

出項目別に評価した結果に対して，それぞれのシステムについての考察を行う．次に両システムの1記事あたりの再現率を比較し，抽出精度を向上させるために構文解析が有効であることを示す．

評価方法として，再現率，適合率を用いる，

$$\text{再現率} = \frac{\text{正しく抽出された情報の数}}{\text{記事に存在する抽出情報の数}}$$

$$\text{適合率} = \frac{\text{正しく抽出された情報の数}}{\text{抽出した情報の数}}$$

5.1 テンプレートを用いた抽出実験

抽出項目	正解	合計	適合率	再現率
販売元	113	116	0.97	0.97
製品種別	17	100	0.22	0.17
製品名	98	140	0.58	0.70
細分類	81	207	0.43	0.39
価格	137	218	1.0	0.63
発売日	63	71	1.0	0.89

表5：テンプレートを用いた抽出結果

「価格」,「発売日」については表記に特徴があるので，1文毎のテンプレートとは別に，表記パターンによるマッチングの処理を加えることで抽出精度を向上させることができた．それ以外の項目では，再現率は,「販売元」,「製品名」以外は抽出そのものに失敗することが多かった．この2つの項目は，直前直後の文字の表記パターンが少ないという理由で抽出が容易であったと考えられる．

多数の表記パターンが予想される「製品種別」,「製品名」,「細分類」の抽出を正しく行うには，動詞の固定パターンの一般化を行ってもテンプレート2433個では不足であると言える．テンプレートを拡充することが重要であるが，処理の高速性が失われることがないよう，固定パターンの部分以外でもテンプレートを一般化する必要がある．

なお，表5のは抽出項目の優先順位が1位についての結果のみを表示している．

5.2 構文解析を用いた抽出実験

抽出項目	正解	合計	適合率	再現率
販売元	100	116	1.00	0.86
製品種別	61	100	0.71	0.61
製品名	79	140	0.73	0.46
細分類	128	207	0.68	0.62
価格	214	218	0.92	1.00
発売日	70	71	0.99	0.99

表6：構文解析を用いた抽出結果

「価格」については他の5項目とは抽出手法が異なっており，4章で示したように他の項目の係り先として表記パターンによる抽出を行っている．

表4-1，4-2で示した格形式のうち「製品種別」「製品名」「細分類」は同じ格が多数存在する．これらについては，抽出自体は成功していても項目を割り当てる時点で失敗しているケースが多く，再現率が低くなった．複数文に存在する項目の出現位置や，前後の係り受け関係による制約を加えることでより正確な割り当てを行うことができると考えている．

抽出失敗の原因としては，並列構造を含むような長い文に対するKNPの解析が正しく行われないことが挙げられる．製品紹介記事に限定した係り受け構造規則を作成してKNPの結果を修正することでより正しい係り受け結果が得られるであろう．その他に，現在の抽出対象は“発売”の表現を含む文章と価格の表記を含む文章に限定しているため，対象とする文の範囲を広げることでより多くの項目が抽出できると考えている．

5.3 テンプレートと構文解析の比較

実験記事112記事について1記事あたりの再現率を表7に示す．括弧内の数値は平均の適合率となっている．

再現率	テンプレート	構文解析
0.80以上	28記事 (0.73)	62記事 (0.80)
0.50〜0.80	39記事 (0.56)	41記事 (0.62)
0.50未満	45記事 (0.31)	9記事 (0.38)

表7：1記事あたりの再現率

この結果から，全体的な抽出精度に関しては構文解析を用いた手法が有効であることが確認できた．

しかし，構文解析を行うことでテンプレートの実験よりも処理時間がかかるという欠点がある．そこでテンプレートで抽出可能な記事に対してはテンプレートを用い，文章構造が複雑で抽出困難な記事に対しては構文解析を適用することで，処理時間をおさえつつ，抽出精度を向上させることができると考えられる．実際にテンプレートで抽出困難な記事が構文解析を用いることで抽出できるかを確認する必要がある．

表8にテンプレートの手法で抽出精度が5割未満であった記事に対して，構文解析の手法を適用した結果を表す．括弧内は表7と同様に平均の適合率を示す．

再現率	構文解析
0.80 以上	15 記事 (0.78)
0.50〜0.80	22 記事 (0.61)
0.50 未満	8 記事 (0.40)

表 8：構文解析を用いた抽出結果
(テンプレートの手法で再現率 0.5 未満の記事)

以上の結果から，テンプレートで再現率が0.50未満であった記事が45記事あったものを構文解析を用いることで8記事に減らすことができ，抽出精度を向上する方法として構文解析が有効であるといえる．

6 まとめ

定型的な文章で書かれていることが多い新聞記事中の複数製品紹介記事を対象として情報抽出を行う2つの手法を提案した．

テンプレートの手法では，複数製品紹介記事の記述パターンの分析をもとにテンプレートを作成して抽出を行った．その結果，複数製品紹介記事にはある程度の定型性があり，その定型性を利用して情報の抽出，項目間の対応付けが行えることを確認した．構文解析の手法では，記事を構文解析した結果から，係り受け関係を用いた抽出を行い，字面だけでは抽出することが困難な複雑な文からの抽出を実現することができた．

今後，テンプレートの手法では抽出精度を更に向上させるためにデータを増やし，より多くのテンプレートを作成する必要がある．しかし，複数製品を紹介した記事は単数の場合に比べて構造が複雑であり，一般性が低いテンプレートを大量に生成されることが予想される．大量のテンプレートとのマッチングには処理時間がかかり，適合率が落ちることになる．そこで重みが大きい，すなわち信頼性が高いテンプレートにマッチするような文は，テンプレートを用いて抽出を行い，それ以外の文に対しては構文解析を用いるというように2つの手法の利点を生かしたシステムの構築することが考えられる．

謝辞

本論文で使用したテキストデータは，「毎日新聞記事データ CD-ROM 版 (1991 年〜1995 年)」を使用した．使用を許可して下さった毎日新聞社に深く感謝致します．

参考文献

[1] 高尾 宜之，永井 秀利，中村 貞吾，野村浩郷：複数製品の紹介記事からの製品情報抽出 -製品記述パターンの分析-，情報処理学会研究報告 99-NL-129, pp. 117-124, 1998

[2] 黒橋禎夫, 長尾真, 日本語形態素解析システム JUMAN 使用説明書 version 3.61, 京都大学大学院工学研究科

[3] 黒橋禎夫，日本語構文解析システム KNP 使用説明書 version 2.0 b6，京都大学大学院工学研究科

[4] まつもとゆきひろ/石塚圭樹 共著，オブジェクト指向スクリプト言語 Ruby :アスキー出版者

[5] 井出 裕二：テンプレートを用いた新聞記事からの製品情報抽出に関する研究, 平成9年度九州工業大学修士論文 (1998)