自然言語処理 129−17
(1999. 1. 21)

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−

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
E-mail: {takao,nagai,teigo,nomura}
@dumbo.ai.kyutech.ac.jp

我々は,抽出すべき情報とその周辺の文字列との関係を記した“テンプレート”を用いて,字面処理による情報抽出処理の研究を行っており,新聞の製品紹介記事を題材とした実験で,テンプレートを用いた抽出処理の有効性を確認してきた.

従来の研究では,1記事中に1個の製品情報が紹介されているとして,テンプレートを作成し,情報抽出を行っていた.しかし,新聞記事中には1記事中に複数の製品を紹介している記事も存在する.この様な記事の場合,従来用いている手法で情報抽出を行っても正しく抽出できなかった.そこで,本論文では,複数の製品が紹介されている記事から情報抽出を行うため製品記述パターンを分析した.その分析をもとに,複数製品情報を扱うためのテンプレートの枠組を提案する.

Information Extraction from Newspaper Articles of Multiple Products

−classification of expression patterns−

Yoshiyuki Takao Hidetoshi Nagai
Teigo Nakamura Hirosato Nomura

Department of Artificial Intelligence, Kyushu Institute of Technology
E-mail: {takao,nagai,teigo,nomura}
@dumbo.ai.kyutech.ac.jp

We have been researching a textual analysis method for information extraction from newspaper articles 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.

In our previous research, it was assumed that information about only one product is described in an article. However, some newspaper articles describe information about multiple products and our system failed to extract correct product information from such articles. To extend our system to deal with multiple products, we explored the articles with multiple products and classified the pattern of expressions that describe multiple information items and correspondence among them.

In this paper, we show the result of our classification and propose the new form of templates which are able to match consecutive sentences and identify the correspondence among information items.

1 はじめに

ネットワークの普及により，膨大かつ多種多様な情報が流通するようになった現在，それらの情報の中から必要な情報だけを人間の手作業で取り出してくることはもはや不可能になりつつある．このような状況下において，計算機可読な文書情報を計算機で自動的に処理し，必要とする情報のみを取り出してくることができれば，氾濫する情報も効率的に管理することができ，情報の価値を落とすことなく適切な利用が可能となる．このように，計算機による情報抽出システムが実現することによって得られるメリットは非常に大きいため，現在盛んに研究されている [5][6][7][8].

筆者らは，大量の計算機可読文書から目的とする情報を高速に抽出する研究を行なっており，これまでに，新聞の新製品紹介記事を対象として，記事の記述形式の定型性を利用した製品情報抽出方式の有効性について報告を行なってきた [1][2][3]．そこで用いている基本的な手法は，抽出すべき項目とその周辺の文字列との関係を記述したテンプレートと入力記事とのマッチングによるものである．

従来の研究では、1記事に複数の製品が紹介されている記事を考慮していなかった．この様な記事は、抽出すべき情報が並列構造により表現され，複数の文にまたがって記述される場合もある．また情報を抽出できたとしても，抽出項目どうしの対応付けが難しく、従来の手法で情報抽出を行っても、正しい情報を抽出できていなかった．

そこで、この様な記事から情報抽出を行うことができるように，記事中の抽出すべき情報に対しタグ付けを行い、それを基に記述パターンを分析した．そして、これを基に複数製品情報に対応したテンプレートの枠組を提案する．

2 テンプレートを用いた情報抽出

2.1 テンプレートの形式

以下では，情報抽出処理のためのテンプレートを定義する．まず，テンプレートの定義のための用語を定める．

抽出項目: 抽出を試みる情報の内容を表すラベル (「製品種別」,「製品名」,「販売元」,「価格」,「発売日」の5項目)
パターンマッチングにおいてはワイルドカードと同様に機能する

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

L を抽出項目，P をパターン，W をワイルドカードとしたとき，テンプレート T を

$$T = C_0 L_1 C_1 L_2 \cdots C_{n-1} L_n C_n$$
$$(C_i = P_0 W_1 P_1 W_2 \cdots P_{m-1} W_m P_m)$$

と定める．なお，文頭の C_0 および文末の C_n は空文字列であってもよい．テンプレートは1文単位で作成し，C と L は必ず交互に現れるものとする．

ワイルドカードを導入することにより，テンプレートをより一般化することができる．テンプレートの一般化の長所は，テンプレート集合の削減につながり，抽出処理に要する時間を短縮することができる点である．

以後，テンプレートを表現する場合，抽出項目は {item} の形式で表す．item は抽出項目名である．ワイルドカードは*で表す．

2.2 1文からの情報抽出

入力文がテンプレートとのパターンマッチングに成功すると，パターンに挟まれた入力文中の文字列が対応する抽出項目の抽出情報として抽出することができる．

図1の例では，テンプレートの中のパターン『は』『の』『「』『」』『を』『に発売する.』が入力文中に存在し，かつ出現順序もテンプレートと同じであるため，4つの抽出項目についての具体的な抽出情報を得ることができた．

2.3 1記事からの情報抽出

テンプレートを用いた1記事からの情報抽出処理は，以下の様な手順で行われる．

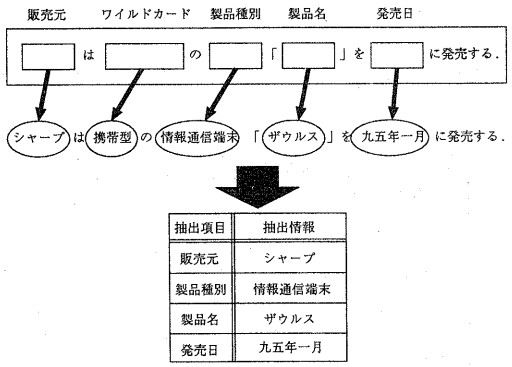

図 1: 1 文からの情報抽出

1. 記事を句点で分割する．
2. 各文に全てのテンプレートをマッチングさせ，マッチした場合は以下の処理を行う．
   (a) 抽出項目に対する制約を用いて，テンプレートとマッチングで得られた文字列をチェックする [3].
   (b) 1 つのテンプレートから得られた全ての文字列が制約チェックを通ったならば，それぞれの抽出項目の解候補に加える．
3. 各抽出項目の解候補に対して優先順位付けを行い，それぞれ抽出情報を決定する．

図 2 に処理の全体の流れを示す．従来，優先順位付けは各候補文字列に対して，その文字列を抽出したテンプレートの個数をスコアとし，そのスコアが最も高いものを抽出情報と決定してきた．

3 記事中の複数製品情報

従来の研究では 1 記事中に 1 つの製品が紹介されている記事を対象に情報抽出を行っていた．本論文では複数の製品紹介記事を対象とするため，抽出情報とその対応関係を記述可能な製品情報の記述形式が必要である．そこで，製品情報の記述形式を定義する．抽出を試みる内容「抽出項目」は以下の 6 種類である．

(「販売元」,「製品種別」,「製品名」,「製品の細分類」,「価格」,「発売日」)

従来の研究では,「製品の細分類」を除く 5

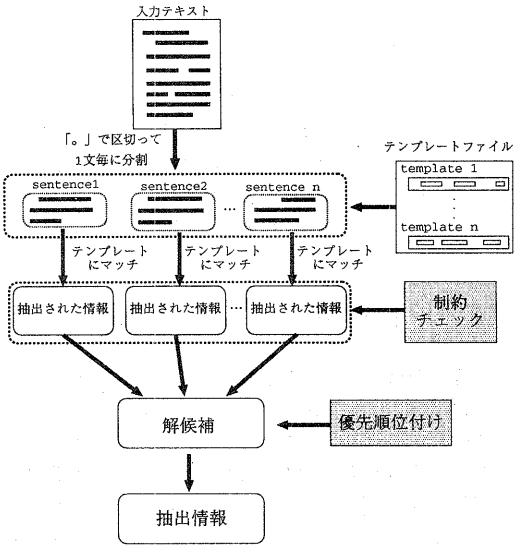

図 2: 1 記事からの情報抽出

種類を抽出項目としていたが，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．次に，1 つの製品情報 i を

$$i =< C, K, N, S, D, P >$$

の 6 つ組と定義する．ただし,C:販売元,K:製品種別,N:製品名,S:製品の細分類,D:価格,P:発売日である．各項目の値は抽出情報であり，抽出情報には「千～二千円」などの範囲値で表現されるものを含んでいる．また，ϕ は値が存在しないことを示す．

さらに，1 記事中の製品情報 I は個々の製品情報の集合とする．

$$I = \{i_1, i_2, \cdots, i_n\}$$

例えば，次の記事の製品情報は以下のように表せる．

◇清涼飲料◇サッポロビール（０３・３５７２・６１１１）の「サッポロ　ビタミン＆レモン」と「サッポロ　ジャンボグルト」＝写真。(中略)《価格・発売時期》どちらも百七円。十七日。

$$
\begin{aligned}
I \;=\; & \{< c_1, k_1, n_1, \phi, p_1, d_1 >, \\
& < c_1, k_1, n_2, \phi, p_1, d_1 > | \\
& c_1 = \text{サッポロビール}, k_1 = \text{清涼飲料}, \\
& n_1 = \text{サッポロ　ビタミン\&レモン}, \\
& n_2 = \text{サッポロ　ジャンボグルト}, \\
& p_1 = \text{百七円}, d_1 = \text{十七日}\}
\end{aligned}
$$

4　複数製品紹介記事の分析

複数製品紹介記事から情報抽出を行うことができるようにするため，このような記事に対しタグ付けを行い記述パターンを分析した．分析には，日本経済新聞 1994 年版の中の製品紹介記事 2005 記事のうち，複数の製品が紹介されている 453 記事を使用した．表 1 に分析記事中の文数と文字数を示す．

	1 記事中の平均	453 記事の合計
文数	6.6 文	3014 文
文字数	293.8 文字	133127 文字

表 1: 分析記事中の文数と文字数

ただし，以下のように意味的には複数の製品を紹介しているが $I = \{i\}$ である場合は複数製品紹介記事ではないと考え，今回の分析の対象外とした．

土谷特殊農機具製作所（帯広市，土谷紀明社長）は今年度から，廃油を燃料に使うストーブを本格的に発売する。(中略)。約百七十平方メートルから五百平方メートルの事務所用で、価格は工事費別で七十万―百七十万円。(以下略)

$$
\begin{aligned}
I \;=\; & \{< c_1, k_1, \phi, \phi, p_1 > | \\
& c_1 = \text{土谷特殊農機具製作所}, \\
& k_1 = \text{ストーブ}, \\
& p_1 = \text{七十万―百七十万円}, \\
& d_1 = \text{今年度})\}
\end{aligned}
$$

4.1　記事のタグ付け

分析対象の 453 記事に，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．表 2 に図中のタグの役割を示す．タグは ⟨*item#*⟩ と ⟨/*item#*⟩ 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。

ウエアの上から装着可能◇⟨*k1*⟩時計⟨/*k1*⟩◇⟨*c1*⟩服部セイコー⟨/*c1*⟩（０３・３５６３・２１１１）の「⟨*n1*⟩スパイアー・ジーコンパクト型ミラー付きウオッチ⟨/*n1*⟩」「⟨*n2*⟩同提げ時計⟨/*n2*⟩」＝写真　《ポイント》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・発売時期》⟨*p1*⟩６０００円⟨/*p1*⟩。⟨*d1*⟩２８日⟨/*d1*⟩。

⟨(*c1*, *k1*, *n1*, *φ*, *p1*, *d1*), (*c1*, *k1*, *n2*, *φ*, *p1*, *d1*)⟩

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 2: テンプレート作成用記事中のタグ

4.2　1 記事中の抽出項目の出現パターン

タグ付けされた記事から，1 記事中でのタグを付けた抽出情報の出現頻度をもとに複数製品紹介記事を分類した．

抽出項目のうち，「販売元」は 453 記事全てに 1 つ以上存在し，複数個存在しても同じ製品を共同で開発・販売するというものが多い．また，「発売日」は記述されない場合が多く，記述されても対応付けが容易である．そのため，分類の評価項目は「製品種別」，「製品名」，「製品の細分類」，「価格」の 4 つの種別とした．ここで，上記の 4 つの各項目の属性値を

＋：任意
ϕ：存在しない
単：1 つだけ存在する
複：2 つ以上存在する

の 4 つとする．表 3 の出現パターンの各属性値は，(販売元，製品種別，製品名，製品の細分類，価格，発売日) に対応している．例えば，出現パターンが (+,ϕ, 単, 単, 複,+) であれば，販売元と発売日が任意で，製品種別がなく，製品名と製

品の細分類が1つずつ,価格が複数個であることを示している.

表3に記事の分類とその頻度を示す.表3より,出現のパターンはかなりのばらつきが見られたが,出現パターン 8,9,11,12 の頻度が高いことがわかる.

出現タイプ	出現パターン	頻度
1	(+, ϕ, 単, 単, 複,+)	1 記事
2	(+, ϕ, 単, 複, ϕ,+)	1 記事
3	(+, ϕ, 単, 複, 複,+)	13 記事
4	(+, 単, ϕ, 複, 単,+)	8 記事
5	(+, 単, ϕ, 複, 複,+)	18 記事
6	(+, 単, 単, ϕ, 複,+)	5 記事
7	(+, 単, 単, 複, ϕ,+)	5 記事
8	(+, 単, 単, 複, 単,+)	54 記事
9	(+, 単, 単, 複, 複,+)	137 記事
10	(+, 単, 複, ϕ, ϕ,+)	3 記事
11	(+, 単, 複, ϕ, 単,+)	74 記事
12	(+, 単, 複, ϕ, 複,+)	96 記事
13	(+, 単, 複, 単, 単,+)	1 記事
14	(+, 単, 複, 複, ϕ,+)	2 記事
15	(+, 単, 複, 複, 単,+)	2 記事
16	(+, 単, 複, 複, 複,+)	9 記事
17	(+, 複, ϕ, ϕ, 複,+)	2 記事
18	(+, 複, 単, ϕ, 複,+)	4 記事
19	(+, 複, 複, ϕ, ϕ,+)	2 記事
20	(+, 複, 複, ϕ, 単,+)	2 記事
21	(+, 複, 複, ϕ, 複,+)	13 記事
22	(+, 複, 複, 複, 複,+)	1 記事

表 3: 1 記事中の出現パターンの分類

4.3 抽出項目の階層関係

次に,タグ付けされた抽出情報の対応づけを行って製品情報 I を作成した.抽出情報の対応付けを行う際,「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3のような階層を設定した.

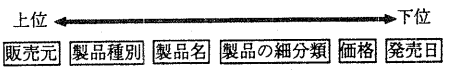

図 3: 抽出項目の階層関係

4.4 抽出項目どうしの対応関係

1記事中の出現パターンが同じでも,抽出項目の抽出情報どうしの対応関係が異なる場合がある.そこでタグ付けされた抽出情報の対応づけを行って,製品情報 I を作成した.項目間の対応は以下のように分類される.

○抽出項目の抽出情報の個数が 1 対 n($n \geq 2$)

上位の抽出項目の抽出情報が1つで,下位の抽出項目の抽出情報が複数の場合.

- タイプ A

(例 1) 県産ブドウのワイン〈島　根〉県経済農業協同組合連合会（ＪＡ島根経済連）の島根ワイナリー（大社町,ＴＥＬ０８５３・５３・５５７７）は,島根県産ブドウを原料にした低アルコールの新ワイン＝写真＝を発売した.甲州種とデラウエア種を使った白の「ディオブラン」と,マスカットベリーＡ種を使ったロゼの「ディオロゼ」の２種.

$$I_1 = \{< c_1, k_1, n_1, \phi, \phi, \phi >, < c_1, k_1, n_2, \phi, \phi, \phi > | c_1 = \text{島根ワイナリー}, k_1 = \text{ワイン}, n_1 = \text{ディオブラン}, n_2 = \text{ディオロゼ}\}$$

○抽出項目の抽出情報の個数が n 対 1($n \geq 2$)

上位の抽出項目の抽出情報が複数で,下位の抽出項目の抽出情報が1つの場合.

- タイプ B

1つの抽出項目が他の抽出項目の複数の抽出情報全てに対応

(例 2) 甲州種とデラウエア種を使った白の「ディオブラン」と, マスカットベリーA種を使ったロゼの「ディオロゼ」の2種. 希望小売価格は７２０ミリリットル入りで１３００円.

I_2 = {< $\phi, \phi, n_1, s_1, p_1, \phi$ >,
< $\phi, \phi, n_2, s_1, p_1, \phi$ > |
n_1 = ディオブラン, n_2 = ディオロゼ,
s_1 = ７２０ミリリットル入り,
p_1 = １３００円}

- タイプC
対応が取れていない場合

(例 3) 積水ハウス北陸は多雪地域向けのアパートの新シリーズ「フレグランス・セフィラ」と「ディアス・セフィラ」を発売した.(中略) 価格は「フレグランス・セフィラ」が三・三平方メートル当たり四十二万円からとなっている.

I_3 = {< $c_1, k_1, n_1, \phi, p_1, \phi$ >,
< $c_1, k_1, n_2, \phi, \phi, \phi$ > |
c_1 = 積水ハウス北陸, k_1 = アパート,
n_1 = フレグランス・セフィラ,
n_2 = ディアス・セフィラ,
p_1 = 四十二万円から }

例2では複数の製品名が存在しているが, それぞれが1つの価格に対応している. 一方, 例3では, 複数の製品名があるにもかかわらず, 片方の製品名の価格だけが記載され, 他方の金額は記載がない. 出現パターン8と11には, 上記のタイプBとタイプCのような違いが存在した. 内訳を表4に示す.

出現タイプ	タイプB	タイプC	計
8	51 記事	3 記事	54 記事
11	67 記事	7 記事	74 記事

表 4: タイプBとタイプCの内訳

○抽出項目の抽出情報の個数が m 対 n($m, n \geq 2$)

抽出項目どうしの抽出情報の個数が複数対複数の場合, それぞれが1対1に対応する場合とそうでない場合が存在する. 例4の場合は「本醸造酒」が「二千円」と「吟醸酒」が「千五百円」と1対1に対応しているが, 例5の場合は「セミソフト」,「ソフト」,「スタンダード」の3種類の製品があり価格帯が「千九百円」と「二千九百円」の2種類がある. つまり, この記事には6種類の製品が紹介されていることになる.

- タイプD
1対1の対応がとれている場合

(例 4) あなたの町だけの祝い酒を承ります——. 浜松地区の有力酒造会社, 花の舞酒造（浜北市, 高田和夫社長）は五月の浜松まつりを記念して町名入り記念ボトル＝写真＝を発売した.(中略) 発売したのは千八百ミリリットルの本醸造酒（二千円）と七百二十ミリリットルの吟醸酒（千五百円）.(以下略)

I_4 = {< $c_1, k_1, \phi, s_1, p_1, \phi$ >,
< $c_1, k_1, \phi, s_2, p_2, \phi$ > |
c_1 = 花の舞酒造, k_1 = 記念ボトル,
s_1 = 本醸造酒, s_2 = 吟醸酒,
p_1 = 二千円, p_2 = 千五百円 }

- タイプE
1対1の対応ではない場合

(例 5) 西友は十五日から, アジア・太平洋地域の十一社が参加するアジア小売連合（ＡＲＡＮ）で共同開発したワイシャツを全国の百四十店舗で販売する.(中略) ワイシャツは西友の技術指導で防菌防臭加工を施しており, セミソフト, ソフト, スタンダードの三種類をそろえた. 価格は千九百円と二千九百円の二種類.(以下略)

I_5 = {< $c_1, k_1, \phi, \{s_1, s_2, s_3\}, p_1, \phi$ >,
< $c_1, k_1, \phi, \{s_1, s_2, s_3\}, p_2, \phi$ > |
c_1 = 西友, k_1 = ワイシャツ,
s_1 = セミソフト, s_2 = ソフト,
s_3 = スタンダード
p_1 = 千九百円, p_2 = 二千九百円 }

分析対象の453記事のうち, 2記事だけタイプEの記述が存在した. 内訳は出現パターン5の18記事のうち, タイプDが17記事, タイプEが1記事である. また, 出現パターン15の1記事は, タイプEである.

4.5 複数製品情報の表現形式

同じ出現パターンにおいても，対応関係の表現形式が異なる場合が存在する。分析対象の記事では，1文中で複数個の同じ抽出項目の抽出情報が記述される場合，出現パターンとして次の2つの記述方法が確認された。

- **対応型** 対応する抽出項目どうしが明示されている場合

> 二百四十五グラムと三百四十グラムの缶入りが百七円，一リットルのＰＥＴボトル入りが二百三十円.

- **列挙型** 同一の抽出項目の異なる抽出情報が列挙して記述されている場合

> 七リットルと十五リットルの二タイプで，価格は三千六百七十五円，七千八百七十五円.

抽出項目の抽出情報の個数がn対1($n \geq 2$)の下のような場合も列挙型と捉えることができる.

> 二百五十，五百グラム入りがあり，各二百七十円.

列挙型の記述の場合，複数文にまたがって抽出項目どうしの関係が対応付けられていることがある.しかし，複数文にまたがっての対応型の記述形式は見られなかった.

4.6 表現パターンと対応タイプの関係

ここで,4.4節で記述した対応関係に基づいて，表現パターンと対応タイプの関係を示す.

○抽出項目の抽出情報の個数が1対n($n \geq 2$)

必ず列挙型の記述形式で，対応タイプはタイプAだけである.

○抽出項目の抽出情報の個数がn対1($n \geq 2$)

列挙型の記述形式であれば，タイプBである.ほとんどの場合n対1の対応であることを示す標識(「ともに」，「すべて」)等が文中に存在する.対応型での記述であれば，タイプCである.

○抽出項目の抽出情報の個数がm対n($m, n \geq 2$)

m,nともに列挙型で$m \neq n$であればタイプEである.それ以外の場合はタイプDかタイプEの判断はできない。ただし，対応関係が$m \neq n$で列挙型の記述形式は分析対象文にはなかった.

5 新しいテンプレートの枠組

複数製品情報を抽出するためのテンプレートの枠組を提案する.

1文中に複数の抽出項目が出現するものに関しては，従来より用いてきたテンプレートを使用する.ただし，そのテンプレートに抽出項目どうしの対応関係を付与しておく.これにより1文内での抽出情報の対応関係が決定できる.

また，複数製品情報の記述において，同じ抽出項目を複数文にまたがって記述していることが多い.そこで，隣接2文にマッチングできるようにテンプレートを拡張する.

5.1 テンプレートの作成

以下のように,テンプレート作成用の文章から，正解データと固定パターンを用いてテンプレートを作成する.正解データには抽出項目どうしの対応関係が記載されているものとする.テンプレートを作成する手順は従来どおりであるが，作成されるテンプレートには抽出情報どうしの対応関係も付与される.

文章 商品名は「まろやかピーチ」「まろやかいちご」。《価格・発売時期》ともに百十円。

正解データ 製品名1：まろやかピーチ，製品名2：まろやかいちご，価格1：百十円(⟨{製品名1}，{価格1}⟩，⟨{製品名2}，{価格1}⟩)

固定パターン 《価格・発売時期》，ともに

テンプレート *は「{ 製品名1}」「{ 製品名2}」。《価格・発売時期》ともに、{価格1}。(⟨{製品名1}，{価格1}⟩，⟨{製品名2}，{価格1}⟩)

5.2 抽出項目の対応付け

5.1節で作成されたテンプレートを用いて情報抽出処理を行う．まず，テンプレートと記事のマッチングで得られる抽出情報およびその対応関係を解候補とする．

次に，解候補に優先順位をつけて抽出項目に対する抽出情報を決定する．優先順位付けは，抽出項目の階層が上位のものから行い，対応関係のついていない抽出項目どうしの対応関係を考慮した優先順位付けを行う．

優先順位付けの際の制約として，以下のようなものを考えている．

- 「販売元」は普通1文目に存在し，「販売元」よりも前に出現できる抽出項目は「製品種別」だけである．
- 最終的な抽出情報が決定した抽出項目に着目し，その抽出情報の(部分)文字列を含むような解候補が存在すれば，それ以外の解候補を排除する．

最終的な抽出情報が決定した抽出項目どうしは，4.6節で示した表現パターンと対応タイプを基にして対応関係を決定する．

6 まとめ

1記事中に複数の製品が紹介されている新聞記事にタグ付けを行い、それを基に記述パターンを分析した．タグ付けの際、製品名が同じでもバージョンや販売方法の違いなどにより、複数の製品情報を記述している記事に対応するため、新しく「製品の細分類」という抽出項目を増やした．

また、1記事中での抽出項目別の出現のパターンと抽出項目どうしの対応関係、その対応関係を示す表現形式の3つの観点から記事の分析を行った．

さらに、その分析を基にして、複数製品情報から情報抽出を行うため、対応関係を考慮した新しいテンプレートの枠組を提案した．今後，この枠組でテンプレートを作成し，情報抽出実験を行う予定である．

謝辞

本論文で使用したテキストデータは，「日本経済新聞記事データ CD-ROM(1994版)」を使用した．使用を許可して下さった日本経済新聞社，および日経総合販売(株)に深く感謝致します．

参考文献

[1] 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, 情報処理学会研究報告 96-NL-115, pp. 83 – 90, 1996

[2] 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, 情報処理学会研究報告 97-NL-118, pp. 7 – 14, 1997

[3] 井出 裕二, 永井 秀利, 中村貞吾, 野村浩郷：単一項目テンプレートによる新聞記事からの製品情報抽出, 情報処理学会研究報告 97-NL-122, pp. 63 – 70, 1997

[4] 柴田 和査：単一項目テンプレートによる新聞記事からの製品情報抽出, 平成8年度九州工業大学卒業論文 (1997)

[5] 江里口 善生, 木谷 強：冨田一般化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

[6] 江里口 善生, 木谷 強：冨田一般化LRパーザを用いた情報抽出, 情報処理学会研究報告 94-NL-102, pp. 9 – 16, 1994

[7] 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

[8] 河合 敦夫, 塚本 雄之, 椎野 努：電子メール文書からの関係情報の自動抽出, 情報処理学会研究報告 94-NL-101, pp. 57 – 64, 1994

[9] 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)