

## 新製品ニュース管理システムのための文と段落のタイプ抽出方法

徳永 秀和 青江 順一

高松高専制御情報工学科 徳島大学工学部知能情報工学科

E-mail : [tokunaga@takamatsu-nct.ac.jp](mailto:tokunaga@takamatsu-nct.ac.jp) , [aoe@is.tokushima-u.ac.jp](mailto:aoe@is.tokushima-u.ac.jp)

現在、インターネット上に多種多彩な新製品のニュースが流通しているので、これらのニュースを有効に活用するシステムの開発は有用である。そこで、我々はXMLを応用してニュースを読みやすくかつ再利用しやすくするシステムを開発している。このシステムのために、まずニュース文の構造を自動的に抽出する技術を開発することにした。本論文ではニュース文を意味タイプに分類する方法について述べる。未知語の名詞が多く存在するというニュース文の特徴を考慮し、述部の解析によりかなり高い精度で文タイプを抽出できる方法を開発した。

## Type Extraction Method of The Statement and Paragraph for The New Product News Management Systems

Hidekazu TOKUNAGA and Jun-ichi AOE

Dept. of Electro-Mechanical Systems Engineering, Takamatsu National College of  
Technology

Dept. of Information Science & Intelligent Systems, University of Tokushima

At present, the news of enormous and various kinds sundry new product is circulating on the Internet. The system that utilizes these news stories effectively is useful. Thereupon, we are developing the system that applied XML that makes it easy to read and easy to manage news. For this system, We develop the technology that extracts the structure of the news statement automatically. We discuss about the method that the news statement is classified into the meaning type in this paper. We developed the method that is able to extract the statement type with fairly high precision by the analysis of the predicate, in consideration of the characteristic of the news statements that many unknown nouns exist.

## 1. はじめに

インターネットの普及により、膨大かつ多種多様な技術情報や新製品情報が流通するようになった。例えば、出版社やプロバイダのホームページには、毎日多くのニュースが掲示されている。さらに、様々な無料や有料のニュースのメール配信サービスが存在している。これらのニュースを全て読み、必要な情報を蓄積し管理する事は、非常に困難な状況になっている。そこで、最近これらの情報から必要な情報のみを抽出する研究が色々行われている。大別すると、多量の記事から興味のある記事だけを取り出す、情報検索やフィルタリング技術[1][2]。もう一つは、要約文の生成や必要情報の抽出などの情報抽出である[3][4]。しかし、これらの技術だけでは、ニュースを有効に活用することはできない。例えば、記事単位でフィルタリングされた場合、全体的には興味の無い記事であるが、ある章だけは必要であるときに対応できない。また、記事からの情報抽出の研究は、製品名や発売日などの細かな情報抽出が中心であり、製品の機能や構成をわかりやすく、読者に提示するような研究が行われていない。

人間は、文章の構造がわかっているならば、必要な情報を斜め読みで抽出することができる。したがって、我々は、XMLを利用することによって、記事の構造を読者にわかりやすく提示できるシステムを開発することにした[5]。また、記事の構造が記述できていれば、記事の中から必要な部分のみを抽出し、他の記事と結合することも可能となる。よって、記事単位でのフィルタリングの問題が解決される。しかし、現在文構造のタグ付けをされたニュースは存在しておらず、人手で新たにタグ付けを行うことは非常に時間がかかる作業となる。

そこで、自然言語処理によりニュース文の構造を明らかにするタグを自動で付ける必要がある。日本語文章の構造解析の研究は、文章の論理構造を抽出するものがいくつか行われているが[6][7]、新製品紹介のニュース文に的を絞ると抽出したい内容が少し異なる。我々は記事構造を抽出する最初のステップとして、ニュースに記述された文を

タイプ分けすることにした。そして、文タイプを利用し、段落の意味付けを行う。本論文では、ニュース中の文タイプの分類とその認識手法について述べる。技術や新製品のニュースといっても様々なものが存在するが、今回は、インターネット関連の記事に絞って解析した。詳細は2章で述べるが、文タイプは10種類とした。また、文タイプ抽出の手法は、未知語が多いニュース文の特徴より、名詞ではなく述部のサ変名詞や動詞、そして助述表現を中心に考案した。抽出ルールの詳細は、3章に述べる。抽出の精度は、ほぼ抽出率が80%以上、正解率が70%以上となった。

## 2. ニュース中の文タイプ分類

解析の対象としたニュース文は、各種雑誌のホームページより、インターネット関連の新技术と新製品の記事を40個収集した。文タイプの決定は、次のような方法で行った。まず、ニュース文を読み、主語としてどのようなものが出てくるか。そして、各主語に対して、どのようなことが述べられているかを解析する。次に、ニュースの読み手は、どのような事を知りたいかを検討する。最後に、この2つの解析結果を結合して、抽出する文タイプを決定した。

### (1) ニュース文の解析

ニュース文を解析した結果、図1に示すような、主語と主語についての事柄が抽出された。

### (2) 読み手が何を知りたいか

新製品ニュースの読み手は、全て主語が新製品であると考えられる情報を欲しがらるものである。ニュース文では、主語がメーカーであっても、読み手は、主語を製品に置き換え受動文として認識する。このような考えに基づいて、図2のような抽出項目を考えた。

### (3) 上記2つの解析結果の結合

ニュース文の解析結果の分類から、読み手が何を知りたいかの分類への写像を考えながら、読み手が何を知りたいかの分類を再構築する方法で文タイプを考案した。最終的な文タイプと、各タイプに含まれた文の数を図3に示す。

主語	メーカー	利用者	システム	著者	世間
主語の動作など	開発 販売 製作 目標	操作 利用 応用	動作 構造	評価	状況

図1. ニュース文の解析結果

製品の入手方法
製品の利用方法
製品の動作
製品の構造
製品の評価

図2. 読み手は何を知りたいか

番号	文タイプの意味	文数
1	メーカーの開発・販売状況 (現在)	78
2	メーカーの開発・販売状況 (将来)	30
3	何ができるか	130
4	どのような分野・応用を狙ったものか	59
5	どのようなものか	54
6	どのようにできているか	58
7	どのように操作・動作するか	118
8	仕様 (価格、動作条件)	23
9	評価	84
10	状況、その他	96

図3. 文タイプ

### 3. 文タイプの抽出方法

本章では、まずニュース文の特徴を述べる。そして、その特徴より文タイプ抽出のために述部の解析を中心とした基本方針を示す。その後、形態素解析を用いて意味符号を生成し文タイプ抽出を行うシステムの構成を示す。また述部処理において助述表現の処理が重要となることより、助述表現の意味抽出方法についても説明する。最後に意味符号を用いた文タイプ抽出のルールを示す。

#### 3-1. ニュース文の特徴と基本方針

ニュース文には、以下のような特徴がある。

(1) 未知語が多い。新製品のニュースであるので、製品名、技術名、メーカー名など常に未知の名

詞が多く現れる。

(2) 多くの文で、主語が省略されている。

(3) 述部にサ変名詞が多く使われている。

(4) 同じ助述表現が多く使われている。

このような特徴より、分類の手がかりとして、名詞の意味情報を使用することを行わず、以下のような情報を利用する。

(1) 述部に出現するサ変名詞の意味、動詞の意味、形容詞の意味、名詞の意味。

(2) 可能や断定などの助述表現。

(3) 動詞と述部直前の格の名詞

(4) 時制や値段など特殊な手がかりとなる名詞や副詞など

これらの意味情報を基に、分類のルールを作成する。

#### 3-2. 文タイプ抽出システム

サ変名詞、動詞、形容詞、副詞、一般的な名詞の意味情報や助述の意味そして手がかり語を抽出するために、青江研究室で開発した形態素解析システム Yuki を利用する。Yuki は図4に示すような辞書を作成することによって、図5のような形態素解析結果をテキストファイルとして出力してくれる。したがって、サ変名詞や動詞の各品詞の意味符号のみでなく、複合語や助述の組み合わせの意味符号を得ることができる。また、述部の抽出は、格助詞と副助詞に区切りの意味符号を付けることで行った。ただし、「のである」のように格助詞や副助詞を含む助述表現については、その格助詞や副助詞は述部の区切りとはならない。この Yuki の出力した意味符号をルールに基づいて分類処理するプログラムによって文のタイプ抽出を行う。処理の流れを図6に示す。

使	つか	0	ア行五段/語幹	v0006006060
でき	でき	0	一段/語幹	j0000010000000000
でき	でき	0	格助詞「で」+一段/未然	j0000010000000000
る	る	0	一段/終止	j0000000000000000
よう	よう	0	助動詞(よう)/連体	j0000000000400000
にな	にな	0	格助詞「に」+補助動詞ラ行/語幹2	j0000000000300000
てい	てい	0	接 終助詞「て」+補助動詞一段/語幹1	j0000000000000200
てい	てい	0	接 終助詞「て」+補助動詞一段/連用1	j0000000000000200

図4. Yuki 用意味辞書

インタフェース	普通名詞	いんたふえーす
を	格助詞「を」	を(z)
使	アワ行五段/語幹	つか(v0006006060)
っ	アワ行五段/連用2「っ」	っ
て	接 終助詞「て」	て(z)
データベース	普通名詞	でーたべーす
と	格助詞「と」	と(z)
データ	普通名詞	でーた
の	格助詞「連体の」	の(z)
やり取り	普通名詞	やりとり
も	副助詞「も」	も(z)
でき	複合語	でき(j0000010000000000)
る	ラ行五段/終止	る(j0000000000000000)
よう	助動詞(よう)/連体	よう
にな	複合語	にな(j0000000000300000)
っ	アワ行五段/連用2「っ」	っ
てい	複合語	てい(j00000000000000200)
る	一段/終止	る(j0000000000000000)
。	句点	まる

図5. Yuki の解析結果の出力

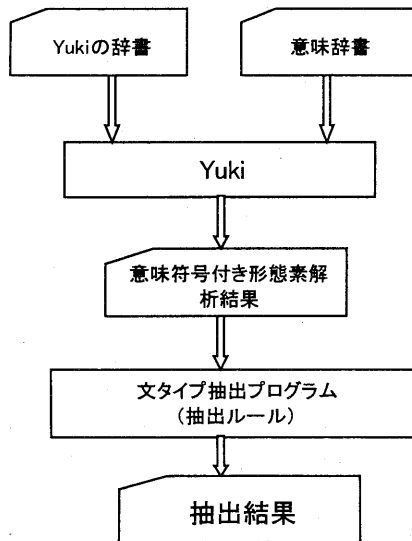


図6. 文タイプ抽出システム

### 3-3. 助述表現の意味抽出

助述表現の意味は、談話理解や文章構造解析の研究においていくつか提案されているが[8][9]、今回の研究のために十分有効なものがなかったため、独自に助述の意味を定義し、意味抽出用の辞書を開発した。定義した助述の意味を図7に示す。助述の意味のグループは、一文の述部に同時には出現しないと思われるもの、そして今回の文タイプ分類に適用しやすいようにグループ化した。助述

表現は「できるようになる」などのように可能、様態、状態変化、現在というふうには、複数の助述の意味が1文中に出現する。したがって、同時には出現しない助述の意味をグループ化し、グループ数と同じ桁の意味符号を割り振ることにより、助述意味符号を単純に加算することによって文の助述全体の意味を判断することができる。例えば、図5の文における助述意味符号はz以下のj符号を加算し、j00000010000300203となる。

### 3-4. 文タイプ抽出ルール

文タイプ抽出ルールは、大別すると以下の4つのパターンとして理解できる。

(パターン1) 述部中のサ変名詞、動詞、形容詞、名詞の意味とほぼ自由な助述表現の組によって抽出できる。

(パターン2) 述部中の動詞、サ変名詞、形容詞、助述表現と述部直前の格の名詞との共起によって抽出できる。

(パターン3) ほとんど助述表現のみによって抽出できる。

(パターン4) その他特殊な単語が手がかりとなり抽出できる。

そこで、サ変名詞、動詞、形容詞、名詞に対しては、以下のような意味符号を割り当てた。

(1) 意味符号の桁数は、文タイプの数と同じ10桁とし、文タイプの抽出に寄与する場合に、そ

桁 値	1	2	3	4	5	6	7
16	要望	願望					
15	意見	見解	断定	言明	定義		
14	推量	伝聞	様態				
13	疑問	呼びかけ					
12	意図	決定					
11	可能						
10	義務	限定	必要性	困難	不都合事態	程度	否定
9	義務否定	限定否定	不必要				
8	使役	使役行為	受動	恩恵受			
7	行動	行動状態	試行	準備			
6	状況説明	状況否定	状態変化				
5	存在	存在状態	不在				
4	家庭	理由	説明				
3	開始	継続	完了				
2	現在 (0)	過去					
1	出来事化	もの化	状態化				

図7. 助述の意味

の文タイプに対応する桁に適当な数字を割り当てる。

(2) サ変名詞はs、動詞はv、形容詞はk、名詞はmの記号を10桁の値の前に付加する。

(3) パターン1で抽出できるものには、1の値を割り当てる。

(4) パターン2で抽出できるものには、6の値を割り当てる。

(5) パターン4の単語に対しては、単独で抽出できるものは1、共起を伴うものは6の値を割り振り、先頭の記号はtとする。

例えば、サ変名詞「販売」は、これのみで文タイプ「メーカーの開発・販売状況（現在）」と判断できるのでs1000000000の符号を振る。また、動詞「備え」は「どのような分野・応用を狙ったものか」と「どのようにできているか」の両方の文タイプに出現し、直前の格の名詞が「特徴」なら「どのような分野・応用を狙ったものか」タイプとなり「ソフト」なら「どのようにできているか」タイプとなる。したがって、v0006060000の符号を動詞「備え」に振る。また、名詞「特徴」は共起のみでなく、述部に単独で現れても「どのような分野・応用を狙ったものか」タイプと認識できるのでm0001000000の符号となる。一方「ソフト」は共起によってのみ「どのようにできているか」タイプと認識される

のでm0000060000の符号となる。

次にルールの詳細を示す。まず文より抽出する意味情報に以下のような記号を割り当てる。

s1：述部のサ変名詞の意味

v1：述部の動詞の意味

m1：述部の名詞の意味

k1：述部の形容詞の意味

s2：述部直前格のサ変名詞の意味

m2：述部直前格の名詞の意味

j：述部の助述表現の意味

t：全文中にある特殊文字の意味

以上の記号は該当する意味を持った単語が存在しない場合に0となるものとする。

さらに、以下のような情報に記号を割り当てる。

m0：述部に普通名詞がある時“真”、ないとき“偽”となる。

次に、意味符号の各桁のマッチング規則を定義する。

\*：どのような値でもマッチする。

?：0以外の値のときマッチする。

数値：同じ数値の時マッチする。

以上の記号と、意味符号のマッチング規則を用いて各文タイプの抽出ルールを以下に示す。なお＝は同値、&&は論理積、||は論理和、!は否定を表す。また、[ ]が一つの抽出ルールである。

(1) 文タイプ1  
 {パターン1} [(s1=1\*\*\*\*\* && !m0) && !(t=0600000000 && !j=\*\*\*\*\*2\*)], [(v1=1\*\*\*\*\* || m1=1\*\*\*\*\*) && !(t=0600000000 && !j=\*\*\*\*\*2\*)]  
 {パターン2} [(s1=6\*\*\*\*\* || v1=6\*\*\*\*\* && m2=6\*\*\*\*\*), [(s1=0 && v1=0 && k1=0) && m2=1\*\*\*\*\*]

(2) 文タイプ2  
 {パターン1} [s1=\*1\*\*\*\*\* || m1=\*1\*\*\*\*\*]  
 {パターン4} [(t=0600000000 && !j=\*\*\*\*\*2\*) && (s1=?\*\*\*\*\* || v1=?\*\*\*\*\*)]

(3) 文タイプ3  
 {パターン2} [v1==\*6\*\*\*\*\* && m2==\*?\*\*\*\*\*], [(s1=0 && v1=0 && k1=0) && m2==\*1\*\*\*\*\*]  
 {パターン3} [j=0\*\*001000\*\*00\*\*0 && (s1=0 && v1=0 && m1=0)], [j=0\*\*001000\*\*00\*\*0 && (s1=\*\*\*\*\*?\*\*\* || v1=\*\*\*\*\*?\*\*\* \* || m1=\*\*\*\*\*?\*\*\*)], [j=0\*\*001000\*\*00\*\*0]

(4) 文タイプ4  
 {パターン1} [s1==\*1\*\*\*\*\* && m1=0], [v1==\*1\*\*\*\*\* || m1==\*1\*\*\*\*\*]  
 {パターン2} [(s1==\*6\*\*\*\*\* || v1==\*6\*\*\*\*\* || k1==\*6\*\*\*\*\* && m2==\*?\*\*\*\*\*), [(s1=0 && v1=0 && k1=0) && m2==\*1\*\*\*\*\*]  
 {パターン4} [t=0001000000]

(5) 文タイプ5  
 {パターン3} [j=050000\*0\*\*0000\*\* || j=040000\*0\*\*0000\*\*) && (s1=0 && v1=0 && k1=0 && m1=0)], [(j=050000\*0\*\*0000\*\* || j=040000\*0\*\*0000\*\*) && (!s1=0 && !j=\*\*\*\*\*2\*)], [j= 0\*000\*00\*\*\*00\*\*2 && m1=0]

{パターン4} [文末が普通名詞 (意味符号がついていない)], [文末が、普通名詞 (意味符号がついていない) と j=0300000000000000]

(6) 文タイプ6  
 {パターン1} [s1=\*\*\*\*\*1\*\*\*\* && !m0 && !j=\*\*\*\*\*1\*\*\*\*\* && !m2==\*1\*\*\*\*\*], [v1=\*\*\*\*\*1\*\*\*\* && !m0 && !m2==\*1\*\*\*\*\* \*], [m1=\*\*\*\*\*1\*\*\*\*]  
 {パターン2} [(s1=\*\*\*\*\*6\*\*\*\* || v1=\*\*\*\*\*6

\*\*\*\*) && m2=\*\*\*\*\*6\*\*\*\*], [(s1=0 && v1=0 && k1=0) && m2=\*\*\*\*\*1\*\*\*\*]

{パターン3} [(j=0000000000100\*\*0 || j=0000000000300\*\*0 || j=0000000000\*010\*\*0 || 0000000000020\*\*0) && (s1=0 && v1=0 && m1=0 && k1=0)]

(7) 文タイプ7  
 {パターン1} [(s1=\*\*\*\*\*1\*\*\* || v1=\*\*\*\*\*1\*\*\*) && j=0\*\*000\*0\*\*0\*0\*\*], [m1=\*\*\*\*\*1\*\*\*]

{パターン2} [s1=\*\*\*\*\*6\*\*\*\* && (m2==\*6\*\*\* || s2=\*\*\*\*\*6\*\*\*\*)], [v1=6\*\*\*\*\* \* && m2=6\*\*\*\*\*], [(s1=0 && v1=0 && k1=0) && ((m2=\*\*\*\*\*1\*\*\* || s2=\*\*\*\*1\*\*\*) && !j=\*\*\*\*\*1\*\*\*\*)]

(8) 文タイプ8  
 {パターン2} [(s1=\*\*\*\*\*6\*\*\* || v1=\*\*\*\*\*6\*\*\*\*) && m2=\*\*\*\*\*6\*\*]

{パターン4} [t=0000000100]

(9) 文タイプ9  
 {パターン1} [k1=\*\*\*\*\*1\* || v1=\*\*\*\*\*1\* || m1=\*\*\*\*\*1\*], [s1=\*\*\*\*\*1\* && j=\*\*\*\*\*1\*\*\*\*\*]

{パターン2} [v1=\*\*\*\*\*6\* && m2=\*\*\*\*\*6\*], [(s1=0 && v1=0 && k1=0) && m2=\*\*\*\*\*1\*]

{パターン3} [j=\*\*\*\*\*17\*\*\*\*\* || j=00000?00000000 || j=0\*0000100\*000000 || j=0\*0000200\*000000 || j=0\*0000300\*000000 || j=0\*0000400\*000000 || j=0\*0000500\*000000]

#### 4. 実験結果と考察

本章では、抽出ルールによる文タイプ抽出実験の結果を示す。そして、抽出ルールで用いた意味辞書を検討し、サ変名詞や動詞および形容詞の意味符号の割り振りが、意味的に妥当なクラスタリングを行っていることを示す。最後に全体的な考察を述べる。

##### 4-1. 抽出実験

実験は、10種類の文タイプを人間が判断し、個別の10個のテキストファイルに保存する。各文タイプの数は図3に示したとおりである。それともう一つ全文を保持したテキストファイルを用意する。そして、各文タイプを保存したファイル

に対して、対応した文タイプ抽出ルールを適用し、文を抽出する。ここで、ルールによって抽出できた文の数をファイル中の文の数で割った値を抽出率として求める。また、全タイプの文を保存したファイルに対して各抽出ルールを適用し、抽出する。この抽出した文のうちの正解数を、抽出した全文数で割った値を正解率として求める。抽出率と正解率の実験結果を図8に示す。また、正解の抽出について、どのルールによって抽出できたかの内訳を図9に示す。

	抽出率 (%)	正解率 (%)
タイプ1	86	89
タイプ2	83	83
タイプ3	86	90
タイプ4	85	71
タイプ5	66	59
タイプ6	77	72
タイプ7	77	83
タイプ8	94	74
タイプ9	86	76

図8. 抽出実験結果

	パターン 1 (個)	パターン 2 (個)	パターン 3 (個)	パターン 4 (個)
タイプ1	55	12		
タイプ2	15		10	
タイプ3		6	105	
タイプ4	32	13		5
タイプ5			28	7
タイプ6	26	8	12	
タイプ7	67	22		
タイプ8		3		14
タイプ9	48	4	12	

図9. ルールパターン毎の抽出数

#### 4-2. 辞書の考察

次に意味辞書について検討する。まず、意味辞書の大きさについて示す。サ変名詞が204個、動詞が104個、形容詞が24個、名詞が63個、特殊な単語が9個である。また助述表現の意味抽出用の辞書は155個である。次に各品詞に割り振られた意味符号が、意味的にクラスタリングされているかどうかを見る。まず、サ変名詞について1の値が割り振られたもの一部を図10に示す。これを見ると、文タイプ1に分類されるものは販売活動や大まかな開発行為に関するサ変名詞である。また、他の列も各々何らかの視点により意味

的にグループ化できる内容になっている。次に、共起をとまなうサ変名詞と動詞を図11に示す。あまり明確に断定できないが、広い概念を示す傾向にあると言えるだろう。最後に共起の対象となる名詞について、比較的抽出数の多いタイプ4とタイプ7のものを図12に示す。タイプ1も12個が共起で抽出されているが、述部に動詞がなく、直前の格にサ変名詞が普通名詞として現れたものがほとんどであるので取り上げない。図12を見ると明らかに2つのグループの名詞は意味的に異なっていることが分かる。ただし、他のタイプ抽出のための共起用名詞については、名詞だけで明確な意味分けをおこなう事は困難であり、共起相手の動詞とのペアでクラスタリングする必要がある。

10桁 目が1	7桁目が 1	5桁目が 1	4桁目が 1	2桁目が 1
公開	サポート	内蔵	測定	専念
開始	重視	付加	アクセス	予想
開催	充実	分離	送受信	実現
開発	強化	存在	認識	削減
販売	検討	導入	把握	解決
発売	確保	構成	操作	進歩
発表	対応	改良	代行	困惑
披露	徹底	実装	入力	注目
出荷		追加	登録	期待
合計 27個	合計 8個	合計 21個	合計 77個	合計 12個

図10. サ変名詞辞書

サ変名詞、動詞	意味符号
利用	s0006066000
指向	s0006000060
投入	s0000060060
使	v0006006060
備え	v0006060000
進め	v6000060000

図11. 共起を伴うサ変名詞と動詞

タイプ4	タイプ7
機能	パソコン
役割	パスワード
ねらい	シンボル
特徴	形
対策	アクション
未対応	—
合計 10個	合計 12個

図12. 共起対象の名詞

#### 4-3. 考察

今回の実験により、述部のみを解析することによりニュース中の文タイプをかなり高い率で分類できることが示された。特にメーカーの活動を記述した文やシステムの動作・操作などは、述部のサ変名詞のみによってかなりの文が判断できている。また、システムの細かな機能は、助述表現の可能によってほとんど判別することができた。

抽出もれや抽出エラーとなった文の特徴を見てみると、「今回富士通が製品として開発したのはこのEJBサーバーと、EJBの規約に沿ったソフト部品などを開発するためのツールAPWORKS。」のように本来述部にくるべき部分が提題として現れているもの。「販売するというもの。」と「得られるものである。」のように助述の出現順序に依存してしまうもの。また、形態素解析の結果からの述部の切り出し、格の切り出しがうまく行えていなかったものなどがあつた。これらの問題は、文解析をより高度にすることによって対応できるものである。一方、「コンパイラなど、周辺のコマンドなどと組み合わせて使用する必要がある。」は“どのようなものか”(文タイプ5)に人間が分類していた文であるが、どの文タイプなのか明確でない文である。これは、文のタイプ分けをどのようにするかという出発点に戻る問題であるが、ニュース文を読みやすくするという我々のシステムの目的からすると、どこに分類されても大きな問題にはならないと考えられる。

#### 5. おわりに

XMLを活用し、新製品情報のニュース文を読みやすく、管理しやすくするシステムを開発するための事前研究として、ニュース中の文をタイプ分けする方法を考案し実験を行った結果について述べた。文のタイプは、ニュース文の特徴と読み手の知識獲得要求の分析より10個の文タイプを考案した。そして、未知語が多く主語が省略された文の多いニュース文に対して、述部のみを解析により高い率で文タイプの抽出が可能であること

が実験で示すことができた。今後、この文タイプを基に段落の意味分けを行い、さらに原因・結果などの簡単な論理関係と重要語のリンク付けを行うことによってニュース文の管理システムを製作していく。

#### 参考文献

- [1]佐藤円、佐藤理史、篠田陽一“電子ニュースのダイジェスト自動生成”情報処理学会論文誌 Vol.36 No.10 pp1371-1378 (1995)
- [2]高尾宣之、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”情報処理学会自然言語研究会 129-17 pp117-124 (1999)
- [3]湯浅夏樹、上田徹、外川文雄“大量文書データ中の単語間共起を利用した文書分類”情報処理学会論文誌 Vol.36 No.8 pp1819-1827 (1995)
- [4]李航、山西健司“線形結合モデルを用いたドキュメント分類”情報処理学会自然言語研究会 119-7 pp37-44 (1997)
- [5]Hidekazu TOKUNAGA and Jun-ichi AOE “Research of The Technical Interpretive Article Database by XML” ICCPOL'99 Proceedings pp281-284 (1999)
- [6]黒橋禎夫、長尾真“表層表現中の情報に基づく文章構造の自動抽出”自然言語学会 Vol.1 No.1 pp3-20 (1994)
- [7]田村直良、和田啓二“セグメントの分割と結合による文章の構造解析”自然言語学会 Vol.5 No.1 pp59-78 (1998)
- [8]福本淳一“著者の主張に基づく日本語文章の構造解析”情報処理学会自然言語研究会 78-15 pp113-120 (1990)
- [9]中挟知延子、島田静雄“字面解析による日本語動詞抽出方法”情報処理学会論文誌 Vol.37 No.2 pp179-187 (1996)