自然言語処理 133−12
(1999. 9. 10)

オントロジ主導による情報抽出の検討

廣田 啓一　佐々木 裕　加藤 恒昭

NTTコミュニケーション科学基礎研究所
〒619-0237 京都府相楽郡精華町光台 2-4
{hirota,sasaki,kato}@cslab.kecl.ntt.co.jp

本稿では，既に提案したオントロジ主導による情報抽出手法ＯＤＩＥ（Ontology-Driven Information Extraction）について詳細な検討を行なう．ＯＤＩＥは，対象とする記事の分野に依存しない情報抽出を実現する新たな手法であり，その実現のために，固有表現抽出による固有用語の認識およびオントロジ辞書を利用した対象分野に特有の分野用語の認識，さらに，これらの用語に対しオントロジ上での語と概念の体系に基づいて記事の中心である事物との関係を探索し，情報として抽出する手法を検討した．本稿では，提案手法について述べるとともに，製品発表に関する新聞記事を対象とした主要情報抽出実験の評価を基に，本手法の有効性を示す．

Design of Ontology-Driven Information Extraction

Keiichi Hirota, Yutaka Sasaki and Tsuneaki Kato

NTT Communication Science Laboratories
2-4 Hikari-dai Seika-cho Souraku-gun Kyoto 619-0237 Japan
{hirota,sasaki,kato}@cslab.kecl.ntt.co.jp

We proposed *ontology-driven information extraction (ODIE)*, a novel approach to extracting information from a text corpus without using domain-dependent *information extraction (IE)* rules. Instead, the ODIE approach employs *ontology* as a semantic guide for selecting key information from the texts. We describe a novel method for achieving ODIE that consists of a marker passing method, the biasing of relation paths, and the extraction of *template* slot names and fillers. Experimental results of the extraction of four pieces of information, such as company names and product names from 250 newspaper articles show high *precision* and *recall*. These results strongly support the feasibility and high potential of ODIE.

1. はじめに

近年，インターネットの飛躍的な発達に伴い，World Wide Web や電子メール，ネットニュース，さらには新聞記事等の電子化された情報が大量に流通している．このような状況の中で，得られた情報から個々の人間にとって必要な情報を選択し，取り出すために大変な労力が必要となりつつある．情報抽出技術は，特定のトピックについて書かれたテキスト情報から主要情報を自動的に取り出す技術であり，このような問題を解決するための手段として期待されている．

情報抽出の研究は，Message Understanding Conference (MUC) を中心にこの１０年間活発に行なわれてきた．MUC は単なる会議ではなく，共通の題材を使った情報抽出システムの性能を比較するためのコンテストの場でもある．対象分野も，海軍のメッセージ（第１回，第２回），ラテンアメリカのテロリズム（第３回，第４回），企業の提携・合併，マイクロチップ製造（第５回），人事異動（第６回）と多岐に渡っている．このような対象分野の変更に対して，MUC 型の情報抽出システムは分野依存のアプローチを取っている．つまり，決められた抽出対象項目を取り出すための機構は，パタンマッチ等を使った分野依存の情報抽出ルール☆により実現されている．例えば，対象分野が企業間の提携についての記事である場合，「会社名」や「提携日」などが抽出対象項目としてあらかじめ決められており，これらの項目を取り出すために分野依存のルールを構築するアプローチを取っている．

我々は，このような対象分野の変更に対して頑健な分野独立のアプローチの１つとして，オントロジ主導による新しい情報抽出手法 ODIE (Ontology-Driven Information Extraction) を提案した [5]．オントロジの持つ高い意味記述能力により，分野に依存する部分をオントロジに集約し，各テキストから情報を抽出する機構を分野独立とする点が，本提案の核となる．

2. 情報抽出技術とオントロジ

2.1 従来手法と提案手法

従来の MUC 型の情報抽出手法は，図 1上部に示すように，（1）テンプレート（抽出すべき情報の項目名と値の空欄からなる表）と（2）情報抽出ルール（抽出対象情報とその周辺の統語的なパタンを表した規則）を用いて，テキストに対して情報抽出ルールによるパタンマッチを行ない，テンプレートを埋めるべき単語を抽出するものが主流である．これらの手法はパタンマッチにより抽出が行なえるため処理が高速であり，かつ適切な抽出ルールを大量に記述すれば，目的とする抽出項目については十分な抽出精度を得る事ができる [2][3][7][9]．

しかし，従来手法には次のような問題点がある．

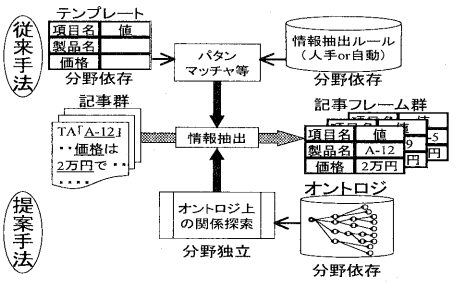

図１　従来手法と提案手法

まず，従来の情報抽出ルールはあらかじめテンプレートに定められた固定的な項目しか取り出せない．次に，情報抽出システムの構築の際に，各テンプレートに対応した分野依存部分の構築のためにコストと時間を要する [1][8]．このような問題を解決するために，データから情報抽出ルールを学習する方法 [11][12] も研究されているが，学習用正解データ作成のコストと時間の問題は解決されていない．

さらに，情報抽出ルールは対象とする文書を検討して作成するため，文書の表現スタイルに強く依存する．例えば新聞記事などの定型的な文書に対しては充分な抽出精度を発揮するが，必ずしも定型的でないネットニュースや電子メールなどに対して，充分な抽出精度を期待できない事になる．

我々はこれらの問題点を考慮した上で，オントロジ主導によりテキスト中の主要情報を表す単語と中心事物との関係を獲得し，抽出項目名とその値との両方をテキストから抽出する，新しい手法 ODIE を提案した（図 1下部）．本手法の特徴は以下のようにまとめられる．

- テンプレートを必要としないため，抽出対象項目が制限されない．
- 分野依存の情報抽出ルールを用いない．
- 対象分野に依存する部分はオントロジにすべて集約されるため，テキストからの情報抽出機構は分野独立である．
- 情報抽出ルールは統語的なパタンによって駆動されるのに対し，本手法はオントロジに与えられた意味関係によって駆動される．
- オントロジの交換のみで，様々な分野への適応が可能である☆☆．

☆ 実際には，有限状態トランスデューサや構文解析用の辞書中のパタンなどルール形式以外の表現をとっているものもあるが，ここでは簡単化のためすべて情報抽出ルールと呼ぶ．

☆☆ 本稿は提案手法の実現可能性の確認が目的であり，単一の対象分野でのみ評価を行なった．オントロジの交換による分野適応の可能性の評価は別の論文に譲る．

2.2 情報抽出を目的としたオントロジ

オントロジとは本来哲学用語であり，「存在に関する体系的な理論（存在論）」という意味を持つ．これに対し，工学分野，特に知識処理の分野においては，「人工物を含めた具体的なものを考察対象として，そこに現われる概念と関係を明示的に示し，明確な意味定義を与えたもの」として扱う [10]．

情報抽出を目的とした場合にオントロジに求められるのは，抽出の対象となる事物がどのような属性，機能，構成要素を持ち，どのような動作をするか，といった，事物の意味定義を与える概念と，これらの概念間の関係の記述である．本提案では，この抽出の対象となる事物を示す概念を**中心概念**と呼び，分野に固有の概念や単語，関連する一般的な概念を中心概念に関係付けて組み立てた，部分的なオントロジを用いる．これを文献 [4] に従って**アプリケーション・オントロジ**と呼ぶ．以降，本稿で用いるオントロジとは，このアプリケーション・オントロジを指すものとする．

本稿におけるオントロジの構成は以下の通りである．また，中心概念をターミナルアダプタ（TA）とするオントロジの例を図2に示す．

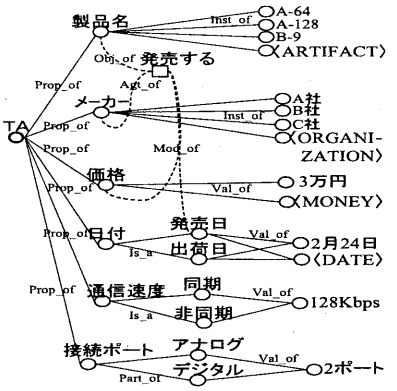

図2 TAに関するアプリケーション・オントロジ例

2.2.1 属性概念

中心概念に対して意味的な定義を与え，中心概念に関連の強い属性や機能などを表す概念を，総称して**属性概念**と呼ぶ．また，動作を示す属性概念を特に**動作概念**と呼ぶ．本オントロジにおける，中心概念及び個々の属性概念を表すノード間を結ぶリンクの関係子を次のように定義する．

Prop_of(property_of): 中心概念に対する属性概念

Is_a(is_a): 属性概念における下位概念

Part_of(part_of): 構成要素となる下位概念

Agt_of(agent_of): 動作の主体となる概念

Obj_of(object_of): 動作の対象となる概念

Mod_of(modification_of): 動作と関係する概念

2.2.2 属性概念のインスタンス

本オントロジにおいて，個々の属性概念はさらに，そのインスタンスとなる語を下位ノードに持つものとする．例えばメーカーという属性概念は，A社，B社などの具体的な会社名を，また価格という属性概念は実際の金額を，下位ノードとして持つ．後者の，金額や日付といった具体的な数値と単位で構成される語を特に**数値概念**と呼び，数値の構成情報と単位の記述による定義を行なう．

属性概念のノードと，そのインスタンス及び数値概念を表すノードを結ぶリンクの関係子を次のように定義する．

Inst_of(instance_of): 属性概念のインスタンス

Val_of(value_of): インスタンスとなる数値概念

3. オントロジ主導による情報抽出手法

本章では，オントロジ主導による情報抽出手法ODIEについて述べる．まず基本手法について述べ，さらに，テキスト中の局所情報を利用するヒューリスティクスを用いた，基本手法に対する改良法について述べる．

3.1 基本手法

基本手法では，（1）テキスト中の主要情報を表す単語の認識，（2）オントロジ上での活性伝播による中心概念との関係列の獲得，（3）関係列の解釈と選択による抽出項目名と値の獲得，という三段階の処理で，テキストからの情報抽出を行なう．

また，情報抽出の対象とする記事に対しては，前処理として形態素解析と固有表現抽出を行なう．

3.1.1 固有表現抽出

企業名や人名といった固有名詞や，金額や日付といった数値的な表現を，**固有表現**と呼ぶ．固有表現は現れ方が多様であり，また人名の「千葉」と地名の「千葉」のように同じ表現でも文脈により固有表

現の種類が異なる場合があるため，形態素解析や辞書のみでテキスト中に出現する固有表現を得る事は難しい．

固有表現を抽出するために，我々は単純置換に基づく日本語固有表現抽出ツール ProCreator を作成した [14]．ProCreator は，多数の単純な置換ルールを逐次的に適用するアプローチをとる事により，現状の技術レベルにおいてかなり精度の良い固有表現抽出ツールとなっている☆．この ProCreator は，対象記事中に出現する固有表現について，製品には ⟨ARTIFACT⟩，日付には ⟨DATE⟩ といった固有表現を表す IREX タグ，および拡張タグ [13] として ⟨REFPERSON⟩ といった照応表現タグなどを付与する．

また，固有表現抽出の延長として，例えば「本製品」は既出の ⟨ARTIFACT⟩ タグの付いた固有表現の照応であるとして，これらを関連づける照応解析も行なっている．

3.1.2 抽出対象語の認識

テキストにおいて主要な情報を表す単語を，**抽出対象語**と呼ぶ．抽出対象語となるのは，固有表現抽出ツールで認識された組織名や人名，日付や金額といった固有表現，および，例えば通信機器の記事であれば通信速度，車の記事であれば排気量といった，対象とする分野に特有の分野用語である．前章で述べたように，このような分野用語についての知識を表現したものがオントロジであり，本提案手法では抽出対象語の認識にこのオントロジを用いる．すなわち，オントロジ上のノードが示す概念を表現する単語が対象となるテキスト中に出現した時，この単語を抽出対象語として認識するものとする．

また，固有表現抽出によるタグに対し，タグに対応するノードを同様にオントロジ上の属性概念のインスタンスとして定義している．抽出された固有表現はタグの種類により，オントロジと関連付けられる．例えば，⟨DATE⟩ タグを付与された固有表現は図 2の ⟨DATE⟩ の位置にリンクされる．これにより，オントロジでは未知の語彙であっても，固有表現であればオントロジ上のノードとして扱う事ができ，抽出対象語として認識できる．

認識した抽出対象語には，テキストにおいてその語が表現する概念の重要さを表す指標として**活性値**を設け，値としてテキストにおける語の出現頻度を与える．なお，ある概念を表現する抽出対象語は複数ある事から，同一の概念を表す語群は同じ単語と見なして出現頻度を計算し，オントロジ上の対応するノードにその出現頻度を活性値として与える．特に Val_of リンクにより下位ノードとなる数値概念については，その定義を満たす個々の値の表現を抽出対象語として考え，値表現ごとに別々のノードを生成して活性値を与える．

3.1.3 オントロジ上での活性伝播による関係列の獲得

認識された各抽出対象語に対して，オントロジ上でノード間の関係を辿る事により，中心概念と語との関係を，概念と関係の経路として得る事が出来る．この経路を表す列を，**関係列**と呼ぶ．本手法ではこの関係列の獲得を，活性伝播を用いて行なう．

オントロジ上の，各抽出対象語に対応するノードから，リンクが張られた上位ノードに向かって活性値を伝播し，現在のノードとリンクと上位ノードとからなる関係列を作る．次に，伝播した先の上位ノードから，さらに上位のノードへと活性値を伝播し，リンクとノードからなる関係列を延ばしていく．伝播の途中でノードが複数のリンクにより複数の上位ノードを持つ場合には，それぞれの上位ノードに対して活性値を伝播し，それぞれに関係列を作る．この活性値の伝播を，中心概念を表すルートノードに到達するまで繰り返す．

このような活性伝播により，活性値の伝播経路に相当する，最下位ノードからルートノードへの関係列が複数生成される．ルートノードに到達した活性値を，その関係列の活性値とする．

3.1.4 関係列の解釈と選択

生成された関係列は，一般に中心概念から属性概念を経て値を表現するノードに至り，ある属性の値が何かという事実情報に対応している．この関係列において，どこまでが抽出項目名を表現し，どこからが値を表現するかを明確にする必要がある．

図 2に示したようなオントロジであれば，末端のノードがその値を表現し，中心概念を示すルートノードと Prop_of で結ばれたノードから末端ノードの直前までが属性を表現すると解釈できる．しかし，「DSU 内蔵」のような真偽値を値とする属性の場合を考えると，必ずしもこのように単純に区切

☆ '99 年 5 月に行なわれた IREX コンテスト [6] において比較的上位の成績であった．

る事ができるわけではない.

そこで，関係列において抽出項目名と値の境界と成り易いリンクを検討し，関係子のリンクによる分割性の強弱を定義づけた．本手法では，この分割性の強弱に従って，リンクの関係子を指標として関係列を項目名と値とに分割する.

関係子の分割性の強弱を次のように規定する（>の左が分割性が強い）.

Inst_of, Val_of > Is_a, Part_of
> Mod_of, Obj_of, Agt_of
> Prop_of

これによりインスタンスや数値概念などの抽出対象語は値となり，属性概念は項目名となる．また真偽値を値とする属性や，対象物などを値とする動作概念も分割する事ができる．なお，関係列中に分割性の強い関係子が複数ある場合には，最も中心概念に近い関係子の位置で二分する.

以上のような方法により，項目名と値を得る方法を**単純手法**と呼ぶ．単純手法で獲得する関係列は，テキスト中の情報の可能性を表現するもので，その全てが正しいわけではない．例えば，一つの語が複数の項目の値として重複して現れたり，一つしか値をとらない項目に対して異なる値を持つ複数の関係列が存在する事がある.

したがって，妥当性の高い関係列のみを取捨選択する必要があり，本提案では，活性値が高い程その関係列は情報として確かであるものとして，個々の項目において最大の活性値を持つ関係列から得られる値を情報として抽出する．また，一つの項目名に対し，最大の活性値を持つ関係列が複数ある場合には，その両方の値を抽出する．このような，関係列の活性値により値の取捨選択を行なう方法を**基本手法**と呼ぶ.

3.2 改良法：活性伝播に対するバイアス

前節で述べた基本手法では，抽出対象語の出現頻度だけでノードの活性値が定まり，それがそのまま伝播されるために，情報の取捨選択を必ずしも適切に行えない．これに対し，情報として適切な抽出対象語ほど活性値を高く伝播するような活性伝播の制御を考え，改良法として導入する事にした．すなわち，抽出対象語のテキストでの用いられ方のヒューリスティクスに基づき，（1）抽出対象語の共起関係，（2）抽出対象語の格関係，を見る事による，活性伝播に対する二つのバイアスを設けた.

このバイアスにより活性値が変化する割合を**変動率**と呼ぶ．活性伝播の際には，ノードの持つ活性値に対し，バイアスによって得られる変動率をかけた値が，上位ノードに伝播する事になる.

3.2.1 共起バイアス

ある抽出対象語が同文中で上位ノードが示す語と共起するならば，これらのノードの間には強い関係があるものとして活性値を強めて伝播し，他の上位ノードとの間には関係はないものとして他のノードへの伝播を禁止する．これを**共起バイアス**と呼ぶ．共起バイアスにより近傍に現れた語同士が強く関係付けられ，これらの語を含む関係列の活性値が高くなる．共起バイアスによる活性値の変動率は，一文中での抽出対象語に対する上位ノードが示す語の出現位置によって，表1のように定める.

表1 共起バイアスにおける出現位置と変動率

出現位置	パターン例	変動率
抽出対象語の後方	...4月21日...発売日...	1.1-1.5☆
抽出対象語の前方	...発売日...4月21日...	1.1-1.5☆
読点，記号を挟んで直後	...4月21日，発売日...	2
助詞，読点を挟んで直後	...4月21日が，発売日...	3
記号を挟んで直前	...発売日：4月21日...	2.5
助詞，読点を挟んで直前	...発売日は，4月21日...	3

例えば「発売日は2月24日です」という文がテキストにある時の共起バイアスの作用例を図3に示す．「2月24日」はVal_ofリンクで，2つの上位ノード「発売日」と「出荷日」に結ばれている．テキスト中の「2月24日」に対し，上位ノードが示す語である「発売日」は助詞を挟んで直前に出現するため，「2月24日」から「発売日」へ活性値を3倍して伝播し，もう一方の上位ノード「出荷日」への伝播を禁止する．その結果「2月24日」から「発売日」への伝播経路である関係列が，高い活性値を持つ.

3.2.2 格バイアス

ある抽出対象語が，テキスト中にどの様な主題役割で現れたかは，語に対応するノードの役割に関する情報となる．これに着目し，オントロジの定義から得られるノードの役割と，テキスト中での抽出対

☆ 抽出対象語と上位ノードが示す語に挟まれた自立語の句数によって計算する．前方に出現する方が関係が強いものと考え，前方出現は $1.5 - 0.1 \times$(句数)で，後方出現は $1.5 - 0.2 \times$(句数)で計算した．なお，どちらの場合も変動率の最低値は1.1とした.

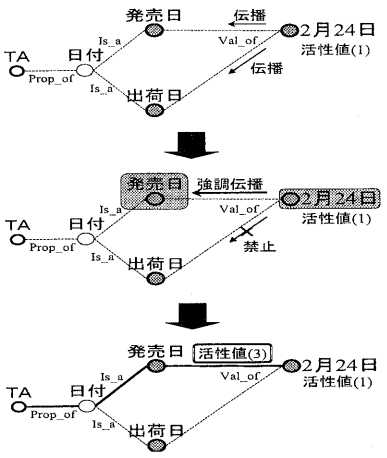

図3 共起バイアスの作用例

象語の役割が一致する時に活性値を強めて伝播し，異なる時に活性値の伝播を弱める．これを**格バイアス**と呼ぶ．

すなわち，ある抽出対象語に対応するオントロジ上のノードに対し，その上位ノードがAgt_ofやObj_ofなど主題役割を表すリンクで動作概念を表すノードと結ばれていて，その抽出対象語がテキスト中の動作概念に対応する用言に対して同じ主題役割を持つ場合に，その活性値を強調して伝播する事で，主題役割の合致度を反映する．逆に異なる主題役割を持つ場合には活性値を弱めて伝播する．

テキスト中の語の主題役割は，表2のような助詞の種類と用言の態による分類表を作成し，抽出対象語に付随する助詞と直後に出現する用言の態から該当する関係子を得る．特に「から」のように複数の主題役割を与え得る助詞には複数の分類を許し，分類表にないその他の助詞については一律にその他と分類した．

表2 主な助詞の種類と用言の態による主題役割の分類

助詞	用言の態	
	能動態	受動態
は	その他	Obj_of
が	Agt_of	Obj_of
に	Mod_of	Mod_of
を	Obj_of	Obj_of
の	Agt_of	Obj_of
で	Mod_of	Mod_of
から	Agt_of/Mod_of	Agt_of/Mod_of

格バイアスによる活性値の変動率は，表3のように定め，主題役割を表す関係子とオントロジ中のリンクの関係子との比較により決定する．

表3 格バイアスにおける主題役割の比較と変動率

主題役割とリンクの関係子の比較	変動率
主題役割がその他である	2
リンクの関係子と対立する (Agt_of ⇔ Obj_of)	0.5
リンクの関係子と一致しない	1.5
リンクの関係子と一致する	3

例えば「A社はA-128を発売いたします．A-128は従来のA-64に対し，...」という文がテキストにある時の格バイアスの作用を図4に示す．「製品名」は動作概念「発売する」とObj_ofリンクで結ばれており，その役割が「発売する」の対象となる事が定義されている．この時，「製品名」のインスタンスである「A-128」は，付随する助詞「を」と能動態の用言「発売する」から，文中での主題役割として分類表によりObj_ofが得られる．主題役割とリンクの関係子を比較すると一致する事から，「A-128」から上位ノード「製品名」へ活性値を3倍して伝播する．一方，「A-64」は動作概念に対応する動詞を持たない事から，活性値は変動せずに伝播する．その結果，「A-128」と「製品名」からなる関係列は，「A-64」と「製品名」からなる関係列よりも高い活性値を持つ．

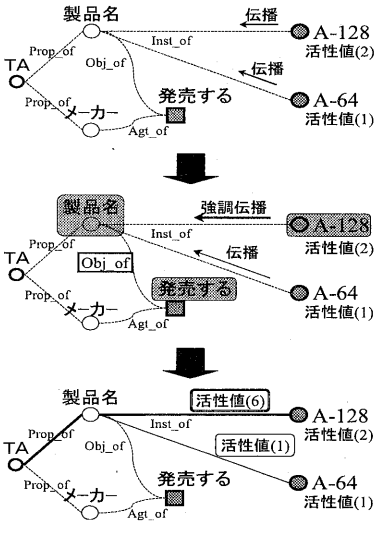

図4 格バイアスの作用例

4. 製品発表記事からの情報抽出実験

今回，CD-毎日新聞94版の1年分の記事の中から製品発表記事250件を対象として，主要情報の抽出実験を行なった．実験で用いるアプリケーション・オントロジの構築に際しては，人手により，記事中の主要な情報項目を属性概念として中心概念と結び付け，抽出対象語となる記事中に出現する単語を固有表現中心に収集して，製品に関する個々の属性概念のインスタンスとした．オントロジの作成のための労力は高々2人日であった．

評価にあたっては，従来の情報抽出手法において主な抽出対象項目として用いられる，製品名，メーカー，価格，発売日の四項目についての正解例を作成し，再現率(Recall) ⋆と適合率(Precision)⋆⋆による評価を行なった．結果を表4，表5に示す．

表4　主要情報の抽出実験評価（再現率）

抽出項目	製品名	メーカー	価格	発売日
単純手法	83.8%	83.7%	96.0%	98.0%
基本手法	81.8%	78.2%	95.3%	93.0%
共起バイアス	81.8%	77.0%	84.5%	93.0%
格バイアス	78.7%	79.0%	95.3%	91.0%
共起 + 格	78.7%	79.4%	84.5%	91.0%

表5　主要情報の抽出実験評価（適合率）

抽出項目	製品名	メーカー	価格	発売日
単純手法	59.0%	44.4%	79.4%	50.6%
基本手法	68.4%	69.3%	82.3%	53.6%
共起バイアス	68.4%	69.5%	87.8%	53.6%
格バイアス	78.0%	79.8%	82.3%	70.7%
共起 + 格	78.0%	80.5%	87.8%	70.7%

また，再現率と適合率の両方を勘案した総合的な評価尺度であるF値(F-mesure) ⋆⋆⋆については，表6のような結果となった．

4.1 結果と考察

まず，単純手法では，抽出対象語に対して候補として可能な関係列を全て抽出するため，再現率は全項目とも高いが，適合率が低い．これに対し，各抽出項目ごとに最も高い活性値を持つ関係列を選択する基本手法では，再現率の減少がほとんどの項目で5%程度と僅かであるのに対し，特に製品名・メーカーに関して適合率の向上が見られた．

さらに，改良法として活性伝播に対するバイアスを導入した事で，適切な関係列を情報として抽出できるようになり，適合率を大幅に向上させる事ができた．バイアスの作用を個別に見ると，共起バイアスで価格，格バイアスで製品名とメーカー，発売日の項目が向上し，両バイアスの共用でメーカーの項目がさらに向上している事から，両バイアスが相互に適切に作用した事がわかる．

一方，表6を見ると，共起バイアスを用いた時，価格の欄でF値が僅かに下がっている．これは，価格の項目が複数の値を正解とするのに対し，共起バイアスを用いると，最も近い共起関係にある抽出対象語の関係列が最大の活性値を持ち，一部の正解しか抽出できずに再現率が低下するためである．

本手法は情報選択後も80%前後の再現率を得，活性伝播に対するバイアスの導入によって，適合率で平均して80%近い結果となった．

また，本実験におけるオントロジ構築は高々2人日の労力で済んでおり，システム構築の際に分野依存のルールの作成や正解例の学習が不要であった事から，従来手法におけるコストと時間の問題が軽減されているといえる．

表6　主要情報の抽出実験評価（F値）

抽出項目	製品名	メーカー	価格	発売日
単純手法	69.2	58.0	86.9	66.7
基本手法	74.5	73.5	88.3	68.0
共起バイアス	74.5	73.1	86.1	68.0
格バイアス	78.3	79.4	88.3	79.6
共起 + 格	78.3	79.9	86.1	79.6

4.2 課　　題

一方，幾つかの課題も明らかになった．まず，余分な抽出対象語の誤認が挙げられる．例えば記事冒頭に現れる記事自体の日付以外に日付に関する記述がない場合，冒頭の日付が発売日と認識される．このような誤った認識を防ぐため，活性値に閾値を設け，確実な情報だけ選択する方法を検討している．

また，複数の値をとる項目（多値属性）の扱いも課題である．提案手法では，最大の活性値を持つ関係列から値を選択するために，複数の正解の内の一

⋆ 再現率 $= \frac{\text{システムが正しく抽出した値の数}}{\text{抽出すべき値の数}}$

⋆⋆ 適合率 $= \frac{\text{システムが正しく抽出した値の数}}{\text{システムが抽出した値の数}}$

⋆⋆⋆ F値 $= \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}}$

部しか抽出できない．さらに，一つの記事に複数の製品が記述されている場合に，個々の製品を区別する事が課題である．

本提案手法は，語の出現に対してオントロジ主導により概念関係を探索し，局所的な共起関係や格関係のみを見ている．したがって，細かい言い回しや表層的な表現に依存しないので，電子メールやネットニュースなどに見られる口語的な記述を取り扱う事ができるものと考えられる．また，共起関係や格関係を活性伝播に対するバイアスの形で扱うため，日本語以外の言語であっても，同様の文法知識をこのようなバイアスの形で与える事で情報の抽出が可能であると考える．今後，提案手法の拡張と実験を通じ，確認していく．

5. ま と め

本稿において，テンプレートや情報抽出ルールを用いる事なく，オントロジ上の活性伝播により情報抽出を行なう，オントロジ主導による情報抽出手法を検討し，製品発表記事からの抽出実験により，手法の評価を行なった．本手法は再現率・適合率ともに高い評価値を得，その有効性を確認できた．

今後の課題として，従来手法ではパターン記述が難しい抽出項目を本手法により抽出可能かどうか，及びオントロジの交換による対象分野の容易な変更が可能かどうかの検討があげられる．また，4.2節で述べた，一つの項目に対する複数の正解の抽出や，一つの記事に複数の対象が記述されている場合の対策，精度の向上も今後の課題である．

謝辞 CD-毎日新聞94版を利用した．コーパスの利用を許可していただいた毎日新聞社殿に深く感謝いたします．

参 考 文 献

[1] Appelt, D., Hobbs, J., Bear, J., Israel, D.J. and Tyson, M.: FASTUS: a finite-state processor for information extraction from real-world text, *Proceedings of IJCAI-93*, pp1172–1178 (1993).

[2] Applet, D., Hobbs, J., Bear, J., Israel, D., Kameyama, M., Kehler, A., Martin, D., Myers, K. and Tyson, M.: SRI International Fastus System MUC-6 Test Results and Analysis, In *Proceedings of the sixth message Understanding Conference(MUC-6)*, pp.237–248 (1995).

[3] Grishman, R.: The NYU System for MUC-6 or Where's the Syntax?, In *Proceedings of the sixth message Understanding Conference(MUC-6)*, pp.167–175 (1995).

[4] Guarino, N.: Semantic Matching: Foraml Ontological Distinctions for Infromation Organization, Extraction, and Integration, *Information Extraction*, LNAI 1299, pp.137–170 (1997).

[5] 廣田啓一，佐々木裕，加藤恒昭: オントロジ主導による情報抽出手法の提案, 言語処理学会第5回年次大会発表論文集, pp.120–123 (1999).

[6] IREX ホームページ. (http://cs.nyu.edu/cs/projects/proteus/irex/)

[7] 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-122-10 (1997).

[8] Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E. and Soderland, S., University of Massachusetts: MUC-4 Test Results and Analysis, in Proc. of *MUC-4*, pp.151–158 (1992).

[9] 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).

[10] 溝口理一郎，池田 満: オントロジー工学序説－内容指向研究の基盤技術と理論の確立を目指して－, 人工知能学会誌, Vol.12, No.4, pp.559–569 (1997).

[11] Riloff, E.: Automatically Generating Extraction Pattern from Untagged Text, *AAAI-96*, pp. 1044–1049 (1996).

[12] Sasaki, Y. and Haruno, M.: RHB^{+}: A Type-Oriented ILP System Learning from Positive Data, *IJCAI-97*, pp.894–899 (1997).

[13] 佐々木 裕: トランスデューサによる日本語固有表現抽出, 言語処理学会第5回年次大会発表論文集, pp.108–111 (1999).

[14] 佐々木 裕，廣田 啓一，加藤 恒昭: ProCreator: 単純置換に基づく日本語固有表現抽出ツール, IREX ワークショップ予稿集 (1999).