特集論文 Technical Papers

オントロジー主導による情報抽出

Ontology-Driven Information Extraction

廣田 啓一* 佐々木 裕* 加藤 恒昭*
Keiichi Hirota Yutaka Sasaki Tsuneaki Kato

* NTTコミュニケーション科学基礎研究所
NTT Communication Science Laboratories, Kyoto 619–0237, Japan.

1999年3月31日 受理

Keywords: ontology, information extraction.

Summary

This paper proposes *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. First, we discuss a relation between information extraction and ontology, then we define *application ontology* for the purpose of ODIE. Next, 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. We also discuss a possibility of a semi-automatic construction of the application ontology and the effect of incompleteness in the ontology on the IE task.

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）と多岐にわたっている．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*1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．

本論文では，最近のオントロジーの研究の発展 [Guarino 97, 溝口 97] に注目し，オントロジーの新しい応用の一つとして，オントロジー主導による新しい

*1 実際には，有限状態トランスデューサ [Roche 97] 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．

情報抽出手法 ODIE (Ontology-Driven Information Extraction) を提案する．オントロジー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ーに集約し，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．

本稿では，以下のような構成で，オントロジー主導による情報抽出を可能にするための構成と手法を提案し，実験による評価結果について述べる．まず，**2** 章では情報抽出技術とオントロジーについて述べる．**3** 章では本論文で想定しているオントロジーを定義する．**4** 章ではオントロジー主導による情報抽出手法を提案する．**5** 章では新聞記事を題材にした実験により，提案手法の評価を行なう．**6** 章ではオントロジーの自動獲得に向けての試みとその要素技術，想定されるオントロジーの不完全さと提案手法について議論する．**7** 章ではまとめを述べ，本論文を締めくくる．

2. 情報抽出技術とオントロジー

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 [Appelt 95, Grishman 95, 井出 97, 松尾 95].

しかし，従来の手法には次のような問題点がある．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．例えば，Umass/MUC-3 システムは情報抽出ルール作成のために 1500 人時を要したと報告されている [Lehnert 92]．SRI の FASTUS では，テロリズムの分野に依存した部分を作成するために 3 週間半を費やしている [Appelt 93]．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 [Riloff 96, Sasaki 97] も研究されているが，学習するための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．

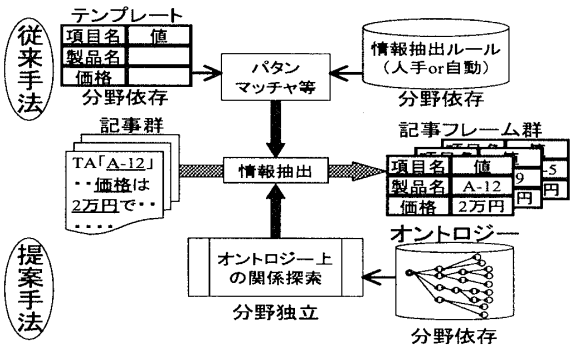


図 1　従来手法と提案手法

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，オントロジー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 を提案する（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ー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ーに与えられた意味関係によって駆動される．
- 各分野毎に，様々なアプリケーションで共通に利用可能な分野知識であるドメインオントロジーが整備されれば，それを利用して情報抽出が可能となる．
- オントロジーの交換のみで，様々な分野への適用が可能である*2.

これまで，情報抽出においてタキソノミ (Taxonomy) やオントロジーを利用した研究は幾つか存在する．文献 [野畑 98] は情報抽出ルールの抽出精度を向上させるために，タキソノミを参照する手法を述べている．文献 [Guarino 97] はテンプレートと抽出対象の意味的な照合をとるためにオントロジーを利用する事を示唆し

*2　本論文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．オントロジー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．

ている．提案手法は，テンプレートや情報抽出ルールを利用する事なくオントロジーを中心に情報抽出する点でこれらの研究と異なっている．

以下，本手法で用いるオントロジーの定義と，手法の詳細について述べる．

3. 情報抽出を目的としたオントロジー

オントロジー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」として扱っている [溝口 97]．

一般に人手によって構築される大規模なオントロジーは，概念を上位下位関係や部分全体関係によって階層的に分類したタキソノミを基本とする．このようなタキソノミを提供するものとして文献 [池原 97, Miller 90] などがある．オントロジーは，更に各概念の意味定義と概念間の詳細な関係を明確に記述したものとなっている [Gruber 91]．このような大規模オントロジーは世界知識の共有を目的としており，汎用的ではあるが，その一方で，個々のドメインに特化した情報を扱うには巨大過ぎる面がある．

情報抽出を目的とした場合にオントロジーに求められるのは，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ーを用いるものとする．これを文献 [Guarino 97] に従って**アプリケーションオントロジー**と呼ぶ．以降，本稿で用いるオントロジーとは，このアプリケーションオントロジーを指すものとする．

このような記事分野によるアプリケーションオントロジーの構築は，理想的には，分野が決まった時に大規模オントロジーから抜き出すか，テキストから自動獲得するかして，人手による作業なしでアプリケーションオントロジーを得る事が望ましいが，前者については今は適当な大規模オントロジーが無く，後者についても検討の段階にある．後者についての試みとその問題点については，**6 章**で述べる．なお，本論文の実験で用いたオントロジーは後者による獲得支援を用いて人手により作成した．

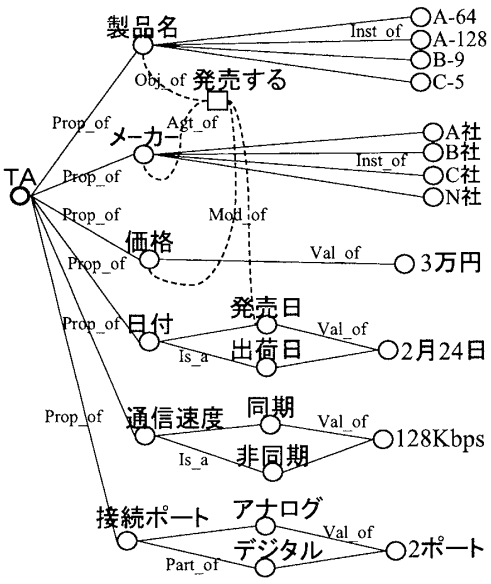


図 2　TA に関するアプリケーションオントロジー例

本稿におけるアプリケーションオントロジーの構成を以下に示す．また，中心概念をターミナルアダプタ（TA）とするアプリケーションオントロジーの例を図2に示す．

3・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．本オントロジーにおける，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

3・2 属性概念のインスタンス

本オントロジーにおいて，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

4. オントロジー主導による情報抽出手法

本章では，提案するオントロジー主導による情報抽出手法 ODIE について述べる．**4・1 節**で，まず基本手法について述べ，さらに，**4・2 節**で，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．

4・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）オントロジー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．

〔1〕 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であって，オントロジー上にその語に対応した概念のノードを持つ単語を，**抽出対象語**と呼ぶ．したがって，対象となるテキスト中にオントロジー上のノードが示す概念を表現する単語が出現した時，この単語を抽出対象語として認識する．

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，オントロジー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．活性値はテキストにおける概念の重要さを表す指標となる．

また，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．

〔2〕 オントロジー上での活性伝播による関係列獲得

認識された各抽出対象語に対し，オントロジー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．

オントロジー上の，各抽出対象語を表すノードから，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〕 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ーであれば，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）．

$$\begin{aligned}\text{Inst\_of, Val\_of} &> \text{Is\_a, Part\_of}\\ &> \text{Mod\_of, Obj\_of, Agt\_of}\\ &> \text{Prop\_of}\end{aligned}$$

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．

4・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．

〔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の伝播を強め，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．

すなわちテキスト中の1文において上位ノードの表す語が抽出対象語の近傍[*3]に現れる時，その上位ノードに活性値を1加算して伝播し，他の上位ノードへの伝播を禁止する．これ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．

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を図3に示す．「2月24日」はVal_ofリンクで，二つの上位ノード「発売日」と「出荷日」に結ばれている．テキスト中の「2月24日」の近傍に「発売日」があるため，「2月24日」から「発売日」へ活性値を1加算して伝播し，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．

〔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

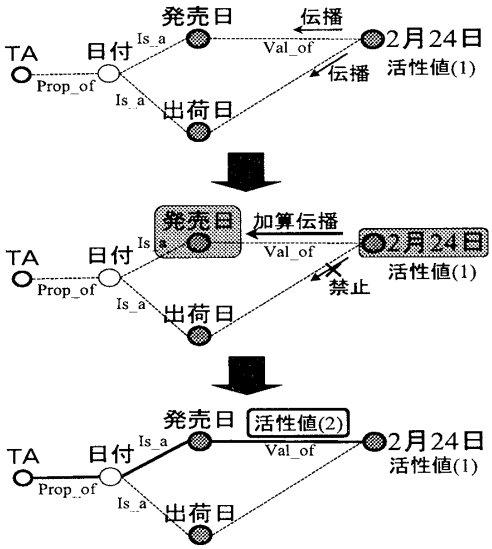


図3 共起バイアスの作用例

なる．ここに着目し，オントロジー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象語の役割が一致する時に活性値の伝播を強め，異なる時に活性値の伝播を弱める．このバイアスを**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ー上のノードに対し，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1加算して伝播する事で，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1減算して伝播する．

テキスト中の語の主題役割は，表1のような助詞の種類と用言の態による分類表を作成し，付随する助詞と直後に出現する用言の態から主体／対象／関係／その他の四つに分類して，オントロジー中の役割と比較した．また，特に「から」のように複数の主題役割を与え得る助詞について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1 主な助詞の種類と用言の態による主題役割の分類

	用言の態	
助詞	能動態	受動態
は	その他	対象
が	主体	対象
に	関係	関係
を	対象	対象
の	主体	対象
で	関係	関係
から	主体/関係	主体/関係

オントロジー中の役割は，動作概念とのリンクの関

*3 実験では，助詞，読点の挿入を許して直前または直後の語とした．

係子によって決定し，Agt_ofが主体，Obj_ofが対象，Mod_ofが関係に対応する．なお，その他に分類された抽出対象語は主題役割の比較を行なわず，活性値を増減させる事なく伝播する．

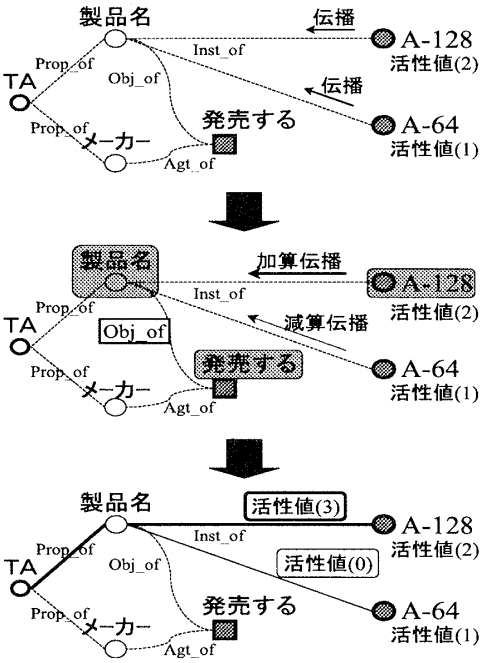


図4　格バイアスの作用例

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，「製品名」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」の文中での主題役割は分類表により「発売する」の対象であるため，役割の一致から活性値を1加算して伝播する．逆に「A-64」は「発売する」の対象でないために活性値を1減算して伝播する．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．

5. 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件を対象とした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ーの構築に際しては，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．オントロジー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)*4と適合率 (Precision)*5 による評価を行なった．結果を表2，表3に示す．

表2　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	100%	100%	100%	100%
基本手法	95.7%	97.2%	99.3%	98.0%
共起バイアス	95.7%	96.8%	88.2%	98.0%
格バイアス	94.5%	98.4%	99.3%	96.0%
共起 ＋ 格	94.5%	98.8%	88.2%	96.0%

表3　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	63.2%	54.9%	80.8%	51.6%
基本手法	76.7%	86.4%	83.6%	53.4%
共起バイアス	76.7%	86.4%	88.6%	53.4%
格バイアス	92.3%	96.6%	83.6%	71.1%
共起 ＋ 格	92.3%	97.7%	88.6%	71.1%

5・1 結果と考察

まず，単純手法では，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100%となるが，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，再現率がほとんどの項目で僅かに下がった一方で，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により，適切な関係列を情報として抽出できるようになった．その結果，候補選択後の再現率をほとんど下げる事なく，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おり，両バイアスが相互に適切に作用している事がわかる．

本手法は情報選択後も90%を越える高い再現率を得，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90%近い結果を得た．学習規模や対象記事が異なるため簡単には比較できないが，従来の抽出手法[井出 97]における製品情報抽出実験の評価値と比べて

*4　再現率 $=$ システムが正しく抽出した値の数 / 抽出すべき値の数

*5　適合率 $=$ システムが正しく抽出した値の数 / システムが抽出した値の数

も遜色ない結果が得られ，手法の有効性がうかがえる．

また，本実験におけるオントロジー構築は高々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．

5・2 課　　題

一方，幾つかの課題も明らかになった．まず，誤まった抽出対象語の認識が挙げられる．例えば，記事の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てしまう．このような誤った認識を防ぐため，活性値に閾値を設けて，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．前処理により，1テキストに1対象のみが記述されるように記事を分割する方法の導入などの拡張が必要である．

本提案手法は，単語の出現に対してオントロジー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．細かい言い回しや表層的な表現に依存しないので，従来の抽出手法では対象とし得なかった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て，確認していく．

6. オントロジーの自動構築に向けて

前章までで，オントロジーを用いた情報抽出手法について述べてきた．本章では，文書の集まりからオントロジーを作成するための支援手法，およびその自動構築に関する可能性と問題点について考察する．

6・1 コーパスからの言語知識獲得とオントロジー

情報抽出を目的とした場合のオントロジー自動獲得は，コーパスからの言語知識獲得の諸技術 [松本 97] と関連づける事ができるが，次の2点で性格を異にする．まず第一に，対象となる固有の分野における文書数が著しく少ない事である．一般に，同一種類の製品に関する発表記事が100件以上あるような状況は期待できない．この数は，例えば，一年分の新聞記事を対象にして行なわれる言語知識の獲得（[春野 95, Pereira 93] 等）と比べるとオーダの違いがある．第二に，コーパスからの言語知識獲得がそのコーパスに含まれている知識を網羅的に取り出そうと試みるのに対し，情報抽出を目的としたオントロジーでは，中心概念と関連した周辺の構造を重点的に取り出す必要がある．

6・2 ノードとリンクの作成

オントロジーを構築するためには，オントロジーのノードとなるような表現を抽出し，それらの表現の役割と関係を明らかにする必要がある．前者には共起（collocation）抽出，固有用語抽出の技術が利用でき，後者は統語的なパタンの抽出とみる事ができる．以下にオントロジーを構成する要素の獲得についての検討を行なう．

〔1〕 ノードとなる表現の獲得

ある分野の文書に頻繁に生じる語の共起，特に語の連続を比較的少ない量の文書から抽出する手法が提案されており [Frantzi 96]，取り出された複合語はその分野で重要な概念を表現すると考える事ができる．例えば，ターミナルアダプタの製品発表記事群にこの手法を適用すると，表4に示すものが重要な複合語として得られる．また，IREXコンテスト [関根 98] の開催により，人名，会社名，時間表現等の固有表現を文書から自動的に抽出する汎用的な手法が確立されつつある [佐々木 99]．これらの手法を用いて，オントロジーのノードとなる表現の獲得が可能である．

表4 製品発表記事から抽出される複合語

重要度	頻度	複合語
162.00	171	ターミナル アダプタ
152.67	81	INS ネット 64
128.86	132	アナログ ポート
106.40	56	ISDN ターミナル アダプタ
105.00	21	非同期 / 同期 PPP 変換 機能
82.00	104	INS ネット
52.00	57	DSU 内蔵
51.76	55	INS メイト
39.33	42	インターネット 接続
39.00	41	ISDN 回線

〔2〕 インスタンスと数値概念の獲得

日本語が複合語内部においても主辞後置である事から，例えば「～株式会社」「～機能」といった得られた複合語を，構成する最右の単語に着目してまとめ，同じ属性概念のインスタンスを集める事が可能である．

また，例えば，「〈クラス表現〉〈開鉤括弧〉〈**複合語等**〉〈閉鉤括弧〉」（強調部分がインスタンス表現）というルールにより，「ターミナルアダプタ『INS メイト V-7DSU』」という表現から，ターミナルアダプタのインスタンス「INS メイト V-7DSU」を得る事ができる．

さらに，数量表現についても「～Kbps」「～ポート」といった助数詞部分に着目してクラス分けが可能であり，Val_of リンクの下位ノードとなる数値概念を得る事ができる．

〔3〕 **属性概念の獲得**

中心概念に関連の強い属性や機能を示す属性概念は，中心概念自体もしくはそのインスタンスと共起して出現する事が多い．例えば「INS メイト V-7DSU の標準価格」のように，中心概念のインスタンス表現が「の」等により連体修飾するパタンで現れる．また，「低価格」「高性能」「DSU 内蔵」といった真偽値を値とするような属性概念が，「低価格多機能のターミナルアダプタ」のようにクラス表現に連体修飾するパタンで現れる．

〔4〕 **動作概念の獲得**

動作概念は「NTT は，～『INS メイト V-7DSU』を平成 x 年 y 月 z 日より販売開始します」のような，中心概念のインスタンス表現を格要素に持つ文から得られる*6．ある用言が中心概念のインスタンス表現を格とする文の主辞である頻度の大小により，その用言を動作概念とすべきかが判定でき，その格要素の集まりから Agt_of, Obj_of 等のリンクが得られる．

6・3 オントロジーの不完全さと情報抽出の精度

前節では，オントロジーを構成する要素を半自動的または自動的に収集する試みについて述べた．最後に，この試みを進めて将来オントロジーの自動獲得が実現した時の，本提案手法の性能を展望する．

自動獲得したオントロジーの最大の問題はノイズの混入である．重要な属性概念を認識できない，インスタンス表現を取得できない，間違った属性概念に対応づけるといった，様々な種類のノイズが考えられる．特に，事物の名称については未知語となる可能性が高く，ノードの欠落や誤ったリンクの発生が予想される．

そこで，**前章**で述べた抽出実験で用いたオントロジーに人為的に不完全さを付与して実験を行ない，ノイズの影響を明らかにする．インスタンスにつながるリンクの追加，削除を無作為に行なったオントロジーを使った実験の結果を図 5に示す．横軸は追加，削除を行なったリンク数，縦軸は F 値*7を表す．

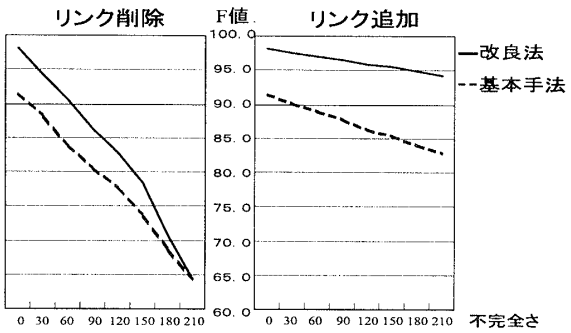


図 5 オントロジーの不完全さによる影響

リンクの追加はインスタンスが複数の属性概念の下位ノードとなっている場合，リンクの削除はインスタンスが未知語として上位ノードを持たない場合に該当する．実験の結果，リンクの削除，すなわちノードが欠落している場合には基本手法，改良法ともに F 値が下がるが，リンクの追加に対してはバイアスを使った改良法の方が基本手法に比べて F 値の低下が緩やかであり，用いたバイアスが誤った情報の抽出を防いでいる事がわかる．

以上の点から，本提案手法は自動獲得されたオントロジーにおいて冗長なノードやリンクがあっても，適切な情報を選択して抽出する事が可能であると考えられる．したがって，オントロジーの自動獲得では余分なノードやリンクを許す一方，取りこぼしを少なくするようなアプローチが重要となる．

また，現在は獲得支援として**前節**で述べた処理を用いているが，シソーラスや大規模オントロジー等の既存知識源を併用する等で，自動獲得への展開を検討中である．

7. 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，オントロジー上の活性伝播により情報抽出を行なう，新しい手法を提案し，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，およびオ

*6 製品発表記事においては，発表される製品が主題となって文中で省略される事が極めて多く，実際にはこのような文は比較的少ない．この事が獲得を困難にしている．

*7 F 値とは，下式で定義される，適合率，再現率の両方を勘案した評価尺度である．

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

ントロジー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．今回行なった実験では，提案手法の実現可能性の確認を目的として，単一の分野で250記事を対象として主要な情報の抽出を行なった．今後，実験規模の拡大により，提案手法のさらなる検討および分野適応の可能性について，評価を進めたい．また，**5・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◇

[Appelt 93] 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*, pp. 1172–1178 (1993).

[Appelt 95] Appelt, 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).

[Frantzi 96] Frantzi, K.T. and Ananiadou, S.: Extracting Nested Collocations, *COLING '96*, pp. 41–46 (1996).

[Grishman 95] 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).

[Gruber 91] Gruber, T.R.: Ontolingua: A Mechanism to Support Portable Ontologies, *Technical report KSL-91-66* (1991).

[Guarino 97] Guarino, N.: Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration, *Information Extraction*, LNAI 1299, pp. 139–170 (1997).

[春野 95] 春野雅彦: 最小汎化とオッカムの原理を用いた動詞格フレーム学習，1995年人工知能学会全国大会（第9回），18-05, pp. 511–514 (1995).

[井出 97] 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, pp. 63–70 (1997).

[池原 97] 池原 悟，宮崎正弘，白井 諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林 良彦: 日本語語彙大系，岩波書店 (1997).

[Lehnert 92] 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).

[松本 97] 松本裕治: コーパスからの言語知識獲得，1997年人工知能学会全国大会（第11回），S3-06, pp. 64–67 (1997).

[松尾 95] 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).

[Miller 90] Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.J.: Introduction to wordnet: An on-line lexical database, *International Journal of Lexicography 3*, No. 4, pp. 235–244 (1990).

[溝口 97] 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).

[野畑 98] 野畑 周，関根 聡: 日本語情報抽出システムの開発と評価，情報処理学会研究報告，98-NL-127-16, pp. 117–124 (1998).

[Pereira 93] Pereira, F., Tishby, N. and Lee, L.: Distributional Clustering of English Words, *ACL '93*, pp. 183–190 (1993).

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

[Roche 97] Roche, E. and Schabes, Y.: *Finite-State Language Processing*, MIT Press (1997).

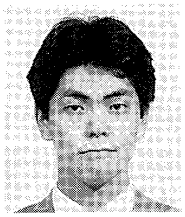
[Sasaki 97] Sasaki, Y. and Haruno, M.: *RHB*$^{+}$: A Type-Oriented ILP System Learning from Positive Data, *IJCAI-97*, pp. 894–899 (1997).

[佐々木 99] 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).

[関根 98] 関根 聡，井佐原 均: IREX：情報検索，情報抽出コンテスト，情報処理学会研究報告，98-NL-127-15, pp. 109–116 (1998).

〔担当委員：北村泰彦〕

著者紹介



廣田　啓一

1995年三重大学工学部情報工学科卒業，1997年同大学院工学研究科情報工学専攻修士課程終了．同年，日本電信電話（株）入社．自然言語理解，情報抽出に関する研究に従事．現在，NTTコミュニケーション科学基礎研究所知能情報研究部に所属．情報処理学会会員．〈hirota@cslab.kecl.ntt.co.jp〉

佐々木　裕（正会員）

1986年筑波大学第三学群情報学類卒業，1988年同大学院理工学研究科修了．同年，日本電信電話（株）入社．現在，NTTコミュニケーション科学基礎研究所知能情報研究部に所属．1995年～1996年サイモン・フレーザー大学（カナダ）客員研究員．主として機械学習，知識処理に関する研究に従事．言語処理学会会員．〈sasaki@cslab.kecl.ntt.co.jp〉

加藤　恒昭（正会員）は，前掲（Vol. 14, No. 3, p.465）参照．