オントロジー主導による情報抽出の実装と評価

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

1 はじめに

我々は，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，オントロジー主導による新しい情報抽出手法 ODIE (Ontology-Driven Information Extraction) を提案した [1][2][3]．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，本提案の中心であった（図1）．

本稿においては，実装したODIEシステムについて述べ，製品発表記事からの抽出実験による本手法の有効性と，固有表現抽出による拡張性を示す．

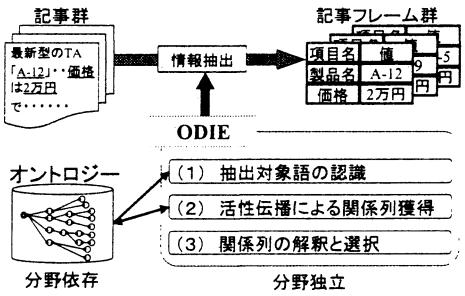

図1 提案手法の概要

2 オントロジー主導による情報抽出

以下、本手法で用いるオントロジー，およびODIEシステムと固有表現抽出について簡単に述べる．

2.1 情報抽出を目的としたオントロジー

本手法で用いるオントロジーは，抽出対象となる事物を示す中心概念に対して，意味的な定義を与える関連の強い属性や機能などを表す属性概念や動作を示す動作概念をノードとして持ち，個々の属性概念はさらにインスタンスとなる語を下位ノードに持つ．また，中心概念およ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として，Prop_of，Is_a，Inst_of，Agt_of などを持つ．オントロジーの例を図2に示す．

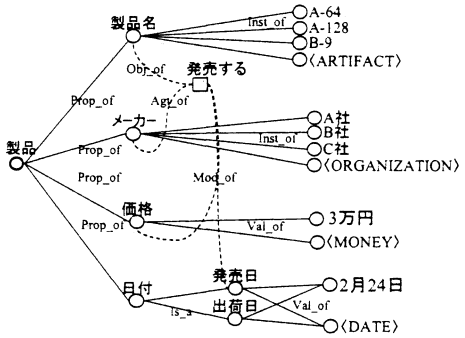

図2 製品に関するオントロジー例

2.2 ODIE システムと固有表現抽出

ODIE システムは，(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 [3]．

本手法では，テキストにおいて主要な情報を表す，組織名や人名，日付や金額といった固有の表現，および，車の記事であれば排気量といった，対象記事の分野に特有の分野用語を抽出対象語と呼び，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に抽出対象語と認識する．このような処理では認識できる単語に限りがあるため，実装にあたっては固有表現抽出技術により，対象範囲の拡張を行なった．

固有表現の抽出には，単純置換に基づく日本語固有表現抽出ツール ProCreator を用いた [4]．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タグを付与する [5]．

このタグの種類に対応するノードを属性概念のインスタンスとして定義し，オントロジーと関連付けた．例えば，〈DATE〉タグを付与された固有表現は図2の

Evaluation of Ontology-driven Information Extraction

Keiichi HIROTA, Yutaka SASAKI and Tsuneaki KATO

NTT Communication Science Laboratories

2-4 Hikaridai,Seika-cho,Soraku-gun,KYOTO,619-0237 Japan

〈DATE〉の位置にリンクされる．これにより，オントロジーでは未知の語彙でも，固有表現であればノードとして扱い，抽出対象語と認識する事ができる．

ODIEシステムは各抽出対象語に対して，オントロジー上での活性伝播により抽出対象事物との関係を得る．活性伝搬の際には，テキスト中での抽出対象語の用いられ方を見る事によるバイアスとして共起バイアス，格バイアスを設け，妥当性の高い情報ほど高い活性値を得るよう活性伝播を制御している．

3 製品発表記事からの情報抽出

ODIEシステムの実装に対し，整備したオントロジーを用いた抽出により手法自体を評価し，次いで新規記事集合を用いて固有表現抽出による拡張性を評価した．

3.1 実験1：整備したオントロジーを用いた抽出

まず，CD-毎日新聞94版1年分から製品発表記事250件を対象に，主要情報の抽出実験を行なった．オントロジーの構築は人手で行ない，主要な情報項目を属性概念として中心概念に関連付け，固有表現中心に記事中の単語を収集して属性概念のインスタンスとした．

抽出結果から，製品名，メーカー，価格，発売日の四項目について，再現率と適合率を評価した．結果を表1に示す．

表1 主要情報の抽出実験（再現率・適合率）

抽出項目		製品名	メーカー	価格	発売日
再現率	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%
適合率	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%

3.2 実験2：新規記事集合からの抽出

新たに毎日新聞記事95年版から77記事を対象として，先の実験で作成したオントロジーを用いて抽出を行なった．形態素解析のみを行なった記事集合と，固有表現抽出を行なった記事集合に対し，実験1と同じ四項目について再現率と適合率を評価した．なお，本実験では共起，格の両バイアスを用いた．結果を表2に示す．

3.3 結果と考察

実験1では，表1のように，再現率で78% ～ 91%，適合率で70% ～ 88%と，ともに平均80%前後の高い結果を得た．また，共起バイアス，格バイアスを用いる事で，再現率を大きく下げる事なく，適合率を向上させる事ができた．価格の欄での再現率の低下は，複数の正解値に対し，最大の活性値を持つ値のみを抽出するためで，このような複数正解の抽出は今後の課題である．

表2 新規記事集合からの抽出実験（再現率・適合率）

抽出項目		製品名	メーカー	価格	発売日
形態素解析	再現率	2.4%	54.4%	77.4%	83.1%
	適合率	5.9%	76.8%	85.7%	68.6%
固有表現抽出	再現率	64.6%	78.5%	78.5%	87.3%
	適合率	52.5%	56.9%	86.9%	72.1%

一方，実験2では，表2から明らかなように，固有表現抽出による単語の認識が有効であり，特に全くの未知の単語である事が多い製品名に対して大きく向上し，他の項目でも精度が向上した．

4 まとめ

本手法はオントロジーを整備した状態で再現率・適合率ともに高い評価値を得，また未知の単語を多く含む新規記事集合からも十分な精度による抽出を実現し，その有効性を確認できた．

本実験においてオントロジー構築は高々2人日の労力で済んでおり，固有表現抽出により未知の単語の処理も十分な精度で行なえる事から，新しい対象分野への移行は容易であると考えられる．

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の抽出や，オントロジーの交換による対象分野の容易な変更についての検討があげられる．また，実装上の問題として，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，精度の向上も課題である．

本実験では，CD-毎日新聞94年版および95年版を利用した．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] 廣田啓一，佐々木裕，加藤恒昭: オントロジー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).

[2] 廣田啓一，佐々木裕，加藤恒昭: オントロジー主導による情報抽出，人工知能学会誌，Vol.14，No.6，pp.1010–1018 (1999).

[3] 廣田啓一，佐々木裕，加藤恒昭: オントロジー主導による情報抽出の検討，情報処理学会研究報告，99-NL-133，pp.85–92 (1999).

[4] 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).

[5] 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).