オントロジ主導による情報抽出手法の提案

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

1　はじめに

近年、インターネットの飛躍的な発達に伴い、新聞記事やネットニュース、電子掲示板、電子メール等の電子化された情報が大量に流通している。このような電子化情報の洪水の中から、利用者にとって有益かつ必要な情報を抽出し、情報を整理して提供するシステムの必要性が強く論じられている。[1]

このような情報提供システムにおける情報抽出機構の構築にあたり、我々は、オントロジ主導によりテキストからの情報抽出を行なう、新しい手法を提案する。本手法は、与えられた中心語を元に組み立てられたオントロジを用いて、テキスト中の主要情報を表す単語を認識し、さらにオントロジ上での活性伝播により中心語と個々の単語との関係を得て、これを抽出項目名（スロット名）と値に変換するものである。なお、中心語とは、情報抽出の研究においては、特定の分野の特定のトピックのテキストが与えられることが前提とされているが、その与えられたテキストのテーマを代表して表す語のことである。例えば、モデムについて書かれた記事から情報抽出を行なう場合には、中心語は「モデム」となる。

本稿では、提案手法について述べるとともに、製品発表記事から主要情報を抽出する実験の結果を通じて、本手法の長所と今後の課題を明確にする。

2　従来の情報抽出手法

従来の情報抽出手法は、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）抽出ルール（主要情報とその周辺の単語の出現パターンを表した規則）を用いて、テキストに対して抽出ルールによるパターンマッチングを行ない、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ーンマッチング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]

しかし、従来の手法には次のような問題点がある。

まず、抽出ルールは予め準備した範囲の項目しか抽出できない。また、抽出ルールは対象とする分野の文書を検討して作成するため、文書のトピックや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。

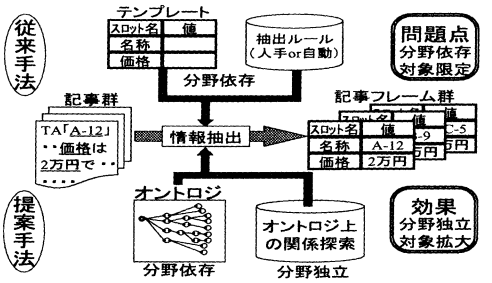

図1　従来手法と提案手法

さらに、対象分野毎に各テンプレートに対応した適切な抽出ルールを人手により作成するには、システムを良く理解した専門家の時間と労力を必要とする。また、抽出ルールを自動あるいは半自動的な手法によって作成する場合でも、既定のテンプレートの抽出項目に対応する十分な量の文書とそれに対する正解例を準備する必要があり、やはり人手による時間と労力を要する。

我々は、このような問題点を考慮した上で、処理の対象とするテキストの分野に対して独立であり、かつ抽出の対象とする情報に制限を受けない手法として、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の両方をテキストから抽出する、新しい手法を提案する。（図1下部）

以下、本手法で用いるオントロジの定義と、手法の詳細について述べる。

3　本手法で用いるオントロジ

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。工学分野においては、「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」として扱われる。[4]

本手法で用いるオントロジを、テキストの分野におけ

Ontology-driven Information Extraction
Keiichi HIROTA, Yutaka SASAKI and Tuneaki KATO
NTT Communication Science Laboratories
2-4 Hikaridai,Seika-cho,Soraku-gun,KYOTO,619-0237 Japan

る中心的な対象を表した語（以下、中心語と呼ぶ）に対して、中心語の意味的な定義を与える概念の体系と各概念に対してその実例となる単語の集合を記述したものと規定する。中心語をターミナルアダプタ（以下、ＴＡと記載）とするオントロジの例を図2に示す。

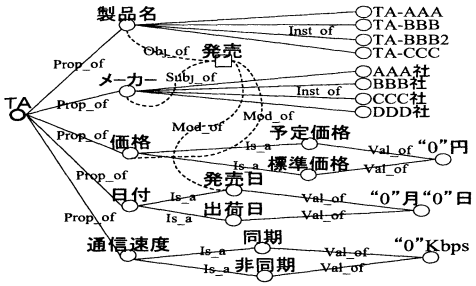

図２　ＴＡに関するオントロジ例

3.1　属性概念

中心語に対して意味的な定義を与える、中心語に関連の強い概念的な用語を属性概念と呼ぶ。また特に動作を示す属性概念を動作概念と呼ぶ。本オントロジにおける、中心語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): 中心語に対する属性概念

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

Subj_of(subject_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。例えば「メーカー」という属性概念は、AAA 社、BBB 社等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。金額や日付など単位を伴う具体的な数値であ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。

属性概念及びそのインスタンス、数値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。

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

3.3　言語表現との対応付け

オントロジ中の各ノードは、そのノードが表す概念を表現する語群と関係付けられている。これらの語を抽出対象語と呼ぶ。テキストにおいて、抽出対象語は主要情報を表す単語である。

4　提案手法の概要

本節では、提案する情報抽出手法の基本手法をのべる。次節で、基本手法に対する改良法を与える。基本手法は、(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語との関係列の獲得、(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。

4.1　抽出対象語の認識

対象となるテキスト中に出現する抽出対象語を認識し、その出現回数に応じた活性値をオントロジ上の対応するノードに与える。また、Val_of リンクの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、値ごとに別々のノードを生成して活性値を与える。

4.2　オントロジ上での活性伝播による情報関係の獲得

各抽出対象語に対し、オントロジ上でノード間の関係を辿り、中心語との関係を獲得する。本手法ではこれを活性伝播を用いて行なう。

抽出対象語を表すノードからリンクさ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を結合した関係列を作る。

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、リンクと上位ノードを結合した関係列を延ばしていく。途中、ノードが複数のリンクを持つ場合は、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。これを、中心語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ルートノードから最下位ノードへのパスが複数生成される。このパスを情報関係列と呼び、ルートノードに到達した活性値を、その情報関係列の活性値とする。

4.3　関係列の解釈と選択

形成された情報関係列は、一般に中心語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。この情報関係列において、どこまでが属性（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、中心語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。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、どのようなリンクが属性（項目名）と値の境界となりやすいかというリンクの関係子の分割性を定義し、これを指標として属性表現と値表現を分離する。

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/Subj_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。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、より中心語に近い関係子の位置で二分する。

このような方法により、項目名と値を得る方法を単純な活性伝播による方法と呼ぶ。単純な方法での情報関係列は、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に重複して現れたり、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選択する必要があり、基本手法では、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、最大の活性値を持つ関係列からの情報を抽出する。最大の活性値を持つものが複数ある場合には、その両方を抽出する。

5　活性伝播に対するフィルタ

前節で述べた単純な活性伝播では、抽出対象語の出現回数だけでノードの活性値が定まり、それが伝播された。これに対し、抽出対象語のテキストでの用いられ方を利用したヒューリスティクスに基づく活性伝播の制御を考え、次のような二つのフィルタを設けた。

(1) 複数上位ノードを持つ場合のフィルタ

ある抽出対象語が複数の上位ノードを持つ事は、その語が複数の属性の値と成り得る事を意味する。しかし、テキスト中の抽出対象語はその内の一つにしか成り得ず、属する上位ノードを決定する必要がある。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に現れた時、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る。

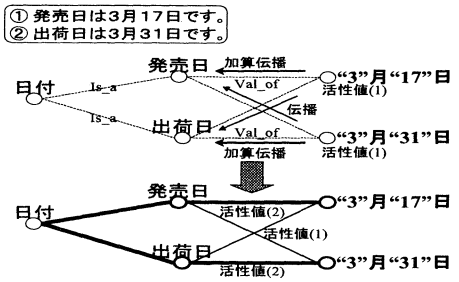

図３　複数上位ノードを持つ場合のフィルタの作用例

例えば図3において、「3月17日」と「3月31日」はVal_ofリンクで、上位ノード「発売日」と「出荷日」に結ばれている。至近に「発売日」があるため「3月17日」から「発売日」へ活性値を１加算して伝播し、他のノード「出荷日」への伝播を禁止する。同様に「3月31日」からは「出荷日」にのみ伝播する。その結果「3月17日」と「発売日」、「3月31日」と「出荷日」からなる関係列は、それぞれ高い活性値を持つ。

(2) 動作概念との関係によるフィルタ

テキスト中に抽出対象語が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。

ある語に対応するノードの上位ノードがSubj_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おり、その語がこの動作概念に対応する用言に対し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関連の強さを反映する。逆に異なる主題役割を持つ場合活性値を１減算して伝播する。

テキスト中の語の主題役割の判定は助詞の種類と近傍の用言から近似的に行ない、主体／対象／関係／その他の４つに分類してオントロジ中の役割と比較した。なお、その他に判定された語は、主題役割の比較の結果によらず、活性値を増減させる事なく伝播する。

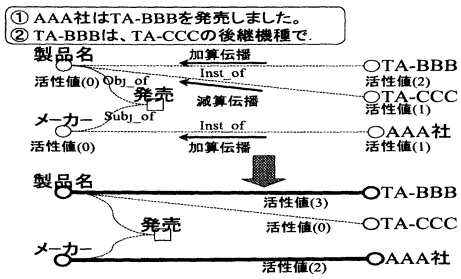

図４　動作概念との関係によるフィルタの作用例

例えば図4において、「製品名」はObj_ofリンクで「発売」と結ばれている。すなわち「製品名」は「発売」の対象である。この時、「製品名」のインスタンスとなる「TA-BBB」の主題役割は、テキスト中において「発売」の対象であるため、活性値を１加算して伝播し、逆に「TA-CCC」からは１減算して伝播する。その結果、「TA-BBB」は「TA-CCC」よりも高い活性値を持つ関係列を形成する。

6　製品発表記事からの情報抽出実験

ターミナルアダプタ（以下、ＴＡと記載）に関する製品発表記事５６記事に対し、抽出対象語となる記事中の単語を収集して、ＴＡに関するオントロジ上の個々の属性概念のインスタンスとし、情報抽出実験を行なった。

評価にあたっては、従来の抽出手法において主な抽出対象として用いられる、製品名、メーカー、価格、発売日、およびＴＡに特有の情報である通信速度の五項目についての正解例を作成し、再現率と適合率による評価を行なった。結果を図5に示す。

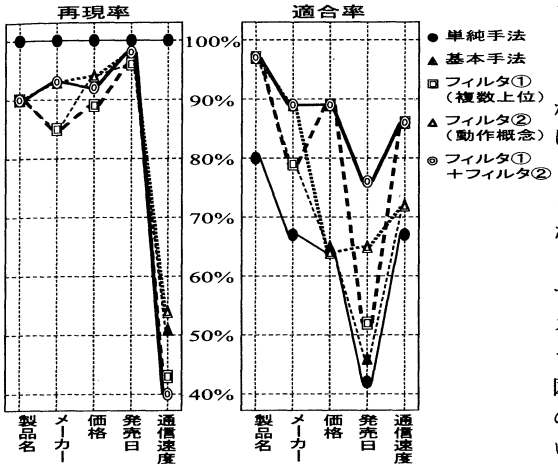

図５　主要情報の抽出実験評価

7　結果と考察

まず、単純な活性伝播による方法では、候補として可能な情報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９０％前後、候補数が多い通信速度で５０％と下がった一方で、製品名・メーカーに関して適合率の向上が見られた。

さらに活性伝播に対するフィルタの導入により、抽出対象語の振る舞いが正しい情報関係列ほど活性値が高くなり、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。フィルタの作用を個別に見ると、複数上位ノードを持つ場合のフィルタで価格と通信速度、動作概念との関係によるフィルタでメーカーの項目が向上し、両フィルタの共用で発売日の項目がさらに向上した事から、両フィルタが相互に適切に作用している事がわかる。

本手法は一部項目を除いて９０％を越える高い再現率を得、活性伝播におけるフィルタの導入により適合率においても平均して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[3]における製品情報抽出実験の評価値と比べても遜色ない結果が得られ、手法の有効性がうかがえる。

その一方で、幾つかの課題も明らかになった。

まず、本手法は活性値によって抽出する関係列を選択しているが、実際は複数の正解がある場合にも活性値の高い関係列しか抽出せず、通信速度の項目に見られるように再現率を低めている。このような、抽出項目に対する抽出値の数の制限は大きな課題である。

一方、適合率において発売日の項目が依然低い理由として、項目名が確定できない抽出対象語と未記述の情報との関係が挙げられる。例えば、書誌情報としてテキスト冒頭に日付があり、本文中に発売の日付に関する記述がない場合、本手法はこの日付を発売日の項目に結び付けて出力する。この判断は人間にとっても難しく、正解とも誤りとも言えない。このような不確定な情報の除去として、活性値に閾値を設けて、確実と判断できる情報だけを抽出する事を検討している。

また、本手法のみならず情報抽出手法全体の問題として、複数製品の区別という課題がある。本手法は、テキスト中の全ての抽出対象語に対し、同じ一つのオントロジ上で中心語との情報関係を求めるため、個々の情報を区別して抽出できない。このような情報を区別した処理のためには、個々の製品に対して個々のオントロジを用い、共通の属性、個別の属性を区別して、並行的に活性伝播を行なうような、手法の拡張が必要である。

8　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、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、問題点としてあげた、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。

参考文献

[1] 八巻他「知識プロバイダ構想の提案」, 第55回情処全大,3AF-2,1997.
[2] 松尾、木本「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」, 情処論,Vol.36,No.8,1995.
[3] 井出他「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」, 自然言語処理,122-10,1997.
[4] 溝口、池田「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.4,1997.