オントロジー構築のための文書からの意味関係抽出

長野 伸一　　溝口 祐美子　　稲葉 真純　　服部 正典
(株) 東芝 研究開発センター

1 はじめに

近年，オントロジーと呼ばれる概念辞書の体系化やデータの整備が進みつつある．WordNetは，汎用の概念辞書として最も著名なものの1つであり，言語処理や知識処理の基盤とされてきた．その日本語版*が公開されたことにより，日本語文書を対象とした研究が進みつつある．また，不特定多数のインターネット利用者により開発された知識ベースであるWikipediaから，汎用Webオントロジーとして構築されたものととして，YAGO[1]，DBpedia[2]などがある．特定の領域向けとしては，例えば生命科学分野での遺伝子オントロジー†などがある．このように，少しずつではあるが着実にオントロジーが整備されてきており，既存のオントロジーを再利用したり，拡張することにより，目的のオントロジーを構築することが可能となってきている．

一方，企業内では，汎用のオントロジーとは異なる用語が用いられており，文書処理を行うための最初の用語体系を，どのように構築するかが課題となっている．初期のオントロジー構築に限らず，構築後の保守までを考慮すると，全てを人手で実施することは大きなコストとなってしまう．また，人手の作業では，オントロジーの品質が均質化されない．そのため，文書からのオントロジー獲得に関する研究が盛んに行われている[3]．

本研究は，企業内文書を対象として，オントロジーの構築・保守に係る品質，コストを制御可能とするオントロジー自動構築技術の確立を目的とする．本稿では，その基本技術として，文書データから意味関係抽出を行う手法を提案する．提案法は，あらかじめ定義した少量の概念対をシードとして，指定した意味関係にある概念対を文書から抽出する．簡単な実験により，提案法の有用性を評価する．

2 提案手法

提案法の概要について述べる．提案法は，あらかじめ定義しておいた概念対をシード(種)として，機械学習技術を用いて大量の文書データから概念および概念間の関係を抽出し，オントロジーを構築する．提案法の特徴は，ブートスラップと呼ばれる方式を採用している点にある．まず，シードとして与えた概念対が共起する構文木を抽出し，概念対の出現を変数に置き換えたものを関係抽出パターンとして獲得する．次に，獲得した関係抽出パターンと，文書中の各文とを，構文木レベルでのパターンマッチングを行い，変数に対応する概念の組を概念対として抽出する．この2つの処理を交互に繰り返し実行することにより，概念対を順次獲得する．なお，文書から関係抽出パターンを抽出する際に，そのパターンが有用な概念対を獲得しうるか否かを分類器を用いて判定し，正例と判定されたパターンだけを取得する．分類器の作成には，構文木の文節を素性とするベクトルを用いる．正例負例の正解ラベルには，シードとして与える概念対のラベルを用いる．

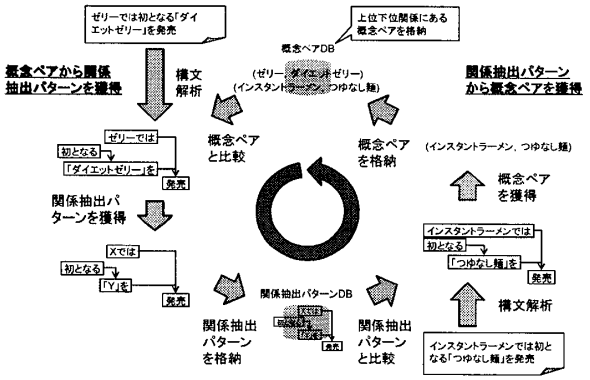

図1: 概念対と関係抽出パターンの獲得の例

提案法による概念対の獲得の例を，図1に示す．本例は，食品・飲料メーカーによるプレスリリース文書から，商品分類，商品名に関する上位下位関係にある概念対を抽出している．シードとして，概念対(ゼリー，ダイエットゼリー)が与えられていると仮定する．上位概念がゼリー，下位概念がダイエットゼリーである．文書から文『ゼリーでは初となる「ダイエットゼリー」を発売」が得られたと仮定する．この文にはシードの概念対が含まれており，両概念それぞれを変数X,Yに置換した構文木を関係抽出パターンとして獲得する．次に，このパターンを用いて他の文とのパターンマッチ

Semantic Relation Extraction from Documents for Ontology Learning
Shinichi NAGANO, Yumiko MIZOGUCHI, Masumi INABA, Masanori Hattori
Corporate R&D Center, Toshiba Corporation
*http://nlpwww.nict.go.jp/wn-ja/
†http://www.geneontology.org/

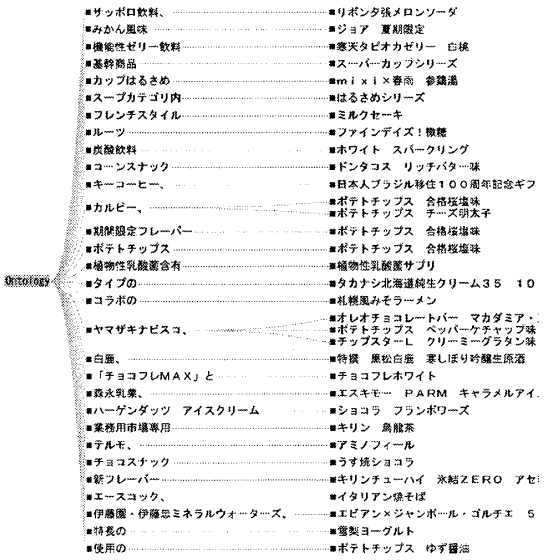

図 2: 意味関係の抽出の例

ングを行う．今，文書から文『インスタントラーメンでは初となる「つゆなし麺」を発売』が得られたと仮定する．この構文木は，上記の獲得パターンと構造が一致するため，変数 X,Y それぞれに対応する概念の組 (インスタントラーメン，つゆなし麺) を概念対として獲得する．食品・飲料メーカーによるプレスリリース文書から抽出した概念対を図 2 に示す．左のノードが上位概念，右のノードが下位概念を表す．

3 評価実験

3.1 実験概要

提案法の有用性を検証する評価実験について述べる．食品・飲料メーカーによるプレスリリース文書 300 件を題材として，商品分野，商品名に関する上位下位関係にある概念対の抽出を行う．文書データには，上位下位関係の正解情報を付与しておき，抽出した概念対に関する適合率，再現率により精度評価を行う．シードとして与える概念対は，数を 10 から 50 まで 10 ずつ増やした 5 つのグループを用意した．選んだシードの質による精度評価の偏りを防ぐため，各グループについて異なる 3 セットのシードを作成し，精度の平均を求める．なお，構文解析には京大の KNP を利用し，辞書はデフォルトのままとした．

3.2 実験結果

精度評価の結果を表 1 に示す．適合率は，概ね 0.2 から 0.4 の範囲であることが確認された．シード数 10 の場合，結果が良好であるが，選択したシードの偏りによるものと思われる．獲得した関係抽出パターンを見ると，比較的単純なものが多く見られた．例えば，パターン『X,「Y」を発売。』は，商品分野と商品名を組とした，シードの概念対から獲得されたものであるが，このパターンからメーカー名と商品名の組も獲得された．このように，誤った意味関係にある概念対も獲得してしまったため，適合率が低下した．提案法は，文の表層だけを見て，同じ構造を持つパターンを一様に扱っているが，文の意味情報を活用してパターンを分類するなどの改良が必要である．

表 1: 抽出精度の評価結果

シード概念対	10	20	30	40	50
適合率	0.806	0.383	0.231	0.288	0.343
再現率	0.004	0.006	0.009	0.012	0.018

一方，再現率は，シード数が増えるにつれて向上することが確認されたものの，獲得数が非常に少なく，極めて悪い結果となった．主な原因として，文書には商品名に関する複合語，未知語が多くあり，構文解析により意図しない形態素や文節が得られたことにある．例えば，商品名「一平ちゃん夜店の焼そば」は，3 つの形態素「一平ちゃん」,「夜店の」,「焼そば」として解析されてしまった．本研究では，概念辞書の自動構築を目的としており，概念獲得の対象である文書を解析するために，構文解析の辞書を事前に整備しておくことはできない．用語の出現に関する統計的指標を用いた複合語抽出を行うなどの改良が必要である．

4 まとめ

オントロジー構築のための文書からの意味関係抽出に関する手法を提案し，簡単な評価実験により，提案手法の有用性の一端を確認した．適合率，再現率ともに実用レベルには未だ遠く，今後は精度向上に向けた方式改良を進める．

参考文献

[1] F.M. Suchanek, et al., Yago - a core of semantic knowledge, Proc. of WWW, 2007.

[2] C. Bizer, et al., DBpedia - a crystallization point for the web of data, Journal of Web Semantics, vol.7, no.3, pp.154-165, 2009.

[3] F. M. Suchanek, et al., Combining linguistic and statistical analysis to extract relations from web documents, Proc. of KDD, 2006.