Vol.2009-NL-194 No.4
2009/11/16

日本語語彙大系を用いたWikipediaからの汎用オントロジー構築

柴木優美[†1]　永田昌明[†2]　山本和英[†1]

日本語語彙大系を上位階層として，日本語 Wikipedia から is-a 関係のオントロジーを半自動で構築する手法を提案する．はじめに，語彙大系の末端の意味属性に，分類基準が同じ Wikipedia のカテゴリを半自動で対応づける．次に，対応づけされた Wikipedia のカテゴリより下位のカテゴリから，語彙大系の知識を利用して自動で is-a 関係の階層構造を構築する．最後に，カテゴリに所属する記事の見出し語をインスタンスとして抽出する．構築した is-a 関係の階層構造からカテゴリを取り出し，その親カテゴリとの is-a 関係が成り立つかどうか，また先祖のカテゴリ全てで is-a 関係が成り立つかどうかをサンプル調査した．その結果，適合率はそれぞれ 92.8 %，82.6 %であった．また，インスタンスの適合率は 98.6 %であった．Wikipedia のカテゴリ 49,543 件のうち 23,289 件（47 %），記事の見出し語 479,231 件のうち 263,631 件（55 %）をオントロジー化することに成功した．本手法により，Wikipedia から高精度で大規模な is-a 関係の汎用オントロジーを構築することができた．

Construction of General Ontology from Wikipedia using a Large-Scale Japanese Thesaurus

Yumi Shibaki,[†1] Masaaki Nagata[†2]
and Kazuhide Yamamoto[†1]

We present a semi-automatic method to construct a generic, large-scale is-a ontology from the Japanese Wikipedia using a Japanese thesaurus, Nihongo Goi-Taikei, as its upper ontology. First, the leaf categories of the Nihongo Goi-Taikei are manually aligned with the Wikipedia categories. Then, their subcategories are made automatically by extracting is-a relations from the Wikipedia category network. Finally, the titles of the articles belong to each Wikipedia category are extracted as its instance. Sample evaluation shows that the precision of immediate is-a relation and that of all is-a relations up to the root are 92.8% and 82.6%, respectively. The precision of instance is 98.6%. From the Wikipedia, 47% categories and 55% article titles are used in the resulted ontology by this method.

1. はじめに

Wikipedia は即時更新性，語彙網羅性に優れた自由に利用できるオンライン百科事典である．Wikipedia は，その知識量の多さと半構造化された文書構造が有用とされ，近年自然言語処理の分野で幅広く利用されている．しかし Wikipedia のカテゴリ階層構造は，カテゴリ間の意味関係やカテゴリの分類基準が厳密に定義されていないため，そのままではオントロジーとしては利用しにくい．そこで本研究では，日本語 Wikipedia から自動で汎用オントロジーを構築することを目的とする．しかし，明確な分類基準の指針がない状態で，分類基準の一貫性をもった大規模なオントロジーを構築するのは難しい．そこで，本手法では既に人手で作成されている日本語語彙大系 1) を上位階層として用いることにした．

以下に，Wikipedia を利用した知識抽出に関する研究について紹介する．

隅田ら 2) は，Wikipedia の記事構造を利用して is-a 関係[⋆1]の単語ペアを獲得する研究を行なっている．この単語ペアは高精度で大規模だが，それぞれの単語ペア同士は独立しているため，日本語語彙大系のような体系的な分類基準は存在しない．オントロジーを構築する研究ではないが，清田ら 3) は図書分類体系と Wikipedia を統合し，情報探索に利用することを提案している．本研究では日本語語彙大系をベースとすることにより，Wikipedia の知識を自然言語処理に応用することが容易になると考えている．

Ponzetto et al. 4) は，英語 Wikipedia のカテゴリ階層から is-a 関係と not-is-a 関係の階層構造を抽出する手法を提案している．桜井ら 5) は，Ponzetto らの手法の一部を利用した手法に独自の手法を加え，日本語 Wikipedia に対し，カテゴリ階層から is-a 関係のオントロジーを抽出する手法を提案している．これらの手法で構築されるオントロジーは，1 つの階層ではなく，いくつもの独立した階層の集合からなっている．本手法は，日本語語彙大系を利用してこれらを 1 つの階層に統合している点で異なる．

Suchanek et al. 6) は YAGO において Wikipedia のカテゴリを WordNet のクラス（synset）の下位クラスとして統合することにより，高精度なオントロジー構築を試みてい

†1 長岡技術科学大学 電気系
Department of Electrical Engineering, Nagaoka University of Technology
†2 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

⋆1 “ is-a 関係 ”とは，B is a (kind of) A が成り立つときの A と B の関係をいう．

る．小林ら 7) は YAGO とは異なる手法で，日本語語彙大系と Wikipedia を統合する手法を提案している．彼らは語彙大系の意味属性に対して Wikipedia のカテゴリを割り当てることで，意味属性の 1 つ下位までの階層構築を行なっている．本手法では Wikipedia の階層構造を利用することにより，統合した部分からさらに is-a 関係の階層構造を構築していく点で異なっている．

2. 本研究で使用する言語資源

2.1 日本語語彙大系

日本語語彙大系（以下, 語彙大系）は，日本語約 30 万語を約 3,000 種類の意味属性で分類した日本最大級のシソーラスである．語彙大系には，約 2,700 の語彙大系クラス[*1]と約 10 万の語彙大系インスタンスを持つ一般名詞の意味体系（**図 1**）と，約 130 の語彙大系クラスと約 20 万の語彙大系インスタンスを持つ固有名詞の意味体系が含まれている．両者は別々の意味体系であるが，固有名詞のクラスは一般名詞のクラスに対応づけされている．意味体系では，多義性がある語彙大系インスタンスはいくつかの語彙大系クラスに分類されている．

2.2 日本語 Wikipedia

Wikipedia は自由に利用できる大規模な Web 百科事典であり，Web 上で XML 形式のダンプデータが公開されている[*2]．Wikipedia のページとそのソーステキストを**図 2**，**図 3** に示す．図のように，見出し語と説明文，その見出し語が分類されるカテゴリが書かれているページを「記事ページ」と呼ぶこととする．記事ページでは，説明文の第一文は見出し語の定義文であることが多い．Wikipedia はページを分類するためのカテゴリがあり（以下,Wikipedia カテゴリ），各記事ページはいくつかのカテゴリに所属している．このカテゴリは主要カテゴリと呼ばれる 9 種類のカテゴリを最上位とした階層構造となっている．この階層構造では 1 つのカテゴリに対し親カテゴリが複数存在することが多く，循環もある．カテゴリ間の関係は多様だが，下位の階層になるほど分類はより具体的になり is-a 関係になりやすい傾向にある．Wikipedia にはカテゴリのページも存在し，所属する記事ページの見出し語の一覧が表示される．

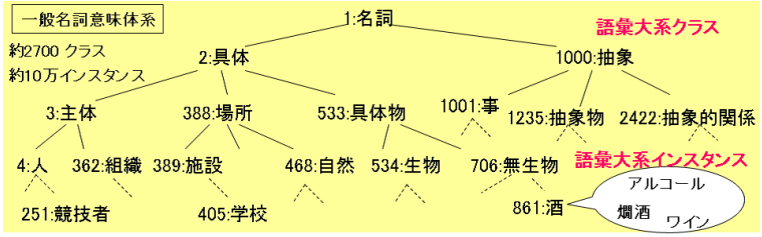


図 1 日本語語彙大系

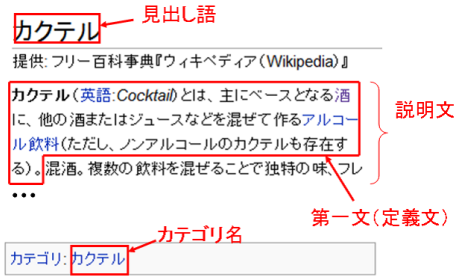


図 2 記事ページのスクリーンショットの一部

```
<title>カクテル</title>
'''カクテル'''（[[英語]]:"Cocktail"）とは、主にベースとなる[[酒]]に、他の酒またはジュースなどを混ぜて作る[[アルコール飲料]]（ただし、ノンアルコールのカクテルも存在する）。・・・

[[category:カクテル|*]]
```

図 3 記事ページのソーステキストの一部

3. オントロジー構築手法

本手法では，以下の手順によりオントロジーを構築する．（各手順は**図 4** に対応する．）

手順 1） 末端の語彙大系クラスに，同じ分類基準を持つ Wikipedia カテゴリ（以下，接点カテゴリ）を半自動で対応づける．

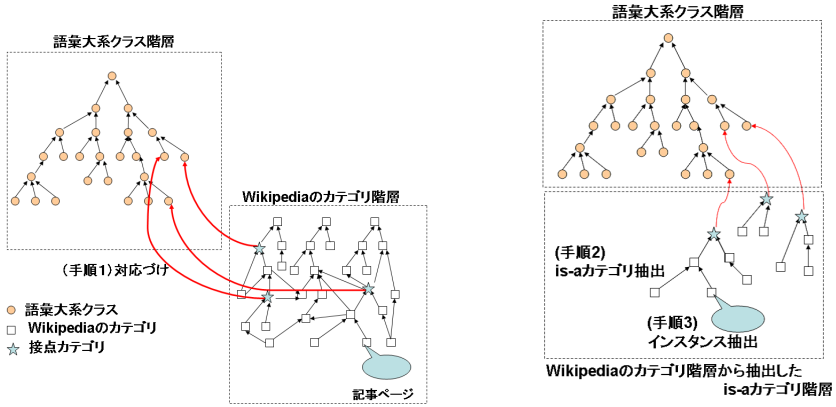
手順 2） 接点カテゴリより下位の Wikipedia のカテゴリ階層から，is-a 関係になっている Wikipedia カテゴリ（以下，is-a カテゴリ）を自動抽出する．

手順 3） is-a カテゴリに所属する記事ページの見出し語からインスタンスとなるものを自動抽出する．

3.1 手順 1：Wikipedia カテゴリの対応づけ

末端の語彙大系クラス 1,921 件に，分類基準が同じ Wikipedia カテゴリ（接点カテゴリ）を人手で対応づける．ただし，51,284 件の Wikipedia カテゴリの中から必要なカテゴリを人手で選ぶのは困難なため，はじめに自動でいくつかの候補を列挙する．

⋆1 本稿では，日本語語彙大系の意味属性を“ 語彙大系クラス ”，分類された単語を“ 語彙大系インスタンス ”と呼ぶ．
⋆2 http://download.wikimedia.org/jawiki/



(a)Wikipedia カテゴリの対応づけ　　(b)is-a カテゴリとインスタンスの抽出

図 4　オントロジー構築手法の概要

3.1.1　接点カテゴリ候補の自動抽出

語彙大系クラスと分類基準が同じ Wikipedia カテゴリの候補を抽出するため,「分類基準が同じものは分類名も同じである可能性が高い」という仮説と「分類基準が同じものは,同じインスタンスまたは下位クラスを持つ」という仮説をもとに,接点カテゴリの候補(以下,接点カテゴリ候補)を自動的に抽出する規則を作成した.Wikipedia カテゴリが以下の候補抽出規則のいずれかに当てはまった場合,そのカテゴリを接点カテゴリ候補として抽出する.

規則 1)語彙大系クラス名と Wikipedia カテゴリ名が完全一致する.

規則 2)語彙大系インスタンス名と Wikipedia カテゴリ名が完全一致する.

規則 3)語彙大系クラスに所属するインスタンス名 3 件以上が,Wikipedia カテゴリの「所属する記事ページの見出し語 3 件以上[⋆1]」または「下位カテゴリ名 3 件以上」と完全一致する.

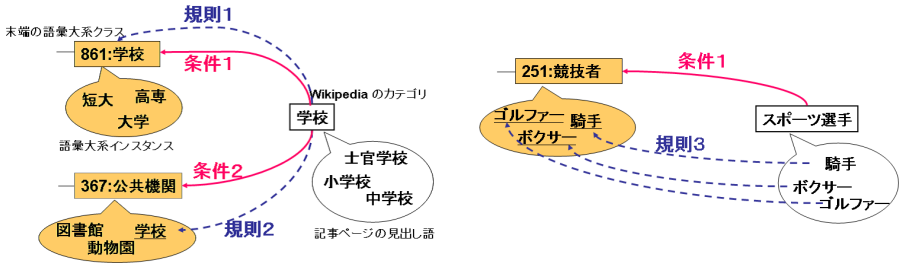
図 5 に,規則別の接点カテゴリ候補の抽出例を示す.

3.1.2　人手による接点カテゴリの選択

本手法では,接点カテゴリ候補のうち以下の 2 つの条件のどちらかを満たすものを接点カテゴリであると定義する.図 5 に人手抽出される条件の例を示す.

条件 1)語彙大系クラスと Wikipedia カテゴリの分類基準が同じ.

条件 2)語彙大系クラスのインスタンスを語彙大系の下位クラスと考えたとき,このクラスと Wikipedia カテゴリの分類基準が同じ.



(a) 候補抽出規則 1,2　　(b) 候補抽出規則 3

図 5　接点カテゴリ抽出例

分類名が同じでも分類基準が異なるのは,分類名が多義である場合が多い.例えば,Wikipedia カテゴリ〈ロケット〉に対して語彙大系クラス [乗り物(本体(移動(空圏)))] と [装身具] が接点カテゴリ候補として抽出されるが,Wikipedia カテゴリ〈ロケット〉は乗り物のロケットを指すので,語彙大系クラス [装身具] とは一致しない[⋆2].また同じインスタンスがあっても分類基準が違う場合もある.語彙大系クラス [平面図形] はインスタンス《正方形》,《三角形》,《楕円》を含むが,明らかに分類基準の異なる Wikipedia カテゴリ〈初等数学〉にも同じ見出し語の記事ページが所属している.Wikipedia カテゴリには,必ずしもカテゴリ名と is-a 関係にある記事ページが分類されているわけではないため,このようなことが起こりやすい.これらを自動判定するのは今後の課題とし,今回は精度を重視し接点カテゴリの選択は人手で行なった.

3.2　手順 2:is-a 関係の Wikipedia カテゴリ抽出

3.2.1　従 来 手 法

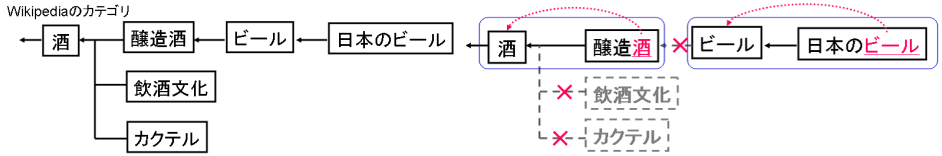
本節ではまず,日本語 Wikipedia のカテゴリ階層から is-a 関係の Wikipedia カテゴリ(is-a カテゴリ)を抽出する従来研究を概説し,次に本研究で提案する手法を述べる.

⋆1 予備調査により,一致数を 3 件以上とした.

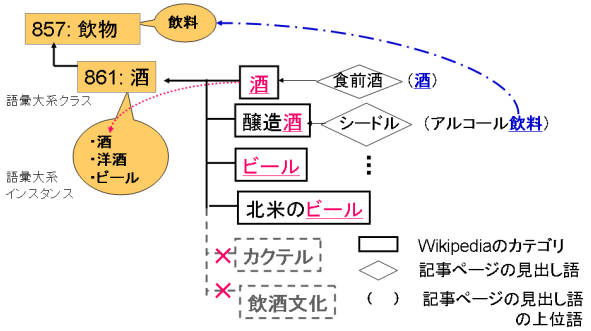
⋆2 本稿では,語彙大系クラス名は [ ],Wikipedia カテゴリ名は〈〉,インスタンス名は《》で囲む.

桜井ら 5) は,「後方文字列照合」と「前方文字列照合部除去*1」という手法により，Wikipedia のカテゴリ階層から is-a カテゴリの階層を抽出している.「後方文字列照合」とは，Wikipedia の上位カテゴリに対しその下位カテゴリ名が“ 任意の文字列＋上位カテゴリ名 ”であったとき両者は is-a 関係であるとする手法である．例えば，Wikipedia カテゴリ〈酒〉の下位カテゴリに〈醸造酒〉が存在した場合，両者は is-a 関係と判定される．**図 6**（a）のような Wikipedia のカテゴリ階層が与えられた場合に，桜井らの手法で構築される is-a カテゴリ階層を図 6（b）に示す．桜井らの手法では，2 つの孤立した階層が抽出される．

小林ら 7) は Wikipedia を利用して，語彙大系クラスより 1 つ下位に接続される is-a カテゴリとそのインスタンスを作成する手法を提案している．桜井らがカテゴリ間の文字列照合のみで is-a カテゴリを抽出していたのに対し，小林らは語彙大系インスタンスや，記事



(a)Wikipedia のカテゴリ階層　　(b) 桜井らの手法による is-a カテゴリ階層



(c) 小林らの手法による is-a カテゴリ階層

図 6　従来手法での is-a カテゴリ階層構築の概要

ページの第一文から抽出した見出し語の上位語*2を用いることで，文字列照合する単語数を増やし網羅性を上げている．図 6（a）の Wikipedia カテゴリ階層に対し小林らの手法を適応したときの手法の概要を図 6（c）に示す．はじめに，語彙大系クラスのインスタンス名と後方文字列が一致する Wikipedia カテゴリを is-a カテゴリの候補として接続する．図 6（c）では,〈酒〉,〈醸造酒〉など 5 つの Wikipedia カテゴリが候補となる．次に，is-a カテゴリの候補に所属する記事ページの見出し語のうち，その上位語の後方文字列が語彙大系インスタンス名と一致するものをインスタンスとして抽出する．図 6（c）の例では〈醸造酒〉に所属する記事ページの見出し語《シードル》の上位語“ アルコール飲料 ”が語彙大系インスタンス名《飲料》と一致するため,《シードル》はインスタンスとなる．is-a カテゴリの候補が 1 つ以上インスタンスを持てば, そのカテゴリは is-a カテゴリとなる．

3.2.2　提 案 手 法

本手法では，手順 1 で抽出した接点カテゴリを頂点とし，それより下位の Wikipedia の階層構造を利用して is-a カテゴリ階層を構築する．**図 7** に本手法での is-a カテゴリ階層を抽出する手法の概要を示す．接点カテゴリより下位の Wikipedia のカテゴリ階層には,〈飲酒文化〉のように is-a カテゴリとしてはふさわしくないカテゴリも存在するため，そのまま全てを is-a カテゴリ階層とみなすことはできない．そこで本手法ではカテゴリ名の後方の文字列が,「自身より上位の階層の単語」に一致した場合，そのカテゴリを is-a カテゴリとみなすこととした．この「自身より上位の階層の単語」のことを，本稿では「上位語候補」と呼ぶ．上位語候補は，以下の単語を指す.（例として Wikipedia カテゴリ〈ビール〉の上位語候補を図 7 に示す.）

（1）　末端の語彙大系クラスとその 2 階層上位までのクラス名*3.
（2）　末端の語彙大系クラスとその 2 階層上位までのインスタンス名.
（3）　自身より上位の is-a カテゴリ名と接点カテゴリ名.

これらの操作を接点カテゴリをスタートとして下位のカテゴリ階層に適応していく．途中で is-a カテゴリとされなかったカテゴリがあった場合，それより下位のカテゴリも is-a カテゴリとならない．小林らの手法は記事ページの上位語を利用しているが，本手法では新たに「カテゴリの上位語*4」を設定した．これにより，カテゴリの上位語でも文字列照合を行なうことで，カテゴリ名が異なっていても is-a 関係を抽出できるようにした．

⋆1 「前方文字列照合部除去」という手法に関しては，本手法では使用していないため説明は省略する．
⋆2 抽出方法については 3.4.1 節で述べる．
⋆3 予備調査により，上位語候補は 2 階層上位までが適切と判断した．
⋆4 抽出方法については 3.4.2 節で述べる．

Vol.2009-NL-194 No.4
2009/11/16

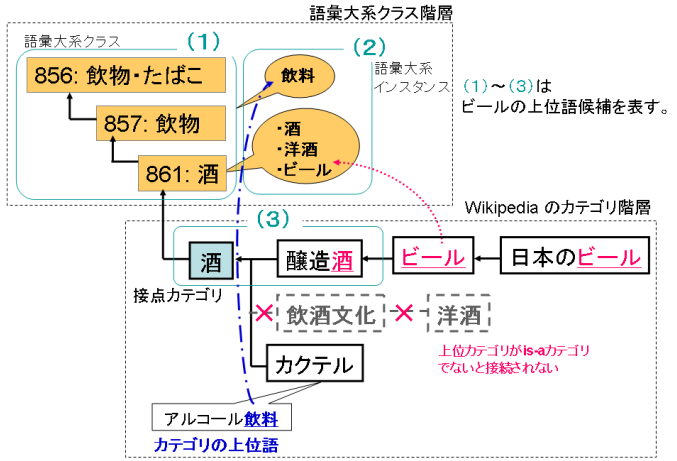


図 7 is-a 関係のカテゴリ抽出手法の概要

最後に，接点カテゴリから下位に構築された is-a カテゴリ階層の中に，同じ語彙大系クラスの他の接点カテゴリがあれば，上位にある接点カテゴリを優先して下位にある接点カテゴリを削除する．

3.3 手順 3：インスタンスの抽出

is-a カテゴリに所属する記事ページの見出し語から，インスタンスとなるものを抽出する．インスタンスの抽出手法は is-a カテゴリの抽出手法と同じである．**図 8** に手法の概要を示す．図 8 のように「is-a カテゴリに所属する記事の見出し語または上位語」の後方の文字列が上位語候補に一致した場合，その所属する記事ページの見出し語をインスタンスとみなす．図 8 の記事ページの見出し語《アースクエイク》,《卵酒》はそれぞれ上位語とカテゴリ名の後方文字列が上位語候補と一致することによって is-a カテゴリ〈カクテル〉のインスタンスとなる．しかし記事ページの見出し語《シェイカー》は見出し語も上位語も上位語候補とマッチしないので，インスタンスとならない．

3.4 Wikipedia からの知識抽出

本節では，本手法で利用する Wikipedia の知識の抽出手法について説明する．

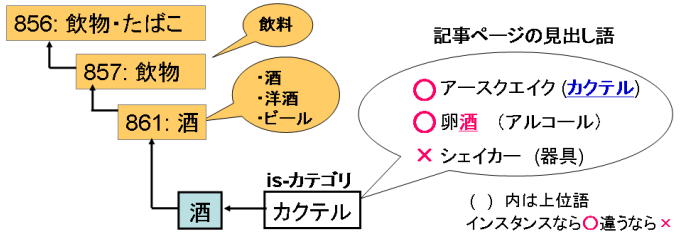


図 8 is-a カテゴリに所属する記事の見出し語からインスタンスを抽出する手法の概要

[上位語] の (一つ | 一種 | 名称 |...)
[上位語](をいい | である |...)
[上位語].

図 9 上位語抽出パターン

3.4.1 Wikipedia の記事ページの知識抽出

本手法で利用する記事ページの知識である「見出し語」,「見出し語の上位語」,「リダイレクト元の見出し語」,「記事ページが所属するカテゴリ名」をタグを利用して抽出する．ただし上位語は Wikipedia の記事ページの定義文（説明文の第一文）から文字列のパターンマッチにより抽出する．隅田ら 2) や小林ら 7) の手法を参考に，独自で上位語抽出パターンを作成し上位語抽出を行なった．この手法を元に作成した上位語抽出パターンの例を**図 9** に示す．

ここで [上位語] は任意の名詞の連続と照合する．例えば，図 3 の定義文からは，見出し語《カクテル》の上位語として“ アルコール飲料 ”を抽出する．図 3 の記事ページからは，上位語“ アルコール飲料 ”が抽出される．リダイレクト元の見出し語とは，記事ページにリダイレクト（転送）するページの見出し語のことである．リダイレクト元の見出し語は同義語や表記ゆれに対応していることが多く (例:《カクテル》と《混合酒》)，文字列照合の際に網羅性を上げられると考える．3.1.1 節の手順 1 の接点カテゴリ候補抽出規則の「記事ページの上位語の見出し語」はこのリダイレクト元の見出し語を含む．

3.4.2 Wikipedia カテゴリの知識抽出

本手法で利用する Wikipedia カテゴリの知識は,「カテゴリ名」,「上位語」,「リダイレクト元の見出し語」,「下位カテゴリ」である．カテゴリのページは記事ページと違いカテゴリを定義するような文が書かれていることは少なく，ページ内から上位語を抽出するのが難

Vol.2009-NL-194 No.4
2009/11/16

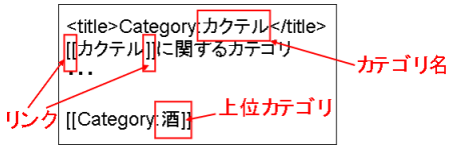


図 10 カテゴリのページのソーステキストの一部

しい．そこで本手法では，カテゴリのページがリンクする記事ページから知識を獲得する．ただし，このリンク先の記事ページは，カテゴリ名の後方の文字列と一致するものとする．例えば，**図 10** の Wikipedia カテゴリのページからリンクする記事ページ《カクテル》の上位語とリダイレクト元の見出し語を，Wikipedia カテゴリ〈カクテル〉の知識として獲得する．これにより，Wikipedia カテゴリ〈カクテル〉の上位語は“ アルコール飲料 ”になり，リダイレクト元の見出し語は“ 混合酒 ”となる．3.1.1 節の手順 1 の接点カテゴリ候補抽出規則の「カテゴリ名」はこのカテゴリのリダイレクト元の見出し語を含む．

4. 実験と考察

4.1 実　　験

2008 年 7 月 28 日時点での日本語 Wikipedia のダンプデータを使用して評価実験を行なった．カテゴリページ数は 49,543 件，記事ページ数は 479,231 件である．本手法では Wikipedia のオントロジーの上位階層に，一般名詞の意味体系を利用する．語彙大系の知識の量を増やすため，固有名詞の意味体系のインスタンスを，対応する一般名詞の意味体系のクラスに追加する．また，語彙大系インスタンスの表記揺れに対応させるため，全てがひらがな（カタカナ）の単語はカタカナ（ひらがな）表記に変換して単語を追加した．例えば，語彙大系インスタンスに《たばこ》は存在しても《タバコ》が存在しなかった場合，同じクラス内に《タバコ》を追加する．

4.2 接点カテゴリ抽出（手順 1）の実験結果

末端の語彙大系クラス 1,921 件とそのインスタンス 108,247 件に対し，3 種類の候補抽出規則を適応させた結果，1 つ以上の接点カテゴリを持つ末端の語彙大系クラスは 719 件（719/1921 = 37.4 %）であった．また全 Wikipedia カテゴリ 49,543 件から 6,301 件の接点カテゴリが候補として抽出され，そのうち人手で接点カテゴリとして抽出されたのは 2,477 件であった．接点カテゴリ候補数と，候補から人手で抽出した接点カテゴリの数を**表 1** に示す．表 1 では，どの規則が接点カテゴリ抽出に有効なのか分かるように，各規則を独立で

表 1 規則別の接点カテゴリ数

規則番号	接点カテゴリ候補抽出規則	接点カテゴリの候補数	人手抽出した接点カテゴリ
1	語彙大系クラス名と Wikipedia カテゴリ名が完全一致する	336	302
2	語彙大系インスタンス名と Wikipedia カテゴリ名が完全一致する	4,310	2,440
3	語彙大系クラスに所属するインスタンス名 3 件以上が Wikipedia カテゴリの「所属する記事ページの見出し語 3 件以上」または「下位カテゴリ 3 件以上」と完全一致する	2,742	713
1-3	規則 1-3 のうち 1 つでも当てはまるもの	6,301	2,477

使用した場合の数値も示している．実際は規則 1〜3 のうち 1 つでも適用された Wikipedia カテゴリが接点カテゴリ候補となる．規則 1 では，候補数のうち 90 %が実際に接点カテゴリとなっているのに対し，規則 2 では 57 %と低い．しかし，全接点カテゴリ 2,477 件のうち 2,440 件（99 %）が規則 2 に適用されていることから，分類基準が同じものは，分類名も同じになりやすいことが分かった．規則 3 は抽出率が 26 %と低い結果となった．規則 3 がないと抽出できない接点カテゴリは，[文具] – 〈筆記具〉[⋆1]，[遊び道具・運動具] – 〈遊具〉など全部で 26 件ある．またリダイレクト元の見出し語との照合を行なわないと抽出できないものは，[物性] – 〈物質の性質〉，[こけ・しだ] – 〈コケ植物〉など全部で 111 件ある．is-a カテゴリ階層構築後はいくつかの接点カテゴリが削除され，最終的には 1,503 件の接点カテゴリが得られた．

4.3 is-a カテゴリ階層構築（手順 2）の実験結果

本手法では，Wikipedia カテゴリ 49,543 件のうち 23,289 件（47 %）を利用し，85,071 件の is-a カテゴリと接点カテゴリを得た．本手法では同じ Wikipedia カテゴリが複数の場所で is-a カテゴリになってもよいとしている．そのため，部分的に似た木構造が何回も出現することがある（ただし循環はしない）．これが原因で，利用した Wikipedia カテゴリ数よりも is-a カテゴリの数のほうが大幅に多い結果となった．各接点カテゴリより下位の is-a カテゴリの葉の深さの平均と，is-a カテゴリ数の関係を示したグラフを**図 11** に示す[⋆2]．点が左下に多く分布していることから，構築したオントロジーは小規模な多くの is-a カテゴリ階層と，大規模ないくつかの is-a カテゴリ階層からなることが分かる．葉の深さ平均

⋆1 本稿では，語彙大系クラスと接点カテゴリの接続を“ – ”で表す．
⋆2 本稿では末端の語彙大系クラスを深さ 0 と考え，接点カテゴリを深さ 1 としている．

Vol.2009-NL-194 No.4
2009/11/16

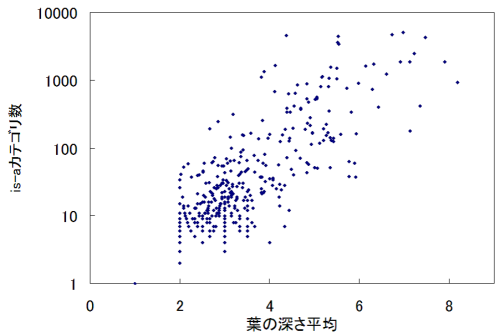


図 11 各接点カテゴリにおける葉の深さ平均と is-a カテゴリ数の関係

が 1〜2 の接点カテゴリが最も多く，全体のおよそ 50 %を占めている．全体で，接点カテゴリ 1 件あたりの葉の深さ平均は 1.96，is-a カテゴリ数は 56.6 件であった．

本実験では，is-a カテゴリ階層について，以下の 2 つの方法で適合率を求めた．

（1） 親-子の適合率
対象とする is-a カテゴリの 1 つ上位の親と is-a 関係にあるかどうか．

（2） 先祖-子孫の適合率
対象とする is-a カテゴリより上位のカテゴリ（先祖のカテゴリ）すべてが is-a 関係の階層となっているか．

深さごとに is-a カテゴリを 100 件ずつランダム抽出（100 件以下なら全て抽出）し，それぞれの適合率を求めた結果と，深さ別の is-a カテゴリ数を**図 12** に示す．is-a カテゴリが深くなるにつれ適合率は下がるが，先祖-子孫の適合率は深さ 1〜5 で 90 %以上，親-子の適合率は深さ 1〜7 で 90 %以上と高い数値である．人手で作成した深さ 1 の接点カテゴリを除いた全 is-a カテゴリの適合率は親-子で 92.8 %，先祖-子孫で 82.6 %となった．

本実験での is-a カテゴリの抽出エラーの例を**表 2** に示す．上位の語彙大系クラスとは is-a 関係が成り立つが，上位カテゴリとは is-a 関係でない関係（Part of など）として下位カテゴリになっている場合，誤った is-a 関係を抽出してしまう（表の 1,2）．しかし，途中で間違った is-a 関係が発生しても，それより下位では is-a 関係が成立した is-a カテゴリ階層が構築されることが多い（表の 3）．そのため，全体として親-子の適合率のほうが高い結果となっている．

4.4 インスタンス抽出（手順 3）の実験結果

Wikipedia の記事ページ 479,231 件中，263,631 件（55 %）の記事ページの見出し語をインスタンスとして抽出した．インスタンスを最も多く持つ is-a カテゴリは〈日本の俳優〉でインスタンス数は 5,632 件であった．一方，インスタンスを 1 件ももたない is-a カテゴリは全体の 20 %を占めている．is-a カテゴリ 1 件あたりのインスタンス数は，17.8 件であった．

以下の手順で作成したテストセットで，インスタンスの適合率と再現率を求めた．はじめに「is-a カテゴリ」とそれに所属する「記事ページの見出し語」のすべての組み合わせを列挙し，そこからランダムに 400 件抽出する．400 件のうち is-a カテゴリの先祖のカテゴリすべてが is-a 関係の階層となっているものを判定し，違うものは破棄する．残ったものに対し，記事ページの見出し語がインスタンスになるかどうかを人手で判定したものをテストセットとする．テストセットの記事ページの見出し語 359 件のうち，インスタンスと

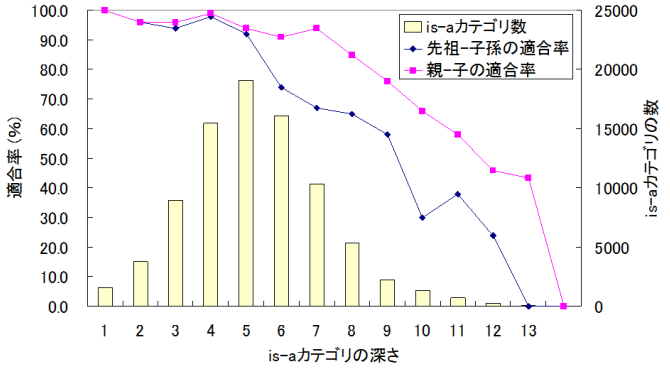


図 12 is-a カテゴリの深さ別の数と適合率

表 2 is-a カテゴリ階層のエラー例

	構築した階層構造	エラー内容
1	← [公共機関] – 〈鉄道〉 ← 〈各国の鉄道〉 ← 〈各国の鉄道駅〉	鉄道駅は鉄道の設備の一部である．
2	← [司法機関] – 〈警察〉 ← 〈各国の警察〉 ← 〈日本の警察〉 ← 〈日本の警察署〉	警察署は警察ではない．
3	← [学問分野・学科] – 〈学問の分野〉 ← 〈理学〉 ← 〈生物学〉 ← 〈生物〉 ← 〈動物〉 ← 〈刺胞動物〉	階層の途中で〈生物〉 is a 〈生物学〉の誤りがある．

表 3 人手抽出ではインスタンスとしなかったにもかかわらず，自動抽出ではインスタンスとしてしまったエラー例

	抽出したインスタンスとその上位の階層構造	エラー内容
1	← [乾物・漬物] − 〈漬物〉 ← 〈日本の漬物〉 ← 《 漬物 》	日本の漬物の下位に漬物があるのはおかしい.
2	← [都市] − 〈都市〉 ← 〈起源別の都市〉 ← 〈宗教都市〉 ← 〈門前町〉 ← 《長野市》	長野市は門前町がある市であり門前町ではない.
3	← [組織] ← [国家] ← (国) − (大陸別の国) ← (ヨーロッパの国) ← (ブルガリア) ← 《ブルガリア正教会》	ブルガリア正教会 はブルガリではない.

人手で判定されたものは 278 件ある．解析の結果，適合率 98.6 % (205/208)，再現率 83.0 % (205/247) と高い数値を得た．

人手抽出ではインスタンスとしなかったにもかかわらず，自動抽出ではインスタンスとしてしまったケースがあった．このエラーの原因は，上位の語彙大系クラスとは is-a 関係が成り立つが，直属のカテゴリとは is-a 関係にならなかったためである．エラー例を**表 3** に示す．一方，人手抽出ではインスタンス記事としなかったにもかかわらず，自動抽出ではインスタンス記事と判定されなかったのは，カテゴリ名や上位語からは，上位語候補とマッチングがとれなかったことが原因である．

4.5 従来手法との比較

桜井らは，2007 年 11 月のダンプデータに手法を適応したところ，親-子の正解率は 91.2 ± 1.63 %（95 %信頼区間を算出）であったとしている．本手法の正解率は 92.8 %なので，桜井らの手法と精度はほぼ同じであるといえる．また，桜井らは 6,672 件のカテゴリが階層構築に利用されたとしている．全カテゴリ数が記載されておらず，本手法とは使用したダンプデータが違うため単純比較しにくい．しかし，本手法で利用したカテゴリ数は 23,289 件と大幅に多いことから，本手法は高精度で大規模なオントロジーの構築に有効であるといえる．また本手法では階層を 1 つに統合している点で，自然言語処理の分野で利用しやすいオントロジーを構築できたといえる．

小林らの手法は Wikipeida のカテゴリ階層を利用していないため，カテゴリは全て末端の語彙大系クラスの 1 つ下位に接続される．そのため,〈醸造酒〉 is a 〈酒〉 というような Wikipedia カテゴリ間での is-a 関係が得られない．また小林らの手法では語彙大系インスタンスにカテゴリ名と照合するキーワードがなければ，条件 1 を満たせずカテゴリは接続されない．本手法では Wikipedia カテゴリに上位語を設定し，上位語でも文字列の照合を行なうことで，カテゴリ名自体が異なっていても is-a 関係として抽出できるようにしている．

5. おわりに

本稿では，日本語 Wikipedia から高精度で大規模な is-a 関係のオントロジーを構築した．サンプル評価の結果，is-a カテゴリ階層の適合率は親-子間で 92.8 %，先祖-子孫間で 82.6 %となった．また，インスタンスの適合率は 98.6 %，再現率は 83.0 %であった．Wikipedia カテゴリ 49,543 件のうち 23,289 件（47 %），Wikipedia の記事ページ 479,231 件のうち，263,631 件（55 %）をオントロジー化することができた．日本語語彙大系を利用することで，従来手法よりも知識が多く，より深い is-a 関係の階層を構築することができた．本手法では半自動で行なっていた接点カテゴリの抽出を自動化させることが今後の課題である．さらに，現在はパターン規則を用いている上位語の抽出や is-a 関係の抽出に機械学習を導入することを検討したい．

参 考 文 献

1) 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
2) 隅田飛鳥, 吉永直樹, 島澤健太郎： Wikipedia の記事構造からの上位下位関係抽出, 自然言語処理, 16(3), pp.3-24, 2009.
3) 清田 陽司, 田村 悟之, 中川 裕志, 増田 英孝: Reference Navigator： 異種オントロジーの統合ブラウジングツール ～図書館の分類体系と Wikipedia カテゴリの対応付け～, 言語処理学会 第 13 回年次大会 ワークショップ「言語的オントロジーの構築・連携・利用」論文集, pp. 35-38, 2007.
4) Simone Paolo Ponzetto, Michael Strube: Deriving a Large Scale Taxonomy fromWikipedia, Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence(AAAI), pp.1440-1445, 2007.
5) 桜井慎弥, 手島拓也, 石川雅之, 森田武史, 和泉憲明, 山口高平: 汎用オントロジー構築における日本語 Wikipedia の適用可能性, 人工知能学会, 第 18 回セマンティックウェブとオントロジー研究会, pp.7-14, 2008.
6) Fabian M. Suchanek and Gjergji Kasneci and Gerhard Weikum. Yago: A Core of Semantic Knowledge unifying wordnet and Wikipedia, Proceedings of the 16th International Conference on World Wide Web(WWW), pp.697-706, 2007.
7) 小林 暁雄, 増山 繁, 関根 聡: 日本語語彙大系と日本語ウィキペディアにおける知識の自動結合による汎用オントロジー構築手法: 情報処理学会研究報告. 自然言語処理研究会報告 2008-NL-187, pp.7-14.