情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2010-NL-198 No.3
2010/9/16

Wikipediaからの大規模な人オントロジー構築

柴 木 優 美[†1]　永 田 昌 明[†2]　山 本 和 英[†1]

Wikipedia を利用し，人に関する大規模な is-a 関係のオントロジーを構築する手法を提案する．本手法では初めに，人を表すカテゴリを機械学習による分類器で判定し，Wikipedia の階層構造をそのまま利用して is-a 関係だけから構成される人のカテゴリ階層を構築する．その後，人を表すカテゴリが付与されている記事から，人を表す記事をインスタンスとして抽出する．機械学習では，カテゴリ名及びカテゴリの周辺の単語が，日本語語彙大系のインスタンスとどのようにマッチするかを素性にした．その結果，人を表すカテゴリを適合率 99.3%，再現率 98.4%，人を表すインスタンスを適合率 98.2%，再現率 98.6% で抽出することができた．

Constructing Large-Scale Person Ontology from Wikipedia

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

This paper presents a method for constructing a large-scale Person Ontology with category hierarchy from Wikipedia. We first extract Wikipedia category labels which represent person by using a machine learning classifier. We then construct a person category hierarchy by detecting is-a relations in the Wikipedia category network. We then extract the titles of Wikipedia articles which represent person. Experiments show that the accuracy of Wikipedia person category extraction is 99.3% precision and 98.4% recall, while that of person instance extraction is 98.2% and 98.6%, respectively.

†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. はじめに

近年，質問応答や要約，含意認識などで，幅広い知識の必要性が高まっている．幅広い分野の一般的知識を記述した汎用オントロジーには，WordNet（Fellbaum et al., 1998）や日本語語彙大系（池原ら, 1997）などがある．しかし，固有名詞も含め，日々生まれる新しい語彙への即時対応が難しいのが現状である．そこで, 即時更新性に優れたオンライン百科事典である Wikipedia を利用した知識資源の作成が注目されている．

DBpedia （Bizer et al. 2009）は Wikipedia の記事中にある infobox から RDF トリプルを抽出する研究を行なっている．これにより，“小泉純一郎”，“出身地”，“神奈川県横須賀市” のような主語，述語，目的語の関係が得られる．Ponzetto et al. （2007） は，英語 Wikipedia において，親子関係にあるカテゴリ同士が is-a 関係[⋆1]か否か，または not-is-a 関係か否かを判定する手法を提案している．YAGO （Suchanek et al. 2007）では，WordNet の synset（語義またはカテゴリのようなもの）に is-a 関係となる Wikipedia のカテゴリを接続し，さらに，分類されている記事をインスタンスとする手法を提案している．

我々は，人に関する is-a 関係のオントロジーを構築することを目指す．なぜなら Wikipedia には人を表す記事やカテゴリ，is-a 関係のカテゴリ階層が多いためである．人に関するオントロジーを構築することができれば，人名検索や固有表現抽出などに活用できる．また，Wikipedia の即時更新性により，人に関する新しい知識をいち早くオントロジーに追加することができる．

人に関するオントロジーは，日本語語彙大系にも含まれている．日本語語彙大系の中には人を表すカテゴリが存在し，それらは is-a 関係の階層構造を持っている．さらに，カテゴリには人を表す一般名詞のインスタンスが付与されている．他にも，人に関するオントロジーを含む，関根の拡張固有表現階層[⋆2]が Web 上で公開されている．これは，人手で構築された固有表現のカテゴリ階層である．このカテゴリ階層は人の名前を収録するためのカテゴリ “人名” を持つ．語彙大系と違い，人名より下位に is-a 関係の階層構造を持たないが，“出身市町村”，“パリ” のような属性と属性値が定義されている．

本手法で抽出する人を表す記事とは，人名や職業名（例:イチロー，ファイナンシャルプランナー）などである．また，人を表すカテゴリとは，人を表す記事を分類するカテゴリ

⋆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．
⋆2 http://sites.google.com/site/extendednamedentityhierarchy/

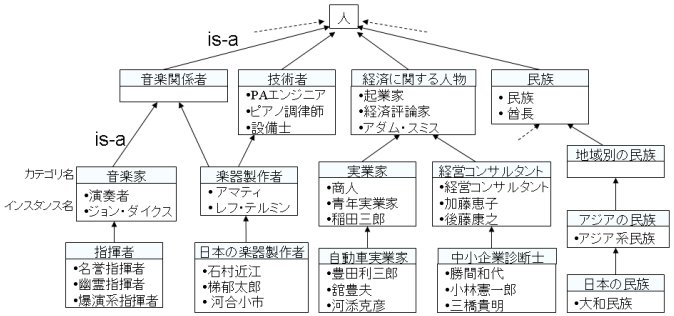

図 1 本手法で構築するオントロジーの一部

(例:スポーツ選手，経済に関する人物）などを指す．本手法で構築するオントロジーの例を図 1 に示す．

本手法では，初めに，機械学習による分類器で Wikipedia の全てのカテゴリの中から人を表すカテゴリ（以下，人カテゴリ）をオントロジーのカテゴリとして抽出する．次に，親子関係のリンクのある人カテゴリを is-a 関係とみなし，人カテゴリ階層とする．最後に，人カテゴリが付与されている記事のうち，人を表す記事（以下，人インスタンス）をルールベースで抽出する．

本稿では，初めに本手法で利用する言語資源と関連研究について説明する．次に，人オントロジー構築手法と実験方法，結果を述べ，最後に考察を述べる．

2. 言 語 資 源

2.1 日本語 Wikipedia

Wikipedia は即時更新性に優れた自由に利用できるオンライン百科事典であり，Web 上で XML 形式のダンプデータが公開されている[⋆1]．Wikipedia の記事の例を図 2 に示す．記事には，見出し語と説明文，記事を分類するカテゴリが書かれている．記事では，説明文の第一文は見出し語の定義文であることが多い．見出し語は固有名詞，カテゴリ名は一般名詞であることが多い．Wikipedia には記事を分類するためのカテゴリがあり，各記事にはいくつかのカテゴリが付与されている．カテゴリは，主要カテゴリと呼ばれる 9 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

⋆1 http://download.wikimedia.org/jawiki

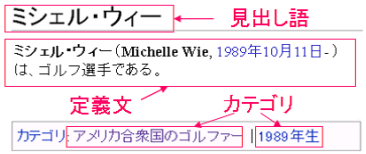

図 2 Wikipedia の記事の例

数存在することが多く，循環もある．カテゴリ間の関係やカテゴリと記事の関係は多様で，is-a 関係でないものもある．例えば，カテゴリ “アメリカ合衆国のゴルファー” と，このカテゴリが付与されている記事 “ミシェル・ウィー” は is-a 関係にあるが，同じく “アメリカ合衆国のゴルファー” が付与されている記事 “PGA ツアー” とは is-a 関係にない．

2.2 日本語語彙大系

Wikipedia から人カテゴリを抽出するため，機械学習による分類器を使用する．Wikipedia のカテゴリ名は一般名詞であることが多い．そのため我々は，カテゴリ名の最後の形態素やカテゴリ名そのものが一般名詞のオントロジーのどのインスタンスとマッチするかが重要な素性になると考えた．そこで，一般名詞の意味体系を収録している日本語語彙大系（以下，語彙大系）を利用することにした．

日本語語彙大系は，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，is-a 関係から構成されるオントロジーである（図 3）．語彙大系には，約 2,700 件のカテゴリと約 10 万件のインスタンスを持つ一般名詞の意味体系が収録されている．多義性があるインスタンスはいくつかのカテゴリが付与されている．例えば，ライターのように writer と lighter の 2 つの意味がある場合，2 つのカテゴリ “筆者” と “家庭用品” にライターが所属することになる．

3. 関 連 研 究

3.1 Ponzetto et al. の手法と桜井らの手法

Ponzetto et al.（2007）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（2008） は，Ponzetto et al. 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．桜井らは,「後方文字列照合」という手法により，Wikipedia のカテゴリ間が is-a 関係かどうかを判定している．後方文字

列照合とは，Wikipedia の親カテゴリに対しその子カテゴリ名が “任意の文字列＋親カテゴリ名” であったとき両者を is-a 関係とする手法である．例えば，カテゴリ “空港” の子カテゴリに “日本の空港” が存在した場合，両者は is-a 関係と判定される．

3.2 小林らの手法

小林ら（2008）は，日本語語彙大系に Wikipedia を統合する手法を提案している．YAGO は英語 WordNet に英語 Wikipedia を統合する手法だが，カテゴリ名が複数形であれば概念を表すカテゴリになりやすい，というような英語依存の手法を利用しているため，そのままでは日本語 Wikipedia に適用することができない．そのため小林らは，YAGO の手法を日本語 Wikipedia に適用できるよう改良した．これら 2 つの手法と提案手法は Wikipedia のカテゴリをオントロジーのカテゴリ，Wikipedia の記事の見出し語をインスタンスとしている点で本手法と似ている．

小林らの手法では，記事の定義文（第一文）からパターンマッチにより見出し語の上位語となる単語を抽出する．パターンマッチの例を以下に示す．

... は、[上位語] の一つである。
... は、[上位語] である。
...[上位語]。

例えば，図 2 の記事の定義文「ミシェル・ウィーは、ゴルフ選手である。」からは，見出し

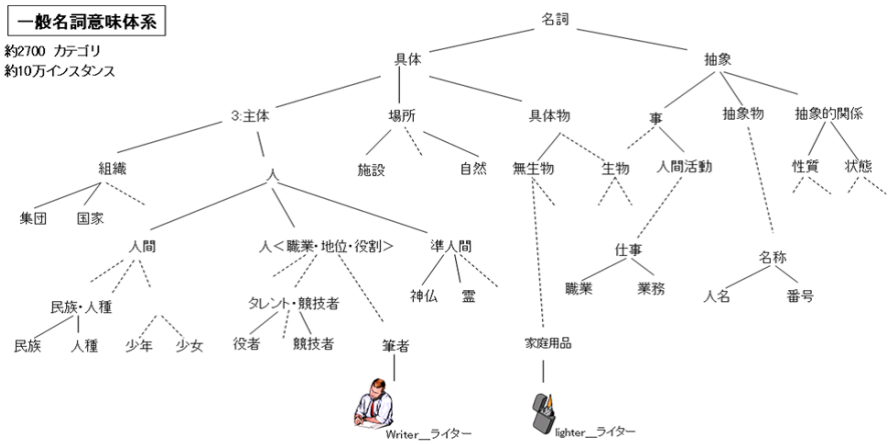

図 3 日本語語彙大系の一般名詞意味体系

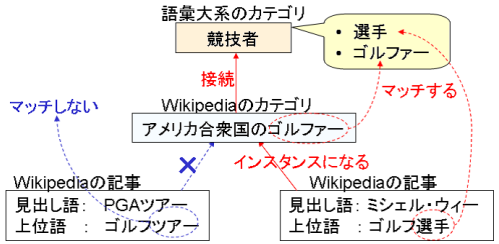

図 4 小林らの手法の概要

語 “ミシェル・ウィー” の上位語として “ゴルフ選手” が抽出される．小林らの手法の概要を図 4 に示す．この図は，語彙大系の人カテゴリ “競技者” に Wikipedia のカテゴリと記事を接続する手順の例である．初めに，語彙大系の人カテゴリのインスタンス名に，最後の文字列がマッチする Wikipedia のカテゴリを，下位カテゴリ候補として対応づける．このカテゴリ候補が付与されている記事の上位語が，語彙大系の人カテゴリ階層のインスタンスとマッチすれば，カテゴリ候補-記事を下位カテゴリ-インスタンスとして抽出する．インスタンスが 1 つも抽出されなかった場合，このカテゴリは破棄される．今回は，語彙大系の人カテゴリを対象にした手法を独自に実装し，提案手法との比較を行なう．

3.3 山下の手法

山下は日本語 Wikipedia から人名記事を抽出するソフトウェア[⋆1]を公開している．山下は，人名記事の抽出のために，“〇年生”（例:2000 年生）というカテゴリを利用している．このカテゴリは，主に人，馬，犬の固有名を分類するためのカテゴリである．山下は，“〇年生” というカテゴリが付与される記事から，馬，犬をパターンマッチにより削除する手法により，人名を抽出している．今回は “年生” のほかに “年没”，“世紀没”，“年代没”，“年生”，“世紀生”，“年代生” も加えて人記事抽出を行なう．山下の手法での記事抽出精度を提案手法と比較する．

4. 人オントロジー構築手法

4.1 予 備 調 査

2008 年 7 月 24 日の日本語版 Wikipedia の記事，カテゴリ，カテゴリ階層を調査した．先ほど述べたように，Wikipedia は特に，人を表す記事やカテゴリが多い．クリーニング後

⋆1 http://coderepos.org/share/browser/lang/perl/misc/wikipejago

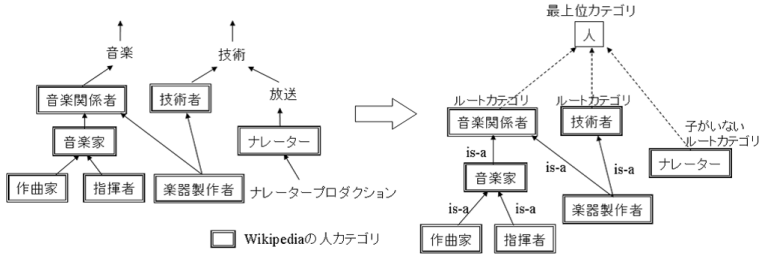

図 5 Wikipedia のカテゴリ階層の一部とそこから構築される人カテゴリ階層の例

の全記事 477,094 件からランダムに 1,000 件を抽出して調査した結果，277 件（27.7%）が人インスタンスであった．予想される全人インスタンス数は約 130,000 件である．さらに，人インスタンス 277 件のうち，97.3%の人インスタンスが人カテゴリに所属していることがわかった．また，277 件のうち，96.0%が人名であった．

次に，クリーニング後の全カテゴリ 39,767 件を調査した結果，8,485 件（21.4%）が人カテゴリであった．人カテゴリのうち，親子関係のある 2 つのカテゴリ 1,000 件をサンプル調査した結果，is-a 関係が 98.7%成り立っていることがわかった．そのため，is-a 関係から構成されるカテゴリ階層が構築しやすいと考えられる．また，全カテゴリ名の 92.1%の最後の形態素が，語彙大系のインスタンスとマッチした．これは，Wikipedia のカテゴリ名の最後の形態素の多くが一般名詞であることを示している．

4.2 人カテゴリ階層の構築

Wikipedia のカテゴリ階層から，機械学習による分類器で人カテゴリを抽出し，カテゴリ階層を構築する．もし親子関係にある 2 つのカテゴリが共に人カテゴリとして抽出された場合，両者を is-a 関係とみなす．なぜなら，親子関係のある 2 つの人カテゴリは，is-a 関係が 98.7%成り立っていたためである（4.1 節）．構築されたいくつかのカテゴリ階層のルートカテゴリを，最上位カテゴリ “人” に接続し，1 つのカテゴリ階層にする．Wikipedia のカテゴリ階層と，そこから構築される Wikipedia の人カテゴリ階層の例を図 5 に示す．

初めに，Wikipedia のカテゴリが人カテゴリであるか否かの 2 値を，SVM で学習した分類器によって判定する．分類対象のカテゴリのカテゴリ名や，周辺の単語を素性に利用する．本手法では以下の 3 要素の組み合わせ 48 種（6 種 × 2 種 × 4 種）を素性にする．

1. 対象カテゴリの周辺単語（6 種）
2. 語彙大系と照合する文字列の範囲（2 種）
3. 語彙大系との照合の様態（4 種）

対象カテゴリの周辺単語とは，以下の 6 種類を指す．

対象カテゴリの周辺単語

A. 対象カテゴリ名
B. 対象カテゴリの親カテゴリ名
C. 対象カテゴリの子カテゴリ名
D. 対象カテゴリの兄弟カテゴリ名
E. 対象カテゴリが付与されている記事の定義文から抽出される上位語[⋆1]
F. 対象カテゴリと同名の記事[⋆2]の定義文から抽出される上位語

以上のようなカテゴリ名や定義文から抽出される上位語は一般名詞であることが多いため，語彙大系のインスタンスとマッチすることが多い．次に，これら A〜F の単語の，以下の 2 種類の文字列の範囲を語彙大系のインスタンスと照合する．

語彙大系と照合する文字列の範囲

I. 単語の全文字列
II. 単語の最後の形態素

本手法では，語彙大系のカテゴリを，人カテゴリとそれ以外のカテゴリの 2 つにわける．今回は，語彙大系の人カテゴリを，“人間”，“人＜職業・地位・役割＞” 以下の全てのカテゴリと，“職業”，“人名” カテゴリを指すものとする（図 3）．

A〜F と I，II を組み合わせた 12 種の文字列に対し，以下の a〜d の頻度の割合を求める．

語彙大系との照合の様態

a. 文字列が，語彙大系の人カテゴリのインスタンスのみにマッチする頻度
b. 文字列が，語彙大系の人カテゴリ以外のインスタンスのみにマッチする頻度
c. 文字列が，語彙大系の人カテゴリ・人カテゴリ以外のインスタンスの両方にマッチする頻度
d. 文字列が，どの語彙大系のインスタンスにもマッチしない頻度

例えば，図 5 において，対象カテゴリが “音楽家” の場合，C-II（対象カテゴリの子カテゴリ名の最後の形態素）は子カテゴリ “作曲家” の最後の形態素 “家” と，“指揮者” の最後の

⋆1 定義文からの上位語抽出パターンは，小林ら（2008）と隅田ら（2009）の手法を元に作成したものを使用した．なお，彼らは Hearst（1992）の手法を参考にしている．
⋆2 カテゴリ “ベーシスト” に記事 “ベーシスト” が分類されている場合，カテゴリ “ベーシスト” の同名記事は記事 “ベーシスト” となる．

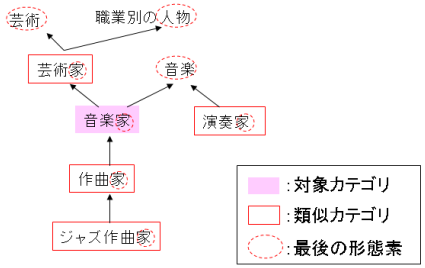

図 6 対象カテゴリが “音楽家” のときの類似カテゴリ

形態素 “者” である．“家” は c，“者” は a に該当するため，a,b,c,d は 1,0,1,0 となる．よって，C-II の a,b,c,d の割合 0.5,0,0.5,0 が素性となる．このような方法で，A〜F，I,II，a〜d の組み合わせ 48 素性を計算し，SVM の素性とする．

4.3 類似カテゴリの抽出

Wikipedia では，記事が 1 つもないカテゴリや，周辺にカテゴリが少ないカテゴリが存在する．そのため，素性を計算するための単語が十分に得られない可能性がある．そこで本手法では，対象カテゴリの周辺にある，対象カテゴリと類似している可能性の高いカテゴリ（以下，類似カテゴリ）の周辺単語も素性に利用する．類似するカテゴリは，対象カテゴリと最後の形態素が一致する親・子・兄弟カテゴリを指す．ただし，親・子カテゴリに関しては，親（子）が類似カテゴリと判定された場合，更にその親（子）も最後の形態素が一致する限り，親子カテゴリをつなげて抽出する．

対象カテゴリが “音楽家” であるときの類似カテゴリの例を図 6 に示す．“音楽家” の類似カテゴリの対象カテゴリの周辺単語 A〜F を “音楽家” の A〜F に加える．例えば，類似カテゴリである芸術家の B（対象カテゴリの親カテゴリ名）である “芸術” と “職業別の人物” も対象カテゴリ “音楽家” の B に加える．

4.4 人インスタンスの抽出

本手法では，抽出した Wikipedia の人カテゴリが付与されている記事の見出し語のうち，以下の 4 つの条件のいずれかを満たすものを人インスタンスとして抽出する．

1. 記事の定義文から抽出される上位語の最後の形態素が，語彙大系の人カテゴリのインスタンスと一致する．
2. 記事の見出し語の最後の形態素が，語彙大系の人カテゴリのインスタンスと一致する．
3. 記事が分類されている Wikipedia のカテゴリ名の後方の文字列が，1 件以上，以下のパターンにマッチする．
   （年没，世紀没，年代没，年生，世紀生，年代生）
   これらのカテゴリは，多くの人名を分類するためのカテゴリである．
4. 記事が分類されている Wikipedia のカテゴリ名が，以下の条件式を満たす．

$$\frac{\text{4.2 節で抽出した人カテゴリの数}}{\text{記事に付与されている Wikipedia のカテゴリ数}} > 0.5$$

これは，Wikipedia の記事が付与されているカテゴリの人カテゴリの割合が高いほど，この記事が人インスタンスである可能性が高いという指標に基づく．閾地 0.5 は予備調査によって求めた．

5. 実　　験

5.1 実験設定

提案手法の評価に，2008 年 7 月 24 日の日本語版 Wikipedia を用いた．初めに，Wikipedia の内部向けの記事（例:“画像:”，“Help:”）をパターンマッチで取り除いた．その結果，477,094 件の記事と 39,767 件のカテゴリを得た．39,767 件のカテゴリに対し，人カテゴリであれば正，人カテゴリ以外ならば負を付与した正解データを作成した．曖昧なカテゴリ名に関しては，以下の基準で判定を行なった．

- 人名（例:マイケル・ジャクソン）には，分類される具体物記事がないため，人カテゴリとしない．（ただし，記事が人名の場合，人インスタンスとする．）
- 職業名（例:警察官）は，人カテゴリとする．
- 家族名，民族名，氏族名（例:ブランデンブルク家，アイヌ人，蘇我氏）は，人カテゴリとする．（語彙大系の分類基準に従っている）
- グループ名（例:カーペンターズ）は人カテゴリとしない．（人インスタンスともしない．）

正解データからランダムに抽出した 2,000 件のカテゴリ（正 435 件:負 1,565 件）を学習用データ，残りの 37,767 件のカテゴリを評価用データとした[⋆1]．評価用の記事は，学習用データのカテゴリに 1 つも分類されていない記事 417,476 件を対象とする．

本稿では，単語の最後の形態素を抽出するために，形態素解析器 Juman6.0[⋆2]を使用した．

⋆1 予備実験により，2,000 件で十分な精度が得られることを確認した．詳細は 6.3 節で述べる．
⋆2 http://www-lab25.kuee.kyoto-u.ac.jp/nlresource/ juman.html

SVM には TinySVM0.09[⋆1]を利用した．SVM のカーネルには線形カーネルを用いた．本手法の有効性を評価するため，小林らの手法と山下の手法との結果を比較した．

5.2 実験結果

評価用データ 37,767 件に対して本手法と小林らの手法を適用し，人カテゴリを抽出した結果の精度を表 1 に示す．小林手法と比較し，本手法は適合率が 6.5 ポイント，再現率は 14.8 ポイント上回っている．

抽出した Wikipedia の人カテゴリから，親子関係にあるカテゴリのペアをランダムに 1,000 件抽出し，両者が人カテゴリであり，かつ is-a 関係が成り立っているかどうかを判定した結果，適合率は 98.3%であった．エラーは，“千葉氏-大須賀氏” のように，Wikipedia のカテゴリ階層において is-a 関係のないカテゴリ同士が親子関係にあるときに発生することが多かった．しかし，このような is-a 関係のない人カテゴリは少なく，全体の 98.7%は is-a 関係が成り立っている（4.1 節）．

人インスタンスを抽出した結果の精度を表 2 に示す．Wikipedia の記事の中からランダムに抽出した 1,000 件の記事のペアについて人インスタンスなら正，人インスタンス以外なら負を付与した評価データ（正 281 件:負 719 件）を作成した．本手法は山下の手法に比べ再現率が 21.0 ポイント上回っている．また本手法の F 値は 3 手法の中で最も高い．

抽出した Wikipedia の人カテゴリと人インスタンスのペア（例: “アメリカ合衆国のゴルファー-ミシェル・ウィー”，“芸術家 - 功労芸術家”）の抽出精度を表 3 に示す．ランダムに抽出した 1,000 件のカテゴリと記事のペアについて，両者が “人カテゴリ - 人インスタンス” であり，かつ is-a 関係が成り立っているものを正とした評価データを作成した（正 296 件:負 704 件）．本手法は小林らの手法より，適合率は 2.1 ポイント，再現率は 11.8 ポイント上回った．

表 1 Wikipedia の人カテゴリ抽出精度

	適合率	再現率	F 値
小林らの手法	92.8% (6727/7247)	83.6% (6727/8050)	88.0%
本手法	**99.3%** (7922/7979)	**98.4%** (7922/8050)	**98.8%**

表 2 Wikipedia の人インスタンス抽出精度

	適合率	再現率	F 値
山下の手法	**100.0%** (218/218)	77.6% (218/281)	87.4%
小林らの手法	96.0% (264/275)	94.0% (264/281)	95.0%
本手法	98.2% (277/282)	**98.6%**(277/281)	**98.4%**

⋆1 http://chasen.org/ taku/software/TinySVM/

表 3 Wikipedia の人カテゴリ-人インスタンスのペアの抽出精度

	適合率	再現率	F 値
小林らの手法	95.9% (259/270)	87.5% (259/296)	91.5%
本手法	**98.0%** (294/300)	**99.3%** (294/296)	**98.7%**

6. 考　　察

6.1 構築した人カテゴリ階層

学習データに利用した Wikipedia の人カテゴリを加えた，全 8,357 件の人カテゴリから構築した人カテゴリ階層は，224 件のルートカテゴリを持つ．そのうち，194 件は子カテゴリを 1 つも持たない（抽出した人カテゴリの 2.3%）．まれに，カテゴリ階層が循環することがある（例:カテゴリ “歴史家” と “歴史学者” は双方にリンクされている）．

6.2 抽出した人カテゴリと人インスタンス

本手法では，機械学習の際に，対象カテゴリだけでなく周辺のカテゴリや記事も素性に利用したことにより，語彙大系のインスタンスとマッチしない，未知の Wikipedia の人カテゴリ名も抽出することができた（例:ヴァイオリニスト，アニメーター）．また，多義性のある Wikipedia の人カテゴリを正しく判定できた．例えば，“ショッピングセンター” の最後の形態素 “センター” には多義性があるが，本手法により，正しく人カテゴリではないと判定できた．小林らの手法では，統合するオントロジーにマッチする単語がないと，人カテゴリを抽出できない．また，パターンマッチを元にした手法のため，多義性による誤りが多い．本手法ではカテゴリの周辺単語を素性にすることで，未知のカテゴリ名や多義性のあるカテゴリ名にも柔軟に対応できる．

本手法では，桜井ら（2008）の手法で抽出できなかった is-a 関係のカテゴリのペアを抽出できている．彼らの手法である後方文字列照合では，“農学者-日本の農学者” のように文字列がマッチするものしか is-a 関係として抽出できなかった．しかし本手法では “ジャーナリスト-スポーツライター” のように，文字列に関係なく is-a 関係を抽出することができる．抽出した is-a 関係の人カテゴリのペアで，最後の形態素がマッチしないものは，全ペ

ア 14,408 件中 5,558 件（38.6%）あった．

6.3 学習データ量

今後更新されていく Wikipedia に対しても本手法を実装できるよう，人カテゴリを判定する分類器に必要な学習データ量を調査した．Wikipedia の全カテゴリ 39,767 件からランダムに抽出した 30,000 件を学習データとし，1,000 件から 30,000 件まで 1,000 件ずつ学習データの量を変化させたときの分類器の精度を評価した．評価用データは残りの 9,767 件のカテゴリである．学習データ量に伴う適合率，再現率，F 値の変化を図 7 に示す．学習データが 1,000 件で F 値が 98.5%，30,000 件で 98.9%とわずか 0.4 ポイントしか違いがなかった．図 7 より，学習データが少なくても高精度な分類器を作成できることが示された．

6.4 素性の効果

本手法では，分類対象とする Wikipedia のカテゴリの他に，周辺にある類似カテゴリも，素性作成に利用している（4.3 節）．類似カテゴリの有効性を確認するため，類似カテゴリを利用しない場合との比較を行なった．また，本手法では，素性ベクトル作成の際に語彙大系を利用していた．この素性の有効性を確認するため，一般的に行われる，単語の頻度を素性ベクトルとする方法との比較も行なった（類似カテゴリは利用する）．学習データ量を変化させたときの提案手法と比較手法の F 値を図 8 に示す．図 8 より，類似カテゴリを利用すると，利用しなかった場合と比較し，学習データ量に関わらず F 値が高いことがわかる．また，単語の頻度を素性にする方法では，学習データ量が少ない場合，他の 2 手法に比べて F 値が大幅に低いことがわかる．学習データを増やしても他の 2 手法の精度を上回らなかった．このことより，類似カテゴリと語彙大系を使用することで，より有効な素性が作成できることが示された．

6.5 人以外のカテゴリに関する考察

本稿では，“人” カテゴリに分野を限定してオントロジー構築を行なった．一方，柴木ら（2009）は，分野を限定せずに Wikipedia のカテゴリ階層から 1 つに統合されたオントロジーを構築している．その際，オントロジーの上位階層に語彙大系を利用している．しかしこの手法では，語彙大系の末端のカテゴリのみに Wikipedia のカテゴリ階層を接続しているため，Wikipedia 全体の約半数のカテゴリしかオントロジーに利用できていない．本手法を “人” 以外の分野でも適用することができれば，より多くのカテゴリを利用してオントロジーを構築することができる．そこで，Wikipedia のカテゴリを調査した結果を報告する．

本手法を適用するためには，Wikipedia のカテゴリを分類するための分野を定義し，その分野を語彙大系に対応付ける必要がある．そこで我々は，Wikipedia のカテゴリと語彙大系双方を調査し，独自に，分類分野と，分類分野に対応する語彙大系のカテゴリを定義した（表 4）．全カテゴリからランダムに抽出した 1,000 件のカテゴリの分類結果も合わせて表 4 に示す．

次に，Wikipedia の親子カテゴリのペアの調査を行なった．本手法では，親子カテゴリの

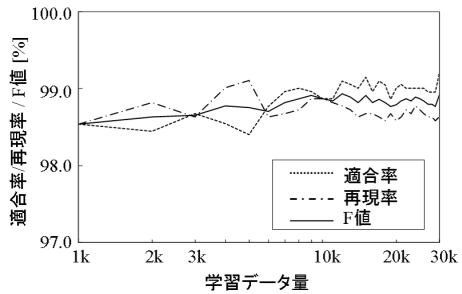

図 7 学習データ量に伴う人カテゴリ抽出の適合率，再現率，F 値の変化

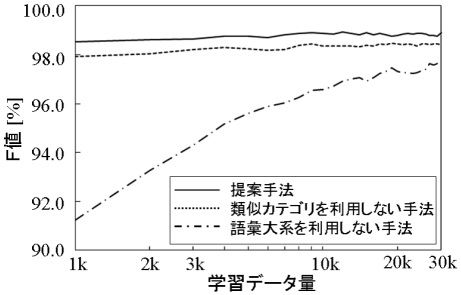

図 8 類似カテゴリと語彙大系を利用した時の素性の効果

表 4 Wikipedia のカテゴリの分類分野と，全体から見た割合

分類分野	対応する語彙大系のカテゴリ	分類するカテゴリの例	全体からみた割合 [%]
人	人，職業，人名	人名，職業，氏族 民族，魔物，マスコット	21.7
組織	組織	スポーツチーム，グループ	4.2
企業	企業，仕事場	会社，銀行，テレビ局	6.7
地名	地域，自然，行政機関，国家	都市名	11.0
地勢	自然	山，川，峠，海，天体	3.0
公共機関・施設	施設，公共機関，建造物	企業名以外の建物，学校，博物館	13.4
生物	生物	動物，植物	2.7
無生物	無生物	物質，道具	5.7
創作物	抽象物（精神）	映画，音楽，番組，文書，言語 技術，宗教，元号，問題	19.4
事	事，抽象物（行為）	動作，感情，法律，スポーツ イベント，天気，色，病気，経済	10.3
抽象的関係	抽象的関係	数，日，時代	0.7
管理カテゴリ	無し	画像ページ，管理ページ	1.1

※ 分類分野 “人” は本稿での人カテゴリの分類基準と少し異なる
※ 1 つのカテゴリは 1 つの分類分野に分類される

表 5 分類分野別の is-a 関係のカテゴリの割合

分類分野	親・子カテゴリのペア数	そのペアが is-a 関係である数	is-a 関係が成り立つ割合 [%]
人	159	156	98.1
組織	21	15	71.4
企業	59	54	91.5
地名	69	60	87.0
地勢	29	29	100.0
公共機関・施設	81	80	98.8
生物	18	17	94.4
無生物	41	41	100.0
創作物	167	144	86.2
事	75	66	88.0
抽象的関係	10	10	100.0
合計	729	672	92.2

両方が人カテゴリである場合，両者にほぼ is-a 関係が成り立つという特徴を生かした手法である．そこで我々は，他分野での親子カテゴリ関係の傾向を調査した．初めに，親子カテゴリの全ペアからランダムに 1,000 件のペアを抽出し，管理カテゴリを含むペアを取り除いた．その結果，990 件のカテゴリを抽出できた．次に，990 件のうち，“格闘家 - レスリング選手” のように両者が同じ分野に分類されるペアを各分野に分類した（is-a 関係でないものも含む）．その結果を，表 5 に示す．全体で 729 件（73.6%) がいずれかの分野に分類された．残りの 261 件は，“競輪 - 競輪場” など，分野が違うカテゴリ同士のペアである．さらに，分類したペアうち，両者に is-a 関係が成り立つペアを調査した．その結果を表 5 に示す．表 5 によると，分野によって，親子カテゴリに is-a 関係がほぼ成り立つもの（地勢，無生物など）と，is-a 関係が成り立たない場合を多く含むもの（地名，創作物など）があることがわかった．特に地名は，“アジア - 日本” などの part-of 関係が多かった．このような分野に本手法を適用するのであれば，is-a 関係でないカテゴリ同士がリンクしているという問題を解決していく必要がある．

7. おわりに

本稿では，日本語 Wikipedia のカテゴリ階層と記事を利用し，高精度で大規模な is-a 関係から構成される人のオントロジーを構築した．今後は，組織，地名など，人以外の分野に対してのオントロジー構築を試みる．いくつかの分野でのカテゴリ階層を統合し，大規模かつ 1 つに統合したオントロジーを構築する予定である．今後，構築した人オントロジーを Web 上に公開する予定である．

参 考 文 献

1) Bizer, C., J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann.: DBpedia - A crystallization point for the web of data, *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 7, No. 3, pp. 154–165 (2009).
2) Bond, F., H. Isahara, Kyoko Kanzaki, and K. Uchimoto.: Boot-strapping a wordnet using multiple existing wordnets. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pp.28–30 (2008).
3) Fellbaum, C.: *WordNet: An Electronic Lexical Database, Language, Speech, and Communication Series*, MIT Press (1998).
4) Hearst, M. A.: Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics (COLING)*, pp. 539–545 (1992).
5) 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).
6) 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 (2008).
7) Ponzetto, S. P. and M. Strube.: Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence (AAAI)*, pp.1440–1445 (2007).
8) 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).
9) 柴木 優美, 永田 昌明, 山本 和英: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:情報処理学会研究報告, 自然言語処理研究会報告 2009-NL-194-4 (2009).
10) Suchanek, F. M., G. Kasneci, and G. Weikum.: Yago: A core of semantic knowledge unifying wordnet and Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*, pp.697–706 (2007).
11) 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).