Wikipedia からの大規模な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]

[1] 長岡技術科学大学 電気系　[2]NTT コミュニケーション科学基礎研究所

{shibaki, yamamoto}@jnlp.org　nagata.masaaki@lab.ntt.co.jp

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ものに汎用オントロジーがあるが，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 を利用したオントロジーの構築が注目されている．その中でも Wikipedia の is-a 関係のリンクに着目してオントロジーを構築している研究が多数存在する．Wikipedia の記事にはカテゴリが付与され，そのカテゴリは他のカテゴリとリンクして階層構造をつくっている．Ponzetto et al.[1] や桜井ら [2] は，親子関係にある Wikipedia のカテゴリ同士の主辞が一致していれば is-a 関係[1]とする手法を提案している．YAGO[3] では，WordNet の synset(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．小林ら [4] は，YAGO の手法を日本語 Wikipedia 用に改良し，日本語語彙大系 [5](以下，語彙大系) に Wikipedia を統合する手法を提案している．この手法では 1 つに統一された階層構造をもつオントロジーを構築できるが，Wikipedia のカテゴリ階層の情報は失われる．我々[6] は，語彙大系の下位に Wikipedia から抽出した部分的な階層構造を接続した．その結果オントロジーは 1 つの階層となり，Wikipedia のカテゴリ階層の情報も反映するが，Wikipedia 全体の約半数のカテゴリと記事しかオントロジーに組み込めない．これらの手法は is-a 関係のリンクの抽出に文字列照合を用いるため，適合率は高いが再現率が低い．また，Wikipedia から 1 つに統一されたオントロジーを抽出するためには既存のオントロジーが必要だった．そこで我々[7] は，機械学習による分類器を用いて Wikipedia の人に関するカテゴリ階層と記事を抽出し，抽出した階層をオントロジーの階層ように扱うことで，is-a 関係の人オントロジーを構築する方法を提案した．

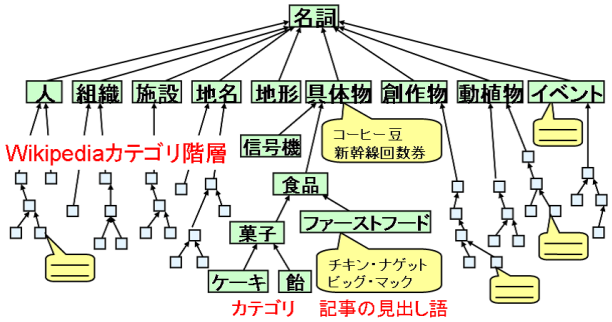

図 1: 本手法で構築するオントロジーの全体像

本研究では人オントロジー構築法を拡張し，Wikipedia のカテゴリ階層と記事を出来るだけそのまま生かし，“人”，“組織”，“施設”，“地名”，“地形”，“具体物”，“創作物”，“動植物”，“イベント” の 9 種類の意味属性を包含した，1 つに統一された is-a 関係のオントロジーの構築を目的とする (図 1)．本手法は，カテゴリ間とカテゴリ-記事間の is-a 関係でないリンクを網羅的に削除し，残ったリンクを is-a 関係とみなすことで，従来手法より多くのカテゴリと記事をオントロジーに組み込むことを目指す．

実験の結果，is-a 関係の精度は，カテゴリ間で適合率 94.8%，再現率 97.0%，カテゴリ-記事間で適合率 95.2%，再現率 96.2%と高精度であった．提案手法により，全カテゴリの 85.0%(約 3 万 4000 件)，全記事の 85.1%(約 37 万件) をオントロジーに組み込めた．

2　Wikipedia のリンクと意味関係

我々は Wikipedia のカテゴリ間，カテゴリ-記事間の関係を調査し，is-a 関係になりにくい場合を以下の 4 種類にまとめた．

1. 意味が抽象的[2]過ぎる単語の場合
   (例) 技術 ← 道具，社会 ← 経済
   抽象的な単語は意味が多様なため，単語間の関係を明確に決定し難い．こういった現象はとくに最

[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．上位下位関係ともいう．以降 is-a 関係にある 2 つの単語を “A←B” と表す．

[2] is-a 関係のオントロジーで上位にある概念

上位階層に多く見られる．抽象的な単語でも is-a 関係になる場合があるが，本手法では適合率も重視しているため対象外とする．

2. 親子が意味的に類似していない
   (例) 筆記用具 ← 万年筆メーカー，植物 ← 草木の神
   単語同士が深く関連していても，意味的に類似していない場合は is-a 関係にならない．
3. 子名の前方が親名と一致する場合
   (例) 火星 ← 火星の衛星，缶 ← 缶コーヒー
   親名の主辞が子名の主辞以外に存在するとき，子と親は part-of 関係やトピック的類似関係にあることが多い．
4. 親が固有名詞の場合
   (例) 少年ジャンプ ←ONE PIECE，新潟県 ← 長岡市
   固有名詞は基本的に下位に単語を持たず，多くは part-of 関係である．こういった現象はとくに最下位階層に多く見られる．

項目 1，2 を判定するために，幅広い分野に適用可能な 9 種類の意味属性 (表 1) に単語を分類することで近似する．どの意味属性にも分類されない単語は抽象的過ぎると判定し，親子が同じ意味属性に分類されなければ，意味的に類似していないと判定し，いずれかを満たす場合には is-a 関係でないとする．項目 3 は，単純な文字列照合で判定可能である．項目 4 は親名が固有名詞かどうかを判定すればよい．

これらの条件で is-a 関係でないリンクを判定したとき，どの程度 is-a 関係を抽出できるのか人手で調査した[3]．その結果，9 種類の意味属性での is-a 関係の精度は，カテゴリ間で適合率 98.9%，再現率 99.3%，カテゴリ-記事で適合率 99.3%，再現率 98.9%であった．適合率を下げる誤りは，親子が同じ意味属性かつ親名が固有名詞でも is-a 関係とならない場合に発生する (例: 血液 ← 血球，千葉県の道路 ← 千葉県の道の駅)．再現率を下げる誤りは，親名が固有名詞でも is-a 関係が成り立つ場合 (例:日本人 ← 帰国子女，最上氏 ← 最上義光) や，子名の前方が親名と一致しても is-a 関係が成り立つ場合 (例：沖縄県営鉄道 ← 沖縄県営鉄道糸満線，映画 ← 映画作品) に発生する．しかし，全体から見ればこれらは少数の例外とみなせるため，結果として is-a 関係を高精度で判定できることを確認した．

3 汎用オントロジー構築手法

3.1 意味属性の設定

我々は Wikipedia のカテゴリを調査し，独自に Wikipedia のカテゴリと記事を分類するための意味属性を定義した．本手法では意味属性を “人”，“組織”，“施設”，“地名”，“地形”，“具体物”，“創作物”，“動植物”，“イベント” の計 9 種類に設定した．抽象的過ぎる概念以外を網羅していること，一般的な上位下位概念の粒度 10 前後の分類[4]とほぼ対応がとれることを考慮して意味属性を設定した．機械学習による分類器が作れるほどのカテゴリと記事数がないものや，語彙大系に対応付けが難しいものに関しては意味属性を設定しても分類精度が落ちるため，今回は対象外とした．サンプル調査の結果，Wikipedia のカテゴリでは全体の 86.3%，記事では 90.4%がいずれかの意味属性に分類された．表 1 に意味属性に対応する主な語彙大系のカテゴリと，単語の分類例を示す．

表 1: 意味属性に対応する主な語彙大系のカテゴリと，分類する単語の例

意味属性	語彙大系 カテゴリ	分類する単語例
人	人，職業	人名，職業，氏族，民族，魔物，マスコット
組織	仕事場，組織	企業，スポーツ団体，政党，公演組織，軍，家系
施設	公共機関，施設	交通路，学校，病院，文化施設，公園，城，競技場
地名	地域，国家	国，都市，場所，地域
地形	自然	山，川，峠，島，天体
具体物	無生物	物質，道具，食べ物
創作物	創作物，出版物	映画，音楽，番組，本，新聞，ソフトウェア，絵画
動植物	生物	動物，植物，体の一部
イベント	出来事，現象	事件，式，行事，社会運動，天気，物理現象，色，病気

※上記の対応する語彙大系のカテゴリは主なもので，実際は平均 130 個のカテゴリを対応づけている．また，語彙大系のカテゴリと意味属性は 1 対 1 で対応する．

3.2 is-a 関係の判定

本手法では，2 節の 4 つの項目を基準として is-a 関係でないリンクを判定する．項目 3 に対しては文字列照合を適用し，項目 4 を解決するための固有名詞判定には MeCab [5]の出力結果を用いた．本節では以降，項目 1，2 を解決するための，カテゴリと記事を意味属性に分類する手法について述べる．

3.2.1 カテゴリ分類

Wikipedia のカテゴリを SVM による分類器を用いて 9 種類の意味属性に分類する．多値分類を行うために one-vs-rest 法を用いた．素性作成にはカテゴリ名または周辺の単語を用い，形態素や品詞を素性にした．また，カテゴリ名の末尾の文字列とマッチする語彙大系のインスタンスに付与された，語彙大系のカテゴリ

[3]2008 年 7 月 24 日の Wikipedia からランダム抽出した 2,500 件調べ．以降の統計量は全て 2,500 件のサンプル調査による．

[4]関根の拡張固有表現 (http://sites.google.com/site/extendednamedentityhierarchy/) の第一階層 (10 カテゴリ) を参考にしている．これは語彙大系のカテゴリの第四階層とほぼ対応がとれる．

[5]http://mecab.sourceforge.net/

名及び表 3 で対応づけた意味属性名も素性にした[6].

本手法のカテゴリ分類では再現率の向上のため，前ステップで得られた出力を学習データに加えるブートストラップ的な手法を用いる．また，直前のステップまでに決定したカテゴリの意味属性をもとにした素性を追加することで，既に意味属性が決定したカテゴリの周辺カテゴリの意味属性を決定しやすくする．

3.2.2 記事分類

カテゴリ分類の後，SVM による分類器を用いて記事を 9 種類の意味属性に分類する．カテゴリ分類器と同様，素性作成には記事名または周辺の単語[7]と語彙大系を利用した．その他に，精度を向上させるために以下の 3 点を工夫した．

1. 記事が分類される可能性の高い意味属性に分類先を絞り，適合率の向上を図る

   記事分類の時点でカテゴリの意味属性は決定しているので，記事に付与されているカテゴリと同じ意味属性の中から記事の意味属性を選ぶ．例えば， 記事 “バーモンドカレー” に付与されているカテゴリが “日本のカレー (具体物)”，“ハウス食品 (組織)” だった場合，“バーモンドカレー” の分類先を，意味属性 “具体物”，“組織” に限定する．
2. カテゴリ名と記事名の類似性を判定し，適合率の向上を図る．

   記事名とカテゴリ名が似ていれば，そのカテゴリの意味属性が優位になるように素性を設計する．例えば，記事 “バーモンドカレー” とカテゴリ “日本のカレー (具体物)” は末尾の形態素が一致 (意味的に類似) するので，“バーモンドカレー” が具体物である可能性が高くなるように素性を設計する．
3. 既に意味属性が決定した記事を元に，分類器で分類できなかった記事を分類し，再現率の向上を図る．

   既に決定したカテゴリの意味属性と記事の意味属性が一致する割合を求め，この割合があらかじめ決めた閾値以上であれば，意味属性が未確定の記事を同じ意味属性に分類する．例えば，カテゴリ “カクテル (具体物)” に分類されている意味属性が決定した記事のうち，3 件が “具体物”(ウーロンハイ，ハイボール，ホットカクテル)，1 件が “人”(バーテンダー) だったとする．このとき，カテゴリと同じ意味属性である “具体物” の割合は 75%である．この割合を閾値とし，割合が閾値以上であれば未確定の記事 (カルピスサワー) を “具体物” に分類する．

[6]素性作成に使用した単語，素性の作成方法の詳細は我々が提案した手法 [7] を参照．

[7]記事に付与されるカテゴリ名，本文の第一文の形態素，第一文から文字列照合で抽出する上位語

3.3 オントロジー階層の再構成

is-a 関係でつながっているカテゴリと記事のまとまりを 1 つの階層と考えると，複数の階層ができることになる．異なる意味属性が付与された親子は is-a 関係でないとみなすため，1 つの階層に含まれるカテゴリと記事は全て同じ意味属性となる．そこで，3.1 節で設定した意味属性を最上位カテゴリとし，下位に同じ意味属性の階層を接続する．その際，階層の中で親を持たないカテゴリ及び記事を接続点とする．図 1 を例にすると，親を持たないカテゴリは “信号機”，“食品” であり，親を持たない記事は “コーヒー豆”，“新幹線回数券” である．

4 実験と考察

4.1 実験設定

2008 年 7 月 24 日時点での日本語 Wikipedia のダンプデータ[8]を使用して評価実験を行なった．カテゴリ数は 40,385 件，記事数は 469,023 件，カテゴリペア数は 85,355 件 (内，is-a 関係は 72.1%), カテゴリ-記事ペア数は 1,173,894 件 (内，is-a 関係は 74.3%) である[9]．カテゴリ分類器の精度評価は，評価データ 2,500 件の 5 分割交差検定で行なった．記事分類のための分類器の学習データと閾値の決定には，Wikipedia の記事に対して関根の拡張固有表現の分類を付与した渡邉らによる NAIST-jene のデータ 11,554 件を利用した[10]．記事分類の評価には，これとは別に用意した 2,500 件の評価データを用いた．is-a 関係判定の評価には，カテゴリペア，カテゴリ-記事ペアそれぞれ 2,500 件の評価データを用いた[11]．本稿では，単語の形態素，品詞を抽出するために，形態素解析器 Juman6.0[12]を使用した．SVM には TinySVM0.09[13]を利用し，カーネルには線形カーネルを用いた．

4.2 実験結果

本手法では初めに，カテゴリと記事を 9 種類の意味属性へ分類した．カテゴリ分類精度は適合率 98.0%，再現率 99.2%，記事分類精度は適合率 97.0%，再現率

[8]http://download.wikimedia.org/jawiki

[9]初めに，Wikipedia の内部向けのカテゴリや記事 (例:“画像:”, “Help:”)，オントロジーのカテゴリとして扱いにくいカテゴリ “1986 年生” などを文字列照合で取り除いた．

[10]NAIST Japanese ENE Dictionary on Wikipedia, http://sites.google.com/site/masayua/p/naist-jene　本手法の意味属性と，関根の拡張固有表現の第一階層は異なる部分があるので，一部修正して使用した．

[11]評価データ 2,500 件× 4 セットは，いずれもランダムサンプリング後のデータに人手で正解を付与したものである．作業者 1 名による正解付与データを用いた．他の作業者 1 名が同じデータに正解を付与した結果，正解の一致率は 98.7%であった．

[12]http://www-lab25.kuee.kyoto-u.ac.jp/nlresource/juman.html

[13]http://chasen.org/ taku/software/TinySVM/

表 2: カテゴリ，記事の意味属性別の is-a 関係判定精度

	カテゴリ間			カテゴリ-記事間		
意味属性	適合率	再現率	割合	適合率	再現率	割合
人	98.4	99.1	28.9	99.3	97.4	39.1
組織	91.1	95.4	10.0	85.0	91.9	7.6
施設	98.5	97.1	13.7	98.2	96.4	15.7
地名	93.2	93.9	8.6	92.6	96.6	6.6
地形	96.9	97.9	6.4	100	100	1.5
具体物	98.9	95.7	6.1	91.9	92.9	4.8
創作物	94.5	96.7	19.8	94.5	98.0	19.8
動植物	100	93.5	3.0	100	94.5	3.1
イベント	94.2	96.1	3.4	84.6	71.0	1.7
全体	94.8	97.0	100	95.2	96.2	100

※ 9 種の意味属性以外の is-a 関係を除いた精度
※ is-a 関係と意味属性が共に正しければ正解とする

96.8%であり，高い分類精度が得られた．カテゴリ名は普通名詞が多いため，意味属性に対応づけた語彙大系のカテゴリ情報との一致を素性にすることで高い精度が得られたと考えられる．記事名は固有名詞が多くカテゴリに比べて分類が困難だが，記事に付与されたカテゴリの意味属性を素性に用いることで高精度な分類ができたと思われる．Wikipedia のカテゴリ全体の 85.0%(34,145 件)，記事全体の 85.1%(369,154 件) が 9 種類の意味属性のいずれかへ分類された．

9 種類の意味属性に限定したカテゴリ間，カテゴリ-記事間の is-a 関係の意味属性別と全体の精度，全体からみた割合を表 2 に示す．is-a 関係の精度は，カテゴリ間で適合率 94.8%， 再現率 97.0%，カテゴリ-記事間で適合率 95.2%，再現率 96.2%と高精度であった．本手法での is-a 関係の判定誤りは，意味属性分類誤りと固有名詞判定誤りの 2 つが主な原因である．意味属性分類誤りによる is-a 関係の判定誤りは，主に “イベント” で多く発生する．イベント名はカテゴリ，記事の種類が多様なため，他の意味属性に比べて分類精度が低い．一方，固有名詞判定誤りが原因で is-a 関係判定精度が下がっている主な意味属性は “組織” と “地名” である．今回は固有名詞の判定に MeCab の出力結果のみを使用しているため，固有名詞判定の再現率が低い．そのため，“恵須取支庁 ← 名好町” のような part-of 関係や，“関東鉄道 ← 竜崎鉄道” のような現企業名と合併前の企業名の関係を is-a 関係とみなしてしまうため，適合率が低い．我々の手法 [7] でも高精度だった意味属性 “人” において is-a 関係判定精度が高い要因は，“人” は他の意味属性より，意味属性判定と固有名詞判定が容易だからである．

カテゴリ間の is-a 関係の判定精度は桜井らの手法と比較し，カテゴリ-記事間の精度は小林らの手法と比較した (表 3，4)．表 3，4 は表 2 の精度と異なり，9 種類の意味属性以外の is-a 関係も含んだ精度である．提案手法のカテゴリ間の is-a 関係の適合率は桜井らの手法より 2.4 ポイント低い．これは，桜井らの手法は is-a 関係を判定するための強力な文字列照合を用いて

表 3: カテゴリ間の is-a 関係判定精度の比較

	適合率	再現率	F 値	抽出数
桜井らの手法	97.6%	57.7%	72.5%	32,516
提案手法	95.2%	82.2%	88.2%	51,093
差分	-2.4	+24.5	+15.7	+18,577

※ 9 種の意味属性以外の is-a 関係も含んだ精度
※意味属性に関係なく，is-a 関係が正しければ正解とする

表 4: カテゴリ-記事間の is-a 関係判定精度の比較

	適合率	再現率	F 値	抽出数
小林らの手法	93.0%	67.9%	78.5%	642,221
提案手法	95.6%	92.4%	94.0%	850,394
差分	+2.6	+24.5	+15.4	+208,173

※ 9 種の意味属性以外の is-a 関係も含んだ精度
※意味属性に関係なく，is-a 関係が正しければ正解とする

いるためだと考えられる．一方小林らの手法と比較すると，is-a 関係の適合率は 2.6 ポイント高い．小林らの手法は文字列照合のみを用いるので，“センター” や “家” など多義性が原因の誤りが発生することが多いが，提案手法は機械学習による分類器を用いているので，多義性をうまく解消できたと思われる．再現率は，どちらの従来手法よりも，提案手法のほうが 24 ポイント以上高い．以上により is-a 関係でないリンクを判定することで，より網羅的に is-a 関係を抽出する提案手法の有効性が示された．

5 おわりに

本稿では，Wikipedia のカテゴリ階層と記事を利用し，幅広い分野を包含した is-a 関係のオントロジーを，高適合率，高再現率で構築した．意味属性分類器の精度向上，高精度な固有名詞判定手法を構築することで，is-a 関係の精度をさらに向上させることが今後の課題である．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] 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).

[2] 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).

[3] 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).

[4] 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).

[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] 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).

[7] 柴木 優美, 永田 昌明, 山本 和英:Wikipedia からの大規模な人オントロジー構築, 情報処理学会研究報告, 自然言語処理研究会報告 2010-NL-198-3 (2010).