2008－NL－187 (2)
2008／9／24

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

小林 暁雄† 増山 繁† 関根 聡‡

†豊橋技術科学大学 ‡ニューヨーク大学

概要:本研究では，日本語を意味により分類・体系化した日本語語彙大系と，大規模な百科事典である日本語ウィキペディアの知識とを結合したオントロジーを自動構築する手法を提案する．提案手法では，オントロジーのクラス階層を日本語語彙大系の一般名詞意味体系と一般語，日本語ウィキペディアのカテゴリから構築し，日本語ウィキペディアの各記事からオントロジーのインスタンスを取得する．

A Method for Automatic Construction of General Ontology by Merging GoiTaikei and Japanese Wikipedia

Akio Kobayashi† Shigeru Masuyama† Satoshi Sekine‡

†Toyohashi University Of Technology ‡New York University

Abstract: In this research, we propose a method to construct an ontology by merging goitaikei(a Japanese lexicon) and Japanese Wikipedia. In this method, ontology classes are extracted from goitaikei and Japanese Wikipedia categories, and ontology instances are extracted from Japanese Wikipedia pages.

1 はじめに

ウィキペディアは大規模な Web 百科事典であり，誰でも Web ブラウザを通じて記事を加筆・修正することができる．この特徴により，一般的な百科事典とは異なり，新しい記事の追加や既存の記事の更新が頻繁に行われている．また，ウィキペディアは記事間の多量なリンク構造や，語彙の一意性といった特徴を持っている．このため，知識抽出のための有用なコーパスとして，人工知能や情報検索，Web マイニングなどの研究分野で最近急速に注目を集めている．

ウィキペディアの各記事は階層構造のカテゴリに分類されている．このカテゴリ分類は，ウィキペディアの方針として，記事や他のカテゴリを分類するためのカテゴリであるか，または記事と関連の深いキーワードであるとしている．二つ目の方針により，記事と関連の深いキーワードもカテゴリとすることができるため，ウィキペディアのカテゴリの階層構造は，概念の上位下位関係を分類する大系としてはあまり適切ではない．

一方，日本語語彙大系 [2] は，約 30 万の日本語の単語が 3000 種類の意味の体系に分類された大規模日本語辞書である．この意味体系は，日本語の意味の上位下位関係を示した木構造になっており，ウィキペディアのカテゴリ階層と比較して，概念の上位下位関係を明確に示している．

本研究では，日本語語彙大系の意味体系の分類方針に沿うように，日本語ウィキペディアの記事とその所属するカテゴリを選定し，日本語語彙大系の意味分類とそれらの情報を結合する．これにより，日本語語彙大系の意味分類に従って明確に分類された階層構造を持つ，大規模な日本語汎用オントロジーを自動構築する手法を提案する．

2 関連研究

ウィキペディアは，その大規模な知識とカテゴリ階層や密なリンク構造などの特徴から，オントロジーを構築するための様々な研究に利用されている．

中山ら [3] は，ウィキペディアのリンク構造から，各記事の重要文を抽出し，その重要文を構文解析した結果から関係文を抽出し，オントロジーとする手法を提案している．

桜井ら [4] は，ウィキペディアのカテゴリ階層から，後方文字列照合と前方文字列除去という二つの手法によって上位下位関係を取得する試みを行っている．

Suchanek らは Yago[5] において，WordNet[1] とウィキペディアから汎用オントロジーを自動構築する手法を提案している．この手法では，ウィキペディアの各記事が直属するカテゴリから，オントロジーのクラスとするのに適したカテゴリを Conceptual Category として選出する．Yago[5] では，WordNet の synset をオントロジーのクラスと見なし，Conceptual Category をそのサブクラ

スとして設定することによって大規模かつ上位下位関係の明確なオントロジーの自動構築を可能としている．しかしながら，Yago[5] の Conceptual Category の選出方法は，各記事に対するカテゴリのうち，主要語として複数形名詞を含んでいるかどうかで判断しており，英語ウィキペディアに限定した手法となっている．また，WordNet との結合の際もこのような Conceptual Category の主要語である複数形名詞に対して，単数形名詞を含む synset に結合するという手法をとっており，結合するシソーラスは英語に限定される．このような Conceptual Category 選択手法は，各カテゴリの主要語が複数形名詞の場合は，分類を表すカテゴリである可能性が高いという英語ウィキペディアの特徴に基づくものである．このため，日本語のウィキペディアと日本語のシソーラスを用いて同様のオントロジーを構築するには，記事に関連したキーワードを示しているだけのカテゴリを取り除きつつ二つの知識を結合する手法を考案する必要がある．

本手法では日本語語彙大系の意味体系と，日本語ウィキペディアの各記事から抽出した上位語，各カテゴリとを照合することにより，各カテゴリのうち，分類を表すカテゴリを選出する．また，照合先の意味分類と各カテゴリとを結合することにより日本語語彙大系の意味体系に沿ったオントロジーの自動構築を行う．

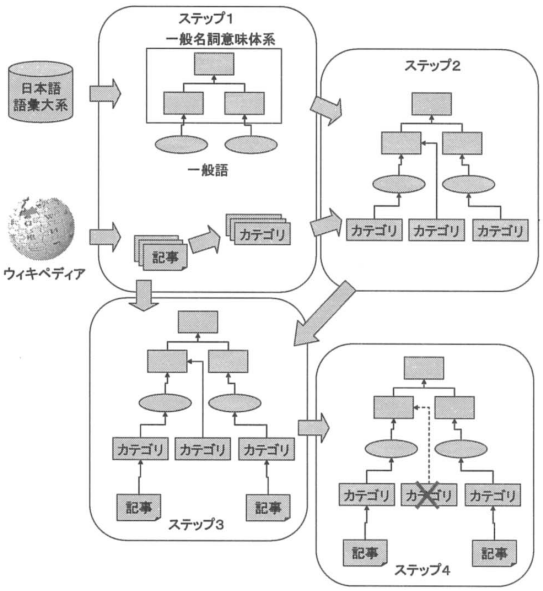

図 1: 提案手法の概要

3 使用するデータ

3.1 ウィキペディア

本手法では，ウィキペディアからは，各記事をオントロジーのインスタンスとして，記事の属するクラスをオントロジーの最下位のクラスとして抽出する．ウィキペディアのコンテンツは MySQL データベース・データを XML 形式とした形で提供されており，不定期で更新されている．このデータを MySQL にインポートし，そこから各記事と，その所属するカテゴリを取得する．

3.2 日本語語彙大系

日本語語彙大系からは一般名詞の意味体系と，一般語の一覧を用いる．日本語語彙大系には他にも固有名詞の意味体系，固有名詞，用言の意味体系などが含まれている．固有名詞の意味体系は行政区画などの概念については詳細な分類が存在するが，作品，出版物，商品などはその他固有名詞というカテゴリの下位に属しており，かなり粗い分類体系になっている．また，一般名詞の意味体系が 2700 種類の意味分類を持つのに対し，固有名詞の意味体系は 130 種類と分類が少ない．このことから，本手法では一般名詞の意味体系と一般語のデータを用いる．

4 ウィキペディア知識の日本語語彙大系への結合

本章では，ウィキペディアと日本語語彙大系の知識の結合手法を提案する．図 1 に本手法の概要を示す．本手法は，日本語語彙大系，日本語ウィキペディアからの知識の抽出，抽出された知識からのクラス階層構築，抽出された知識からのインスタンス生成，不適切なカテゴリ，インスタンスの除去の 4 ステップで構成される．以下に各ステップについて詳述する．

4.1 知識抽出

本節では本手法に用いた各コーパスの知識とその抽出手法の概要について述べる．

4.1.1 ウィキペディア記事・カテゴリの抽出

ウィキペディアは MySQL ダンプデータを公開している．このダンプデータには MySQL テーブルによって分類されたウィキペディアの全記事情報が含まれている．本手法では，このダンプデータの中から各記事を抽出する．各記事には，それぞれその記事が示す概念の直属するカテゴリが記述されている．このカテゴリも併せて抽出する（図 2）．

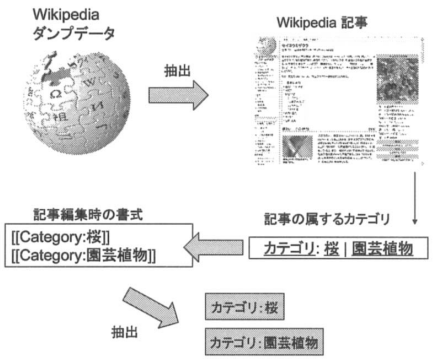

図 2: ウィキペディアダンプデータからの記事・カテゴリ抽出例.

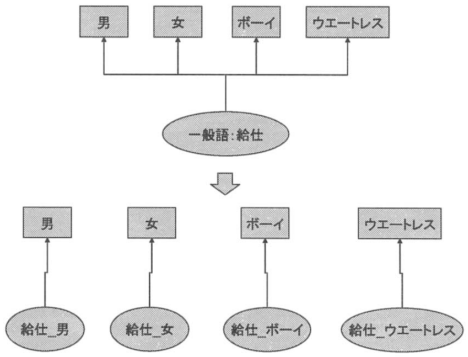

図 3: 日本語語彙大系の対応する意味分類による一般語の分割例.

4.1.2 日本語語彙大系知識の抽出

日本語語彙大系からは，一般名詞の意味体系と一般語の一覧から知識を抽出する．一般語は単一の一般名詞の意味に対応する場合と，複数の一般名詞の意味に対応する場合が存在する．複数対応している場合，対応先の意味に対する一般語をそれぞれ別々の単語として抽出する（図 3）．

4.2 クラスの生成

本節では前節で抽出された知識から，オントロジーのクラス階層を構築する手法について述べる．

具体的には，図 4 のように，4.1.1 において抽出された日本語ウィキペディアのカテゴリを 4.1.2 において抽出された日本語語彙大系の知識に結合したクラス階層を構築する．

4.1.2 によって抽出された知識は日本語語彙大系の一般名詞意味体系に一般語を各意味分類の下位に位置づけ，生成するオントロジーのクラス階層として扱う．

各知識源の結合方法はウィキペディアカテゴリ名と日本語語彙大系の知識（クラス名）とを後方文字列照合することによって行われる．照合先クラスの存在するウィキペディアカテゴリは日本語語彙大系から生成されたクラス階層の照合先クラスの下位に位置づけられる．照合先のないカテゴリは破棄される．

4.3 インスタンス生成

本節では，日本語ウィキペディアから抽出された各記事を，前節までに生成されたクラス階層のインスタンスとして設定する (クラスとインスタンスの間に instanceOf 関係を設定する) 手法について述べる．この手法は，各記事を，その属するカテゴリから得られたクラスのインスタンスとして設定するステップと，設定されているクラスとインスタンス間の instanceOf 関係が正しいかどうかを判定し，正しくない場合はこの関係を破棄するステップの 2 ステップからなる．以下に各ステップについて詳述する．

4.3.1 記事とカテゴリ間の対応関係からのインスタンス生成

4.1.1 において，各記事と，その記事の属するカテゴリを抽出した．各カテゴリは，前節において，オントロジーのクラスとして設定されている．そこで，対応するカテゴリに属している記事を，それらのカテゴリから生成されたクラスに対応するインスタンスとして設定する（図 5）．

4.3.2 instanceOf 関係の誤りの除去

ウィキペディアのカテゴリは，記事を分類するためのものか，あるいは記事に関連したキーワードのいずれかである．例えば，”千葉港”という記事に対して”千葉市”という関連したキーワードがカテゴリになっていた場合には，前小節の手法によって，”千葉市”というクラスが”千葉港”というインスタンスのクラスになっている．しかしながら，これらのクラスはインスタンスを分類するためのクラスではないため，この関係が正しいとは言い難い．そこで，このような，記事とそれに関するキーワー

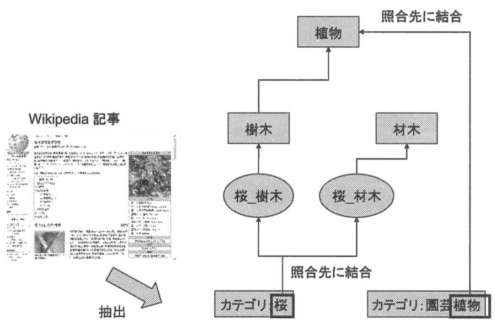

図 4: ウィキペディアカテゴリと日本語語彙大系知識との結合例.

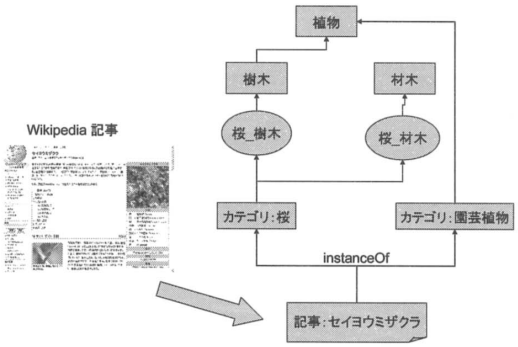

図 5: ウィキペディアカテゴリと日本語語彙大系知識との結合.

ドであるカテゴリから生成されたクラスとの instanceOf 関係を除去する手法を考案した.

この除去手法では, まず各記事の第一文から, 記事の上位語 (あるいは上位語を含む文) をパターンマッチングによって抽出する. パターンは [6] を参考に作成した 27 種類 (パターン例:表 1) を用いる. このような上位語について, 4.2 と同様に, 日本語語彙大系から生成されたクラス階層との後方文字列照合を行う. 照合先と, 現在属しているクラスとの間に上位下位関係が無い場合, このクラスとインスタンス間の instanceOf 関係を除去する. 例として, 図 6 では, パターンマッチングにより定義文から抜き出した上位語"ヨーロッパ、北西アフリカ、西アジアに自生するサクラ属の植物"とクラス階層とを照合し, クラス"植物"と上位下位関係にないクラス"桜_材木"のサブクラスとしての"カテゴリ:桜"との instanceOf 関係を除去している ("カテゴリ:桜"は"桜_樹木"のサブクラスとしても抽出されており, "桜_樹木"は"植物"の下位のクラスであるため, 最終的には"セイヨウミザクラ"は"カテゴリ:桜"のインスタンスとして設定される).

表 1: 上位語を抽出するパターン例

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

4.4 不要なクラス・インスタンスの破棄

本手法では, 不要なクラス, インスタンスを破棄する. クラスについては, インスタンスを一つも持たないクラスは不要であるとして破棄する. また, 前節までの手法により, どのクラスとも関係を持たなくなった, クラス階層から参照不可能なインスタンスを破棄する.

5 評価実験

前章で解説した結合手法について, 実験を行った.

構築されたオントロジーの規模は, インスタンス数が 179,399 件, instanceOf 関係は 283,206 件, ウィキペディアより得られたクラス数が 19,426 件となった. 日本語語彙大系より得られたクラスは日本語語彙大系の一般名詞の意味体系と一般語を分割したものを合わせて 95,053 件となった.

以下に手法の全ステップの内, クラスの結合ステップと, インスタンスの生成ステップの 2 ステップについて実験を行い, その評価, エラー解析について詳述する. 記事抽出ステップについては, コーパスからオントロジー構築に必要なデータを抽出してくるステップであり, 各コーパスの提供しているデータのうち必要な項目を選択しているだけであるため, 評価は行わない. また, 不要なクラス・インスタンスの除去ステップについては, 除去後のクラス・インスタンスの件数を以下の各ステップより得られたクラス・インスタンス数の最終的な値として併記した. また, 以下の各ステップにおける正解率の評価は, この最終的なクラス・インスタンスに対して行っている.

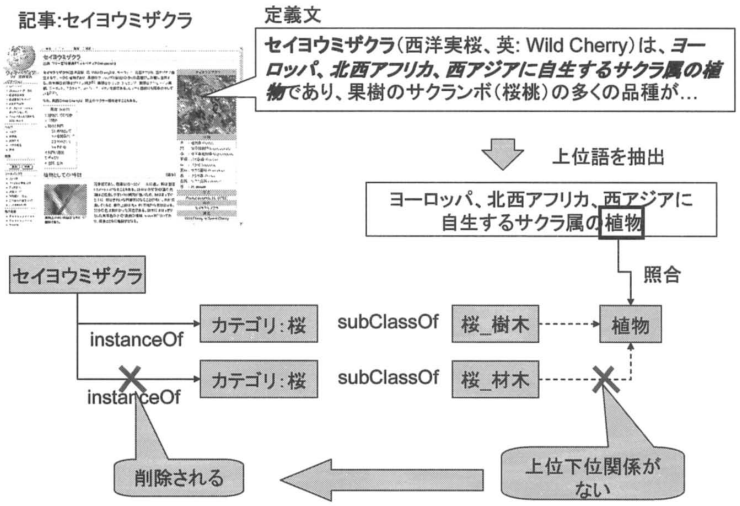

図 6: 不適切 instanceOf 関係の除去手法例.

5.1 クラスの結合実験

本節では，ウィキペディアから得られたクラスと日本語語彙大系から得られたクラスの結合手法の実験について詳述する.

5.1.1 クラス結合実験概要

節 4.2 において提案した手法の評価実験を行った.

この実験では，日本語語彙大系から得られたクラス 95,053 件に対し，ウィキペディアのカテゴリ 43,071 件をクラスとして結合を行い，その結果を評価する.

日本語語彙大系より得られたクラス階層は，日本語語彙大系の一般語を図 3 のように対応する意味分類毎に分割したクラスと，意味分類自体からなるクラスによって構成されている．よって，このクラス階層の正解率は日本語語彙大系の一般名詞意味分類と同一であるため，本手法では正解率 100%であると信頼し，評価を行わない.

日本語語彙大系から得られたクラスと日本語ウィキペディアから得られたクラスとの結合の正解率については，本手法により自動で行われているため，評価を行う．評価手法は以下のように人手により行う.

調査に使用したデータは，ウィキペディアからのクラスと日本語語彙大系からのクラスの対応一覧から 1,000 件のサンプルをランダムに抽出したものである．これらのサンプルを別々に抽出したデータを 3 名により調査し，その結果の平均を求めた.

5.1.2 クラス結合実験結果

クラスの結合については，記事から抽出されたウィキペディアのカテゴリ 43,071 件のうち，日本語語彙大系より得られたクラス階層に照合先を持たなかったために，クラスとして設定されなかった（破棄された）カテゴリは 5,090 件であった.

また，この中からインスタンスを持たないものを除去した結果，最終的な結合されたクラスは 19,426 件であった．人手によるサンプルの評価はこの最終的な結合されたクラスに対し行った．その結果，約 93%の正解率であった.

破棄されたカテゴリとして，表 2 にいくつか例を示す.

表 2: 破棄されたカテゴリ例

カテゴリ名	属している記事例
日本のインターチェンジ　あ	あきる野インターチェンジ
漫画作品　た	∀ガンダム
北海道鉄道（初代）	ニセコ駅
日本の国会議員 (1890-1947)	一条実輝
古典期	マヤ
中央アメリカ	ロスアルトス_(中央アメリカ)

表 2 の 1，2 行目のような五十音順の分類に分けられたカテゴリ”日本のインターチェンジ　あ”などが最も多く，

次に括弧書で注訳が付与されたもの（表 2 中 3 行目，”北海道鉄道（初代）”等）が多く破棄されていた．括弧書により除去されたカテゴリの中には表 2 の 4，5 行目の”埼玉県の市町村（廃止）”や”日本の国会議員 (1890-1947)”といった，オントロジーのクラスとして利用できそうなものも含まれていた．このことから，括弧が付与されているカテゴリについては，照合の際に括弧書部分を含まないカテゴリ名として扱うなどの場合分けの導入を検討する必要がある．6 行目のカテゴリ”マヤ”，7 行目のカテゴリ”中央アメリカ”は日本語語彙大系の一般語，一般名詞意味分類共に”マヤ”，”アメリカ”が含まれていないため削除された．

5.1.3 クラス結合実験のエラー解析

ウィキペディアから抽出されたクラスと日本語語彙大系の意味体系から得られたクラスとを結合する際に発生したエラーの一部を表 3 に示す．

表 3: クラス結合時に発生したエラー例

ウィキペディアからのクラス	意味体系からのクラス
1999 年のシングル	シングル_身の上
仮面ライダーシリーズの登場キャラクター	キャラクター_類型
スクウェア・エニックスのゲームソフト	ソフト_方法
豊富町	町_道路
ソビエト連邦の人物	人物_男
学校法人の理事長	長_度量衡
日本の海軍工廠と造船所	造船所_工場
2002 年のコンピュータゲーム	ゲーム_スポーツ

エラーの原因は大きく二つに大別される．

一つは日本語語彙大系から得られたクラスにそもそも正しい結合先が存在しない場合である．表 3 中の 1-3 行目のエラーがこれに該当する．

もう一つは，日本語語彙大系の一般語が，一般名詞意味体系の複数の意味に対応している場合である．

さらに，後者の場合は，インスタンス生成の際の，不適切な instanceOf 関係除去において，インスタンスとなる記事の照合するクラスの種類によって，以下の二つに原因が分別される．

- インスタンス生成の際の意味体系の照合先と，クラス結合の際の照合先が同じであり，複数の照合先の中から正しいものが選択できない．
- 不適切な instanceOf 関係除去の際の除去漏れ

照合先が同一であった場合，本手法では日本語語彙大系の一般語の分類の先頭にあるものを用いている．このため，不適切なクラスが結合先として選ばれる可能性があり，それに起因するエラーが発生している．表 3 中 6，8 行目がこれに該当する

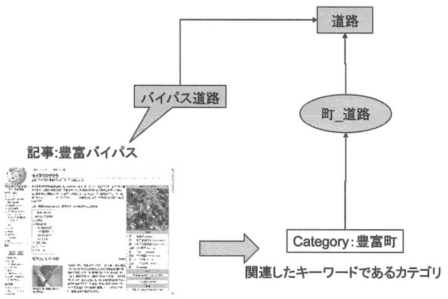

図 7: 不適切 instanceOf 関係除去漏れによる結合エラー例．

6 行目のエラーに対しては，日本語語彙大系の意味分類中に人物の地位を示す”長”という分類が存在し，かつ一般語の一覧にも同様の語彙が存在するため，そちらを選択すべきであるが，この問題は結合先の曖昧さから発生しており，この曖昧性解消については更に深く追求する必要がある．

また，8 行目のエラーに対しては，日本語語彙大系の意味分類中に”ゲーム”という分類が存在し，”スポーツ”とは別の分類体系となっており，こちらに分類されるべきであると考えられる．

一方，不適切な instanceOf 関係の除去漏れに起因するエラーとしては，表 3 中 4，5 行目のエラーが該当する．4 行目については，”豊富町”カテゴリに属する記事のうち，道路に関する記事（”豊富バイパス”など）が含まれており，意味体系から取得されたクラス間の上位下位関係，”町_道路”と結びついてしまっている（図 7）．5 行目のエラーに関しては，日本語語彙大系の一般語”人物”が”人物_男”，”人物_幹部”，”人物_人間”といったクラスに分割されており，その内で，”人物_人間”を選択すべきところをそれ以外を選択してしまったエラーである．このエラーでは，”ソビエト連邦の人物”中にある男性についての記事があった為に発生しているが，もしなんらかの組織の幹部にあたる人物についての記事が存在していた場合，”人物_幹部”が”ソビエト連邦の人物”の上位クラスとして選択されていた可能性もある．このことから，この問題を解消するためには，より正確に上位クラスを選択するための手法が必要であると考えられる．

また，ウィキペディアのカテゴリ名の記述において，表 3 中 7 行目の”日本の海軍工廠と造船所”カテゴリのように，二つの概念が併記されたカテゴリについては，後

者（”造船所”）の対応するクラスの下位に設定されるエラーが発生していた．

5.2 インスタンス生成実験

本節では，ウィキペディアの記事をインスタンスとして設定する手法の実験について詳述する．

5.2.1 インスタンス生成実験概要

節 4.3 において提案した手法の評価実験を行った．この実験では，ウィキペディアから得られた記事をインスタンスとしてオントロジーのクラスとの間に instanceOf 関係を設定し，その後，そこから不適切な instanceOf 関係の除去を行い，その結果を評価する．

5.2.2 インスタンス生成実験結果

本手法により生成されたインスタンス数は 305,465 件となった．また，この手法により設定された instanceOf 関係は 283,206 件となった．さらに，生成されたインスタンスのうち，どのクラスとも instanceOf 関係を持たないものを除いた結果，最終的なインスタンス数は 179,399 件となった．

ウィキペディアの記事数が約 40 万件であることから，インスタンス数は約半数まで減少している．これは，不適切な instanceOf 関係の除去において，例えば，”インターネット・バブル”という記事は，”インターネット”，”流行語”，”デジタル革命”といった関連したキーワードにのみ分類されており，このような記事をインスタンスとしたときに，その属するクラスとの間の instanceOf 関係がすべて除去された為発生している．特に，分類を示すクラスが一つしかないインスタンス（”セイヨウショウロ”instanceOf”食用キノコ”等）は，その instanceOf 関係が誤って除去されてしまうことにより，不要なインスタンスとして削除される事になる．このような原因によってインスタンス数はウィキペディアの全記事数の半数程度になっている．

また，不適切な instanceOf 関係の除去を行った後の最終的なインスタンスとクラス間の instanceOf 関係の正解率を調査した．調査方法はクラスの正解率を調査したときと同様である．その結果，99%の正解率が得られた．

5.2.3 インスタンス生成実験のエラー解析

instanceOf 関係のエラーは，不適切な instanceOf 関係除去の際の漏れに起因するエラーである．表 4 にその例を示す．これらのエラーに起因するエラーが前節のクラス間の結合にも発生している．

2 行目の例については，図 7 のように，日本語語彙大系の意味分類において，”町”という一般語が”道路”に分類されており，かつ”豊臣町”というカテゴリの中に道路に関する記事”豊臣バイパス”が存在したため（更に，”豊臣バイパス”の定義文からこの記事の上位語として”バイパス道路”がパターンマッチングによって取得されたことで，”豊臣バイパス”と”道路”の間には上位下位関係があると判断された），”豊臣町”は”道路”を意味する”町”であり，”豊臣バイパス”のようなインスタンスを分類するためのクラスとしてふさわしいと解釈されてしまったことによるエラーである．このようなエラーは，クラス結合におけるエラーを併発させる原因になっている．特に”町_道路”にまつわるエラーについては同様のエラーが散見された．このことから，”町”については行政区画の意味で使用される場合が殆どであるという統計情報を，なんらかの事前分布を用いることによって取得する必要がある．

表 4: 不適切な instanceOf 関係の除去漏れによるエラー例

インスタンス	クラス
TOSS	教育
豊臣バイパス	豊富町
百葉箱	気象学教育
女子大生	学校文化
あっち向いてホイ	じゃんけん
Webcat	検索

6 まとめと今後の課題

本手法では，日本語語彙大系と日本語ウィキペディアからオントロジーを構築するために，日本語ウィキペディアの記事，記事の直属するカテゴリと日本語語彙大系の意味体系を照合する手法を提案した．これにより，約 18 万件のインスタンスと日本語語彙大系の意味分類を含んだ約 2 万件のクラスによるオントロジーを自動構築することができた．インスタンスとクラス間の接続の正解率は 99%と高精度であった．また，日本語ウィキペディアから抽出されたクラスと，日本語語彙大系の意味体系から取得されたクラスを結合させた際の正解率は約 93%と高精度であった．

しかし，本提案手法には未だ解決すべき課題がいくつか存在していることが実験結果より明らかになった．

構築されたオントロジーの規模に関しては，高精度ではあるが，インスタンス数がウィキペディアの全記事数の約半数程度に減少してしまっている．これを解決するために，手法により除去されてしまった記事・カテゴリを再びオントロジーに取り入れる手法の開発が必要である．特に，不要な instanceOf 関係の除去において削除されてしまった記事が多く，除去手法の改良は今後の最も重要な課題である．

日本語ウィキペディアから抽出されたクラスと，日本語語彙大系の意味体系から取得されたクラスの結合においては，まず，日本語語彙大系の一般語において，現在では一般的な語であっても登録されていないため，誤った結合が発生してしまうという問題がある．また，日本語語彙大系において，一般語がどの意味を指しているのかが不明であるという意味体系の対応先のあいまい性に起因する分類の問題が発生している．これについては，なんらかの事前分布を用いることにより，統計的に有為な対応先を選択する手法の導入を検討している．事前分布については Google N-gram[7] の単語の共起情報などを検討している．

不適切な instanceOf 関係の除去手法の漏れについては，さらに別な情報（他の百科事典や固有名詞の辞書など）を用いた，より厳密な除去手法を開発する必要がある．

また，一般語をウィキペディアの曖昧さ回避のページから日本語語彙大系に追加することにより，日本語語彙大系の一般語の語彙を拡張することを検討している．

今後は上記の問題点を解決していくとともに，各記事の属性が記述された Infobox から属性を自動的に抽出し，各記事の示す概念を表現する精度を向上させていく予定である．

謝辞

日本語語彙大系を研究用データの形で提供いただいた日本電信電話株式会社に深謝の意を表す．

参考文献

[1] C. Fellbaum editor. *WordNet: An Electronic Lexical Database.* MIT Press, 1998.

[2] 日本語語彙大系，岩波書店，1999

[3] 中山浩太郎：ウィキペディアマイニングによる大規模 Web オントロジの実現，第 22 回人工知能学会全国大会

[4] 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ウィキペディアの適用可能性，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，2008

[5] Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum: Yago: a core of semantic knowledge, Proceedings of the 16th international conference on World Wide Web, ACM, pp.697-706, 2007.

[6] 隅田飛鳥，吉永直樹，鳥澤健太郎，萬成賢太郎：ウィキペディアからの大規模な上位下位関係の獲得，言語処理学会第 14 回年次大会，pp.769-772，2008

[7] 大規模日本語 n-gram データの公開：http://googlejapan.blogspot.com/2007/11/n-gram.html