1I2-R-4-6

日本語 Wikipedia オントロジーと日本語 WordNet の統合

Integrating Japanese Wikipedia Ontology and Japanese WordNet

森田 武史[*1] 玉川 奨[*2] 山口 高平[*2]
Takeshi Morita Susumu Tamagawa Takahira Yamaguchi

[*1] 青山学院大学 社会情報学部 School of Social Informatics, Aoyama Gakuin University
[*2] 慶應義塾大学大学院 理工学研究科 Graduate School of Science and Technology, Keio University

Japanese Wikipedia Ontology (JWO), which we have constructed semi-automatically from Japanese Wikipedia, has problems with a lack of upper classes, and appropriate definitions of class-instance relationships and properties. The purpose of our research is to complement the upper classes in JWO by integrating JWO and Japanese WordNet (JWN). In order to accomplish this, we develop tools which support the user to refine the class-instance relationships, to identify JWO classes to be aligned with JWN synsets, and to align the JWO classes with the JWN synsets through user interaction. In addition, we integrate JWO and JWN by using a domain ontology development environment, DODDLE-OWL.

1. はじめに

日本語 Wikipedia から自動構築したオントロジーである日本語 Wikipedia オントロジー(JWO)[玉川 10]には，上位クラスの欠如，クラスーインスタンス関係の誤り，継承を考慮したプロパティ定義がクラスになされていないなどの問題があった．本研究では，JWO と日本語 WordNet(JWN)を統合することにより，JWO における上位クラスを補完することを目的とする．本目的を達成するために，クラスーインスタンス関係の洗練およびアライメント対象クラスを同定するためのツール，JWO におけるクラスと JWN における Synset(同義語セット)のアライメントを支援するためのツールを開発する．また，領域オントロジー構築支援環境 DODDLE-OWL[Morita 08]を用いて JWO と JWN を統合する．

2. 日本語 Wikipedia オントロジーの問題点

日本語 Wikipedia オントロジー(JWO)は，日本語 Wikipedia における様々なリソース(カテゴリツリーや一覧記事など)から，概念および概念間の関係(is-a 関係やクラスーインスタンス関係など)を抽出することにより構築した，高精度かつ大規模な汎用オントロジーである．

現状の JWO には，上位クラス(抽象的なクラス)が欠如しているという問題がある．JWO では，主にカテゴリ階層からクラス階層を自動構築しているが，Wikipedia におけるカテゴリは記事を分類するために作成されるため，記事の分類に寄与しない抽象的なカテゴリは，わずかにしか作成されていない．また，カテゴリ階層は，「学問」，「技術」，「自然」，「社会」，「地理」，「人間」，「文化」，「歴史」，「総記」の 9 つの「主要カテゴリ」がルートカテゴリとなっている．これらのルートカテゴリは，オントロジーにおけるクラス階層の上位クラスとしては，不十分だと考えられる．上位クラスが欠如しているため，現在，JWO におけるクラス階層には約 3,000 個のルートクラスが存在し，断片的な Is-a 関係の集合体となっている．

本研究では，オントロジーアライメント(OA)の技術を用いて，JWO と日本語 WordNet(JWN)を統合することにより，JWO における上位クラスの補完を試みる．JWN は，独立行政法人情報通信研究機構(NICT)が開発した日本語の意味辞書であり，人手で作成された英語の意味辞書である Princeton WordNet 3.0 に準じているため，抽象的な概念(クラス)が数多く含まれている．

また，本研究では全自動で JWO と JWN の統合を行うのではなく，計算機では高精度に処理ができない箇所については，ユーザに最終的な判断を委ねる方針を取っている．例えば，OA 技術により獲得した JWO のクラスと JWN の Synset の対応候補をランク付けしていくつかユーザに提示し，Synset の定義文や上位・兄弟・下位 Synset などを参照しながら，最終的な対応付けはユーザが行うようにしている．

3. 日本語 Wikipedia オントロジーと日本語 WordNet の統合

3.1 概要

日本語 Wikipedia オントロジー(JWO)と日本語 WordNet(JWN)の統合の手順を図 1 に示す．主に，以下の手順に従って統合を行う．

1. クラスーインスタンス関係の抽出
2. クラスーインスタンス関係の洗練とアライメント対象クラスの同定
3. JWO のクラスと JWN の Synset のアライメント
4. DODDLE-OWL を用いた JWO と JWN の統合
5. 冗長なクラスーインスタンス関係の除去

以下では，各手順の詳細について説明する．なお，本稿では，2010 年 11 月時点の JWO と JWN ver.1.1 を利用している．

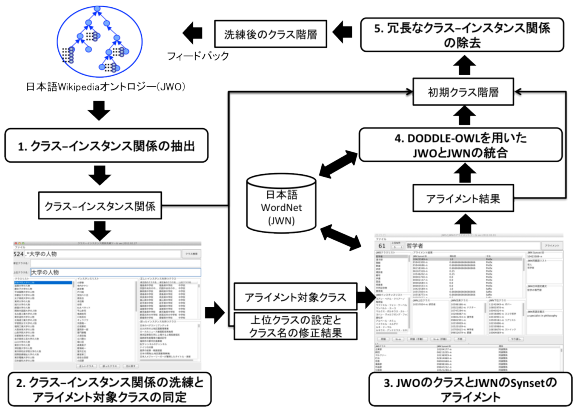

図 1: JWO と JWN の統合手順

連絡先：森田 武史，青山学院大学 社会情報学部，
〒252-5258 相模原市中央区淵野辺 5-10-1，
Tel: 042-759-6119，E-mail: t_morita@si.aoyama.ac.jp

3.2 クラスーインスタンス関係の抽出

本研究では，JWO と JWN を統合するために，JWO におけるインスタンスを持つクラスと JWN の Synset のアライメントを試みる．JWO における上位クラスの補完という点では，JWO のルートクラスと JWN の Synset をアライメントする方法も考えられる．しかしながら，JWO のルートクラスは約 3,000 個と数が多いことと，JWO の上位クラスには，is-a 関係の誤りが多く含まれているため，本研究では JWO におけるインスタンスを持つクラスのみを JWN と統合することとした．JWO におけるインスタンスを持つクラス数は 3,010 個であった．また，クラスーインスタンス関係数は，434,939 であった．

JWO では，主に一覧記事からクラスーインスタンス関係を抽出しているが，一覧記事は人手で編集されているため，記事の掲載漏れが生じる可能性がある．その場合，インスタンスのタイプが欠落する．例えば，JWO には「～県出身の人物」クラスが存在するが，それらのインスタンスは，「日本の俳優」や「数学者」といったような，より特化したタイプも同時に持つことが多い．しかしながら，一覧記事への掲載漏れにより，特化したタイプを持たないインスタンスも数多く存在している．本研究では，この問題を解決するために，Wikipedia における infobox から抽出したインスタンストリプルより，インスタンスのタイプの補完を試みる．インスタンストリプルの中には，JWO におけるクラス名と同名のインスタンスを目的語とするトリプルが存在する．これらのトリプルの中には，クラスーインスタンス関係とみなせる関係も含まれていると仮定した．クラス名と同名のインスタンスを目的語とするトリプル数で順位付けしたプロパティ一覧を表 1 に示す．これらのプロパティを持つトリプルを確認したところ，表 1 に太字で示した「職業」，「種類」，「種別」プロパティについては，クラスーインスタンス関係と同様の関係を表していることがわかった．そこで，これらのプロパティを「rdf:type」プロパティと同様とみなして，クラスーインスタンス関係の抽出を行った．インスタンスを 10 個以上持つクラスーインスタンス関係を抽出したところ，203 個のクラス，27,821 のクラスーインスタンス関係を抽出することができた．

一覧記事から抽出したクラスーインスタンス関係とインスタンストリプルから抽出したクラスーインスタンス関係を合わせると，3,185 個のクラス，462,247 のクラスーインスタンス関係が抽出できた．

表 1: クラス名と同名のインスタンスを目的語とするトリプル数で順位付けしたプロパティ一覧

順位	プロパティ名	トリプル数	順位	プロパティ名	トリプル数
1	国籍	12083	12	テレビドラマ	2957
2	言語	9076	13	地方	2410
3	国	8951	14	生国	2305
4	ジャンル	8867	15	FIFAワールドカップの成績	2300
5	製作国	8140	16	運動部	2030
6	スタッフ	7473	17	施設	1915
7	業種	6713	18	所属政体	1859
8	部活動	5795	19	駅周辺	1573
9	本社所在地	5516	**20**	**種類**	**1530**
10	**職業**	**4429**	21	出身有名人	1511
11	出身地	3175	**22**	**種別**	**1350**

3.3 クラスーインスタンス関係の洗練とアライメント対象クラスの同定

JWO のクラスーインスタンス関係は，自動抽出しているため，誤りが一定数含まれている．誤ったインスタンスを持つクラスを JWN の Synset とアライメントしないように，あらかじめ，クラスーインスタンス関係の洗練を行う．また，JWO におけるインスタンスを持つクラスの中には，ある地域のスポーツ選手やある地域出身の人物など，ハイブランチなクラスが多く含まれている．JWN の Synset とアライメントを行う際には，例えば，「日本のスポーツ選手」，「アメリカのスポーツ選手」などのクラスがあった場合，「スポーツ選手」のみ JWN とアライメントを行うようにすることで，アライメント対象クラス数を削減したい．同時に，不適切なクラス名については，修正も行いたい．これらの問題を解決するために，図 2 に示すクラスーインスタンス関係の洗練およびアライメント対象クラスを同定するためのツールを実装した．

本ツールは，クラスーインスタンス関係を入力として，画面左側のクラスリストにインスタンスを持つクラス一覧を表示する．クラスを選択するとそのクラスが持つインスタンスを 100 個，画面中央のインスタンスリストに表示する．クラスーインスタンス関係は一覧記事やインスタンストリプルからあるパターンに従って抽出されているため，100 程度のクラスーインスタンス関係を確認し，誤りが含まれていなければ残りもおおよそ正しいとみなす．ユーザは，クラスーインスタンス関係が正しければ，「正しいクラス」ボタンを，誤っていれば「誤ったクラス」ボタンを押して，インスタンスを持つクラスの正誤を判定していく．その際に，不適切なクラス名については，修正できる．また，ハイブランチなクラスについては，上位クラス名を設定できる．さらに，正規表現を用いて特定のパターンを持つクラスを一括して検索し，まとめて上位クラスを設定することも可能である．図 2 では，「大学の人物」と後方文字列照合するクラスを検索し，上位クラスとして「大学の人物」を設定している．複数クラスを選択した状態で「正しいクラス」ボタンを選択することで，一括して上位クラスの設定ができる．

本ツールを用いて 3.2 節で抽出したクラスーインスタンス関係の洗練およびアライメント対象クラスの同定を試みたところ，1 人のユーザで約 7 時間かかった．洗練後の全クラスは 2,947 個，クラスーインスタンス関係数は 449,186，上位クラスを設定したクラスは 2,558 個，修正したクラスは 37 個，アライメント対象クラスは 736 個であった．ここで，アライメント対象クラスは，正しいインスタンスを持つクラスの中で，上位クラスが設定されている場合にはそのクラスを，設定されていない場合には，元のクラスを対象とした．ただし，修正がある場合には修正後のクラスをアライメント対象クラスとした．

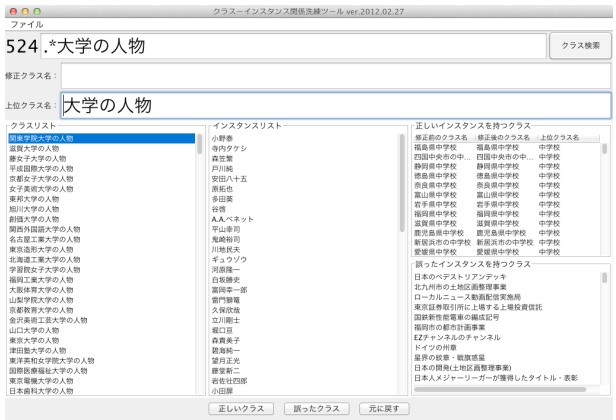

図 2: クラスーインスタンス関係の洗練およびアライメント対象クラスを同定するためのツール

3.4 日本語 Wikipedia オントロジーのクラスと日本語 WordNet の Synset のアライメント

JWO と JWN を統合するために，オントロジーアライメント(OA)の技術を適用する．OA とは，異なるオントロジーにおいて意味的に関連(同値関係，包含関係，排他的関係など)するエンティティ(クラス，プロパティ，個体など)間の対応を見つけることを意味する．OA における概念の類似性の尺度として，語類似度，語リスト類似度，概念階層類似度，構造類似度など，これ

までに様々な手法が提案されてきており，その有効性が確認されている[市瀬 08]. 通常，OA は，比較的類似するオントロジー間への適用が想定されている. しかしながら，本研究で統合を試みる JWO と JWN は，大きく構造が異なっている. 例えば，JWO におけるクラスにはインスタンスが存在するが，JWN の Synset にはインスタンスが存在しない. そのため，インスタンスを利用した OA 手法は本研究では適用できない. また，JWO には，具体的なクラス(インスタンスに近いクラス)が多く存在し，抽象的なクラスが少ないのに対して，JWN は，その逆である. そのため，概念階層類似度を用いた手法も大きな効果が期待できない. 以上より，本研究では，主に語類似度を用いて JWO と JWN の統合を試みる. 語類似度を用いた文字列に基づく 4 種類の手法として，プレフィックス，サフィックス，編集距離，n グラムを用いた.

語類似度を用いた手法の精度を高めるためには，JWO におけるクラスの同義語セットが必要となる. しかしながら，現状の JWO におけるクラスの同義語セットは，リダイレクトリンクのみから抽出しており，精度は 7 割程度と低く，多くのクラスには同義語の定義がなされていない. そのため，本研究では，JWO におけるクラスの URI のローカル名と JWN の Synset について，語類似度を用いた文字列に基づく 4 種類の手法を適用する.

JWO におけるクラスの同義語セットを利用していないために，アライメントの精度は高くないことが予想される. そのため，ユーザとのインタラクションを通して，JWO におけるクラスと JWN における Synset のアライメントを支援するツールを実装した(図 3). 本ツールでは，3.3 節で説明した方法で同定した JWO のクラスと JWN の Synset のアライメントを動的に行うことができる. JWO のクラスを選択し，アライメントボタンを押すと，語類似度を用いた文字列に基づく 4 種類の手法により求めた類似度の高い上位 N 件(ユーザにより指定可能)の JWN の Synset が表示される. Synset を選択すると，同義語セットと日本語および英語の定義文，上位，兄弟，下位クラスが同時に表示される. ユーザは，アライメント結果に対応する Synset が存在しない場合には，上位，兄弟，下位クラスからもアライメント対象 Synset を選択できる. また，同値関係または Is-a 関係ボタンを押すことで，JWO のクラスと JWN の Synset の対応関係を保存することができる.

表 2 に JWO のクラスと JWN の Synset のアライメント結果を示す. 表 2 の中で，「同値関係」と「Is-a 関係」は，アライメント対象の JWO のクラスと JWN の Synset とのアライメント結果の中から，同値または Is-a 関係を選択した数を示している.「同値関係(手動)」と「Is-a 関係(手動)」については，JWO のクラスのローカル名によるアライメント結果の中には，同値または Is-a 関係が存在しなかった場合を意味している. ユーザが類似するキーワードを入力し，再度アライメント実行したか，または，選択した Synset の上位，兄弟，下位クラスの中から同値または Is-a 関係を選択した結果となっている.「不明」は，対象となる JWO のクラスに対応する JWN の Synset が発見できなかったことを意味する. なお，本ツールを用いて，1 人のユーザが約 6 時間かけて，736 個の JWO のアライメント対象クラスと JWN の Synset のアライメントを行った.

3.5 DODDLE-OWL を用いた日本語 Wikipedia オントロジーと日本語 WordNet の統合

JWO と JWN を統合するために，本研究では領域オントロジー構築支援環境 DODDLE-OWL[Morita 08]を用いる. DODDLE-OWL は，領域専門文書を入力として，WordNet や EDR 電子化辞書などの汎用オントロジーを参照オントロジーとして，領域オントロジーを構築可能なツールである. 本研究では，JWO のクラスと JWN の Synset のアライメント結果を DODDLE-OWL に入力し，対応する JWN の Synset から JWN のルートクラスまでのパスを抽出し，合成する. また，入力した JWN の Synset 間の位相関係(祖先・親子・兄弟関係)を保持することに貢献しない，抽出したパスに含まれる JWN の Synset を削除する. これにより，JWO のインスタンス分類に特に貢献するクラス階層が構築できる. さらに，DODDLE-OWL の多重継承除去機能により，構築したクラス階層から多重継承を除去する. その後，3.4 節で述べた，JWO におけるクラスと JWN における Synset のアライメントを支援するためのツールの出力結果より，is-a 関係として対応づけた JWO のクラスを対応する JWN の Synset の下位クラスとして追加する. また，3.3 節で述べた，クラス−インスタンス関係の洗練およびアライメント対象クラスを同定するためのツールの出力結果より，上位クラスを設定していたクラスを，設定した上位クラスの下位クラスとして追加する. 同時に，元の JWO におけるクラス−インスタンス関係より，対応するクラスにインスタンスを追加する. 最終的には，クラス数 3,453 のクラス階層が構築できた.

図 3: JWO におけるクラスと JWN における Synset のアライメントを支援するためのツール

表 2: JWO のクラスと JWN の Synset のアライメント結果

関係名	関係数
同値関係	489
同値関係(手動)	90
Is-a関係	17
Is-a関係(手動)	135
不明	5

3.6 冗長なクラス−インスタンス関係の除去

JWO と JWN を統合したオントロジーには，344,934 個のインスタンスが含まれている. 各インスタンスは一つ以上のタイプを持つが，その中には，冗長な定義も含まれている. 例えば，「東野圭吾」インスタンスには，「小説家」，「日本の小説家」，「推理作家」，「生年別推理作家_1950 年代」，「大阪府立大学の人物」，「大阪府出身の人物」という六つのタイプが定義されている. ここで，「日本の小説家」は「小説家」クラスのサブクラスであり，「生年別推理作家_1950 年代」は「推理作家」クラスのサブクラスとして，クラス階層で定義されている.「小説家」と「推理作家」クラスは，クラス階層および「日本の小説家」と「生年別推理作家_1950 年代」クラスと「東野圭吾」インスタンスの関係から，推論により導出することが可能なため，冗長なタイプであるといえる. 本研究では，上記の例で示したような冗長なタイプ(クラス−インスタンス関係)を以下の手順により，削除する.

1. 各インスタンスのタイプセットを取得
2. 該当タイプの上位クラスセットを取得し，該当タイプ以外のタイプが上位クラスセットに含まれていた場合には冗長なタイプとみなす
3. 冗長なタイプセットを元のタイプセットから削除

上記の方法により，449,186 のクラス−インスタンス関係から 4,587 の冗長なタイプを削除することができた．(削除後のクラス−インスタンス関係 444,599)

図 4 に JWO と JWN を統合したクラス階層の一部(「生物」と「クリエイター」クラスのサブクラス)を示す．図 4 のクラスを示すアイコンの中で，背景色が白色のアイコンはインスタンスを持たないクラスを表し，背景色が青系，緑系，赤系のアイコンは，インスタンスを持つクラスを表している．

図 4: JWO と JWN を統合したクラス階層の一部(「生物」と「クリエイター」クラスのサブクラス)

4. 関連研究

[山田 11]では，Wikipedia から獲得した上位下位関係と，Web テキストから獲得した語句間類似度情報を併用することで，網羅的かつ高精度に上位下位関係を獲得する手法を提案している．質問応答などの自然言語処理アプリケーションのための上位下位関係の獲得を試みており，「A は B の一例である」が成立する A と B も上位下位関係に含めている．そのため，日本語 Wikipedia オントロジー(JWO)における is-a とクラス−インスタンス関係が混在した関係を獲得していると考えられる．また，抽象的な上位下位関係が獲得できているかどうかについては言及されていないため，本研究の目的を解決するために直接利用可能かどうかは不明である．しかしながら，JWO では獲得できていない is-a 関係やクラス−インスタンス関係を多く獲得できていると考えられるため，今後，JWO の拡張に利用可能かどうかを検討したい．

[小林 08]では，Wikipedia のカテゴリを日本語語彙体系の知識に結合したクラス階層の構築を試みている．Wikipedia のカテゴリ名と日本語語彙体系の知識(クラス名)とを後方文字列照合することによって自動的に結合している．照合先のないカテゴリは破棄している．

[柴木 08]では，日本語語彙体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している．語彙体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ている．最初に，1. 語彙体系クラス名と Wikipedia カテゴリ名が完全に一致する，2. 語彙体系インスタンス名と Wikipedia カテゴリ名が完全に一致する，3. 語彙体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「下位カテゴリ名 3 件以上」と完全一致する，という 3 つの規則により，接点カテゴリ候補を抽出する．最終的な接点カテゴリの選択は人手で行なっている．

本研究では，日本語 WordNet(JWN)と JWO におけるインスタンスを持つクラスの結合を試みている．完全照合や後方文字列照合だけでなく，語類似度を用いた文字列に基づく 4 種類の手法(プレフィックス，サフィックス，編集距離，n グラム)を用いて，JWN の Synset と JWO のクラスの照合を行い，ユーザとインタラクションを取りながら，結合を行うことができるようにしている．

5. おわりに

本稿では，日本語 Wikipedia オントロジー(JWO)と日本語 WordNet(JWN)を統合することにより，JWO における上位クラスの補完を行った．クラス−インスタンス関係の洗練およびアライメント対象クラスを同定するためのツール，JWO におけるクラスと JWN における Synset のアライメントを支援するためのツールを開発し，これらのツールと領域オントロジー構築支援環境 DODDLE-OWL を用いて JWO と JWN を統合する手法を提案した．

今後は，本研究で構築したクラス階層を利用して，ユーザとインタラクションを取りながら，クラス−インスタンス関係の洗練，継承を考慮したプロパティ定義，欠落している中位クラスの挿入など，JWO のためのオントロジーデバッギングツールを開発する予定である．

参考文献

[市瀬 08] 市瀬龍太郎: オントロジーマッピングに有効な特徴の抽出,第 22 回 AI 学会全国大会,2E1-1, 2008.

[小林 08] 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, 2008.

[柴木 08] 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 自然言語処理研究会, 2009.

[玉川 10] 玉川奨, 桜井慎弥, 手島拓也, 森田武史, 和泉憲明, 山口高平: 日本語 Wikipedia からの大規模オントロジー学習, 人工知能学会論文誌, Vol.25, No. 5, pp.623-636, 2010.

[Morita 08] Takeshi Morita, Noriaki Izumi, Naoki Fukuta, Takahira Yamaguchi, "DODDLE-OWL: Interactive Domain Ontology Development with Open Source Software in Java", IEICE Transactions on Information and Systems, Special Issue on Knowledge-Based Software Engineering Vol.E91-D No.4 pp.945-958, 2008.

[山田 11] 分布類似度と Wikipedia から獲得した構造情報を利用した上位下位関係獲得: 山田一郎, 鳥澤健太郎, 風間淳一, 黒田航, 村田真樹, StijnDe Saeger, Francis Bond, 隅田飛鳥, 橋本力, 情報処理学会論文誌 Vol.52(12), pp.3435-3447, 2011.