様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ８ 日現在

機関番号：３２６０１
研究種目：若手研究(B)
研究期間：2012～2014
課題番号：２４７００１５０

研究課題名（和文）ＷｉｋｉｐｅｄｉａとＷｏｒｄＮｅｔの統合に基づくプロパティ付きクラス階層の構築

研究課題名（英文）Building up a class hierarchy with properties by integrating Japanese Wikipedia Ontology and Japanese WordNet

研究代表者

森田　武史（Morita, Takeshi）

青山学院大学・社会情報学部・助教

研究者番号：５０５９０１７１

交付決定額（研究期間全体）：（直接経費）　　1,400,000円

研究成果の概要（和文）：これまで研究開発してきた日本語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)には，上位クラスの欠如およびクラスに適切なプロパティ定義がなされていないという問題があった．本研究では，JWOと日本語WordNet(JWN)を統合することにより，プロパティ付きクラス階層を構築することを目的とする．プロパティ付きクラス階層を有する大規模なオントロジーの自動構築を試みている研究は，国内外の関連研究には見受けられないため，独創的であると考えられる．構築したオントロジーについて，定量的および定性的な評価を行い，妥当性を検証した．

研究成果の概要（英文）：Previously, we constructed the Japanese Wikipedia Ontology (JWO) via a semi-automatic process using the Japanese Wikipedia, but it had problems due to a lack of upper classes and appropriate definitions of properties. Thus, the aim of the current study was to complement the upper classes in JWO by refining and integrating JWO and Japanese WordNet (JWN) to build a class hierarchy with defined properties based on the considerations of property inheritance. To achieve this, we developed tools that help users to refine the class-instance relationships, to identify the JWO classes that need to be aligned with JWN synsets, and to align the JWO classes with the JWN synsets via user interaction. We also integrated JWO and JWN using a domain ontology development environment, DODDLE-OWL. We also propose a method for building a class hierarchy with defined properties by elevating the common properties defined in sibling classes to higher classes in JWO.

研究分野：セマンティックWeb

キーワード：オントロジー　セマンティックWeb　オントロジー学習　WordNet　日本語Wikipediaオントロジー　Linked Open Data

１．研究開始当初の背景

セマンティック Web は，ソフトウェアが意味理解可能な辞書（オントロジー）に基づいて，Web コンテンツにソフトウェア可読なメタデータを付与することによって，ソフトウェアが Web コンテンツを理解し，推論することを可能にしようという試みである．セマンティック Web の実現により，現状のキーワード検索を超えた意味検索やアプリケーションを横断したデータの統合および再利用などが可能となる．しかしながら，セマンティック Web 標準技術である，RDF，RDFS，OWL により記述されるセマンティック Web コンテンツ(SWC)の構築には多大なコストを要している．

上記の問題を解決するために，日本語 Wikipedia から同義語，クラスーインスタンス関係，Is-a 関係，Infobox トリプル，プロパティの定義域および値域の関係を抽出する手法を提案してきた．ここで構築されたオントロジーを日本語 Wikipedia オントロジー(JWO)と呼ぶ．

JWO には，上位クラスの欠如およびクラスに適切なプロパティ定義がなされていないという問題があった．

２．研究の目的

本研究では，JWO と日本語 WordNet(JWN)を統合することにより，JWO の上位クラスが欠如している問題を解決し，適切な定義がなされたプロパティ付きクラス階層を構築することを目的とする．

３．研究の方法

本研究は，(1)JWO と JWN の統合および(2)プロパティ付きクラス階層の構築の主に二つの手法により実現する．

(1) JWO と JWN の統合は，(1)-①クラス-インスタンス関係の抽出，(1)-② クラス-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，(1)-③ JWO のクラスと JWN の Synset のアライメント，(1)-④ DODDLE-OWL を用いた JWO と JWN の統合，(1)-⑤ 冗長なクラス-インスタンス関係の除去の手順により行う．

(2) プロパティ付きクラス階層の構築は，(2)-① インスタンストリプルとインスタンスのタイプからのプロパティ定義域の抽出，(2)-②クラスの深さの算出 ，(2)-③リーフクラスからルートクラスへのプロパティのリフト，(2)-④明示的に定義された継承プロパティの除去の手順により行う．

４．研究成果

(1) JWO と JWN の統合の手順を図 1 に示す．以下の(1)-①から⑤に，統合手順の詳細を示す．なお，本研究では，2010 年 11 月時点の JWO と JWN ver. 1.1 を利用している．

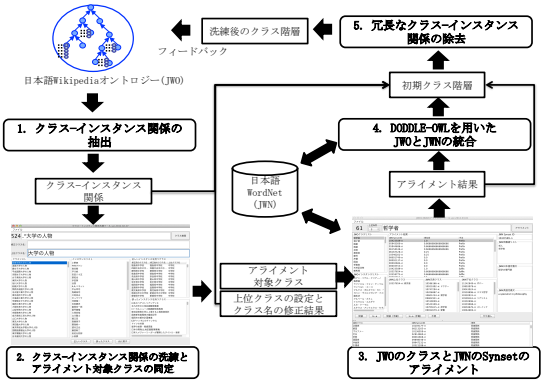

図 1: JWO と JWN の統合手順

(1)-①クラス-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-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-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-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を確認したところ，「職業」，「種類」，「種別」プロパティについては，クラス-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-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-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-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-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-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-インスタンス関係が抽出できた．

(1)-②クラス-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-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-イ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-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．

本ツールを用いてクラス-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-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．

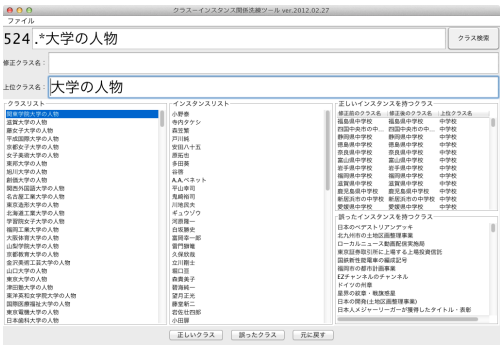

図 2: クラス-インスタンス関係の洗練およびアライメント対象クラスを同定するためのツールのスクリーンショット

(1)-③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． 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)．

表 1 に JWO のクラスと JWN の Synset のアライメント結果を示す．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．

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のスクリーンショット

表 1: 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

(1)-④ 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 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，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，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．また，506 個の上位クラスが JWN により補完できた．

(1)-⑤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）
(1)-⑥ 最終的に 506 個の上位クラスが JWN により補完できた．図 4 は，統合したクラス階層の上位クラスの一部（「物理的な存在がある実体」クラスのサブクラスの一部）を示している．

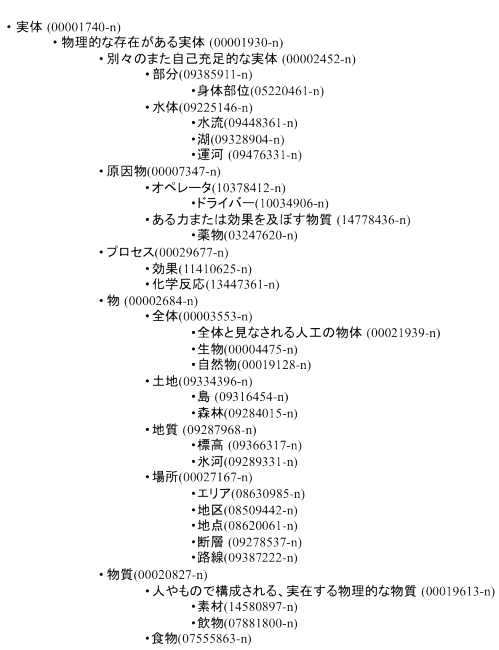

図 4: JWO と JWN を統合したクラス階層における「物理的な存在がある実体」クラスのサブクラスの一部

(2)プロパティ付きクラス階層の構築は以下の(2)-①から④の手順により行う．

(2)-①インスタンストリプルとインスタンスのタイプからのプロパティ定義域の抽出

本研究では，プロパティ継承を考慮したプロパティ定義域の定義を行うために，JWO におけるインスタンストリプルとその主語リソースのタイプから，そのタイプ（クラス）が持つプロパティを抽出する．具体的には，インスタンストリプル S-P-O が存在し，S のタイプが T の時，クラス T はプロパティ P を持つ，換言すると，プロパティ P の定義域を T とする．

(2)-② クラスの深さの算出

適切に下位クラスから上位クラスにプロパティをリフトするためには，リーフクラスからルートクラスへと順にプロパティをリフトする必要がある．そこで JWO におけるすべてのクラスについてルートクラスまでのパスを抽出し，対象クラスからルートクラスまでの距離を「深さ」として算出する．

(2)-③ リーフクラスからルートクラスへのプロパティのリフト

リーフクラスからルートクラスに向かって，兄弟クラスが共通して持つプロパティを上位クラスへリフトする．その際，すべての

兄弟クラスが持つプロパティには「complete」というラベルを付与し，下位クラスからは削除する．二つ以上の兄弟クラスが持つプロパティについては，「candidate」というラベルを付与し，同様に下位クラスから削除する．また，対象クラスしか保有しないプロパティには「default」というラベルを付与し，対象クラス固有のプロパティとする．これにより，冗長な定義域の定義を削除し，固有プロパティと継承プロパティの識別が可能となる．

(2)-④ 明示的に定義された継承プロパティの除去

最後に，クラス階層の中で明示的に定義された継承プロパティを除去する．プロパティのリフトを行った後には，いくつかのクラスについて，それらのスーパークラスに定義されているプロパティが明示的に定義されていることがある．これらのプロパティは，推論により導出することが可能な冗長なプロパティであるため，削除する．

(2)-⑤ 構築結果

表 2 に，構築したプロパティ付きクラス階層におけるクラス数，プロパティ数，クラス-インスタンス関係数を示す．

表 3 に，プロパティリフトによるプロパティ定義域の定義数の変化を示す．表 3 より，冗長なプロパティ定義域の定義が削除されたことがわかる．

図 5 に，JWO と JWN を統合したクラス階層の一部を，表 4 に，プロパティリフト結果の例を示す．図 5 および表 4 より，「School」や「University」などのクラスがプロパティリフト前に持っていた「location」や「school type」などのプロパティは，より上位のクラスである「Organization」や「Educational Institution」クラスに適切にリフトされたことがわかる．また，図 5 のクラス名の後に括弧書きで「JWO」と記載されたクラス以外については，JWN により補完されたクラスを表しており，JWO と JWN の統合により，抽象的な上位クラスが補完された様子が図 5 より確認できる．

なお，本研究で開発したプログラムおよび構築したプロパティ付きクラス階層は，GitHub(https://github.com/t-morita/JWO_Refinement_Tools)にて，オープンソースとして公開している．

表 2: 構築したプロパティ付きクラス階層におけるクラス数，プロパティ数，クラス-インスタンス関係数

クラス数	3,462
JWO から抽出したクラス数	2,787
JWN から抽出したクラス数	675
プロパティ数	4,357
クラス-インスタンス関係数	444,597

表 3: プロパティリフトによるプロパティ定義域の定義数の変化

	プロパティ定義域の定義数	**プロパティを持つクラス数**
プロパティリフト前	143,500	2,929
プロパティリフト後	18,678	1,690

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

①Takeshi Morita, Yuka Sekimoto, Susumu Tamagawa and Takahira Yamaguchi, "Building up a class hierarchy with properties by refining and integrating Japanese Wikipedia Ontology and Japanese WordNet", Web Intelligence and Agent Systems: An International Journal, 査読有, Volume 12, Number 2, pp. 211-233, IOS Press (2014)
DOI: 10.3233/WIA-140293

〔学会発表〕（計 6 件）

①Takeshi Morita, Yuka Sekimoto, Susumu Tamagawa, Takahira Yamaguchi, "Building up a Class Hierarchy with Properties from Japanese Wikipedia", Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01, pp. 514-521,マカオ(中国),2012 年 12 月 7 日
②森田 武史, 玉川 奨, 山口 高平, "オントロジーアライメントを用いた日本語 Wikipedia オントロジーと日本語 WordNet の統合", 第 28 回 セマンティックウェブとオントロジー研究会 SIG-SWO-A1202-07 , 鯖江公民館（福井県鯖江市), 2012 年 10 月 5 日

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）
○取得状況（計 0 件）

〔その他〕
ホームページ等
https://github.com/t-morita/JWO_Refinement_Tools

６．研究組織

(1)研究代表者

森田 武史 (MORITA, Takeshi)
青山学院大学・社会情報学部・助教
研究者番号：50590171

(2)研究分担者

なし

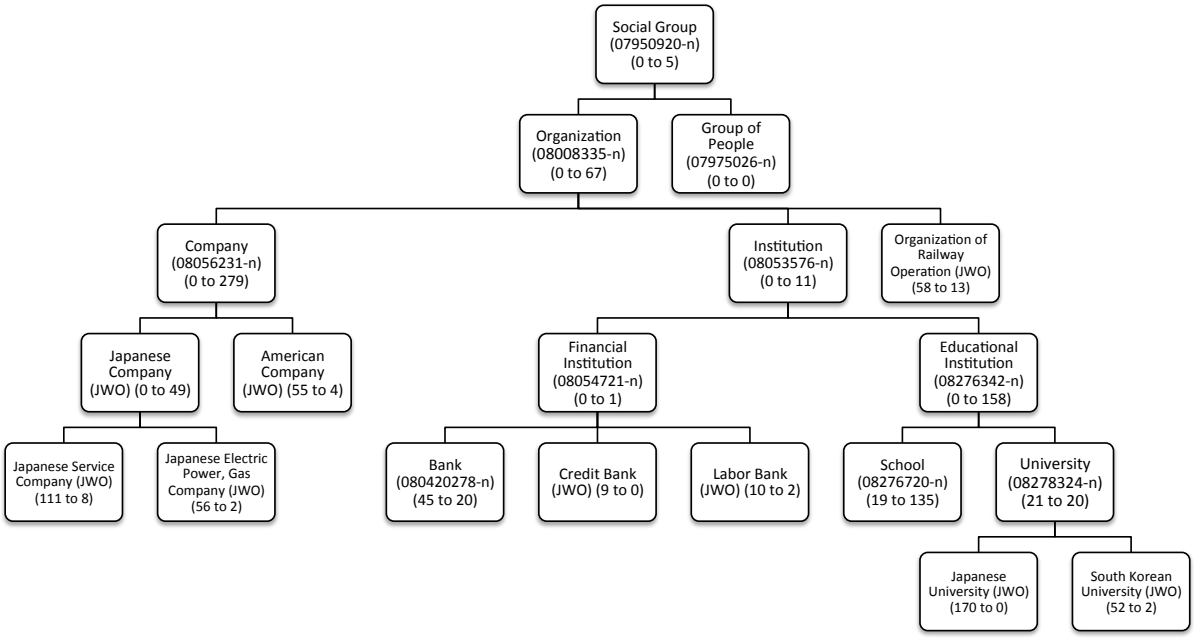

図 5 JWO と JWN を統合したクラス階層の一部

表 4 プロパティリフト結果の例

クラス	クラスが持つプロパティ （プロパティリフト前）	クラスが持つプロパティ （プロパティリフト後）
Social Group (07950920-n)		film, founded, official site, outline, other
Organization (08008335-n)		corporate development, employees, founder, location, location of central office, shareholder
Company (08056231-n)		affiliate company
Institution (08053576-n)		date of foundation, institution personnel
Organization of Railway Operation (JWO)	shareholder	
Financial Institution (08054721-n)		bank code
Educational Institution (08276342-n)		educational policy, school type
American Company (JWO)	affiliate company, corporate evelopment, date of foundation, employees, founder, location of central office	
Japanese Service Company (JWO)	affiliate company, corporate evelopment, date of foundation, employees, founder, location, official site	
Japanese Electric Power, Gas Company (JWO)	affiliate company, corporate development, date of foundation, employees, founder, location of central office	
Bank (080420278-n)	central bank code, date of foundation, founder, location of central office	central bank code
Credit Bank (JWO)	bank code, date of foundation, founder, location of central office	
Labor Bank (JWO)	bank code, date of foundation, founder	
School (08276720-n)	institution personnel, location, school type	
University (08278324-n)	institution personnel, location, school type	
Japanese University (JWO)	educational policy, institution personnel, location, school type	
South Korean University (JWO)	institution personnel, location, school type	