様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　　元 年　　５ 月 ３１ 日現在

機関番号：３２６１２
研究種目：若手研究(B)
研究期間：2015～2018
課題番号：１５Ｋ１６０９０

研究課題名（和文）大規模多言語RDFグラフからのクラススキーマ階層の構築および洗練

研究課題名（英文）Construction and Refinement of Class Schema Hierarchy from Large Scale Multilingual RDF Graph

研究代表者

森田　武史（Morita, Takeshi）

慶應義塾大学・理工学部（矢上）・准教授（有期）

研究者番号：５０５９０１７１

交付決定額（研究期間全体）：（直接経費）　　1,700,000円

研究成果の概要（和文）：近年，Linked Open Data(LOD)と呼ばれるウェブ上でソフトウェア可読なデータを公開・共有するための技術が注目を集めている．LODが普及することにより，ウェブは巨大な知識ベースとして機能し，アプリケーションを横断したデータ統合・再利用が可能となる．LODを統合するためのハブとしてはDBpediaやYAGOが有名だが，LODの構造を規定するオントロジーの定義が不十分という問題がある．本研究では，英語版Wikipediaの本文情報を利用して，DBpediaやYAGOを補完可能な大規模オントロジーを構築する手法を提案し，その有用性を示した．

研究成果の学術的意義や社会的意義
本研究で構築した大規模オントロジーは，ナレッジグラフを活用した質問応答システム，ドメインオントロジーを構築するための参照リソース，ナレッジグラフの推論を用いた意味検索，アプリケーションを横断したデータ統合・再利用等への活用が期待でき，実用性の高いものである．また，評価結果より，本研究で構築した大規模オントロジーは，現在，LODのハブとして活用されているDBpediaやYAGOを補完できる可能性があり，今後，DBpediaやYAGOを活用したアプリケーションへの適用も期待できる．

研究成果の概要（英文）：In recent years, technology for publishing and sharing software-readable data on the web called Linked Open Data (LOD) has attracted attention. With the widespread use of LODs, the web will become a huge knowledge base, enabling data integration and reuse across applications. Although DBpedia and YAGO are famous as hubs for integrating LOD, there is an issue that the definition of the ontology that defines the structure of LOD is insufficient. In this research, we proposed a method to construct large-scale ontology from the text information of English Wikipedia. From the evaluation, we demonstrated that the ontology could complement DBpedia and YAGO.

研究分野：オントロジー

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

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

1. 研究開始当初の背景

近年，Linked Open Data (LOD)と呼ばれるウェブ上でソフトウェア可読なデータを公開・共有するための技術（または，公開・共有されたデータ）が注目を集めている．LOD が普及することにより，ウェブは巨大な知識ベースとして機能するようになり，自然言語による質問応答やアプリケーションを横断したデータ統合および再利用が可能となる．

LOD のハブとしては DBpedia (http://dbpedia.org/)が有名である．DBPedia は，Wikipedia における Infobox，外部リンク，カテゴリなどの半構造情報資源から大規模な RDF グラフを構築している．DBpedia は，英語版 Wikipedia を核として，多言語リンクを活用することにより，多言語の RDF グラフを構築している点も特徴的である．

現状の LOD の問題点として，オントロジーの定義が不十分であることがあげられる．既存 LOD の多くはボトムアップにデータから構築されるため，オントロジーに基いて構築されていないものやオントロジーの定義が不十分なものが多く，問題となっている．LOD のハブである DBpedia についても，クラス階層，インスタンスのタイプ，プロパティの定義域と値域の定義については，手動で行っており，誤りも多く含まれている．また，YAGO では DBpedia と英語版 WordNet を統合して，DBpedia におけるクラス階層やインスタンスのタイプの拡充を試みているが，カテゴリと記事の関係をクラスとインスタンスの関係と見なしていることから，誤りも多く見受けられる．既存研究では，クラス階層におけるプロパティ継承を考慮したプロパティの定義域と値域の定義（クラススキーマ階層構築）を支援している研究は少ない．DBpedia における RDF グラフを用いた質問応答システムを構築する際に，オントロジー（クラススキーマ階層など）が適切に定義されていなければ，推論機能を活用することができない．また，今後，LOD がさらに普及した際にも，オントロジーに基づいて LOD が構築されていなければ，データ統合を行う際にも支障が出ることが予想される．そのため，LOD のハブである DBpedia や YAGO を拡充および洗練することは意義があると考えられる．

2. 研究の目的

本研究では，YAGO や DBpedia の問題点を踏まえ，英語版 Wikipedia の本文情報（一覧記事，定義文，見出しなど）を利用し，クラススキーマ階層，プロパティ，インスタンストリプルを含んだ大規模な汎用オントロジー（RDF グラフ）を構築することを目的とする．また，構築したオントロジーを DBpedia や YAGO と比較し，それらを補完できる可能性を示す．

3. 研究の方法

これまでに，日本語 Wikipedia における様々なリソース（カテゴリーツリー，一覧記事，Infobox，Infobox テンプレート，定義文，見出し，リダイレクトリンク）から Is-a 関係，クラスインスタンス関係，プロパティ定義域，プロパティ値域，プロパティタイプ，同義語，インスタンストリプルなどの関係を抽出する手法を提案し，日本語 Wikipedia オントロジー（以下，JWO）を提案してきた[引用文献①,②]．JWO は DBpedia や YAGO とは異なる手法を用いて自動抽出しているが，言語依存した処理が多く，多言語コミュニティでは利用できないという問題があった．

本研究では，JWO の構築手法のうち，文字列照合や形態素解析，スクレイピングといった言語依存している処理を含む手法を，英語の性質を踏まえた上で変更し，英語版 Wikipedia でも利用できるようにする．

4. 研究成果

(1) Is-a 関係抽出

英語版 Wikipedia のカテゴリ階層に対する文字列照合，Infobox テンプレート名とカテゴリ名の照合，見出し・リスト構造のスクレイピングの 3 種類の方法を用いて Is-a 関係を抽出する．

①カテゴリ階層に対する文字列照合

カテゴリ階層とは記事の分類を目的として人手で構築された階層的なカテゴリのことを指す．カテゴリ階層には，Is-a 関係と見なせないものも多く存在するため，本手法では文字列照合として後方文字列照合と後方文字列照合部除去を利用して Is-a 関係の抽出を行う．後方文字列照合とはカテゴリ階層を構成する親カテゴリと子カテゴリを比較し，子カテゴリ名が「任意の文字列＋親カテゴリ名」となっているものを抽出する手法である．例えば，親カテゴリ「Directors」，子カテゴリ「Woman directors」というカテゴリ階層が存在する場合，この親子カテゴリを Is-a 関係として抽出する．後方文字列照合部除去とは名詞の後ろの修飾部が一致しているものを抽出し，照合部を除去する手法である．ここで名詞の後ろの修飾部に限定したのは，名詞を後ろから就職する場合には名詞の意味が限定されることが多く，正しい Is-a 関係が抽出されやすいためである．例えば，親カテゴリ「Organizations based in Japan」，子カテゴリ「Companies based in Japan」というカテゴリ階層が存在する場合，「Companies」 is-a「Organizations」という Is-a 関係が抽出できる．

②Infobox テンプレート名とカテゴリ名の照合

Infobox とはテーブルを利用して Wikipedia の記事の属性と属性値を整理して表示するものである．Infobox における属性名は Infobox テンプレートにより定められている．本研究では以下の手順により，Infobox テンプレート名とカテゴリ名を照合し，Is-a 関係を抽出する．

1. カテゴリ名とテンプレート名の文字列照合（照合したカテゴリをルートカテゴリと呼ぶ）
2. ルートカテゴリ以下に存在するカテゴリ階層と，照合したテンプレートを持つ記事が所属するすべてのカテゴリ名との照合
3. ルートカテゴリ以下に存在するカテゴリ階層の中で 2.で照合したものを Is-a 関係とする
4. 子クラスが”任意の文字列＋前置詞＋親クラス”の形式になっている場合にはこれを削除する

③ 見出し・リスト構造のスクレイピング

Wikipedia の記事には見出しが存在するが，この見出しはダンプデータ上では「=」で囲まれて記載され，「=」の数によってその見出しのレベルが決まる．見出しに「分類」や「種類」を意味する単語が含まれていた場合，その内部に存在するリスト構造と見出しの関係を Is-a 関係として抽出する．英語 Wikipedia の場合，「Classification」，「Taxonomy」，「Type」が含まれる見出し直下のリスト構造をスクレイピングすることにより，Is-a 関係を抽出する．

(2) クラスインスタンス関係抽出

英語版 Wikipedia の一覧記事および記事の定義文からクラスインスタンス関係を抽出する．

① 一覧記事からのクラスインスタンス関係抽出

一覧記事とはある基準にしたがって，関連する物事が列挙された記事である．英語版 Wikipedia では，一覧記事における記事名は「List of …」の形式となっている．一覧記事名と一覧記事に列挙されている項目の関係をクラスインスタンス関係として抽出する．

② 定義文からの上位・下位関係抽出

定義文とは Wikipedia 記事の第一文に記載されている，その記事タイトル名の定義や意味が書かれている文のことを指す.この定義文を StanfordParser を用いて形態素解析・係り受け解析を行うことにより，クラス・インスタンス関係を抽出する．ここでは定義文の第一文の動詞が be 動詞であり，かつ主語が記事タイトルと一 致しているものを抽出し，形態素解析を行う．例えば記事 「Novelist」では本文の冒頭に「A novelist is an author or writer of novels, though often novelists also write in other genres of both fiction and non-fiction.」と記載されている．ここから本手法を用いることで 「author」，「writer」を上位語，「novelist」 を下位語として抽出できる．また 抽出した上位下位関係から (1)で抽出したクラスの集合を用いて，クラスインスタンス関係を抽出する．具体的には上位語・下位語ともにクラスの集合に含まれていた場合は Is-a 関係，上位語のみがクラスに含まれていた場合はクラスインスタンス関係に分類する．

(3) Is-a 関係とクラスインスタンス関係の洗練・統合

① Is-a 関係の洗練・統合

(1)で抽出した Is-a 関係には具体的なクラスが多く，上位クラスが不足する傾向がある．そこで，DBpedia オントロジーを用いて，上位クラスの補完を行う．DBpedia オントロジーは人手で作られたクラス階層であり，クラス数は少ないが，Infobox テンプレートを元に作られているため，Wikipedia から構築されるオントロジーとは親和性が高いと考えられる．DBpedia オントロジーと本手法で抽出した Is-a 関係は，完全文字列照合または(1)①で用いた文字列照合の手法を用いて統合する．

② Is-a 関係とクラスインスタンス関係の統合

Is-a 関係抽出とクラスインスタンス関係抽出では，入力となるリソースが異なるため，完全文字列照合では照合しないことが多い．例えば，一覧記事からは「Japanese film of 1966」というクラスが抽出できるのに対し，カテゴリ階層からは「Japanese films」は抽出できるが，「Japanese films of 1966」は抽出できない．この場合，「Japanese films of 1966」 is-a 「Japanese films」という Is-a 関係を作成することにより，クラスインスタンス関係を統合することができる．

(4) インスタンストリプルの抽出

本研究では Infobox および記事のリスト構造からインスタンストリプルを抽出する．

① Infobox からのトリプル抽出

Infobox が有する「記事のタイトル-項目-値」という三つ組を「インスタンス-プロパティ-プロパティ値」として抽出する．DBpedia は DBpedia Information Extraction Framework (DBIEF)を提供しており，これを用いることで英語 Wikipedia からインスタンストリプルが抽出できる．

② 記事のリスト構造を用いたトリプル抽出

Wikipedia 本文中に存在するリスト構造に着目し，「記事名-リスト構造の見出し-リスト構造の各値」をトリプルとして抽出する．抽出手順は以下の通りである．

1. ダンプデータから記事ごとにカテゴリと見出し語を抽出
2. 1 から出現頻度が 10 以下の見出し語を除去
3. 2 で得た見出し語をプロパティ名として，記事ごとにリスト構造の各値を抽出

(5) 評価

2018 年 9 月 20 日時点の英語 Wikipedia ダンプデータを用いて，英語版 Wikipedia オントロジーを構築し，各手法により抽出された関係から 1,000 個の標本をランダムに抽出して評価を行なった．

表 1 に構築したオントロジー全体の関係数と精度を示す．表 1 より，Is-a 関係抽出については ，見出し・リスト構造のスクレイピングによる抽出手法の精度は 6 割程度で低いものの，それ以外は比較的高精度で抽出できたことがわかる．クラスインスタンス関係については 8 割以上の精度で抽出できている．インスタンストリプルについては，Is-a 関係と同様に見出し・リスト構造を用いた手法の精度が低いことがわかる．

表 1: 構築したオントロジー全体の関係数と精度

| 関係 | | 抽出数 | 精度 |
|---|---|---|---|
| Is-a 関係 | 文字列照合 | 425,172 | 96.9±1.07% |
| | Infobox テンプレート照合 | 153,437 | 93.8±1.49% |
| | 見出し・リスト構造 | 16,190 | 52.9±3.00% |
| クラスインスタンス関係 | 一覧記事 | 2,544,524 | 92.7±1.61% |
| | 定義文 | 2,951,927 | 95.7±1.26% |
| インスタンストリプル | Infobox | 1,227,520 | 85.7±2.17% |
| | 見出し・リスト構造 | 5,563,421 | 82.0±2.38% |
| | DBIEF | 33,134,711 | － |

(6) 既存オントロジーとの比較評価

①YAGO との比較

表 2 より，抽出方法については，YAGO では WordNet のクラスと Wikipedia のカテゴリを対応付けることで階層を構築し，そのカテゴリに属している記事をインスタンスとして抽出している．この手法はカテゴリを持つ記事全てがインスタンスとなる可能性があるため，多くのインスタンスを抽出することが可能である．一方で本研究ではカテゴリ階層だけでなく，見出しや定義文，一覧記事中の項目といった本文中の情報も利用しており，精度は下がるものの，記事が存在しないインスンタンスやクラスの抽出が可能という点で差別化が出来ている．

表 3 より，Is-a 関係については，抽出数は本研究の方が多く，精度もほぼ同じという結果になった．一方で，クラスインスタンス関係については精度・抽出数ともに YAGO に劣る結果となった．精度が低いのは一覧記事などの本文情報から抽出を試みているためである．一方で一覧記事にしか書かれていない関係があった場合には，YAGO で抽出できない関係を抽出できる．

②DBpedia オントロジーとの比較

Is-a 関係については DBpedia オントロジーとの統合を行っているため，比較することはできないがクラスインスタンス関係については YAGO と同様の比較ができる．DBpedia のクラス階層は Wikipedia の Infobox テンプレートを元にして構築されており，Infobox テンプレート名がクラスに手作業でマッピングされている．クラスインスタンス関係はこの情報を元に，記事が所属している Infobox テンプレートをクラス，記事タイトルをインスタンスとして抽出している．この手法では Infobox が定義されていない記事はインスタンスとして抽出できない．

表 4 に，本研究で正しく抽出することができたクラスインスタンス関係 1,000 個について，その関係が DBpedia オントロジーに存在するかどうかを判定した結果を示す．DBpedia はクラス階層を手作業で記述しているため，上位クラスが多い．一方，本研究で抽出したクラスは下位概念であることが多いため，多くのクラスインスタンス関係が DBpedia オントロジー中に存在しないという結果になった．また YAGO と同様に記事自体は存在せず，一覧記事内にのみ記載されている様な情報も DBpedia オントロジーでは得ることができない情報の特徴である．

最後に YAGO でも DBpedia オントロジーでも抽出できない関係がどれだけ存在するかを調べる．表 4 に DBpedia オントロジー及び YAGO に存在しない関係について調べた結果を示している．表 4 より，5 割以上のクラスインスタンス関係が YAGO にも DBpedia オントロジーにも存在しない関係である事がわかる．

表 2: 本研究と YAGO および DBpedia の抽出方法の比較

| | 本研究 | YAGO | DBpedia |
|---|---|---|---|
| Is-a 関係 | 文字列照合 | WordNet を利用した Wikipedia カテゴリによる記事分類 | 手作業 |
| | Infobox テンプレート照合 | | |
| | 見出し・リスト構造スクレイピング | | |
| クラスインスタン関係 | 一覧記事スクレイピング | | クラスと Infobox テンプレートの照合 |

表 3: 本研究と YAGO および DBpedia の抽出数・精度の比較

| | 本研究 | | YAGO | | DBpedia | |
|---|---|---|---|---|---|---|
| | 抽出数 | 精度 | 抽出数 | 精度 | 抽出数 | 精度 |
| Is-a 関係 | 534,788 | 94.0% | 367,040 | 93.4% | 745 | － |
| クラスインスタンス関係 | 4,963,968 | 92.4% | 8,414,398 | 97.7% | 5,214,242 | － |

表 4: 本研究と YAGO・DBpedia との抽出した関係数の比較

| | YAGO にない関係数 | DBpedia にない関係数 | YAGO にも DBpedia にもない関係数 |
|---|---|---|---|
| Is-a 関係 | 520 | － | 520 |
| クラスインスタンス関係 | 547 | 856 | 505 |

(7) まとめ

本研究では，JWO の構築手法を英語版 Wikipedia に適用し，Is-a 関係，クラスインスタンス関係，インスタンストリプルといったリソース及びリソース間の関係について抽出した．また，構築したオントロジーの精度を評価し，DBpedia や YAGO との比較をすることにより，それらを補完できる可能性を示すことができた．

＜引用文献＞

① 玉川 奨，森田 武史，山口 高平，”日本語 Wikipedia からプロパティを備えたオントロジーの構築”，人工知能学会論文誌 Vol.26 No.4 pp.504-517，2011，DOI: 10.1527/tjsai.26.504

② 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，DOI: 10.1527/tjsai.25.623

5．主な発表論文等
〔雑誌論文〕（計 4 件）
①杉山 岳弘, 大野 祐, 牧山 宅矢, 森田 武史, 小薬 洋昭, 手嶋 秀之, 山口 高平, ”高速道路からの立ち寄り観光推薦アプリの開発および実証実験”, 観光と情報, 第14巻, 第1号, pp.27-42 (2018) 査読有
②Takeshi Morita, Shunsuke Akashiba, Chihiro Nishimoto, Naoya Takahashi, Reiji Kukihara, Misae Kuwayama, Takahira Yamaguchi, “A Practical Teacher-Robot Collaboration Lesson Application Based on PRINTEPS”, The Review of Socionetwork Strategies, Springer, Vol. 12, Issue 1, pp 97-126, DOI: 10.1007/s12626-018-0021-x (2018) 査読有
③Takeshi Morita, Kodai Nakamura, Hiroki Komatsushiro, Takahira Yamaguchi, “PRINTEPS: An Integrated Intelligent Application Development Platform based on Stream Reasoning and ROS”, The Review of Socionetwork Strategies, Springer, Vol. 12, Issue 1, pp 71-96, DOI: 10.1007/s12626-018-0020-y (2018) 査読有
④ Takeshi Morita, Chie Iijima, Wataru Okawara, Yoshitaro Enomoto, Takahira Yamaguchi, “Implementing Mobility Service with Japanese Linked Data”, International Journal of Computational Intelligence Studies, Vol. 5, No. 3/4, pp.267-288, DOI: 10.1504/IJCISTUDIES.2016.083573 (2016) 査読有

〔学会発表〕（計 6 件）
①Tokio Kawakami, Takeshi Morita, Takahira Yamaguchi, “Building up Ontologies with Property Axioms from Wikipedia”, The Joint International Workshop on PAOS 2018 and PASSCR 2018, Vol-2293, pp. 1-12 (2018)
②Tokio Kawakami, Takeshi Morita, and Takahira Yamaguchi, “Building Wikipedia Ontology with More Semi-Structured Information Resources”, The 7th Joint International Semantic Technology Conference (JIST2017), LNCS, vol.10675, pp.3-18, DOI: 10.1007/978-3-319-70682-5_1(2017)
③川上 時生, 森田 武史, 山口 高平, ”半構造情報資源を用いた Wikipedia オントロジーの構築”, 第 43 回セマンティックウェブとオントロジー研究会, SIG-SWO-043-06 (2017)
④川上 時生, 森田 武史, 山口 高平, ”英語版 Wikipedia オントロジー構築とYAGO との比較評価”, 2017 年度人工知能学会全国大会（第 31 回）, 1N2-OS-39a-3 (2017)
⑤Hiroshi Asano, Takeshi Morita, and Takahira Yamaguchi, “Development and Evaluation of an Operational Service Robot Using Wikipedia-based and Domain Ontologies”, 2016 IEEE/WIC/ACM International Conference on Web Intelligence, pp.511-514, DOI: 10.1109/WI.2016.0086 (2016)
⑥淺野 泰史, 森田 武史, 山口 高平, “ドメインオントロジーと日本語 Wikipedia オントロジーの統合に基づく質問応答ロボットの開発と評価”, 第 12 回 情報システム学会 全国大会・研究発表大会, P016 (2016)

〔図書〕（計 1 件）
①“人工知能学大事典”, 人工知能学会（編集）, 共立出版 (2017) ISBN-13: 978-4320124202
森田 武史, 第 18 章 知識工学とセマンティックテクノロジー, 18-13 大規模知識モデリングと集合知アプローチ, pp. 1275-1276.
森田 武史, 第 18 章 知識工学とセマンティックテクノロジー, 18-21 オントロジー学習（オントロジー半自動構築）, pp. 1292-1293.

〔産業財産権〕
○出願状況（計 0 件）
○取得状況（計 0 件）

〔その他〕
なし

6．研究組織

(1)研究分担者
なし

(2)研究協力者
なし