人工知能学会研究会資料 SIG-SWO-048-07

日本語 Wikipedia オントロジーの構築

Construction of Japanese Wikipedia Ontology

中川嵩教 [1*]　小板橋佳晃 [1]
吉岡真治 [1,2]
Takanori Nakagawa[1]　YoshiakiKoitabashi[1]　Masaharu Yoshioka[1,2]

[1] 北海道大学
[1] Hokkaido University
[2] 理研 AIP
[2] RIKEN AIP

Abstract: Wikipedia is a largest online encyclopedia that covers varieties of topics using structured documents (e.g., infobox for describing metadata, and classication using Wikipedia category). There are several efforts to extract structured knowledge, such as DBpedia, YAGO2, and Japanese Wikipedia ontology. However, it was not used in consideration of the characteristics of the Wikipedia category. So far, we have analyzed and organized the Wikipedia category. In this study, We make a Japanese Wikipedia ontology from the rearranged Wikipedia categories, in publishing as LinkedOpenData, we will report what we have defined and consider future use.

1　はじめに

Wikipedia[1]は世界最大のインターネット百科事典であり、様々な形で知識源として活用する方法が検討されている。具体的には、各記事に関する属性情報（所在地、生年月日、所属…）などを記述したインフォボックスから属性情報を抽出する DBpedia[Bizer 09] の研究や、カテゴリの情報を用いて、様々な単語間の類似性を計算する方法 [Ponzetto 07, Taieb 13] や、オントロジーの構築に役立てる YAGO2[Hoffart 13] や、日本語 Wikipedia オントロジー [玉川 10] の研究などが存在する。また、我々は、これまでに、このカテゴリの性質に関する分析 [藤原 12, Yoshioka 14] を行っており、これらの研究の中で、Wikipedia のカテゴリの持つ特殊性を指摘し、その特殊性を考慮した利用方法が必要であることを提案し、Wikipedia カテゴリの種類とカテゴリ間の関係を分類することで、日本語 Wikipedia カテゴリの整理を行ってきた [中川 18, 中川 19]。

本研究では、整理された日本語 Wikipedia カテゴリを元に、Wikipedia カテゴリの特殊性を考慮した日本語 Wikipedia カテゴリオントロジーを構築し、Linked Open Data(以下 LOD と略記) として公開した。

*連絡先：北海道大学大学院情報科学研究科
〒 060-0814 札幌市北区北 14 条西 9 丁目
E-mail:t.nakagawa@kb.ist.hokudai.ac.jp
[1]https://www.wikipedia.org/

2　Wikipedia カテゴリ

2.1　Wikipedia カテゴリの階層構造

Wikipedia において、カテゴリとは、膨大な記事群を様々な観点から分類するための索引であり、各記事には、それぞれに一つ以上のカテゴリが付与される。また、このカテゴリは、さらに詳細なカテゴリと関連付けることにより、カテゴリは階層的な構造となっている。このカテゴリ階層については、基本的には、下位カテゴリに属する記事は、上位カテゴリにも属するという包含関係が成立すること想定されていた。よってカテゴリ階層は、知識工学で用いられる概念階層と似た性質を持つと考えた様々な利用方法が提案されている。しかし、このカテゴリ階層の構造は、Wikipedia に登録される記事の増加に伴い、より詳細なカテゴリ分類が求められることになり、結果として、新しいカテゴリ階層の作り方に関する考え方が導入されることとなった。そのため、必ずしも、包含関係が成り立たない形でカテゴリ階層が存在している。次小節では、包含関係に注目した際に注意すべきであるカテゴリの種類について述べていく。

2.2　カテゴリの種類

Wikipedia カテゴリには「大学」「企業」といったクラスのような役割のカテゴリと「北海道大学」「トヨ

タ自動車」のような具体的な事象を表すようなカテゴリが存在する。英語版 Wikipedia では前者のようなカテゴリを set カテゴリ、後者のようなカテゴリを topic カテゴリと呼んでいる。また Wikipedia カテゴリは、1 つのカテゴリに大量のページが付与されると、カテゴリからページの閲覧を行う際に不都合であるという考えから、様々な基準により、より詳細なカテゴリへ分割するということが行われている。これにより作られるカテゴリの多くは「北海道の大学」「日本の企業」のように、Set と Topic の組み合わせであることから Set-and-Topic カテゴリと呼ばれている。

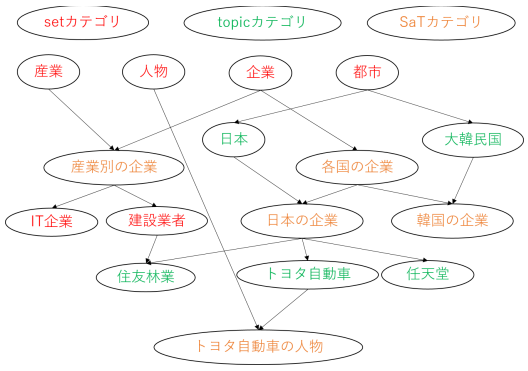

図 1: 分割のためのカテゴリを含むカテゴリ階層

2.3 Wikipedia カテゴリの再整理

前小節で述べた Set-and-Topic カテゴリにおいては、その親カテゴリに Set や Topic を持つものが多く、このことが、下位カテゴリに属するページが上位カテゴリに属するという包含関係を満たさないような関係を作り出す大きな原因の一つになっていると考えた。この問題を解決するために、我々は Wikipedia カテゴリの知識工学的な再整理を行ってきた [中川 18, 中川 19]。具体的に行ったカテゴリの種類の分類とカテゴリ間の関係の分類について下記で説明する。

2.3.1 カテゴリの種類の分類

関係の分析をする上で、まずはそれぞれのカテゴリの種類を分類することが必要であると考えた。英語版 Wikipedia では、カテゴリを set、topic、SaT の 3 種類に分類していたが、SaT カテゴリには「産業別の企業」のような Set-and-Set の形のカテゴリや、「日本のサッカー」のような Topic-and-Topic の形をしたカテゴリが存在することから [中川 19] では、SaT カテゴリの中で、set の性質を持つカテゴリを、制約付き set という意味で CS カテゴリ (ConstrainedSet カテゴリ) と名付けた。topic の性質を持つカテゴリは、topic としていたが、制約付き topic という意味で CT カテゴリ (ConstrainedTopic カテゴリ) と名付けることとした。以上の 4 種類にカテゴリを分類した。

2.3.2 カテゴリ間の関係の分類

カテゴリの種類の分類をふまえ、カテゴリ間の関係を、包含関係が成り立つかを特に意識して、次の 5 種類の分類を提案した [中川 19]。

クラス-サブクラス 主に、Set カテゴリ間で定義される「作家」→「著作家」のような概念間の包含関係を持つ関係。

制約詳細化 Set から CS、Topic から CT のように制約が付与される場合や、CS から CS、CT から CT の時に、「アジアの企業」→「日本の企業」のように制約部分が詳細化される関係。

topic 包含関係 Topic から Topic の関係の内、「イギリス」→「イングランド」のような上位のトピックが下位のトピックを含む（地理的な包含関係や、グループとメンバー等）関係。

クラス付加 CS カテゴリとその親カテゴリの関係の内、「北海道大学」→「北海道大学の人物」のように、CS カテゴリにおける Set の制約条件として働くカテゴリとの関係。

Instance of Set や CS カテゴリから Topic に対するカテゴリで、「日本の国公立大学」→「北海道大学」のように、下位カテゴリが上位カテゴリの概念の具体例となるような関係。

また、これらの分類基準に当てはまらないものは「その他」として扱うこととした。

3 Wikipedia カテゴリオントロジーの提案

これまでに行ってきた Wikipedia カテゴリの分析結果をふまえ、Wikipedia カテゴリの階層構造を適切に分類し、知識工学的観点からの利用を支援するための Wikipedia カテゴリオントロジー (Wikipedia のカテゴリを構成要素に分解すると共に、その関係を明示化するためのオントロジー) を構築する。

この目的のために、2.3 節で述べた Wikipedia カテゴリに関する分類結果を以下のような形式で整理し、RDF トリプルとして表現する枠組を提案する。

3.1 カテゴリの分類

まず、最初に、カテゴリのタイプについては、Set, Topic, CS, CT に対応する以下の 4 つのカテゴリを表す語彙を定義する。これらの語彙は、http://wcontology.org/ の URI で公開するオントロジーの中核となる語彙であるため、http://wcontology.org/core# をつけた名前を正式名称とする[2]。

- setCategory :「大学」「企業」といったクラスのような役割のカテゴリ
- topicCategory :「北海道大学」「トヨタ自動車」のような具体的な事象を表すようなカテゴリ
- constrainedSetCategory :「アジアの企業」「産業別の企業」のような SaT カテゴリの中で、set の性質を持つカテゴリ
- constrainedTopicCategory:「日本のサッカー」「各年の日本」のような SaT カテゴリの中で、topic の性質を持つカテゴリ

また、カテゴリについては、各言語ごとへの拡張を考慮し、日本語 Wikipedia にもとづくカテゴリについては、jwcor:Category:[カテゴリ名][3] の形で Resource を表現する。

このとき、カテゴリとカテゴリのタイプについては、RDF のクラス-インスタンス関係を表す rdf:type[4] の関係で表現する。

例えば、「アジアの企業」が constrainedSetCategory という情報は、
主語：jwcor:Category:アジアの企業
述語：rdf:type
目的語：wco:constrainedSetCategory
と表現される。

3.2 カテゴリの親子関係の分類

次に、2.3 節で述べた 5 つのカテゴリ間の関係の分類をベースに、以下の 7 つの語彙を定義した。

- classificationCriteria :「日本の経済」→「日本の経済_(都市別)」のように「制約詳細化」の中で、子カテゴリが親カテゴリに分類の基準が加わったのみで、集合としては同じであるような関係。
- specifiedConstraints :「各国の企業」→「日本の企業」のように制約が詳細化されることにより分割が行われる関係。
- addConstraints :「企業」→「日本の企業」のように制約が付け加わることで分割が行われるような関係。
- classSubclass :「企業」→「多国籍企業」のように子カテゴリが親カテゴリのサブクラスとなるような関係。
- usedForConstraint :「トヨタ自動車」→「トヨタ自動車の人物」のように親カテゴリが子カテゴリに対し分割の基準となるような関係。
- instanceOf :「豊田市の企業」→「トヨタ自動車」子カテゴリが親カテゴリの概念の具体例となるような関係
- narrower: 「イギリス」→「イングランド」のような topic 包含関係を、別の分類にしようと考えたが、あいまいな事例も多く、上記の分類に含まれないが、下位のカテゴリに属するページが上位のカテゴリに属すると判断できる関係については、narrower と分類する。

ただし、包含関係が維持される classificationCriteria、specifiedConstraints、addConstraints、classSubclass は narrower の特殊形と考えられることと、上記の分類に当てはまらないカテゴリ間の親子関係があることを考慮して、一般的な親子関係を表す語彙として、categoryRelation を定義し、narrower,usedForConstraint, instanceOf については、この Relation との間に、rdfs:subPropertyOf[5] の関係を定義した。また、narrower の特殊形である 4 つの関係と narrower についても、同じく rdfs:subPropertyOf の関係を定義した。例えば、「企業」→「日本の企業」の関係についての、RDF トリプルの表現としては、
主語：jwcor:Category:日本の企業
述語：wco:addConstraints
目的語：jwcor:Category:企業
となる。

4 公開データの作成

2017 年 10 月 20 日取得の Wikipedia カテゴリデータ (カテゴリ総数 184,481 件、親子カテゴリペア 451,074 件) を前章で述べたカテゴリタイプと親子関係タイプに分類すると以下のようになった。

[2]以下では、http://wcontology.org/core を wco のネームスペースで参照する。
[3]以下では、http://ja.wcontology.org/resource/ を jwcor のネームスペースで参照する。
[4]以下では、http://www.w3.org/1992/02/22-rdf-syntax-ns を rdf のネームスペースで参照する。
[5]以下では、http://www.w3.org/2000/01/rdf-schema# を rdfs のネームスペースで参照する。

表 1: カテゴリタイプ

setCategory	14,982 件
topicCategory	40,520 件
constrainedSetCategory	118,521 件
constrainedTopicCategory	10,458 件

表 2: 親子関係タイプ

categoryRelation	narrower	classificationCriteria	1,834 件
		specifiedConstraints	132,169 件
		addConstraints	7,869 件
		classSubclass	78,621 件
		other(narrower)	17,753 件
		usedForConstraint	36,509 件
		instanceOf	63,377 件
		other(categoryRelation)	112,942 件

「other(narrower)」に関しては、「イギリス」→「イングランド」のような topic 包含関係が主に存在し、これらは今後名前をつけて分類していく予定である。また「other(categoryRelation)」は未分類のものや、一度分類をしたが適切かが疑問となり、一度保留としているものが多く含まれる。これらを適切に分類していくことは今後の課題である。

今回作成したデータと DBpedia の接続としては owl:sameAs[6] の関係を利用した。例えば、「北海道大学」のカテゴリの接続は
主語：jwcor:北海道大学
述語：owl:sameAs
目的語:http://ja.dbpedia.org/resource/Category:北海道大学
となる。

ただし、DBpedia との接続には、一つ問題点がある。DBpedia においては、あるカテゴリとその下位カテゴリの関係は全て skos:broader[7]となっている。これは、Wikipedia のカテゴリ間の階層関係に全て包含関係が成り立つと考えているために、この関係が設定されていると考えられる。そのため、単純に sameAs で定義することは、この関係も引き継ぐことになるため、あまり、適切ではないと考える立場もある。しかし、本研究では、これは、DBpedia の不適切な取り扱いだと考えているため、将来的には、DBpedia に修正を提案することが必要であると考えている。
今回作成したデータには http://ja.wcontology.org/から SPARQL Query を記述することにより、アクセスすることができる。例えば「学校」サブクラスには
select distinct * where ?s wco:classSubclass jwcor:学校 .
というクエリでアクセスできる。

[6]以下では、http://www.w3.org/2002/07/owl#を owl のネームスペースで参照する。
[7]skos

5 まとめ

本研究ではこれまでの Wikipedia カテゴリに対する分析・分類を元に、日本語 Wikipedia カテゴリオントロジーを構築し、DBpedia のカテゴリデータとリンク付けすることにより LOD としてデータを公開した。これにより、整理した Wikipedia カテゴリデータを共有することが可能となった。今後は、フィードバックを元にデータの更なる洗練、そしてデータの応用ということを行っていきたい。特に、Wikipedia データの活用という面では、Wikipedia のページを利用したものが多いため、ページとカテゴリの関係付けを行っていきたい。

参考文献

[Bizer 09] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S.: DBpedia - A crystallization point for the Web of Data, *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 7, No. 3, pp. 154 – 165 (2009)

[Hoffart 13] Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia, *Artificial Intelligence*, Vol. 194, No. 0, pp. 28 – 61 (2013)

[Ponzetto 07] Ponzetto, S. P. and Strube, M.: Knowledge Derived from Wikipedia for Computing Semantic Relatedness, *J. Artif. Int. Res.*, Vol. 30, No. 1, pp. 181–212 (2007)

[Taieb 13] Taieb, M. A. H., Aouicha, M. B., and Hamadou, A. B.: Computing semantic relatedness using Wikipedia features, *Knowledge-Based Systems*, Vol. 50, No. 0, pp. 260 – 278 (2013)

[Yoshioka 14] Yoshioka, M.: Analysis of Japanese Wikipedia Category for Constructing Wikipedia Ontology and Semantic Similarity Measure, in *Information Retrieval Technology 10th Asia Infomation Retrieval Societies Conference, AIRS 2014, Kuching, Malaysia, December 3-5, 2014 Proceedings*, pp. 470–481, Springer-Verlag GmbH (2014), LNCS8870

[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?F 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)

[中川 18] 中川 嵩教, 吉岡 真治?F 知識工学者のための日本語 Wikipedia のカテゴリ階層構造の再整理, 人工知能学会全国大会論文集, Vol. JSAI2018, pp. 2F402–2F402 (2018)

[中川 19] 中川 嵩教, 小板橋佳晃 吉岡 真治?F カテゴリの親子関係の種類に基づく Wikipedia カテゴリの再整理 (2019)

[藤原 12] 藤原 嵩大, 吉岡 真治?FWikipedia の階層関係を分析するためのカテゴリパターンの提案, 2012 年度人工知能学会全国大会 (第 26 回) 論文集 (2012), CD-ROM 2C1-NFC2-4