オントロジーマッチングを用いた知識グラフの構築

Extending knowledge graph with ontology matching

上松 大輝 [1] 趙 麗花 [1] Natthawut Kertkeidkachorn [1] 市瀬 龍太郎 [2,1]

Hiroki Uematsu[1] , Lihua Zhao[1], Natthawut Kertkeidkachorn[1], Ryutaro Ichise[2,1]

[1] 産業技術総合研究所

[1] National Institute of Advanced Industrial Science and Technology

[2] 国立情報学研究所

[2] National Institute of Informatics

Abstract: In this paper, we proposed methods that develop knowledge graph using ontology matching. Wikipedia, DBpedia, and other Linked Data resources are almost clustered by systematic ontologies, but some resource does not have ontologies it should be linked. "Structuring Wikipedia" project categorizes Wikipedia resources using Extended Named Entity (ENE). Since, DBpedia resources are based on Wikipedia, we use ENE for categorizing DBpedia resources.

背景

DBpedia に代表される Linked Data として公開されたエンティティは, Wikipedia の Resource や, Linkded Data 作成者が意図した分類に基づいており，体系化された分類がなされていない．また，すべての Resource に適した属性が付与されているわけはなく，同一カテゴリの Resource だとしても，付与されている属性にばらつきが存在している状況である．

一方，関根ら[1]は「Wikipedia の構造化」プロジェクトにおいて，関根の拡張固有表現[2][3]を用いて Wikipedia エンティティを構造化するタスクへの協力を広く求めている．

そこで，本論文では Wikipedia エンティティを Linked Open Data として公開されている，日本語版 DBpedia の Resource から，拡張固有表現を用いて再分類し新たな知識グラフとして利活用可能とすることを目的として，各 Resouce が持つプロパティやオントロジーを，拡張固有表現が持つエンティティとマッチングすることで，再分類する．

Resource の分類

本論文では，「Wikipedia の構造化」プロジェクトに基づいて，関根の拡張固有表現（以下 ENE）を使って DBpedia の Resource を分類する．ENE には，図 1 に示すとおり 154 件の ENE が定義されている．DBpedia の Resource を，ENE に基づいて分類するに

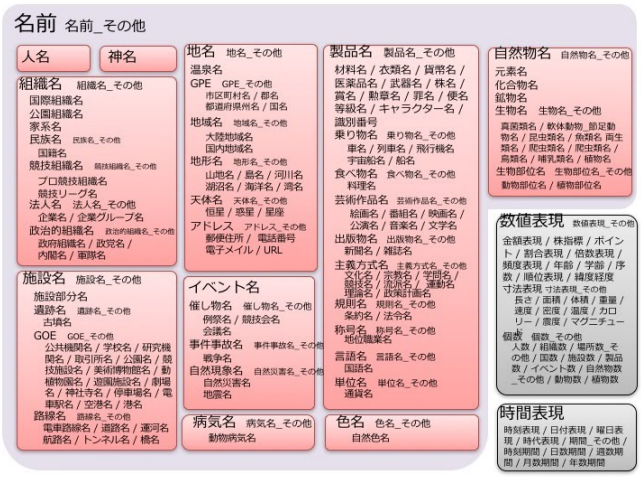

図 1. 関根の拡張固有表現定義

際して，ENE に定義された日本語，英語それぞれの表記を DBpedia 内の Resouce が持つトリプルに接続されたオントロジーとマッチングさせて，そのオントロジーを持つ Resource を ENE に当てはめることで分類を行う．分類にあたって，日本語/英語それぞれの ENE とのマッチングと, Link 関係を基に推論する幾つかの手法を検討した． 以下の節で，それぞれの手法と，その結果について述べる．

表 1. 日本語 ENE とのマッチング

ENE	Resource 数
人名	15140
組織名	312
民族名	6
法人名	360
企業名	418
内閣名	96
地名	8
市区町村名	395
郡名	960
国名	2938
地域名	85
島名	2038
電話番号	1541
施設名	6501
学校名	708
空港名	900
港名	324
路線名	2042
道路名	9328
製品名	32
罪名	272
キャラクター名	2
車名	5669
列車名	421
番組名	95887
流派名	14
イベント名	11
化合物名	136
鉱物名	324
生物名	46

日本語表記 ENE を用いた分類

ENE に定義された拡張固有表現のうち，日本語で表記された ENE を DBpedia のオントロジーとマッチングさせ，そのオントロジーを持つ Resource を該当の ENE として分類した．表 2 に結果を示す．

対象とした 154 件の ENE のうち，ENE をオントロジーとしたときに 30 件の ENE が日本語版 DBpedia が持つオントロジーとマッチングした．残りの 124 件を確認すると，ENE は○○名として定義されており，例えば，「国際組織名」は（http://ja.dbepdia.org/property/国際組織名）と読み替えているが，このようなオントロジーは存在せず，（http://ja.dbepdia.org/property/国際組織）となる．そこで，同様に日本語表記された ENE から，「名」を削除して正規化した ENE を用いて，先ほどマッチングしなかった ENE124 件を対象にした結果を表 1 に示す．

表 2. 日本語 ENE(正規化)とのマッチング

日本語 ENE（正規化）	Resource 数
名前	148556
国際組織	860
国籍	15833
政党	1718
河川	1665
湖沼	2
星座	2085
研究機関	513
武器	411
賞	932
等級	4
映画	256
音楽	37474
出版物	20
宗教	3411
競技	1071
法令	2
称号	1468
言語	41706
国語	4
例祭	5858
戦争	822
色	31089

「名」を削除することで，さらに 23 件の ENE とオントロジーのマッチングができた．日本語表記された ENE をそのままオントロジーとしてマッチング，または「名」などの表記を省くことで，154 件中 53 件の ENE と共通のオントロジーが見つかり，日本語版 DBpedia の Resource を ENE で分類することができた．しかし，これは全体の 3 割ほどでの分類できているものの，マッチングできなかった 101 件の中には「公園名」や「都道府県州名」，「電車駅名」などの主要な ENE が分類されておらず，また DBPedia 内にもこれらの ENE に割り当てることのできる Resource が存在するため，この Resource を抽出するために，英語表記 ENE を用いて再分類を行った．

表 3. 英語 ENE とのマッチング

日本語 ENE	英語 ENE	Resource 数
山地名	Mountain	10325
公園名	Park	5602
美術博物館名	Museum	8493
動植物園名	Zoo	1
電車駅名	Station	58910
医薬品名	Drug	1
飛行機名	Aircraft	7349
船名	Ship	1
文学名	Book	8572
競技会名	Game	1
真菌類名	Fungus	722
昆虫類名	Insect	404
魚類名	Fish	1508
爬虫類名	Reptile	280
鳥類名	Bird	640
哺乳類名	Mammal	185
病気名	Disease_Other	3910

英語表記 ENE を用いた分類

同様に，ENE の英語表記と DBpedia のオントロジーをマッチングさせ Resource の分類を行った結果を表 3 に示す.

17 件の ENE とマッチングしたオントロジーが，それぞれ Link された Resource が見つかった.

ここで, DBpedia の Resource は Linked Data の形式で記述されているため，日本語版 DBpedia が持つ情報と同様の情報，つまり sameAs Link で接続された多言語の DBpedia の Resource などが LoD クラウドに存在するはずである．日本語版 DBpedia にて定義された情報のみでは分類できなかった Resource を分類するため，sameAs Link をたどって，各 Resource の分類を試みた．図 2 に.「公園名」を例とした SPARQL Query を示す.

```
SELECT DISTINCT COUNT(?s)
WHERE{
  ?s owl:sameAs ?same .
  ?same ?p <http://dbpedia.org/ontology/Park> .
}
```

図 2. sameAs Link を使用した Query

sameAs Link を使用して, 足りない情報を補完する手法は Lihua らによって実装されており，プロパティとオブジェクトから推論して，情報を補完している. 本論文では, Resource が持つオントロジーは ENE によって定義されるため, sameAs Link で接続された他言語の Resource に含まれるオントロジーを利用して，ENE とのマッチングを行う．実際に，日本語表記の ENE と，英語表記で単純マッチできなかった 84 件を対象に sameAs Link を使用してマッチングさせた結果を表 4 を示す.

表 4. SameAs Link を用いたマッチング

日本語 ENE	英語 ENE	Resource 数
海洋名	Sea	80
恒星名	Star	1531
惑星名	Planet	4027
運河名	Canal	76
トンネル名	Tunnel	12
橋名	Bridge	852
食べ物名	Food	2583
新聞名	Newspaper	1026
雑誌名	Magazine	967
通貨単位名	Currency	489
地震名	Earthquake	382

表 5. 英語 ENE(正規化)とのマッチング

日本語 ENE	英語 ENE	Resource 数
競技リーグ名	SportsLeague	90

今回は，日本語版 DBPedia から DBpedia.org への SameAs Link をたどることで，これまでの手法でマッチングしなかった ENE が 11 件分類可能となる．さらに, 英語表記の ENE は, 複数単語の場合に「_」を使用して接続されるが, 表 5 に示すとおり 1 件のみであるが, これを外すことで新規に ENE と紐付けることができる.

これまでの処理で, 約 54%の ENE とオントロジーのマッチングができ，ENE を基に DBpedia の Resource を分類することが可能となった.

ENE の具体例からの推論

ENE とオントロジーのマッチングを利用して分類を行ったが, それでも分類されない ENE が存在している．例えば，温泉名という ENE がそれに当たる．日本語版 DBpedia 内では「温泉」というオントロジーは存在せず, 同様に英語表記 ENE で示されている「Spa」というオントロジーも存在しないためである．実際，温泉に関する Resource を探すと，ENE にて例

表 6. 温泉名の例が持つトリプル

Type
http://www.w3.org/2002/07/owl#Thing
http://dbpedia.org/ontology/HotSpring
http://dbpedia.org/ontology/Location
http://dbpedia.org/ontology/NaturalPlace
http://dbpedia.org/ontology/Place
http://schema.org/Place
http://www.wikidata.org/entity/Q177380

として挙げられている遠刈田温泉（https://ja.dbpedia.org/resource/遠刈田温泉）や，福地温泉（http://ja.dbpedia.org/resource/福地温泉）はそれぞれ存在する．例として挙げられている 5 つの温泉地名を利用して，各 Resource が持つ type を抽出する Query を図 3 に示す．

```
SELECT DISTINCT *
WHERE {
  OPTIONAL {
    <http://ja.dbpedia.org/resource/月ヶ瀬温泉>
rdf:type ?type .
  } OPTIONAL {
    <http://ja.dbpedia.org/resource/遠刈田温泉>
rdf:type ?type .
  } OPTIONAL {
    <http://ja.dbpedia.org/resource/白馬温泉>
rdf:type ?type .
  } OPTIONAL {
    <http://ja.dbpedia.org/resource/福地温泉>
rdf:type ?type .
  } OPTIONAL {
    <http://ja.dbpedia.org/resource/湯の山温泉>
rdf:type ?type .
  }
}
```

図 3. 具体例から抽出する Query

表 6 は，図 3 の Query を発行して取得した Type の一覧である．location に関するオントロジーとともに，HotSpring のオントロジーが Link されている．これらの Resource は，例で定義されているように「温泉名」という ENE が割り当てられることになるが，DBpedia には「温泉名」というオントロジーは存在せず，「HotSpring」という別の表記のオントロジーが存在し Link されているため，ENE の例から推論する手法を検討した．

表 7. 具体例から抽出したオントロジー

日本語 ENE	オントロジー	Resource 数
温泉名	http://dbpedia.org/ontology/HotSpring	2382
地形名	http://dbpedia.org/ontology/WorldHeritageSite	3144
神社寺名	http://dbpedia.org/ontology/Temple http://dbpedia.org/ontology/ReligiousBuilding	6707
絵画名	http://dbpedia.org/ontology/Artwork	5199
動物病気名	http://dbpedia.org/ontology/Disease	7259

1. ENE に定義された例を Resource として検索
2. 各 Resource が持つ Type を抽出
3. 抽出した Type のうち，ENE に対してユニークなものを抽出

上記手順で，マッチングしていない 72 件の ENE のオントロジーを設定した結果が表 7 である．

温泉名や神社寺名などは，英語表記された ENE において Spa や Worship_Place と定義されており，オントロジーで使用されている HotSpring，Temple とは異なっているために，これまでの手法では Resouce に該当するオントロジーが見つけられなかったことがわかる．

考察

ENE を用いて，日本語版 DBpedia の Resource の再分類を行ったが，約 56%の ENE に対して Resource を割り当てることができた．しかし，約 44%の ENE は該当するオントロジーが見つけられていない．それぞれの手法でマッチングできなかった ENE を幾つか抽出したものを表 8 に示す．

ENE として採用されてている表記と，オントロジーとの間に違いがあること，また，同じ概念の Resource は存在するものの適切なオントロジーが設定されていない，といった場合があることがわかった．

DBpedia の Resource で，同じカテゴリの概念であるにも関わらず，共通したオントロジーが Link されていない場合は多々存在する．このような場合に，

表 8. マッチングしなかったオントロジー

日本語 ENE	例	英語 ENE
神名	アテネ, インドラ, ゼウス, 大国主命, 帝釈天	God
プロ競技組織	読売ジャイアンツ, ＡＣミラン, 鹿島アントラーズ, ニューヨーク・ヤンキース	Pro_Sports_Organization
劇場名	明治座, ボリショイ劇場, パリ・オペラ座, メトロポリタン歌劇場	Theater
電車路線名	関西本線, 山口線, 東海道本線, 釧網本線, 宝成線	Railroad
自然災害名	伊勢湾台風, 諫早豪雨, 雲仙普賢岳噴火災害, 寛永の飢饉	Natural_Desaster

機械的にオントロジーを付与するような研究はすでに多く行われている[4][5]. 今回は，すでに公開されている Resouce を使用して分類を行ったが, DBpedia のオントロジー拡充を行った上で，ENE での分類を行ったり, Resource に付与した ENE から近傍のオントロジーを抽出し, それらを基に ENE に該当するオントロジーを抽出するなど，より複雑な手法の検討が必要である.

また，ENE はそれぞれの属性情報を持つが，その属性ともオントロジーマッチングを行うことで，マッチしたオントロジーをプロパティとして持つ Resource を探し出し，ENE に基づいて分類することも可能となる．今後，属性情報とのマッチングも含めた手法も検討する必要がある.

まとめ

本論文では，関根の拡張固有表現に沿って，日本語版 DBpedia のデータを分類し，新たな知識グラフを構築するための手法を提案した．本手法を用いることで, 日本語版 DBpedia 以外の Linked Data の分類も行うことが可能である．また，分類ができなかった Resource や，Resource が存在するにも関わらず，オントロジーとのマッチングができなかった拡張固有表現については，具体例からの推論処理を行う際に, より多くの具体例を別のソースから収集したり，拡張固有表現の表記ゆれを考慮するなど，新たな手法を検討していく.

また，本手法で分類し，作成された知識グラフについては，今後 Web 上で公開する予定である.
（http://ri-www.nii.ac.jp/ENEmatching）

参考文献

[1] 関根聡, 小林暁雄, 安藤まや, 乾健太郎: 拡張固有表現に基づく Wikipedia 項目の分類と構造化, 第 43 回 SWO 研究会, 2017

[2] Satoshi Sekine: Extended Named Entity Ontology with Attribute Information, Proceedings of the International Conference on Language Resources and Evaluation (LREC'08), 2008

[3] Satoshi Sekine, Chikashi Nobata: Definition, Dictionary and Tagger for Extended Named, Proceedings of the International Conference on Language Resources and Evaluation (LREC'04), 2004

[4] Lihua Zhao, Rumana FerdousMunne, Natthawut Kertkeidkachorn, and Ryutaro Ichise: Missing RDF Triples Detection and Correction in Knowledge Graphs, Proceedings of the 7th Joint International Semantic Technology Conference (JIST2017), pp. 164-180, Gold Coast, Australia, Nov 10-12, 2017.

[5] Lihua Zhao, Natthawut Kertkeidkachorn, Ryutaro Ichise: Knowledge Discovery from Linked Data, The 31st Annual Conference of the Japanese Society for Artificial Intelligence, 2017.