森羅と日本版DBpediaにおける属性の取り扱いに関する比較分析

小板橋佳晃[1] 吉岡真治[2]

[1] 北海道大学 [2] 北海道大学院情報科学
研究科, 理研 AIP
yk-wwm-aa@eis.hokudai.ac.jp
yoshioka@ist.hokudai.ac.jp

1 はじめに

Wikipediaは世界最大の知識を有する百科事典であり、その知識を活用して計算機で扱うことができる知識とする研究が数多く行われている。主に、インフォボックス（ページの主題についての要約情報を提供することを目的とした、記事の右上に配置する形の規定フォーマットの表）やカテゴリといったデータに注目して構造化を行うプロジェクトとして、ページを単位とした固有物に対してメタデータを付与するDBpedia[1]や、それをオントロジーとして整理するYago[2]の研究などが代表的である。これに対し、近年、Wikipediaの本文から自然言語処理技術などを活用して、より詳細なメタデータを抽出する森羅[3]プロジェクトが開始されている。このプロジェクトでは、特定のカテゴリー（人名、空港、地名など）に対し、必須の属性を定義し、その属性に対応する値をWikipedia中のインフォボックスだけでなく、本文のデータも利用して、積極的に情報抽出を行っている。

本研究では、日本語版Wikipediaを対象として作成された知識源である日本語版DBpediaと森羅におけるメタデータの対応関係を具体的な事例を通して分析することで、各々のデータベースの特性について議論を行う。具体的には、森羅で定義されている属性と属性値のペアに相当する情報が日本語DBpediaにどれくらい存在するかを比較する。

2 森羅と日本語版DBpedia

森羅とは、関根らにより定義された200種の拡張固有表現のカテゴリを用いて、Wikipediaのデータを整理するプロジェクトである。これらのカテゴリのうち、末端に属するカテゴリには、そのカテゴリに属する固有物が持つべき属性のリストが定義されており、森羅では、Wikipediaのページ情報から、これらの属性に対する値を自動的に抽出する研究を行っている。

これに対し、DBpediaは、主に、Wikipediaのインフォボックス、リンク、カテゴリなどに構造的に記述されている固有物に関連する情報から、対応する固有物のメタデータを抽出している。DBpediaにも、森羅と同様に、概念を分類するためのオントロジーが定義されており、109種の概念クラスが定義されるとともに、想定される属性のリストが定義されている。しかし、これらの属性について、全てが抽出の対象になっている保証はなく、英語版のDBpediaには存在するが、日本語版には存在しないような属性と属性値のペアなども存在する。

また、その詳細度についても、拡張固有表現の企業クラスとDBpediaの会社クラスの属性の個数を比較すると、前者は34個に対し後者は20個と必須となる属性に関する考え方が違うことがわかる。その上、DBpediaにおいては、必ずしも、概念クラスに対応する形の属性のみが抽出されるわけではなく、想定したパターンに応じた様々な属性を含むメタデータが抽出されている。また、そのメタデータの表現の一貫性も保証されていないため、同様の属性に複数の名前でメタデータが付与されている例も存在する。

本研究では、これらの現状を踏まえた形で森羅の属性と属性値の情報と、日本語版DBpediaにおける同様の属性に関する情報の対応関係について分析することをその目標としている。この情報が得られることで、例えば、各国語版DBpediaとの連携などを行う際の情報として活用が期待される。

森羅では、自動抽出システムによる属性と属性値のデータの自動抽出を行っているが、その結果をそのま

ま使うと、誤った情報などを含む可能性があり、正確な対応関係を考えるうえで問題がある。よって、本研究では、森羅が提供している人手で作成した正解データであるトレーニングデータに存在する属性と属性値の関係を対象として、その関係に対応する日本語版 DBpedia のメタデータの組み合わせを対応付けることにより、二つの知識源の比較を行う。

3 分析手法

基本構想を図 1 に示す。具体的には、トレーニングデータに存在する属性と属性値の組み合わせを使い。DBpedia の RDF トリプル subject と object を指定することで森羅の属性と DBpedia のメタデータにおける対応する predicate を発見する。しかし、予備実験を行ったところ、object に格納されているデータが以下に述べるような様々な形式が存在し、単純に SPARQL のみを用いた分析では不十分であることがわかった。例えば、熊本空港に対して、所在地を表すものは DBpedia だと ”熊本空港の位置（熊本県）” のように”()”で追加情報が入っているものや、object の形式として「http://ja.dbpedia.org/resource/???」のような形式のもの,「”リュッゲ空軍基地@ja」のような形式のものが存在する。

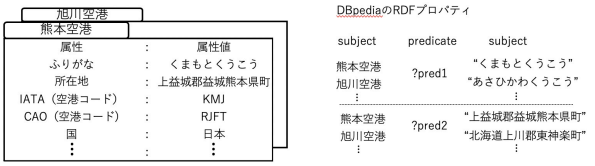

図 1: 基本的な構想

3.1 分析手順

上記の問題を考慮して、本研究では、object を文字列として扱った際に、候補となる文字列を含んでいる場合に、対応関係がつくという考え方に基づき、以下のような手順で実験を行った[1]。

1. トレーニングデータの下処理
   トレーニングデータからページ名と属性値のペアーを取得。この際、”(”,”)”,”[”,”]”,”?”,”+”は、DBpedia の object とパターンマッチングを行う際、不一致となるため除去した。トレーニングデータも Wikipedia から抽出した情報であり、Wikipedia から抽出した情報は、誰でも情報の編集や追加が可能な Wikipedia の特性があるので、整合性が取れていないことが多いからである。

2. 日本語版 DBpedia による predicate 候補の生成
   カテゴリに属する全ての森羅で指定されているページについて、DBpedia のページを subject としてもつ RDF トリプルを検索することにより、predicate と object のペアを抽出する。このペアの object とページに付与された属性値を比較し、object 中に対応する文字列を含む場合に、森羅の属性名と DBpedia の predicate の対応関係の候補とする。

3. 一対多での検索
   予備実験の結果、上記で得られた候補には、同じ森羅の属性に対して、複数の異なる predicate が対応する場合が散見された。これは、DBpedia の属性定義がテンプレートからの情報抽出を考慮したボトムアップな属性を利用していることがあるためである。そのため、複数の predicate の組み合わせで情報抽出を行うことにより、再現率の向上がみこまれる。たとえば、空港カテゴリの別名には、以下の 4 種の predicate が、先の処理で求められたが、これについては、両方から情報を抽出して結果とするという考え方に基づいて、その組み合わせのデータを作成する。

”http://www.w3.org/2000/01/rdf-schema#label”
”http://www.w3.org/ns/prov#wasDerivedFrom”
”http://xmlns.com/foaf/0.1/isPrimaryTopicOf”
”http://xmlns.com/foaf/0.1/name”

4. 精度、再現率、F 値の計算
   これらの候補の良さを判定するために、精度、再現率、F 値を利用する。ここでは、トレーニングデータの集合を V, 検索結果の集合を E, 正解集合を C とする。

   - 精度 (pre) $= |C|/|V|$
   - 再現率 (rec) $= |C|/|E|$
   - F 値 $= 2 \times$ pre $\times$ rec / （pre + rec）

   ただし、複数の属性値を持つ場合には、一つでもその属性値を見つけることができた場合に、正

[1]description の様に、文書の形式で与えられたものについても、現在では、情報源としては取り出せる可能性があるということで、対応関係がつくとしているが、これについては、今後検討が必要である。

解と判断した。具体的には、以下のトレーニングデータがある。
固有物：”小松飛行場”、属性：”別名”、属性値：”こまつひこうじょう”,”Komatsu Airport”,”Komatsu Airbase”,”小松空港”
ここで、小松飛行場を subject としてもつ RDF トリプルを検索し、属性値を文字列として含む object に限定すると、以下のようなデータが得られた。
”小松飛行場（Komatsu Airbase）”
”小松飛行場（こまつひこうじょう）は、石川県小松市にある共用飛行場である。防衛省が管理しており、航空自衛隊小松基地（英: JASDF Komatsu Airbase）と民間航空（民航）が滑走路を共用する飛行場で、特に後者においてはターミナルビルなどの施設の通称として小松空港（こまつくうこう、英: Komatsu Airport）と呼ばれている。航空交通管制は航空自衛隊が行なっている。”
前者の結果に対しては”小松空港”と”Komatsu Airbase”の 2 つが、後者の結果に対しては”こまつひこうじょう”と”小松空港”,”Komatsu Airport”が正解になるがそれぞれ正解は 1 つとカウントしている。

4 実験

4.1 データセットと対応関係の抽出

森羅 2019 版より提供される森羅 2018 版のカテゴリに、30 カテゴリを加えた計 35 のカテゴリーに関するトレーニングデータを使用した。以下が、35 カテゴリである。それぞれ最低 200 項目から最大 1000 項目のデータが与えられている。詳細に関しては、森羅のホームページを参照されたい。
人名, 市町村名, 企業名, 空港, 化合物,GPE_その他, 都道府県州郡名, 国名, 大陸地域名, 国内地域名, 地名_その他, 温泉, 地形名_その他, 山地名, 島名, 河川名, 湖沼名, 海洋名, 湾名, 組織名_その他, 国際組織名, 公演組織名, 家計名, 民族名_その他, 国籍名, 球技団体名, 競技リーグ名, 競技連盟名, 非営利団体名, 企業グループ名, 政治的組織名_その他, 政府組織名, 政党名, 内閣名, 軍隊

これらのカテゴリについて、上記の手続きにより、対応関係を計算し、表 1 には、再現率 > 0.7 を満たすものをであるものを、表 2 には、精度 >0.5 を満たすものを示した。

4.2 考察

まず、表 1 をもとに再現率が高いものにものに関して考察する。「IATA（空港コード） 」と「ICAO（空港コード） 」は、再現率と精度が共に高い値をとっている。それぞれの predicate を見てみると、”http://dbpedia.org/ontology/iataLocationIdentifier”と”http://dbpedia.org/ontology/icaoLocationIdentifier”であり固有に振り与えられる ID のようなものであることがわかる。しかし、同じ性質をもつと予測される「CAS 番号」は、精度が低かった。DBpedia には「CAS 番号」に対応すると思われる”http://dbpedia.org/ontology/casNumber”という predicate が存在したが、値の抽出に失敗しているものが多かった。これは、日本語の Wikipedia の化合物に対して、CAS 番号を入力するテンプレートが存在するものの、実質的にその値が与えられていない化合物が多く存在するため d ではないかと考えている。また上記カテゴリ以外には、「別名」が 6 つと「活動内容」が 1 つ見つかった。別名に対応する predicate は”http://dbpedia.org/ontology/abbreviation”
”http://xmlns.com/foaf/0.1/nick”
”http://dbpedia.org/ontology/alias”
のいずれかであった。どれも別名を包含する意味のついた predicate であるため再現率が高くなったと考えられる。

次に、表 2 をもとに精度が高いものに関して考察する精度に関して、今回の実験では object に正解の文字列が含まれていることを条件として実験を行った。また、object の中には、Wikipedia のページの冒頭部分がそのまま格納されているものもあるため、predicate を多くすることで精度は高い値が出ると予測していたが、予測と反するものだった。これは、森羅の属性値の書き方に一貫性がないことと DBpedia に predicate の決め方や object の記述の仕方に一貫性がないことが原因であると考えられる。「IATA（空港コード）」と「ICAO（空港コード）」に関して、表 1 の結果も踏まえると一貫して predicate が定義されていると言える。また、「母都市」に関しては、「所在地」のような属性より意味的に制限があり、object と文字列が一致しやすいからだと考えられる。

カテゴリ	属性	$\|T\|$	$\|E\|$	$\|C\|$	精度	再現率	F 値
空港	IATA（空港コード）	557	367	364	0.654	0.992	0.788
	ICAO（空港コード）	564	435	372	0.66	0.855	0.745
国際組織名	別名	817	20	18	0.022	0.9	0.043
政治的組織名	別名	410	15	14	0.034	0.933	0.066
競技連盟名	別名	505	26	26	0.051	1.0	0.098
人名	別名	1363	111	106	0.078	0.955	0.144
非営利団体名	活動内容	952	12	9	0.009	0.75	0.019
公演組織名	別名	154	8	8	0.052	1.0	0.099
球技団体名	別名	282	26	24	0.085	0.923	0.156
化合物	CAS 番号	544	52	52	0.096	1.0	0.174

表 1: recall

カテゴリ	属性	$\|T\|$	$\|E\|$	$\|C\|$	精度	再現率	F 値
空港	IATA（空港コード）	557	7896	388	0.697	0.049	0.092
	ICAO（空港コード）	564	1758	382	0.677	0.217	0.329
	母都市	148	422	82	0.554	0.194	0.288

表 2: precision

5　おわりに

本研究では、森羅の手作業で作成した訓練データに基づいて、その属性が日本語版 DBpedia でどのように取り扱われているかを、対応する DBpedia の predicate としてどのようなものが存在するかという観点から分析を行った。分析の結果から、現時点では、DBpedia の持つ情報は対応関係を議論するのに十分ではないことが判明した。これは、適切な predicate が複数存在し、どの predicate が最も適切なのかがはっきりしない場合や、predicate が存在しても、テンプレート中に記載がなかったり、テンプレートをうまく処理するためのルールが定義されていないために、情報が存在しないといった問題があることが示唆された。今後は、インフォボックスにより多くの情報を持つとともに、大規模な抽出ルールをもつ英語版 DBpedia とそれにたいする翻訳の情報（言語観陸などを利用）を用いた分析などを行うことで、上記の問題の影響を緩和した検討を行っていきたい。

謝辞

本研究の一部は、JSPS 科研費 18H03338 の助成と北海道大学国際連携研究教育局ビッグデータ・サイバーセキュリティグローバルステーションの支援を受けた。ここに記して謝意をあらわす。

参考文献

[1] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Soren Auer, Christian Becker, Richard Cy-ganiak, and Sebastian Hellmann. DBpedia - acrystallization point for the web of data. WebSemantics: Science, Services and Agents on theWorld Wide Web, Vol. 7, No. 3, pp. 154 - 165,2009.

[2] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Arti

cial Intelligence, Vol. 194, No. 0, pp. 28 61, 2013.

[3] 関根聡, 小林暁雄, 安藤まや.Wikipediak 構造化プロジェクト「森羅 2018」. 言語処理学会第 25 回大会

[4] Vrandečić, D. and Krötzsch, M. Wikidata: a free collaborative knowledgebase. Commun. ACM57, pp. 78-85.