1N4-OS-10b-4

日本語 WordNet と IPAdic 辞書の RDF 化と DBpedia リンク

RDFization of Japanese WordNet/IPAdic and Linking to DBpedia Japanese

小出 誠二 *1 武田 英明 *2 加藤 文彦 *1 大向 一輝 *2
Seiji Koide Hideaki Takeda Fumihiko Kato Itsuki Omukai

*1情報・システム研究機構 Research Organization of Information and Systems
*2国立情報学研究所 National Institute of Informatics

In the previous paper, we reported the RDFization of Japanese WordNet, whereas there were no outbound links to DBpedia because of no DBpedia in Japanese at that time. In this paper, we have made the links to DBpedia Japanese, which is recently built after the previous paper. In addition, we have also RDFized IPAdic Japanese dictionary, which contains enough vocabulary and enriched description of Japanese part of speech but no meanings, to enable the unification to Japanese WordNet, which is not enough for Japanese vocabulary but contains enriched description of meanings. The result of making links from IPAdic to DBpedia is also reported. Note that all of these links are made by exact matching in word strings. Linking by meanings among Japanese WordNet, IPAdic dictionary, and DBpedia Japanese will be developed in the near future, based on this result.

1. はじめに

既報 [小出, 他] にて，日本語 WordNet [Isahara 08] の RDF 化について，主に LOD の視点から報告したが，その RDF 化されたエントリに DBpedia へのリンクはなかった．今回，日本語 Wikipedia から情報を抽出した DBpedia(以後 DBpedia Japanese と呼ぶ) が得られたので，LOD 化の実践として，RDF 化された日本語 WordNet を単純な文字列マッチングにより DBpedia Japanese にリンクした．

IPAdic 辞書はフリーで使うことのできる形態素解析用日本語辞書である．IPAdic には意味記述はないが，人名や地名の固有名詞があり，これらについては DBpedia とのマッチングが有効である．日本語語彙を網羅的に LOD 化することを目的に，IPAdic 辞書を RDF 化し，それをやはり単純な文字列マッチングにより DBpedia Japanese にリンクした．

RDF 化の対象となった日本語 WordNet は英語 WordNet [Fellbaum] をベースに日本語化されたものであるため，英語と共通する一般的あるいは抽象的意味の語彙については十分であっても，日本語独自の語彙についてはそうではないという問題があった．日本語 WordNet には意味記述があり，NICT の努力によりバージョン 1.1 では用例も日本語化されている．しかし日本語特有の語彙については残念ながらまだ欠けるところがある．また，日本語としての品詞情報も十分ではない．一方，IPAdic には日本語辞書として十分な語彙と品詞情報はあっても，それらに意味は付与されていない．両者を直接リンクすることで，日本語 WordNet の語彙を充実させ，IPAdic 語彙に意味を与えることが可能になる．将来日本語語彙の事典化やボトムアップ的オントロジー開発について LOD 的アプローチの道が開かれると考える。

本報では，次節にて DBpedia Japanese について述べ，第 3 節において日本語 WordNet の LOD 化について述べる．第 4 節にて IPAdic 辞書の LOD 化について述べ，第 5 節において日本語 WordNet と IPAdic の融合について述べる．第 6 節では LOD の視点から，これら LOD の公開とオープンライセンスについて記述する．第 7 節で関連研究について触れ，最後に，第 8 節にて結論とする．

連絡先: 小出誠二，情報・システム研究機構，koide@nii.ac.jp

2. DBpedia Japanese

DBpedia Japanese は日本語の Wikipedia からスクラッチで生成されたものである．その生成方法は英語 DBpedia [Bizer] のそれと変わるところはない．日本語 Wikipedia のインフォボックスには他国にはない日本特有の情報が多くあり，DBpedia Japanese に対する我々の努力のほとんどは，Wikipedia インフォボックスを対象にしたオントロジーマッピングの作業に集中した．我々はこれまで，博物館収蔵品と生物種に関する LOD を LODAC プロジェクト *1 として実施してきたが，この DBpedia Japanese の開発と提供も，この活動の一環として行われたものである．

DBpedia Japanese には現在 69,833,233 トリプルがあり，表 1 に示すようなオントロジーマッピングの現状である．おおまかに言って，英語 DBpedia の半分強のマッピング割合であり，現在もこの数値を改善中である *2．

3. 日本語 Wordnet の LOD 化

3.1 日本語 WordNet の RDF 化

現在最新の日本語 WordNet(WN-ja, v1.1) は Princeton の英語 WordNet 3.0 をベースに日本語化が行われたものである．実際，その中には Princeton の WordNet 3.0 が含まれる．したがって，日本語 WordNet(WN-ja) の RDF 化には当然のこととして，Princeton WordNet 3.0 の RDF 化が含まれる．詳細は既報 [小出, 他] にゆずるが，我々は WordNet2.0 に対する W3C のワーキングドラフト [Assem, et al.] に従い，スキーマについては WordNet2.0 のための wnfull.rdfs をそのまま WordNet3.0 および WN-ja にも流用して，WordNet2.1 にて導入された，wn21schema:instanceHyponymOf/instanceHypernymOf のみを新しくスキーマ定義した．一方，WordNet におけるプロパティ以外のすべてのインスタンスは，たとえ内容が更新されていなくても WordNet3.0 での内容という意味で，wn20instances から wn30instances の名前空間に更新し，日本語部分については wnja11instances の名前空間におけるインスタンスとした．

*1 http://lod.ac
*2 最新の情報は http://mappings.dbpedia.org/server/statistics/ja/ を参照されたい．

表 1: DBpeida のオントロジーマッピング実績

	日本語		英語	
Wikipedia テンプレートのマップ割合	4.30%	(72 of 1,675)	5.45%	(343 of 6,292)
Wikipedia プロパティのマップ割合	2.53%	(1413 of 55,819)	3.63%	(5,859 of 161,584)
Wikipedia テンプレートの実生起のマップ割合	50.34%	(238,157 of 473,066)	84.59%	(2,002,599 of 2,367,449)
Wikipedia プロパティの実生起のマップ割合	37.99%	(2,661,059 of 7,004,462)	57.29%	(24,247,880 of 42,323,468)

WordNet には語表記を表す語 (Word) と，意味を表す同義語集合 (Synset) と，両者を 1 対 1 につなぐ語義 (Word Sense) がある．RDF においては語と語義は `wn20schema:word` と `wn20schema:sense` でリンクされ，語義と同義語集合も別のプロパティでリンクされる．たとえば，`wnja11instances:word-`銀行の語義として，`wnja11instances:wordsense-`銀行`-noun-1` および `wnja11instances:wordsense-`銀行`-noun-4` があり，前者は `wn30instances:synset-bank-noun-9` につながり，後者は `wn30instances:synset-depository_financial_institution-noun-1` につながっている．しかしながら，日本語のついた同義語集合はない [*3]．

3.2 日本語 WordNet の DBpedia へのリンク

日本語 WordNet を DBpedia にリンクするにあたって，名詞 (厳密には英語の WordNet の Noun に属する語義につながる日本語) のみをとりあげた．その理由は DBpedia にあるリソース [*4] は一部を除きすべて名詞に該当するものと考えたからである．WordNet の名詞と DBpedia のプロパティについても，参考までにリンクを取った．共通する語が見られたからである．

WordNet には意味記述があり，本来はこの意味記述を参考にして，DBpedia Japanese のリソースへのリンクが貼られるべきであるが，今回は単純な文字列の完全一致したものを `skos:closeMatch` でつなぐことをした．たとえば，`wnja11instances:word-`銀行 は `<http://ja.dbpedia.org/resource/`銀行`>` につなげられた．

漢字を多く含む日本語には多義語が多いという特徴がある．WN-ja には `wnja11instances:wordsense-`縁`-xx` という語義が relation から border まで全部で 19 個ある．すなわち，19 の異なる意味が `wnja11instances:word-`縁 にリンクされている．一方，DBpedia では `<http://ja.dbpedia.org/resource/`縁`>` というリソースがあり，これに `dbpedia-ja:`縁起 から `dbpedia-ja:`血縁 まで 5 個の disambiguates のリンクがある．今回の実現では，語表記に着目して `wnja11instances:word-`縁 と `<http://ja.dbpedia.org/resource/`縁`>` がリンクされたが，理想的には WordNet の語のレベルではなく，語義のレベルで DBpedia の disambiguattes のリンク先にリンクすべきであろう．将来の課題としたい．

表 2 は WN-ja から DBpedia Japanese へのリンク数とその割合，表 3 は DBpedia Japanese から WN-ja へのリンク数とその割合を示している．今回の実施では，文字列の厳密な一致をとったので，両者のリンク数はもちろん一致する．WN-ja における 50 %の日本語エントリで DBpedia Japanese へのリンクが得られた．

表 2: WN-ja から DBpedia-ja へのリンク数

DBpedia	# of linked	# of WN nouns	rate
resources	33,017	65,788	50.1%
properties	1,245	65,788	1.9%

表 3: DBpedia-ja から WN-ja へのリンク数

DBpedia	# of linked	# of IRIs	rate
resources	33,017	1,456,158	2.3%
properties	1,245	16,020	7.8%

4. IPAdic 辞書の LOD 化

4.1 IPAdic 辞書の RDF 化

IPAdic バージョン 2.7.0[*5] についてその RDF 化を行った．WordNet のスキーマをそのまま流用できれば好都合であるが，IPAdic の記述と日本語 WordNet を混在させてしかも区別したいとなると，そうはいかない．プロパティ `wn20schema:word` や `wn20schema:sense` をそのまま IPAdic にも流用すると，IPAdic の語や語義の実現も，定義域や値域の制約から `wn20schema:Word` や `wn20schema:WordSense` のインスタンスとなる．RDF ではインスタンスは複数のクラスの実現になり得るので，RDF 意味論では `wn20schema:word` や `wn20schema:sense` の定義域や値域に `ipadic27schema:Word` や `ipadic27schema:WordSense` を追加することは可能であるが，複数の定義域や値域はシステムによっては混乱のもととなることが予想される．そこで wnfull.rdfs に記述された構造と定義をそのまま流用し，ただ名前空間だけを `wn20schema` から `ipadic27schema` に変更した．すなわち，`wn20schema:word` や `wn20schema:sense` ではなく `ipadic27schema:word` や `ipadic27schema:sense` を定義し，これを使うことにした．それらの定義域や値域はもちろん `ipadic27schema:Word` や `ipadic27schema:WordSense` である．

また，IPAdic では品詞情報やその接続コストに加えて，読みや発音情報も記述してある．これらの情報もすべて RDF 化した．読みや発音は語に付くのではない．語の表記が同じでも読みがことなれば違う意味を指すことが普通である．たとえば，IPAdic には「縁」に固有名詞の「ユカリ」と一般名詞の「エン」,「エニシ」,「フチ」,「ヨスガ」,「ユカリ」の五つがある.「エン」と「エニシ」は同義かもしれないが，これらと「フチ」,「ヨスガ」は明らかに異なる意味である．したがって，`ipadic27schema:yomi` の定義域は `ipadic27schema:WordSense` である．

実のところ，IPAdic には意味の記述がない．そこで今回は複数の読みを分解してそれぞれを `ipadic27schema:WordSense` のインスタンスとして，IPAdic における語義を作成した．たとえば，`ipadic27instances:wordsense-`縁

*3 用例 (gloss) には英語の日本語訳が付けられている．
*4 DBpedia にはそのほかにインフォボックス用のプロパティと Wikipedia ページ用のページ情報がある．
*5 `http://chasen.naist.jp/snapshot/ipadic/ipadic/doc/ipadic-ja.pdf`

-noun-1，ipadic27instances:wordsense-縁-noun-2 などである．

4.2 IPAdic の DBpedia へのリンク

日本語 WordNet の場合と同様に，単純な文字列の完全一致により，IPAdic の名詞のみについて DBpedia のリソースへのリンクをとった．ipadic27instances:word-縁 は <http://ja.dbpedia.org/resource/縁> にリンクされる．Wikipedia に人名の「縁」はないようである．「エン」，「エニシ」，「ユカリ」の読みの ipadic27instances:word-縁 に相当するのは <http://ja.dbpedia.org/resource/因縁> かもしれないが，IPAdic には言葉の意味については一切の情報がないので IPAdic の作者の意図はわからない．今回の実現では文字列レベルでの完全一致であるが，今後は，とりあえず作成した ipadic27instances:wordsense-縁-noun-1，ipadic27instances:wordsense-縁-noun-2 などと，「因縁」というような異なる語表記の語も含めて，IPAdic の同義語集合 (Synset) をつくるのと，語義の DBpedia へのリンクが課題である．

表 4 は IPAdic から DBpedia Japanese へのリンク数とその割合，表 5 は DBpedia Japanese から IPAdic へのリンク数とその割合を示している．

表 4: IPAdic から DBpedia-ja へのリンク数

DBpedia	# of linked	# of IPAdic nouns	rate
resources	54,735	197,489	27.7%

表 5: DBpedia-ja から IPAdic へのリンク数

DBpedia	# of linked	# of IRIs	rate
resources	54,735	1,456,158	3.8%

5. 日本語 WordNet と IPAdic の融合

先に述べたように，日本語 WordNet と IPAdic の両者を融合することで，互いの欠点を補って，意味も含んだ日本語の電子化辞書として充実したものにできる可能性がある．その第一歩として，これまでと同様なやり方で，語のレベルで互いのリンクをとってみた．結果は表 6 および 7 に示すとおりである．

表 6: WN-ja から IPAdic へのリンク数

IPDdic	# of linked	# of WN nouns	rate
Noun	27,384	65,788	41.6%

IPAdic の語彙数は WN-ja の 3 倍ある．WN-ja を参考に，IPAdic に意味を付与していくというアプローチが妥当と思われる．

6. LOD としての公開

RDF 化された日本語 WordNet は CC-BY ライセンスとして公開され，Data Hub[*6] に登録された．ダンプファイルは LODAC のサイト [*7] からもダウンロードできる．一つは日本語 WordNet の RDF ファイル，もう一つがその DBpedia Japanese へのリンクファイルである．

表 7: IPAdic から WN-ja へのリンク数

WNja	# of linked	# of IPAdic nouns	rate
noun	27,384	197,489	13.9%

なお，DBpedia Japanese 自身も Data Hub に CC-BY-SA で登録済みであり，我々の DBpedia リポジトリ [*8] において，オンライン（IRI により参照解決可能，dereferenceable）および SPARQL エンドポイントで利用可能であり，ダンプファイルも得ることができる．この DBpedia Japanese のレポジトリには，英語の DBpedia リンクと同様に日本語 WordNet へのリンクも存在している．また英語 DBpedia からもリンクされていることを付け加えておく．

RDF 化された IPAdic も日本語 WordNet と同様な処置で公開する予定である．

7. 関連研究

玉川らはオントロジーの自動構築を目的に，日本語 Wikipedia の各種情報を利用して，大規模な汎用オントロジー構築手法について研究を行って，構築したオントロジーを日本語 Wikipedia オントロジーと呼んで公開している [玉川 10, 玉川 11, 玉川 13]．rdfs:subClassOf や SKOS の各種プロパティよりももっと制約のゆるい曖昧さを許容するプロパティを独自に定義しているのが特徴である．

森田ら [森田，他] はオントロジーアライメントの技術を用いて日本語 WordNet の情報を参考にして，日本語 Wikipedia オントロジーの改善を行っている．

英語 WordNet と英語 Wikipedia を用いたオントロジー構築に YAGO [Suchanek] がある．WordNet の hypernym/hyponym 関係は rdfs:subClassOf に類似であるが，RDF クラス構造から言えば，W3C のスキーマにおいてすべての語，語義，同義語集合はクラスではなくインスタンスである．つまりセマンティックウェブにおけるオントロジーとするには，WordNet からオントロジー的 Is-A 関係を抽出して rdfs:subClassOf 関係を作らなければならないが，WordNet の hypernym/hyponym 関係がオントロジー的に十分整合しているとは言い難い [*9]．そこで YAGO では独自のオントロジー構築を目指している．

上記の研究のいずれも，Wikipedia から直接情報を得ていて，DBpedia は用いていない．玉川らには，DBpedia Japanese の提供が最近であるという事情も推察される．一方，彼らは最近 DBpedia Japanese へのリンクを生成している．YAGO も玉川らも Wikipedia のカテゴリを参考にしているが，DBpedia のインフォボックスを用いてはいない．DBpedia Japanese を提供している我々としては，必要とあらばオントロジー的な整合性に問題がある Wikipedia インフォボックスデータの改善も行って，インフォボックスのオントロジーマッピングを用いて，どのようにオントロジー構築が可能かを探るというのが立場である．その際，既存の WordNet や研究も大いに参考になるであろう．

*6 http://datahub.io/ja/
*7 http://lod.ac/dumps/wordnet/20121228/
*8 http://ja.dbpedia.org/
*9 特に WordNet 2.1 で導入された instanceHyponym/instanceHypernym 関係では実際には問題が多い．

8. まとめ

日本語 WordNet と IPAdic の二つの辞書の RDF 化を行った．電子化辞書にはこれ以外のものもあるが，いずれも商用もしくは有料であり，LOD には利用できない．LOD の視点から言えば，DBpedia にリンクすることが LOD クラウドに存在するために決定的であり，今回 RDF 化した辞書も DBpedia へのリンクを取った．すでに日本語 WordNet については掲載されている．

今回の RDF 化においては，日本語 WordNet のみならず IPAdic においても W3C の WordNet に対する RDF 化指針に沿って行い，将来の日本語 WordNet と IPAdic 辞書の融合を可能とした．

日本語 WordNet，IPAdic 辞書，DBpedia Japanese とそれらの融合が，将来日本における LOD とオープンナレッジのインフラストラクチュアとして発展することを期待している．

参考文献

[小出, 他] 小出誠二, 武田英明, 大向一輝:WordNet 日本語化への LOD アプローチ, 第 26 回セマンティックウェブとオントロジー研究会, SIG-SWO-A1103-05 (2011)

[Isahara 08] Isahara, H. et al.: Development of Japanese WordNet, The 6th Edition of the Language Resources and Evaluation Conference (LREC-2008), Marrakech (2008).

[Fellbaum] Fellbaum, C.: *WordNet: An Electronic Lexical Database*, MIT Press (1998).

[Bizer] Bizer, C., J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann: DBpedia - A Crystallization Point for the Web of Data, J. Web Semantics, **7** (3), pp.154–165 (2009).

[Assem, et al.] van Assem, M., Gangemi, A., Schreiber, G., (eds.): RDF/OWL Representation of WordNet, W3C Working Draft, `http://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/` (2006)

[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: 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論文特集「2009 年度全国大会近未来チャレンジ」, Vol.25, No.5 pp.623–636 (2010)

[玉川 11] 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特集論文「近未来チャレンジ」, Vol.26, No.4 pp.504–517 (2011)

[玉川 13] 玉川奨，香川宏介，森田武史，山口高平: 日本語 Wikipedia オントロジーの構築と利用, 第 29 回セマンティックウェブとオントロジー研究会, SIG-SWO-A1203-01 (2013)

[森田，他] 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 (2012)

[Suchanek] Suchanek, F. M., G. Kasneci, and G. Weikum: YAGO: A Large Ontology from Wikipedia and WordNet, Web Semantics: Science, Services and Agents on the World Wide Web, Vol.6, No.3, Pages 203–217 (2008)