人工知能学会研究会 資料 SIG-SWO-041-13

日本語 Wiktionary の LOD 化

Transformation of Japanese Wiktionary to Linked Open Data

鵜飼孝典 [1*] 小林賢司 [2]

[1] (株) 富士通研究所, [2] 富士通株式会社
[1] Fujitsu Laboratories Ltd. [2] Fujitsu Ltd.

Abstract: It requires a larger dictionary to make an application to interact more naturally with the users. Japanese WordNet is one of the free dictionaries, which includes thesaurus. It has RDF formed dataset that links to other resouces such as DBpedia, so it is useful for our supposing application. However the WordNet is insufficient because of the small volume of vacabulary, lack of parts of speech and derivative relations, and so on. Japanese Wiktionary, which is another free dictionary, has the parts of speech and the derivative relations. We have built a glammar ontology based on the structure of the Wiktionary to express of missing informations in the WordNet ontology. In this article, we combined the Wiktionary with the WordNet and Japanese DBpedia. The combined data is used as one unified Japanese machine readable dictionary. And from the point of the view of Japanese DBpedia, DBpedia is searchable using gramatical elements such as synonymous words. The number of the links between the Wiktionary and WordNet is 13,620 and the number between the Wiktionary and DBpedia is 9,654. The dataset is published using LOD4ALL.

1 はじめに

人工知能分野の活発化に伴い，対話技術や質問応答技術の発展が一層期待されている．その中で我々は，アプリケーションがテキストあるいは音声を入力とした自然言語から，正しく意味を解釈し，自然な回答を出力できることを目指している．そのためには，語句の意味や概念だけでなく，活用や用法，読み，発音などを含んだ，構造化された機械判読可能な辞書データが必要となる．例えば,「目的地まで歩きたい」という入力に対し,「徒歩なら 10 分かかります」と返答するには，動詞 “歩く” の連用形が “歩き” であり，それを名詞化した “歩き” という派生語と “徒歩” が同じ概念である，という知識を保持する必要がある．

現在，公開されている代表的な日本語辞書データとしては，日本語 WordNet[1] や日本語 Wiktionary が挙げられる．日本語 WordNet は，シソーラスであり，語句の意味，同義語，上位／下位概念などがまとめられている．また，データは RDF(Resource Description Framework) で提供され，DBpedia など他のリソースともリンク付けられている．RDF は, Web において情報を記述するグラフベースのデータモデルに基づいた形式的言語であり, Web の情報をソフトウェアによる自動処理などに使われることを想定して作られている [2]. そのため，上記アプリケーションを想定する本研究にとって，日本語 WordNet を利用することは都合が良い．しかし，語彙量は十分ではなく，品詞や活用形の情報が貧弱，派生語の登録有無が曖昧，などの問題がある．

一方，日本語 Wiktionary は，日本語 WordNet と異なり，RDF 化されておらず，概念体系について乏しいが，品詞や活用形，漢字・読み表記関係，訳語，発音，語源など，日本語 WordNet では不足している情報を多く持つ．また，日本語 WordNet が登録していない語句も持っており，更には活用による派生語や，漢字・読み表記関係まで考慮すると，日本語 WordNet 以上の語彙量を抽出できる可能性がある．

本研究は，日本語 Wiktionary のデータを RDF 化し，日本語 WordNet のデータ（オントロジ）と統合することによって，日本語 WordNet で不足している語彙を拡充することを目的とする．これまでに我々は，以下の 2 つを行なった [3].

- 日本語 Wiktionary から抽出したデータを元に，品詞・活用形，派生語関係，表記関係について RDF 化した “文法オントロジ” を作成する．
- 文法オントロジを活用し，日本語 WordNet オン

* 連絡先:(株) 富士通研究所
〒211-8588 神奈川県川崎市中原区上小田中 4-1-1
E-mail: ugai@jp.fujitsu.com

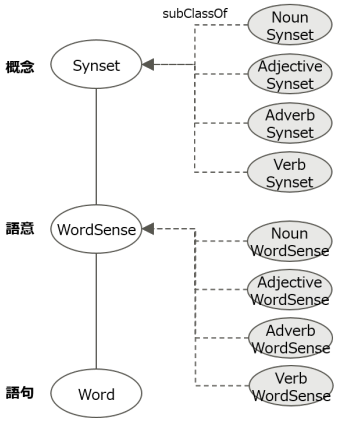

図 1　WordNet オントロジのクラス公理概要

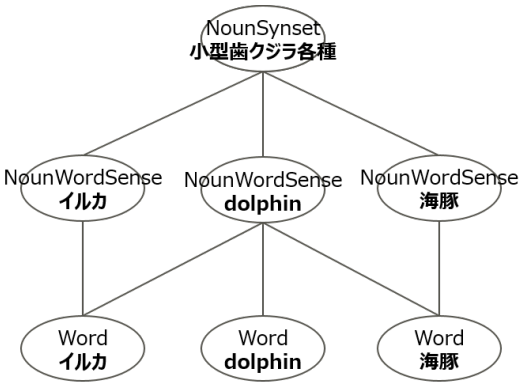

図 2　WordNet オントロジのインスタンス例

トロジにおいて欠落した語句や品詞・活用形，派生語関係，表記関係を補完する．

上記を行った結果，語彙量としては，日本語 Wiktionary が持つ 31,302 語に対して，全体の語彙量は約 1.16 倍となった．

本稿では，先に作成した日本語 Wiktionary のデータと日本語 WordNet を結合し，さらに日本語 DBpedia とも結合する．これにより，日本語 WordNet と日本語 Wiktionay を一体とした日本語辞書として用いることができるようになる．また, 日本語 DBpedia においては, DBpedia が持たない同音異義語など文法的な関係を用いたリンクを使った探索が可能になる．

本稿の構成は次の通りである．第 2 節では，日本語 WordNet とその問題点について述べる．第 3 節では，日本語 Wiktionary とこれまでに我々が日本語 Wiktionary から抽出したデータについて述べる．第 4 節では，今回作成した日本語 WordNet, 日本語 DBpedia へのリンクについて述べ，第 5 節では，考察を述べ，第 6 節で，関連研究について述べる．最後の第 7 節でまとめと今後の課題について述べる．

2　日本語 WordNet

日本語 WordNet は，プリンストン大学で開発された Princeton WordNet をベースとして，日本語向けに開発された WordNet である．WordNet は RDF データも公開されており，日本語 WordNet もこれに則して記述している．語彙量としては，約 9 万語が収録されている．

WordNet オントロジのクラス公理概要を図 1 に示す．WordNet オントロジは，概念 (Synset)，語意 (WordSense)，語句 (Word) の 3 クラスから成る．概念は，同義語となる語句の集合を表すクラスである．概念間では上位・下位，類似などの関係も定義される．語意は，語句の 1 つの意味を示し，概念と語句を紐付けるクラスである．語句は，少なくとも 1 つの意味を持つ表記を示すクラスである．また，概念上の品詞として，名詞・動詞・形容詞・副詞の 4 種に分類されており，概念または語意のサブクラスとして表現される．各国語固有の品詞が登録される際は，その 4 種いずれかに分類される．例えば，日本語における形容動詞や連体形は，形容詞に分類される．

インスタンス例を図 2 に示す．“イルカ”，“dolphin”，“海豚” という語句は，“小型歯クジラ各種” という共通の名詞概念を表す語意を持つ．なお，この例では，“dolphin” の 1 つの語意を表す語句として，“イルカ” や “海豚” も含まれる．

2.1　問題点

自然な対話や質問応答を実現するに当たって，日本語 WordNet を使用するには以下の問題点がある．

語彙量が少ない
: 日本語の一般的な辞書における収録語彙量は 20 万語上あることから，より多くの語が登録されることが望まれる．例えば，岩波書店発行の「広辞苑第六版」においては，約 24 万語が収録されている．

品詞・活用形の情報が不足している
: 自然言語の解析や生成を行うには，言語に合わせた，形容動詞／助動詞など詳細な品詞情報や，形容詞や動詞など用言に分類される品詞の活用形情報が必要となる．しかし品詞については，概念品詞の 4 種に留まり，活用形については分類すらされていない．

派生語の登録が曖昧となっている

WordNet では，概念的な品詞が異なる場合は，派生語であっても区別して登録される．例えば，形容詞 “美しい” と，その副詞的用法となる “美しく” は，両方とも登録される．しかし，実際には，派生語が揃って登録されていない語句がある．

漢字・読み表記関係がない

日本語は，ある語句に対して，主に漢字や平仮名の組合せで異なる表記を行えるため，表記が異なっても検索できることが望ましい．しかし，日本語 WordNet では，漢字・読み表記の関係にある語句は，同じ概念の語句として登録されているが，両者の関係は特に定義されていない．

3 日本語 Wiktionary

日本語 Wiktionary は，Wiki を使った参加編集型の辞書サービスである Wiktionary の日本語版である．1 語句に対して 1 ページで，活用や用法，漢字・読み表記関係，訳語，発音，語源など，日本語 WordNet では不足している語句のより詳細な情報を収録している．また，WordNet ほどではないが，上位語や下位語，類義語など関連語も収録しており，その面でも今後活用できると考えられる．収録される日本語は約 3 万語 (ページ) であるため，日本語 WordNet と比較すると見劣りするが，ページ内には前述した関連語や漢字・読み表記関係など，他の語句の情報も多く含まれているため，潜在的な語彙量は豊富と言える．

3.1 文法オントロジ

本節では, 我々の先の研究で作成した文法オントロジについて述べる．

前述した通り，日本語 Wiktionary は，活用後の語句や，漢字･読み表記関係まで考慮すると，日本語 WordNet の問題を補うことが可能であると考えた．

語句の品詞・活用形，派生語関係，漢字・読み表記関係を表すクラス・プロパティ公理を定義し，日本語 Wiktionary から抽出したデータを当てはめた文法オントロジを作成する．品詞・活用形については，学校文法に習った表現とし，また語意クラスが示すのは概念的な品詞であるため，新規に語句タイプを示すクラスを定義した．派生語関係は，文法的に派生語の生成が可能な関係を示すプロパティ，漢字・読み表記関係は，語意との関係を考慮した表記関係を示すプロパティをそれぞれ定義した．

そして, 文法オントロジと日本語 WordNet オントロジを照らし合わせ，品詞・活用形や派生語関係，表記関係の補完を行った．品詞・活用形については，双方のオントロジが持つ語句と，その語句を持つ語意の概念的品詞が一致すれば，リンクを付与する．派生語関係についても同様に，派生語関係にある語句を持つ語意が存在し，その語意の概念的品詞が一致すれば，その語意間にリンクを付与する．例えば，文法オントロジにおいて，動詞 “走る” の活用形 (連用形名詞) は “走り” であるとする．ここで, “走る” の動詞的語意と “走り” の名詞的語意の表記を確認し，一致していれば，双方の語意をリンク付けする．これによって，ある語意に対して，別の品詞に言い換えた語意へと辿ることが可能となる．表記関係の補完については，日本語 WordNet オントロジにおける同一概念に属す語意の語句が，文法オントロジが持つ漢字・読み表記関係にある語句に一致するか，または，漢字語句を持つ語意に対して，読み表記となるかな語句とリンク付けする．

4 外部リンクの作成

4.1 日本語 WordNet とのリンク

我々が日本語 Wiktionary から抽出したものと追加した派生語をあわせると 50,657 語になり日本語 WordNet が持つ 93,834 語と共通するのは 13,620 語であった．

このリンクは, 図 2 の Word のレベルで行なった. 日本語 Wiktionary から抽出した語の Word クラスのエンティティと日本語 WordNet の Word クラスのエンティティをラベルが同じものをリンク付けした.

4.2 日本語 DBpedia とのリンク

日本語 Wiktionary のデータから日本語 DBpedia とのリンクは, 文献 [4, 5] に示されるものと同様の方法を用いた. すなわち, 日本語 Wiktionary の Word クラスのエンティティから日本語 DBpedia のエンティティのラベルが同じものをリンク付けした. 9,654 語がリンク付けされた.

5 考察

図 3 は, “走る” という語について日本語 Wiktionary からリンクをたどって, 日本語 WordNet と日本語 DBpedia のデータを表示したものである.

左側が 日本語 Wiktionary のデータで, リソースの部分から 8 つの意味が記述されていることがわかる. 下の部

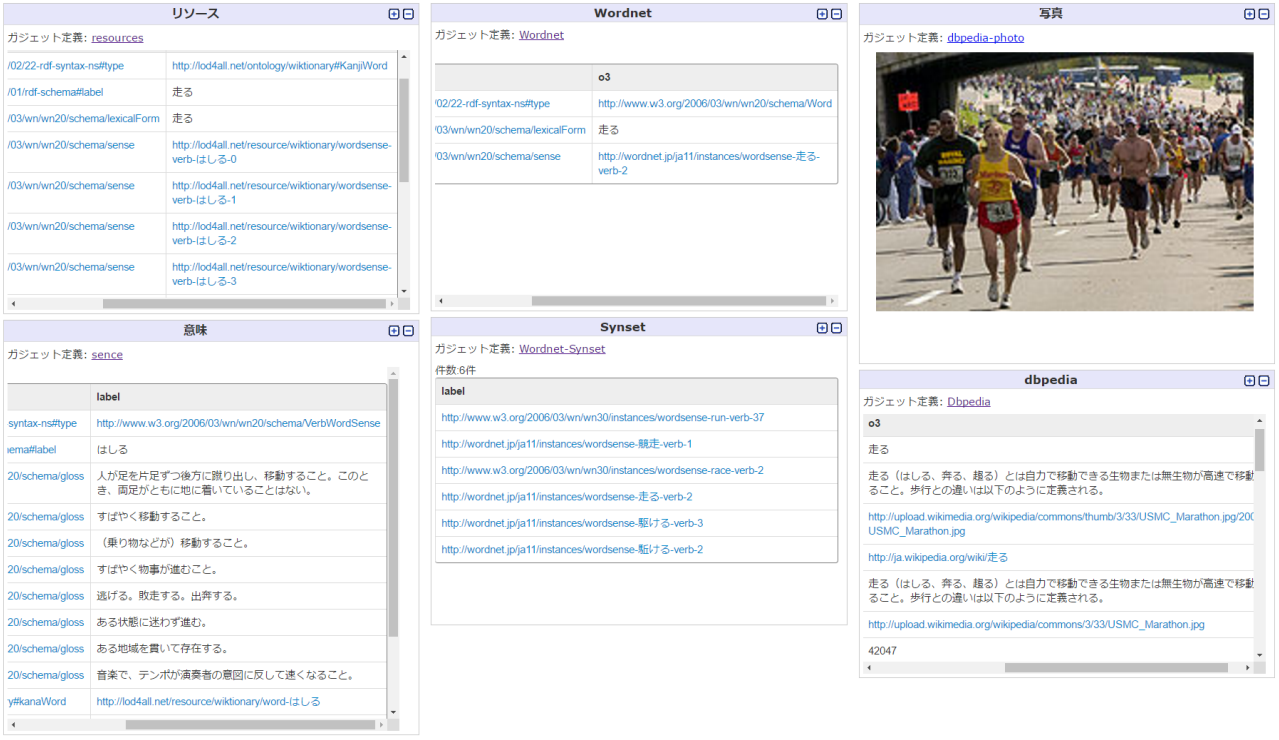

図 3　“走る” のデータの表示例

分から, 8 つの中の最初が, “人が足を片足ずつ後方に蹴り出し、移動すること。このとき、両足がともに地に着いていることはない。” であることがわかる. 真中が, 日本語 WordNet のデータで, 上の部分から意味が 1 つ定義されていて, 下の部分から, 同じ定義を持つ語が, “競争”, “走る”, “駆ける”, “駈ける” の 4 つあることがわかる. 右側が日本語 DBpedia の情報で, 上に DBpedia で使われている画像を表示し, 下の部分にその他の情報を表示している.

このリンク関係を用いると, 日本語 Wiktionary と日本語 WordNet のデータは派生語や読みなどの文法情報を相互補間した日本語辞書として利用することができる. 日本語 DBpedia においては, DBpedia が持たない同音異義語など文法的な関係を用いたリンクを使った探索が可能になる.

本研究により，日本語 Wiktionary から抽出した文法オントロジを活用することによって，日本語 WordNet と日本語 Wiktionary の語意や語句について，用法・活用，派生語関係や表記関係を補完することが出来た．しかし，語意同士のリンク付けは行なっていない. これは，派生語や別表記の語意を定義していない点もあるが，日本語 WordNet と日本語 Wiktionary が持つ語意を独立した扱いとしていることも要因として大きい．つまり，本来であれば，同じ語意は同じリソースに統合されているのが望ましいと考えるが, 現状は別のリソースとして登録してしまっている．これは，現状の抽出情報だけでは，語意の同定が困難であったためである．

6　関連研究

WordNet と Wiktionary を統合する技術として，[6][7] がある．これらは，WordNet と Wiktionary 間で登録されている語句に対し，意味や関連語関係などから類似度を算出することによって，語彙の同定，及び統合を行っている．また，日本語 Wiktionary ではなく日本語 Wikipedia を日本語 WordNet と統合する研究 [8][9] もある．しかし，いずれも各国語由来の品詞や活用形，それを使用した派生語関係，表記関係については表現していない．

派生語の生成について，[10] では，コーパスから収集した派生語用例を生成規則の形で記述し，その適用確率を学習している．適用確率を使用することにより，派生語らしく，使用頻度も高い語が受理される．ただし，派生語は，語幹を成す名詞と接尾語との連接に限られており，本稿のような活用語は対象としていない．しかしながら，コーパスや確率の使用による判定は，本稿の派生

語の精度を上げる可能性があるため，今後検討していきたい．

また，オープンな辞書データとして，IPADIC[11] がある．現在，公式では開発が進められておらず，また類義語や反意語など関連語関係を持たないため，本稿では Wiktionary を採用した．日本語 WordNet と IPADIC をリンク付けする研究も行なわれた [5] が, 現在までに成果は公開されていない．

7 おわりに

本研究では，先に作成した日本語 Wiktionary のデータと日本語 WordNet を結合し，さらに日本語 DBpedia とも結合した．これにより，日本語 WordNet と日本語 Wiktionay を一体とした日本語辞書として用いることができるようになった. また, 日本語 DBpedia においては, DBpedia が持たない同音異義語など文法的な関係を用いたリンクを使った探索が可能になった．

我々が日本語 Wiktionary から抽出したものと追加した派生語をあわせると 50,657 語になり日本語 WordNet が持つ 93,834 語と共通するして, リンク付けできたのは 13,620 語であった．

また, 日本語 Wiktionary の Word クラスのエンティティから日本語 DBpedia のエンティティのラベルが同じもの, 9,654 語がリンク付けされた．

本研究で作成したデータは, LOD4ALL[12][*1]にて公開している．

今後の課題として，現状は語意の統合が不十分であるため，語意の説明文や関連語などを使用して，WordNet と Wiktionary 間の語意を同定，あるいは生成し，統合する方法について引続き検討する．また，派生語については，実際には登録すべきでない使用されない語も含まれるため，コーパスや他の辞書データと連携するなどして，検証する必要がある．

参考文献

[1] Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. Development of the Japanese WordNet. In *LREC*, 2008.

[2] Graham Klyne and Jeremy Carroll. Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C recommendation, W3C, February 2004. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

[3] 小林賢司, 鵜飼孝典, 井形伸之, 西野文人. 日本語 WordNet の語彙拡充のための文法オントロジの作成と活用. 人工知能学会研究会資料 SIG-SWO-040-04, pp. 1–7, 2016.

[4] 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.

[5] 小出誠二, 武田英明, 加藤文彦, 大向一輝. 日本語 wordnet と ipadic 辞書の rdf 化と dbpedia リンク. 2013 年度人工知能学会全国大会,1N4-OS-10b-4, pp. 1–4, 2013.

[6] Francis Bond and Ryan Foster. Linking and Extending an Open Multilingual Wordnet. In *ACL (1)*, pp. 1352–1362, 2013.

[7] John McCrae, Elena Montiel-Ponsoda, and Philipp Cimiano. Integrating WordNet and Wiktionary with lemon. In *Linked Data in Linguistics*, pp. 25–34. Springer, 2012.

[8] 山田一郎, 呉鍾勳, 鳥澤健太郎, 黒田航, 風間淳一, 村田真樹. Wikipedia を利用した日本語 WordNet への用語追加の検討. 言語処理学会第 16 回年次大会発表論文集, pp. 948–951, 2010.

[9] 森田武史, 玉川奨, 山口高平. 日本語 Wikipedia オントロジーと日本語 Wordnet の統合 (学習およびその応用). 知識ベースシステム研究会, Vol. 96, pp. 9–14, 2012.

[10] 市丸夏樹, 中村貞吾, 宮本義昭ほか. シソーラスと確率文法による派生語解析. 情報処理学会論文誌, Vol. 36, No. 4, pp. 849–858, 1995.

[11] Masayuki Asahara and Yuji Matsumoto. IPADIC version 2.7. 0 User's Manual (in Japanese). NAIST. *Information Science Division*, 2003.

[12] A. Naseer, T. Kume, T. Izu, and N. Igata. LOD for All: Unlocking Infinite Opportunities. In *The Semantic Web Challenge 2014, The 13th International Semantic Web Conference*, 2014.

[*1] http://lod4all.net/