DBpedia and Wordnet in Japanese

Seiji Koide [a,b], Hideaki Takeda [a,*], Fumihiro Kato [a,b], Ikki Ohmukai [a], Francis Bond [c], Hitoshi Isahara [d], Takayuki Kuribayashi [d]

[a] *National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan*
E-mail: {takeda, i2k}@nii.ac.jp
[b] *Transdisciplinary Research Integration Center, ditto, E-mail: {koide,fumi}@nii.ac.jp*
[c] *Division of Linguistics and Multilingual Studies, Nanyang Technological University,*
50 Nanyang Avenue Singapore 639798, E-mail: bond@ieee.org
[d] *Toyohashi University of Technology, 1-1 Hibarigaoka, Tenpakucho, Toyohashi, Aichi 441-8580, Japan,*
E-mail: isahara@tut.jp, kuribayashi@lang.cs.tut.ac.jp

Abstract. Both WordNet and Wikipedia are valuable language resources covering wide domains so that an RDF version of WordNet and DBpedia play important roles in the LOD cloud. Combining them provides the basic resources for our linguistic and ontological knowledge. However, the conversion to RDF should be carried out differently for each resource because of each own lineage and characteristics. The idea of LOD should be useful to connect them. We built and published RDF of the Japanese Wordnet and DBpedia Japanese and furthermore provided the basic links between both. We expect that they will be used as the infrastructure to enrich and link other Linked Data datasets in Japan.

Keywords: Wordnet, DBpedia, LOD, ontology, Japanese electronic dictionary

1. Introduction

Research in information processing of Japanese texts has started relatively earlier than other non European languages and yielded significant contribution to digitalization in Japan in spite of the disparity in coding, structure, and vocabulary than others. Whereas the original development for text processing had produced remarkable products, it often amounted to be incoherent to development in other languages. For example, there is the electric dictionary called EDR Electronic Dictionary[1] originally developed from the scratch [1].

WordNet [2], which is the most widely used electric dictionary[2] and ontology, has not had the counterpart instead in Japan for some time. The Japanese Wordnet launched finally in 2008 [3].

DBpedia [4] also took a time to have its counterpart in Japan. DBpedia Japanese was launched in 2012 by National Institute of Informatics (NII).

Today, various datasets are interconnected to each other under the concept of Linked Data. In particular cross-media data such as encyclopedia and dictionary plays a hub to connect data in various fields. We have been realized the need for cross-media Linked Data resources in Japanese, and then we have made the conversion of NICT Japanese Wordnet to RDF and also made the connection between the Japanese Wordnet and DBpedia Japanese.

In this paper, we introduce the RDF version of the Japanese Wordnet in Section 2, DBpedia Japanese in Section 3, and the connection between them as Linked Data in Section 4. Section 5 describes the open licence of DBpedia and Wordnet from the views of LOD. Finally we conclude at Section 6.

[*] Corresponding author. E-mail: takeda@nii.ac.jp.

[1] See `http://www2.nict.go.jp/out-promotion/techtransfer/EDR/J_index.html`. It is still continued updating but not open-free.

[2] Princeton's WordNet has become the global standard of multilingual electric dictionaries today. See `http://www.globalwordnet.org/`

2. Japanese Wordnet in RDF

In this section, we describe the current state of the Japanese Wordnet and its conversion to RDF.

2.1. The Japanese Wordnet (WN-ja)

The efforts for multilingual wordnet has been made worldwide based on the Princeton's English WordNet so far. In 2008, the Japanese Wordnet (WN-ja) was developed and released by the National Institute of Information and Communications Technology (NICT) in Japan [3,5]. Currently WN-ja is built using the structure of the English WordNet. Synsets(concepts) are given Japanese words in addition to the English ones. We used version (1.1) of the Japanese Wordnet with 187,000 senses (word-synset pairs), 57,000 concepts (synsets) and 94,000 unique Japanese words. For up-to-date information on the Japanese Wordnet see `nlpwww.nict.go.jp/wn-ja`.

The first version of the wordnet was made by automatically translating the English and Euro wordnets, and then hand correcting the most common entries [6]. Since then, Japanese definitions and examples have also been added. Further, the Wordnet is linked to other resources: the Suggested Upper Merged Ontology (SUMO) [7], the Japanese semantic lexicon GoiTaikei [8], and a collection of illustrations taken from the Open ClipArt Library [9].

Currently, the Japanese Wordnet is being extended in two main areas: the first is to add more orthographic variants and concepts to the Japanese Wordnet, either by adding Japanese to existing English synsets or by creating new synsets [10]. The second is to link the synsets to more text examples [11].

2.2. W3C Working Draft and OWL Conversion of English WordNet 2.1 and More

In 2006, W3C published the Working Draft for the representation in RDF of WordNet 2.0 [12], in which the OWL representation of WordNet and an OWL schema for WordNet were introduced. Then, we applied the proposal to WordNet 2.1 with the two extended pointer properties for the new version, i.e., `instanceHypernymOf` and `instanceHyponymOf` in 2006 [13,14], and subsequently for WordNet 3.0.

Up to now, several attempts to represent Princeton's WordNet in OWL were made along with updating the WordNet. Whereas the team members of the W3C Working Draft had actually converted WordNet 2.0 to OWL representation [15], and then from the viewpoint of Linked Open Data (LOD), de Melo and Weikum has made the word search web pages [16]. The team members of W3C Working Draft has also RDFized the Princeton's WordNet 3.0 in the related activity of Europeana[3].

2.3. Conversion of the Japanese Wordnet to RDF

The latest WN-ja is built on Princeton's English WordNet 3.0. Appropriate Japanese words are added and linked to synsets via wordsenses as usual in the WordNet manner. Thus, according to the W3C proposal of RDF representation of WordNet, we have made the the conversion of WN-ja to OWL [17]. Here, `word-銀行` is made and linked to both `wordsense-銀行-noun-1` and `wordsense-銀行-noun-4`. Furthermore, the former is linked to `synset-bank-noun-9` (as building) and the latter is linked to `synset-depository_financial_institution-noun-1`. However, because of starting at the English WordNet, the Japanese vocabulary is not comprehensive, and Japanese specific concepts are still not completed.

3. DBpedia Japanese

Japanese DBpedia is the DBpedia generated from the Japanese Wikipedia. It is an internationalization of DBpedia where all softwares used to build it from Wikipedia are those developed for English DBpedia. Therefore most of the building process were done without any extra efforts except ontology building.

Ontology building in DBpedia is creating concepts in DBpedia ontology and mapping them to infoboxes in Wikipedia. Since many original infoboxes are used in the Japanese Wikipedia, we may add new concepts and mappings to enhance DBpedia. Actually, DBpedia Japanese is built and maintained in the activity of LODAC Project[4] where various data resources such as museums and biology are published as LOD. The ontology mapping is also carried out in this activity.

Currently it contains 69,798,971 triples and the statistics for ontology mapping is shown in Table 1. Roughly speaking, the rate for ontology mapping is a half of English DBpedia. There is still room to improve.

[3] `http://semanticweb.cs.vu.nl/lod/wn30/`

[4] http://lod.ac

Table 1

Statistics for Ontology Mapping in DBpeida

	Japanese		English	
rate of all templates in Wikipedia are mapped	2.63%	(44 of 1,675)	5.26%	(331 of 6,292)
rate of all properties in Wikipedia are mapped	1.70%	(951 of 55,819)	3.51%	(5,665 of 161,584)
rate of all template occurrences in Wikipedia are mapped	43.50%	(205,764 of 473,066)	84.04%	(1,989,576 of 2,367,449)
rate of all property occurrences in Wikipedia are mapped	33.93%	(2,376,652 of 7,004,462)	53.31%	(22,561,308 of 42,323,468)

4. Linking DBpedia Japanese and the Japanese Wordnet

Since both WordNet and Wikipedia are the most famous comprehensive languages resources, there are many studies how combining them can contribute to build better languages resources (e.g., Yago [18]). We have also investigated how WN-ja can be enriched by using Wikipedia [19]. Nonetheless, we here link entities in both datasets literally, i.e., link entities which share the same strings since we want to provide the basic dataset for Japanese language resources.

4.1. WordNet Schema and Instance Data

Every word, wordsense, and synset in WordNet becomes an instance for WordNet Schema. For example, word-bank is an instance of Word, wordsense-bank-9 is an instance of NounWordSense, and synset-bank-noun-9 is an instance of NounSynset. In addition, WordNet itself includes ontological ambiguity as concepts and instances. For instance, synset-European_Central_Bank-noun-1 is not linked via instanceHyponymOf but linked via hyponymOf to synset-central_bank-noun-1, although European Central Bank is an instance of *central bank* from the ontological view. These facts suggest that we cannot take the WordNet hyponym/hypernym hierarchy over as a proper ontology from the viewpoint of OWL, syntactically and semantically.

4.2. DBpedia Resource and Names

On the other hand, <http://ja.dbpedia.org/resource/銀行> is an abstract resource that denotes a concept *bank*. The resource of DBpedia exists as an identifier in DBpedia for a thing and it is also an IRI in order to link actual web pages and other various data like infobox in Wikipedia. Therefore, it is a misuse to link IRIs of words in Wordnet and IRIs of DBpedia resources by owl:sameAs. It would result in forcing both exactly the same entity in the world. Furthermore, every entities in DBpedia is an instance of owl:Thing, but it is actually an OWL individual without such a definition that specifies it to owl:Class. Therefore, we used skos:closeMatch[5] to link directly both <http://wordnet.jp/ja11/instances/word-銀行> and <http://ja.dbpedia.org/resource/銀行>.

Note that <http://ja.dbpedia.org/resource/欧州中央銀行> (European Central Bank) is linked <http://ja.dbpedia.org/resource/銀行> only with wikiPageWikiLink in this case.

4.3. Statistics

We selected only nouns from WN-ja. One reason is to avoid needless ambiguity: Japanese verbs presented by Chinese characters in the Japanese wordnet 1.1 are represented as their stem (*noun*) without any inflection (i.e. not *noun* する "do noun"). This makes them indistinguishable from nouns.[6] As we are only linking on lemmas, this would add unnecessary ambiguity.

On the other hand, there are categorically three types of IRIs in DBpedia-ja, i.e., resource, property, and Wikipedia. All Wikipedia URLs are linked from resource IRIs. Therefore, we selected resource type IRIs and property type IRIs for linking. By the way, one more reason of selection of nouns of WN-ja is that resource IRIs of DBpedia may be categorized to noun from the views of part of speech.

Table.2 shows the statistics of linking data of WN-ja to DBpedia-ja, and Table.3 shows the statistics of linking data of DBpedia-ja to WN-ja. In this attempt of linking by words in WN-ja and resource and property names in DBpedia-ja, we made the connection by literally exact matching. Therefore, the mapping is ex-

[5] skos:closeMatch is a subproperty of skos:semanticRelation, and then its domain and range are skos:Concept. Since skos:Concept is typed to owl:Class, then <http://wordnet.jp/ja11/instances/word-銀行> and <http://ja.dbpedia.org/resource/銀行> are entailed as an instance of skos:Concept. This is more acceptable than owl:sameAs and its entailment

[6] In the most recent version of the Japanese Wordnet (in subversion) they are explicitly marked as *noun*+する.

actly one by one and inversely equivalent in this case.

Table 2

Number of Linked Data from WN-ja to DBpedia-ja

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

Table 3

Number of Linked Data from DBpedia-ja to WN-ja

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

5. Publishing as LOD

RDFized the Japanese Wordnet is published as LOD with CC-BY and has been registered to the Data Hub. It is now available the dumped files (one for the Japanese Wordnet itself and the other for links to DBpedia Japanese. See `http://lod.ac/dumps/wordnet/20121228/`).

DBpedia Japanese is also published as LOD with CC-BY-SA and has been registered to the Data Hub. It is available either by online (dereferenceable IRIs), SPARQL Endpoint, and the dumped file (See `http://ja.dbpedia.org/`). It contains links to the Japanese Wordnet and English DBpedia, and it is linked from English DBpedia.

6. Conclusion

Wordnet and Wikipedia are precious language and ontology resources covering wide domains. The RDF representations of Wordnet and DBpedia play important roles in the LOD cloud. Combining them can provide the basic resources for our knowledge. We expect that the RDF representations of the Japanese Wordnet and DBpedia Japanese will be used as the infrastructure to enrich and link other Linked Data datasets in Japan.

References

[1] T. Yokoi: The EDR Electronic Dictionary, Commun. ACM, **38** (11), pp.42–44, ACM (1995).

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

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

[4] C. Bizer, 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).

[5] F. Bond, et al.: Enhancing the Japanese WordNet in The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP 2009, Singapore (2009).

[6] F. Bond, et al.: Japanese WordNet 1.0, 16th Annual Meeting of the Association for Natural Language Processing, A5-3, Tokyo (2010).

[7] I. Niles, A. Pease: Towards a Standard Upper Ontology, Proc. 2nd Int. Conf. Formal Ontology in Information Systems (FOIS-2001), Maine (2001).

[8] S. Ikehara, et al.: *Goi-Taikei — A Japanese Lexicon*, Iwanami Shoten, Tokyo, (1997).

[9] J. Phillips: Introduction to the Open Clip Art Library, `http://rejon.org/media/writings/ocalintro/ocal_intro_phillips.html` (2005.)

[10] K. Kuroda, et al.: Orthographic Variants and Multilingual Sense Tagging with the Japanese WordNet, 17th Annual Meeting of the Association for Natural Language Processing, A4-1, Toyohashi (2011).

[11] F. Bond, et al.: Japanese SemCor: A Sense-tagged Corpus of Japanese, Proc. 6th Global WordNet Conf. (GWC 2012), pp.56–63, Matsue, (2012).

[12] M. van Assem, A Gangemi, and G. Schreiber: RDF/OWL Representation of WordNet, W3C Working Draft 19 June 2006, `http://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/`.

[13] S. Koide, T. Morita, T. Yamaguchi, H. Muljadi, H. Takeda: OWL Expressions on WordNet and EDR, The Japanese AI Society, SIG-SWO-A601-03 (2006).

[14] S. Koide, T. Morita, T. Yamaguchi, H. Muljadi, H. Takeda: RDF/OWL Representation of WordNet 2.1 and Japanese EDR Electronic Dictionary, ISWC2006, Poster (2006).

[15] M. van Assem, A. Gangemi, G. Schreiber: Conversion of WordNet to a Standard RDF/OWL representation, LREC-2006, (2006).

[16] G. de Melo, G. Weikum: Language as a Foundation of the SemanticWeb, ISWC2008, Posters & Demos (2008).

[17] S. Koide, H. Takeda, I. Ohmukai: An LOD Approach toward WordNet Japanization, The Japanese AI Society, SIG-SWO-A1103-05 (2011).

[18] F. M. Suchanek, G. Kasneci, and G. Weikum: YAGO: A Large Ontology from Wikipedia and WordNet, Web Semantics: Science, Services and Agents on the World Wide Web, **6** (3), Pages 203–217 (2008).

[19] K. Kuroda, F. Bond and K. Torisawa: Why Wikipedia needs to make Friends with WordNet in The 5th International Conference of the Global WordNet Association (GWC-2010), pp 9–16, Mumbai (2010).