RDFization of Japanese Electronic Dictionaries and LOD

Seiji Koide
Research Organization of Information and Systems
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo
koide@nii.ac.jp

Hideaki Takeda
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo
takeda@nii.ac.jp

Abstract

This paper describes the practice and the reality of OWL conversion of Japanese WordNet and Japanese dictionary IPAdic. The outcomes of OWL conversion are linked to DBpedia Japanese dataset using lexical word matching. The difficulty originating from the specialty of Japanese, which is shareable by non-English languages, is focused. The potential of LOD in linguistics is also discussed. The goal of our study on Linguistics by LOD is to provide an open and rich environment in linguistics that propels multi-lingual studies for linguistics researchers and bottom-up style ontology buildings for ontologists.

1 Introduction

The traditional study of linguistics in Japanese is somehow domestic and not open so far to unrelated people. Linguistics by Linked Open Data (LLOD) has a potential to break this tradition and to open linguistic resources to broad researchers unlimited within linguistics. However, Japanese linguistic LOD embraces special difficulties that arise from specialties of the nature of Japanese. These difficulties are not only limited to Japanese but also common to non-English languages.

In this paper, we describe the practice and the reality of OWL conversion of Japanese WordNet and Japanese dictionary IPAdic. To make the outcomes into LOD, we linked the entities of them to DBpedia Japanese and made them accessible on WWs.

In the next section, we summarize what is LOD and address the benefit of LLOD along with the introduction of DBpedia Japanese. Our work of RDFization of Japanese WordNet and linkage to DBpedia Japanese are described in Section 3. Section 4 introduces the RDFization of IPAdic and the linkage to DBpedia Japanese. Section 5 presents the publication of our work as LOD. Related work is discussed in Section 6, and Section 7 finally gives the summary and the discussion for future work.

2 LOD and DBpedia

2.1 Linguistic LOD and Five Stars

In Linked Open Data (LOD), Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a five-star deployment scheme. In this view, there was no LOD resource for Japanese linguistics up to this study. EDR (Yokoi, 1995) by Japan Electronic Dictionary Research Center and lately NICT, GoiTaikei (Ikehara, et al., 1997) by NTT, and a Japanese corpora by National Institute for Japanese Language and Linguistics are provided in machine readable forms but not in free use. However, the property of Japanese WordNet (Isahara, et al., 2008), IPAdic/NAIST-jdic (Matsumoto, et al., 1999), and UniDic (Den, et al., 2008) is in free use.

Based on the five-star scheme for LOD, we can deduce the condition of making LOD of a domain as follows.

1. Are materials in the domain open (free in use)?
2. Is the structure of materials disclosed being sufficient for RDFization?
3. Is it possible to name the components by controllable URIs?
4. Is it possible to make linkage to other resources?

Therefore, Japanese WordNet, IPAdic/NAIST-jdic, and UniDic deserve the conversion to LOD.
RDF/OWL data format in order to let them turn data resources in LOD, namely making URIs of all components in dictionaries with controllable domain names and letting them enable to be referenced on the webs (i.e., dereferenceable). Whereby, we can enjoy Japanese linguistic resources in the new paradigm of LOD.

We propose the benefit of LLOD as follows.

- Enables the sharing of linguistic resources.
- Enables the comparison of linguistic resources among them over silos of different dictionaries in their own definitions.
- Enables the usage of linguistic resources with other non-linguistic resources (e.g., DBpedia).
- Enables the development of ontologies starting at the lexical level for multiple vocabulary sets.

2.2 DBpedia Japanese as LOD Hub

DBpedia Japanese is a database generated from Japanese Wikipedia using DBpedia Information Extraction Framework (DIEF). Although there was significant delay in the deployment of DBpedia Japanese, it was launched in 2012 by our colleagues at National Institute of Informatics (NII). Since then, all LOD resources in Japan are being linked to the DBpedia Japanese and it has become the hub of LOD-cloud in Japan as English DBpedia (Bizer, et al., 2009) is in the world. In Japan, there are currently 23 data sets linked directly or indirectly to DBpedia Japanese, which contains 77,445,359 triples, at the time of writing this paper.

3 RDFization of Japanese WordNet and Links to DBpedia Japanese

3.1 Practice of RDFization

In addition to RDF syntax and RDF semantics, we have discovered some pragmatics on RDFization in LOD. General ones over diverse domains are described in Heath and Bizer (2011). In this section, we describe more specific practices in RDFization of Japanese resources.

3.1.1 Normalization of UNICODE

As known by the popular picture of Semantic Web Layer Cake, UNICODE is the proper character encoding set of Semantic Web and LOD. However, it is not known that strings in an RDF graph should be in Normal Form C (NFC) of UNICODE. Otherwise, serious problems may happen in Japanese and other non-English languages. For example, ‘ö’ that is located in Basic Plane 0 is encoded to U+00F6 but it is also printed by octets U+006F (Latin small letter o) + U+0308 (combining dieresis). Then, we may miss string matching “Gödel” between one that consists of U+00F6 and the other that consists of U+006F + U+0308. The same thing can happen in case of Plato (Πλατων) in which ‘ά’ may be U+03AC, or the combination of U+03B1 (Greek small letter alpha) and U+0301 (combining acute accent). In Japanese, ‘δ’(U+304C) may be represented by {δ+’}, and ‘δ’(U+3077) may be represented by {δ+’}. The normalization of NFC solves this ambiguity of character strings in UNICODE.

3.1.2 Supplementary Ideographic Plane in UNICODE

Several extended kanji characters are located in Supplementary Ideographic Plane of UNICODE, which is implemented by surrogate pairs, and these extended kanji characters has been used for Japanese person names before the age of electronics. For example, ‘吉’ (U+20BB7) is very similar to basic kanji ‘吉’ (U+5409), and ‘丈’ (U+4E08) is similar to basic kanji ‘丈’ (U+4E08), but many computer systems cannot print out the extended kanji characters in Supplementary Ideographic Plane. Then, Wikipedia titles a page for a boxer to "辰吉丈一郎" instead of his proper name "辰吉丈一郎", and then guides us to the page, even if we, on top of Wikipedia, search a page with the proper name "辰吉丈一郎”. We must take care of extended kanji characters with surrogate pairs in data resources.

3.1.3 URI vs. IRI

N-Triples is a line-based, plain text format for encoding an RDF graph, but the character encoding

---

6 https://github.com/dbpedia/extraction-framework/wiki/The-DBpedia-Information-Extraction-Framework
7 http://www.w3.org/TR/rdf-mt/
9 http://www.w3.org/TR/rdf-testcases/

in string is designated to 7-bit US-ASCII. So, non-
ASCII characters must be made available by \-escape sequences, such as ‘\u3042’ for Japanese
hiragana ぁ’ (U+3042). 10

RDF/XML syntax 11 designates %-encoding for
disallowable characters that do not correspond to
permitted US-ASCII in URI encoding, in spite that the
UNICODE string as UTF-8 is designated to the
RDF/XML representation. Therefore, the dis-
page/ %E8%BE%B0%E5%90%89%E4%B8%AA%E4% B8%8E%E9%8E%BE in RDF/XML syntax.

Turtle 12 and JSON-LD 13 allow IRIs. We expect
every platform for Semantic Web and LOD can
process format files of Turtle and JSON-LD, and
then the revised edition of RDF/XML will allow
IRIs in near future.

At the end, we will be able to choose URIs if we
focus on the international usability of the data,
or IRIs if we take care of domestic understand-
ability. The RFC3986, the standard of URI, says
for the design of URI, “a URI often has to be re-
membered by people, and it is easier for people to
remember a URI when it consists of meaningful
or familiar components.” This statement can be
rephrased with replacing IRI for URI.

3.2 RDFization of English WordNet

The WordNet (Fellbaum, 1998) is a collection of
sets of synonymous words or synsets, in which
each synset, a set of synonymous words, is associ-
ated with semantic properties and values such as
hypernym, hyponym, holonym, meronym, etc.

In 2006, W3C issued W3C Working Draft on
RDF/OWL Representation of WordNet (van As-
sem, et al., 2006a), and then the authors of the
draft actually made the conversion of WordNet to
the RDF/OWL representation language for Word-
Net 2.0 (van Assem, et al., 2006b).

In the data files of English WordNet, each line
of synsets includes the synonymous words with a
sense number associated to the polysemous word
for this sense. Thus, the W3C Working Draft of
WordNet reflects this many to many relation be-
 tween synsets and polysemous words by setting
word senses.

10 Hiragana are characters that represent Japanese syllables. A syllable is composed of a consonant plus a vowel.
11 http://www.w3.org/TR/REC-rdf-syntax/
12 http://www.w3.org/TR/turtle/
13 http://www.w3.org/TR/json-ld/

After the W3C proposal for OWL conver-
sion of WordNet, the Princeton WordNet was
updated to version 2.1, in which new relations of
instanceHypernym and instanceHyponym has
been introduced, and now the latest version is
3.0. In following the updates of WordNet, the
RDF schema for WordNet 2.0 should be reused
to 2.1 and 3.0, according to one of rules for
the best practice in LOD. Only for two new
properties, wn21schema:instanceHypernymOf
and wn21schema:instanceHyponymOf should be
defined in WordNet 2.1. On the other hand,
the namespaces of every instance of words, 
word senses, and synsets may be updated to
wn21instances or wn30instances, de-
pending on the version numbers in order to dis-
tinguish the version of data, even if the content
of an entry was not updated in a new version.

3.3 RDFization of Japanese WordNet

The latest Japanese WordNet is built on top of
Princeton’s English WordNet 3.0 by adding ap-
propriate Japanese words to the content of Prince-
ton WordNet 3.0 on the framework of the Word-
Net. A polysemous Japanese word is related
to more than one English synset via Japanese
word senses as usual in the WordNet manner.
Thus, we set up the namespace for Japanese
WordNet to wnja11instances. According
to the W3C proposal for OWL conversion of
WordNet, we converted Japanese WordNet to
OWL. Here, wnja11instances:word-犬 (dog) is
made and linked to both wnja11instances:word
sense-犬-noun-1 and wnja11instances:word
sense-犬-noun-2. Furthermore, the former is
linked to wnja11instances:synset-spy-noun-1
and the latter is linked to wnja11instances:
synset-dog-noun-1. Japanese word ‘犬’ means
“dog” and “spy”, but does not mean “frump” in
English. However, because of depending on the
English WordNet framework, the Japanese vocab-
uary is not comprehensive yet, and Japanese spe-
cific concepts are still not completed.

3.4 Linking Japanese WordNet to DBpedia

Japanese
Since both English WordNet and English
Wikipedia are the most famous comprehensive
language resources, there are many studies
how the combination contributes to build better
language resources. We have also investigated
how Wikipedia Japanese can enrich Japanese
WordNet. The result of investigation suggests that it is not easy to build clean hypernym/hyponym relationship by merging two ontologies that are independently built. We think the reason is partly from inaccurate ontology buildings of the Japanese WordNet Developers, and partly from immature methodology of ontology building.

English WordNet itself includes ontological ambiguity between 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 regarded as an instance of concept central bank from the ontological view. White House as an executive department of American government is also not defined as instance of executive department but White House as residence is defined as an instance of residence. These facts suggest that English WordNet adopts some tacit knowledge of instances and classes. However, there is no explicit explanation about it, and it is not common in the community of ontology. Thus, we have no accurate and rational method on a firm foundation to merge WordNet to another ontology, whereas we have several similarity-based studies on ontology merging. They show much room for improvement. On the other hand, it is well known that DBpedia and its terms in the infoboxes are not sufficient to conceive of the infoboxes as ontology.

Therefore, we have here simply linked entities between Japanese WordNet and DBpedia Japanese not ontologically but literally, i.e., we link word noun entities of WordNet to DBpedia resources using property skos:closeMatch, where words in WordNet and resource names in Wikipedia share the same strings. Starting at the literal connection, the way of re-arranging and merging two ontologies will be studied step by step in bottom-up style, from lexicality to meaning, morphology to semantics, and linguistics to ontologies.

In linking Japanese WordNet to DBpedia Japanese, we decided to use only nouns of Japanese WordNet. One reason is that most resources in DBpedia are categorized as nouns, whereas there are categorically three types of IRIs in DBpedia, i.e., resource, property, and page of Wikipedia. Therefore, we selected resource IRIs for candidates of linking.

The other reason is to avoid needless ambiguity. Japanese verbs are categorized into several types of conjugate forms. One type verb is composed of one or more (typically two) kanji characters (root) + “する” (conjugational suffix) for positive14, e.g., “散歩する” (stroll), etc. Then, these roots are mostly nouns. It is obvious that a Japanese noun and a Japanese verb that shares morphemic root with the noun should be discriminated. However, Japanese WordNet does not distinguish them and then marks part-of-speech ‘verb’ to morphemic roots. Thus, word “散歩” is marked as noun and verb. This ambiguity will create needless links, if we link verbs in Japanese WordNet to DBpedia in addition to nouns.

Table 1 shows the statistics of linking data of Japanese WordNet to DBpedia Japanese, and Table 2 shows the statistics of linking data of DBpedia Japanese to Japanese WordNet. The lexically exact mapping produces one by one and inversely equivalent matching between both.

### 4 RDFization of IPAdic and Links to DBpedia Japanese

#### 4.1 OWL Conversion of IPAdic

In the RDFization of IPAdic 2.7.0, we encountered one typical problem in RDF, that is, the domain and range problem. Every property in RDF restricts the class of its subject and object of a given triple in a context. For instance, a property of wn20schema:sense designates an instance of wn20schema:Word for subject and an instance of wn20schema:WordSense for object, and vice versa on wn20schema:word. In the conversion of IPAdic, the adoption of properties defined in WordNet 2.0 schema will result in forcing the classification to WordNet classes on IPAdic entries. Therefore, we newly defined a schema, in which properties of IPAdic which

\[14 \text{and } + “しない” \text{ for negative} \]
are similar to WordNet but whose namespace is different from WordNet. In other words, we, instead of wn20schema:word and wn20schema:sense, defined and used ipadic27schema:word and ipadic27schema:sense, of which the domain and range are ipadic27schema:Word and ipadic27schema:WordSense.

In addition, we reflected the information of parts of speech, connection costs, lemmas, and word readings of IPAdic into the schema. In this RDFization process, we recognized that a lemma and a reading represented by katakana for a kanji word should be assigned to a sense but not the word. Thus, we defined the domain of ipadic27schema:reading as ipadic27schema:WordSense in order to reflect such Japanese sense structure in IPAdic, whereas there is no description of senses or means. We generated entities of word senses from words in order to enable the assignment of lemmas and readings to them.

4.2 Linking IPAdic to DBpedia Japanese

The outcomes of the conversion of IPAdic are linked to DBpedia Japanese with literal matching between noun words in IPAdic and resource names of DBpedia. In spite of the creation of word senses in the IPAdic, the connection of IPAdic entries as sense is suppressed, because there is no explicit evidence on senses in IPAdic for connecting to DBpedia Japanese. The connection from word senses of IPAdic to DBpedia is left as work in near future.

Table 3 shows the number of links and the rate from IPAdic to DBpedia Japanese, and Table 4 for the number of links and the rate from DBpedia Japanese to IPAdic.

Table 3: IPAdict Link Number to DBpedia-ja

<table>
<thead>
<tr>
<th>DBpedia</th>
<th># linked</th>
<th># IPAdic nouns</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>resources</td>
<td>54,735</td>
<td>197,479</td>
<td>27.7%</td>
</tr>
</tbody>
</table>

Table 4: DBpedia-ja Link Number to IPAdic

<table>
<thead>
<tr>
<th>DBpedia</th>
<th># linked</th>
<th># IRIs</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>resources</td>
<td>54,735</td>
<td>1,456,158</td>
<td>3.8%</td>
</tr>
</tbody>
</table>

5 Publishing as LOD

As a means of registration at the Data Hub, DBpedia Japanese has been published as the Japanese hub of LOD with CC-BY-SA license. It is available from our site to access the data dereferenceably, make a query at a SPARQL endpoint, and dump the zip files. This DBpedia Japanese includes the links to Japanese WordNet in lexical level.

Japanese WordNet and IPAdic have also been published under a CC-BY-SA license, same as DBpedia Japanese, from our sites. The dump files are also available at our repository.

It is critical as LOD to make all entities dereferenceable. We acquired the domain names wordnet.jp and ipadic.jp to obtain controllable domain names for Japanese WordNet and IPAdic, and then SPARQL endpoints are opened with http://wordnet.jp/ and http://ipadic.jp/ in addition of making the entries dereferenceable.

6 Related Work

As described so far in this paper, this work is the first attempt of LOD on Japanese linguistic resources. However, several studies in Semantic Webs related to dictionaries and ontologies have been completed before the advent of LOD. Koide, et al. (2006) performed OWL conversion of EDR and Princeton WordNet 2.1 according to the W3C working draft on OWL conversion. The converted files were open and down-loadable but there was no dereferenceable web site and no SPARQL endpoint, as things in the pre-LOD age.

An LOD site for words and characters in multilingualisms were open by de Melo and Weikum (2008).

YAGO (Suchanek, et al., 2008) is the first substantial study of automatic ontology construction from two comprehensive English resources, Wikipedia and WordNet. YAGO conceives of Wikipedia as knowledge about facts. Then, a semantic model like RDFS, which is closed within DBpedia (called YAGO model), is used for capturing facts in DBpedia with reifying the fact.

15Truly, we can set only classes and properties newly required, and add them to an existing set of WordNet properties, since RDF semantics allows that an instance is classified into multiple classes. However, it will be easy to cause misunderstanding and misusage by users.

16Katakana is a Japanese syllabary like hiragana but it is often used to represent loanwords and imitative words.

17http://datahub.io/


20The elemental model in Semantic Webs must be open.
Each synset of WordNet becomes a class of YAGO. The Wikipedia category hierarchy is abandoned, and only the leaves are used for the factual information extraction. The lower classes extracted from Wikipedia conceptual category are connected to higher classes extracted from WordNet. Therefore, YAGO takes care of the quality of types of individuals and there is no way to improve the ontology of WordNet. The automatic ontology construction in higher classes and the merging of multiple-ontologies that may contain inconsistency is still an open problem.

Ontology alignment is critical to obtain one united resource from two inconsistent resources with different coverages, different ontological structures, and different semantics. There are many studies on ontology alignment up to now. However, these studies show immaturity on science and methodology of ontology building. Currently, similarity of lexical texts, synonym sets, and hypernym/hyponym tree structure is only a way to merge multiple linguistic resources. Hayashi (2012) proposed a new method to compute cross-lingual semantic similarity using synonym sets.

7 Conclusion and Feature Work

In this paper, we described the practice, reality, and difficulty of RDFization on two distinct Japanese dictionaries, Japanese WordNet and IPAdic, together with the benefit of and the expectation to LLOD. In this LLOD attempt, the linkage is realized on the surface level of lexicality. The linkage between word senses of WordNet and disambiguated DBpedia resources will be studied in near future, and the connection from word senses of IPAdic to DBpedia, too.

The power of LOD resides in the nature of openness and commonality. Thus, LLOD is the nature of linguistics because of the commonality of linguistics. We believe that the outcomes of LLOD will be infrastructure in each society of countries and the international world in future.

References


