Building Linked Open Data of the Life Science Dictionary

Yasunori Yamamoto and Shoko Kawamoto

Database Center for Life Science, Bunkyo, Tokyo, Japan
{yayamamo,shoko}@dbcls.rois.ac.jp

Abstract. There is a growing need for efficient and integrated access to databases provided by diverse institutions. Using a linked data design pattern allows the diverse data on the Internet to be linked effectively and accessed efficiently by computers. In addition, providing a dictionary to translate words into another language in Resource Description Framework (RDF) is useful to cross a language barrier such as English and Japanese when we want to access datasets in multiple languages. Here, we built a Linked Open Dataset of the Life Science Dictionary (LSD) with links to DBpedia. LSD consists of various lexical resources including English-Japanese / Japanese-English dictionaries and a thesaurus using the MeSH vocabulary. The latest version of LSD contains 110 thousand English and 120 thousand Japanese terms. Since we believe that LSD is a useful language resource in the life science domain to process Japanese and English text data seamlessly, linking LSD to DBpedia enables us to find related knowledge more easily and therefore contributes to the life science research community.

Keywords. Multi-lingual linked data, Dictionary, Linked Open Data

1 Background

To link heterogeneous databases and provide users with access to them in an integrated manner, publishing datasets following the linked data design pattern [1] has increasing appeal to database developers and users. It enables us to access raw data using the World Wide Web approach such as Uniform Resource Identifier (URI) and Hypertext Transfer Protocol (HTTP). To follow that pattern, we express every datum using Resource Description Framework (RDF).

In addition, there are lots of non-English resources on the Internet, and the number of non-English RDF datasets is increasing [2]. For example, the National Diet Library (NDL) of Japan provides an RDF version of Web NDL Authorities [3]. In this situation, there are growing needs of cross language RDF resources that can play a role to link monolingual RDF data sets of different languages [2]. One example is DBpedia [4], which has made the contents of Wikipedia available in RDF. Wikipedia [5] is an open, collaboratively developed encyclopedia project, and DBpedia is the largest hub on the Linked Open Data (LOD). However, it is not necessarily reliable as a translation dictionary in a specific domain. For example, Wikipedia has 149 pages in the

category of "World Health Organization essential medicines" in English, but has only 56 in Japanese [6].

Life Science Dictionary (LSD) [7] consists of various lexical resources including English-Japanese / Japanese-English dictionaries and a thesaurus using the Medical Subject Headings (MeSH) vocabulary, the NLM controlled vocabulary thesaurus used for indexing articles for PubMed. It also contains co-occurring data that show how often a pair of terms appears in a MEDLINE entry. LSD has been edited and maintained by the LSD project since 1993. Members of this project are experts in the domain. In this situation, we built an RDF version of LSD and made a set of links from LSD words to their corresponding DBpedia titles. We aim at using this dataset as a complement to DBpedia in the life science domain.

2 Methods and Results

We used the latest version (Mar. 2011) of LSD that contains 110k English and 120k Japanese terms. Besides, we used the English titles of DBpedia version 3.7 in N-Triples (labels_en.nt.bz2). There are 8,826,375 titles.

We made links from LSD to DBpedia titles by using a series of string match methods from the simplest (exact match) to more sophisticated ones such as cosine similarity-based match. For each DBpedia title, our linking process looks for its corresponding LSD word by applying the following methods in that order: exact match, Fingerprint Key Collision (FKC) method, bi-gram FKC, tri-gram FKC [8], and cosine similarity-based match [9,10]. Once a match is found, the linking process stops working on that DBpedia title and takes a next one. This means that if the process finds an exactly literal word in LSD, it does not use any more sophisticated methods. As for the application of the cosine similarity-based match, we set its threshold relatively lower (70%) to find prospective terms broadly. On the other hand, to filter out undesirable lexically approximate matches such as interleukin-1 and interleukin-2, we made several pattern matching-based filtering rules.

As a result, we obtained 81,065 links and represented them using the skos:exactMatch predicate [11]. In addition, of 155,888 English terms in LSD that have their Japanese translations, about a half of them (79,345) have been linked. Although the number of the DBpedia entries that have links to their corresponding Japanese WikiPedia entries is 390,994, only 9,816 of them have been linked to the LSD English terms. This indicates that the coverage of DBpedia as a translation dictionary in the life science domain is very limited currently.

3 Discussions and Conclusions

We built a linked open dataset of LSD with links to DBpedia. This is a first trial and we found that some exactly matched words are not semantically identical such as single letter words. For those words, DBpedia often has entries for disambiguation, and we should take them into consideration. We have published the LSD-LOD along with its ontology and hope them to be used widely to utilize multilingual resources in

life science seamlessly. It is freely accessible through the SPARQL endpoint http://purl.jp/bio/10/lsd/sparql under the Creative Commons Attribution-NoDereivs 2.0 Generic (CC BY-ND 2.0) license.

Acknowledgements. We thank Dr. Shuji Kaneko for permitting us to release LSD under CC BY-ND 2.0. This work is funded by the Integrated Database Project, Ministry of Education, Culture, Sports, Science and Technology of Japan and National Bioscience Database Center (NBDC) of Japan Science and Technology Agency (JST).

References

1. Linked Data - Connect Distributed Data across the Web, http://linkeddata.org/
2. Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gomez-Perez, A., Buitelaar, P., McCrae, J.: Challenges for the multilingual Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web. 11, 63–71 (2012)
3. Web NDL Authorities, http://id.ndl.go.jp/auth/ndla/
4. Lehmann, J., Bizer, C., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia—a crystallization point for the web of data. Journal of Web Semantics, 7 (3), 154–165 (2009)
5. Wikipedia, http://wikipedia.org/
6. http://en.wikipedia.org/wiki/Category:World_Health_Organization_essential_medicines
7. Kawamoto, T., Ohtake, H., Fujita, N., Takekoshi, Ugawa, H., Takeuchi, H., Kaneko, S.: Life Science Dictionary: statistical and collocational analyses of life science English. 20th IUBMB International Congress of Biochemistry and Molecular Biology and 11th FAOBMB Congress, Kyoto (2006)
8. Clustering In Depth, http://code.google.com/p/google-refine/wiki/ClusteringInDepth
9. Cohen, W. W., Ravikumar, P., Fienberg, S. E.: A comparison of string distance metrics for name-matching tasks. IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03), 73-78 (2003).
10. Okazaki, N., Tsujii, J.: Simple and Efficient Algorithm for Approximate Dictionary Matching. Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China, 851 - 859 (2010)
11. Miles ,A., Bechhofer, S.: SKOS-Simple Knowledge Organization System Reference, W3C Recommendation (=2009)