Multilingual Linking Open Data

Key-Sun Choi, Eun-Kyung Kim, Dong-Hyun Choi

Web Science and Technology Division, Computer Science Department, KAIST, Korea

Ontology Population and Enrichment for Semantic Evolution of LOD

Linking Open data is a syntactic linking of datasets, but it cannot acknowledge whether linked objects are really the same, or the other equivalent objects cannot be linked simply because they have different strings for labeling the objects.

Ontology population is to link the instances in the real world to the already-made ontology, and enrichment means to mine the relations between those instances, which is defined in the ontology schema.

It is not easy to view the whole LOD space to be one set of ontology and even will make controversy about why it is necessary for their use in real application. But almost all of LOD is DBpedia-centric linked structure through their exact match of naming for URI, and always there are problems such that the redundancy of presence of objects cause the missing to aggregate all of relevant data, the non-provability of semantic interoperability/equivalency for linking and even among non-linked objects will cause the inconsistency of the whole space of linked open data. This would like to pursuit the elegant ecosystem to maintain the LOD semantics by incorporating the ontology population and enrichment mechanism to the evolving LOD and even incrementally converging status of LOD to LOD2, so-called semantically stable LOD.

Basic strategy is as follows: it should be recognized that DBpedia and its mother Wikipedia should be a backbone of LOD and its following evolving semantic version and finally their ontology structure. The first step is to make a taxonomy backbone based on Wikipedia structure. Then it will be changed to ontology structured with their bootstrapping of semantic annotation and infobox-based template structure. The second step is an ontology enrichment using the Wikipedia text in order to strengthen its ontology structure. The third ontology population is to accept each dataset as one user ontology of local aggregate of LOD. After transforming each local aggregate to a possible ontology structure with taxonomy, it will be populated from Wikipedia-based ontology. The fourth step is to enrich the local LOD-ontology as well as feedback to the Wikipedia-based ontology for the overall stability of semantic LOD.

Multi-lingual Synchronization

It is necessary to explore an automated approach to synchronize multilingual-linked articles in Wikipedia. Two linked articles in two different languages have different amount of information in Wikipedia. Even some entries in one language have no entry in the other language. Synchronization means translation between two languages plus information balancing for their equivalent degree of information. Even if the full translations between two languages' linked articles are successfully accomplished, the resultant translated content to the other language should be readjusted with the already existing article in that language. Even they have obsolete information in one language. So the next stage is to find the temporal ordering of facts and events that are represented in sentences of article, and to generate the up-to-date information-synchronized articles in two languages.

There are many challenges for massive translation among multilingual-linked articles with synchronizing the information content. This approach will envision the role of infobox in each article of Wikipedia in order to improve the quality of infobox keeping their data categories in a semantically interoperable way. Although the current shape of infoboxes are governed by "templates", their templates are not enough organized to give a standardized concise guide to put the article content to its infobox. One solution is to give an ontology structure for templates of infoboxes and provide a multilingual standard for their data categories. We have a proposal called OntoCloud. If it is set up, the next steps are to find the indirect translation through the infobox pivot among different languages. Here we need machine learning between Wikipedia article's sentences and their infobox. For their alignment between each infobox row and sentences, we will try to find the relevant information extraction from Wikipedia article to its infobox as well as the methods to generate relevant sentences from infobox table. Here is a common minimal exchange of information through infobox based on already-made templates. Of course, one Wikipedia articles can require several templates of infoboxes. Multilingual word nets are also crucial resources to link them.

Ontology structure based on templates of infobox has relationship with category structure in Wikipedia. But because the categories in Wikipedia is user-given tags, their category structure are not the one of ontology and even more their taxonomy is far beyond what we expect in WordNet. We will investigate the inter-feedback between the template ontology and the category structure for their persistent eco-system that can be evolved with keeping their collective intelligence advantage.

Section 3. Issues

Issues raised are summarized as follows:

1. Is it possible and necessary to construct a backbone ontology for LOD?
2. Is it also good to align each dataset of LOD to the backbone ontology?
3. Infobox Schema Management:
    a. ontology structure for templates of infoboxes
    b. Multilingual standard for data categories in templates
4. Multilingual Infobox alignment:
    a. Fining semantic correspondences between multilingual contents
    b. Multilingual thesauri are used to link between different lingual sources
5. Infobox Population: filling the missing and unknown contents
    a. Extracting relevant information from other sources such as Wikipedia articles or external resources
    b. Generating up-to-date information-synchronization in multilingual environment