An Approach for Supplementing the Korean Wikipedia based on DBpedia

Eun-kyung Kim, DongHyun Choi, Jihye Lee, JinHyun Ahn, and Key-Sun Choi

Semantic Web Research Center, CS Department, KAIST, Korea, 305-701
{kekeeo, cdh4696, jhlee20, jhahn, kschoi}@world.kaist.ac.kr

Abstract. In this paper, we try to supplement an information-poor language knowledge base, Korean Wikipedia, to help effectively enrich information written in different languages. We propose an approach for transLating infoboxes that would enable complementing Wikipedia from English to Korean.

1 Introduction

Wikipedia is a Web-based, free-content encyclopedia community which has grown rapidly into one of the largest reference web sites. Wikipedia is a multilingual project which has more than 14,000,000 articles in more than 260 languages[1]. However, Wikipedia still lacks sufficient support for non-Latin languages. For example, English Wikipedia currently contains 3,227,911 articles and Korean Wikipedia contains only 130,629 articles. In addition, smaller languages can not produce articles as fast as larger Wikipedias such as English or German, because the number of editors and users is too low. Due to the differences in the number of articles between English and non-Latin languages in Wikipedia there need to be a supplementation across them automatically. The key features of this approach are two-fold: (1) to translate English infoboxes into Korean infoboxes is an essential first step toward a supplementation system. (2) to construct an Ontology schema based on infoboxes that could improve the generating new templates.

2 Korean DBpedia/Wikipedia Supplementation using Translation and Ontology

Most Wikipedia pages contain an infobox which is the most relevant information for a given concept. We have mainly focused on the translating infoboxes. The translation is often useful to spread information between closely related articles in different languages. The dictionary based translation is easy to set up and just requires access to a bilingual resource. We use bilingual word-pairs which are originally created for English-to-Korean translation through interlanguage-links[1].

[1] http://stats.wikimedia.org/

DBpedia[2] is a community which harvests the information of infoboxes. The infobox extraction algorithm detects such templates and recognizes their structure and saves it in RDF triples. We execute the translation from English DBpedia to Korean. A comparison of datasets as follows:

- English Triples in DBpedia: 43,974,018
- Korean Dataset (Existing Triples/Translated Triples): 354,867/12,915,169

We can get translated Korean triples over 30 times larger than existing Korean triples. However, large amount of translated triples have no predefined templates in Korean. There may be a need to form a template schema to organize the fine-grained template structure.

Thus we have built the template ontology, OntoCloud [2], from DBpedia and Wikipedia, which was released on Sept, 2009, to efficiently build the template structure. It consists of the following steps: (1) extracting templates of DBpedia as concepts in an ontology, for example, the *Template:Infobox_**Person*** (2) extracting attributes of these templates, for example, ***name*** of *Person.* These attributes are mapped to properties in ontology. (3) constructing the concept hierarchy by set inclusion of attributes, for example, *Book* is a subclass of *Book_series.* Because all attributes of *Book_series* belong *Book* class as follows:

- **Book_series** = {name, title_orig, translator, image, image_caption, author, illustrator, cover_artist, country, language, genre, publisher, media_type, pub_date, english_pub_date, preceded_by, followed_by}.
- **Book** = {name, title_orig, translator, image, image_caption, author, illustrator, cover_artist, country, language, genre, publisher, media_type, pub_date, english_pub_date, preceded_by, followed_by, **pages, isbn, oclc, dewey, congress**}.

This means that "Book_series" is more generalized concept.

The ontology building process is useful for effectively align similar types of templates can be grouped into classes, for example, the *Template:infobox_baseball _player* and *Template:infobox_asian_baseball_player* describe baseball player. Moreover different format of properties with same meaning can be normalized, for example, '*birth_place*', '*birthplace and age*' and '*place birth*' are mapped to '*birthPlace*'. Today, OntoCloud v0.2 has 1,927 classes, 74 object properties and 101 data properties.

As future work, we will consolidate the ontology schema, and then generate new articles using OntoCloud and triples at infoboxes. These automated articles is based on infobox, so it can be treated as a summary.

References

1. E. Adar, M. Skinner, and D. S. Weld, Information arbitrage across multi-lingual Wikipedia. ACM, 2009, pp. 94-103
2. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, DBpedia: A Nucleus for a Web of Open Data. ISWC+ASWC 2007, November 2008, pp. 722-735

[2] http://swrc.kaist.ac.kr/OntoCloud