Metadata Synchronization between Bilingual Resources: Case Study in Wikipedia

Eun-kyung Kim
Korea Advanced Institute of Science and Technology
Yuseong-gu, Guseong-dong
Daejeon, Republic of Korea
kekeeo@world.kaist.ac.kr

Matthias Weidl
Universität Leipzig
Department of Computer Science
Johannisgasse 26,
D-04103 Leipzig, Germany
mam07jct@studserv.uni-leipzig.de

Key-Sun Choi
Korea Advanced Institute of Science and Technology
Yuseong-gu, Guseong-dong
Daejeon, Republic of Korea
kschoi@world.kaist.ac.kr

ABSTRACT

In this paper, we present a conceptual study aimed at understanding the impact of international resource synchronization in Wikipedia and DBpedia. In the absence of any information synchronization, each country would construct its own datasets and manage it from its users. Moreover the cooperation across the various countries is adversely affected. The solution is based on the analysis of Wikipedia infobox templates and on experimentation such as term translation.

Categories and Subject Descriptors

H.3.m [**Information Systems**]: Miscellaneous

General Terms

Experimentation

Keywords

Semantic Web, Multilingual, Wikipedia, DBpedia

1. INTRODUCTION

Wikipedia is an international project which is a web-based, free-content encyclopedia and is written collaboratively. Wikipedia has made tremendous effect in the web. It has grown rapidly into one of the largest reference web sites. The main advantage of using Wikipedia is its wide coverage of concepts and languages. Wikipedia currently comprises more than 260 languages. However, Wikipedia still lacks sufficient support for non-English languages. The 22% articles in all Wikipedias belong to the English language version. Although English has been accepted as a global standard to exchange information between different countries, companies and people, the majority of users are attracted by projects and web sites if the information is available in their native language as well. One can also assume that the proportion of users on the Internet who do not speak English will continue to rise.

Due to the differences in the amount of information between English and non-English languages in Wikipedia, there needs to not only avoid information loss, but also enrich informations.

WWW2010, April 26-30, 2010, Raleigh, North Carolina.

In this paper we presents our efforts to create a multilingual system for translation-based information synchronization. Our work is based on idea that Wikipedia as a multilingual corpus and especially translation results between different editions provide valuable multilingual resources to the web. To explore this problem, we developed *Metadata Synchronization*, a platform for translation-based data synchronization between English Wikipedia and Korean Wikipedia. It aims to translate infoboxes from the English Wikipedia into Korean and insert it into the Korean Wikipedia. Because Wikipedia offers a number of structural elements, in particular, the infobox template is used to express structured information about a condensed set of important facts relating to the article[10]. The infobox is manually created by authors that create or edit an article. As a result, many articles have no infoboxes and other articles contain infoboxes which are not complete. Moreover, even the interlanguage linked articles do not use the same infobox template or contain different amount of information. Interlanguage links are links from any page describing an entity in one Wikipedia language to a page describing the same subject in another language. This problem raises an important issue about multi-lingual access on the Web.

The rest of this paper is organized as follows: In section 2, we describe related work. The framework and details of the proposed approach are given in section 3. Section 4 and 5 discuss the experimentation and results. At the end, we summarize the obtained results and point our future work.

2. RELATED WORK

Wikipedia represents a valuable source of knowledge to extract semantic information between concepts. [9] focuses on research that extracts and makes use of the concepts, relations, facts and descriptions found in Wikipedia, and organizes the work into four broad categories: applying Wikipedia to natural language processing; using it to facilitate information retrieval and information extraction; and as a resource for ontology building.

In [5], the authors suggests a methodology that semantic information can be extracted from Wikipedia by analyzing the links between categories. They try to provide a semantic schema for Wikipedia which could improve its search capabilities and provide contributors with meaningful suggestions for editing the Wikipedia pages.

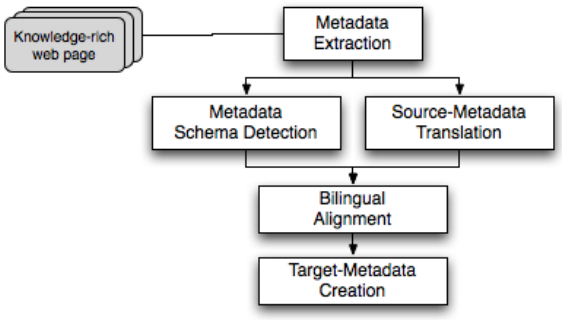

Figure 1: The workflow of the proposed metadata synchronization approach.

DBpedia[1][2] is a large, on-going, project which concentrates on the task of converting Wikipedia content into structured knowledge, and make it usable for the Semantic Web. An important component of DBpedia is harvesting of the information present in infoboxes. The infobox extraction algorithm detects such templates and recognizes their structure and saves it in RDF triples. DBpedia reached a high-quality of the extracted information and offers datasets of the extracted Wikipedia in 91 different languages. However, DBpedia still lacks sufficient support for non-English languages. First, DBpedia only extracts data from non-English articles that have an inter-language link to an English article. Therefore all other articles cannot be queried by DBpedia. Another reason is that the Live Extraction Server [7] only supports the timely extraction of the English Wikipedia. Due to the differences in the number of resources between English and non-English languages in DBpedia, there needs to be a synchronization among them.

[4] presents a method for cross-lingual alignment of template and infobox attributes in Wikipedia. The alignment is used to add and complete templates and infoboxes in one language with information derived from Wikipedia in another language.

In the area of cross-language data fusion, another project has been launched[6]. The goal is to extract infobox data from multiple Wikipedia editions and fusing the extracted data among editions. To increase the quality of articles, missing data in one edition will be complemented by data from other editions. If a value exists more than once, the property which is most likely correct will be selected.

3. FRAMEWORK OF METADATA SYNCHRONIZATION APPROACH

Our metadata synchronization algorithm consists of two consecutive steps: a metadata extraction step and a metadata translation step. Figure 1 illustrates the workflow of our approach. In order to illustrate our approach, we examine a case study in Wikipedia by using the Infobox as metadata. Wikipedia is considered to be one of the most successful collaborative editing communities on the web, we consider it as an interesting example to discuss multilingual synchronization.

In particular, Web pages pose main problems to multilingual resource:

[1]http://dbpedia.org/

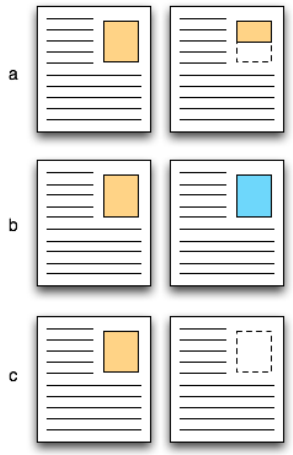

Figure 2: The form of the three template-pairs between different languages' infoboxes. (a) describes S-group, (b) describes D-group and (c) describes M-group.

- Because the number of pages in the non-English ares smaller than the English community.
- Because languages differ in grammar as well, it is obvious that every language uses its own kind of formats and templates.
- The majority of users are attracted by web sites if the information is available in their native language as well.

To explore this problem, we try to synchronize between the knowledge-poor web pages (The Korean Wikipedia: it consists of about 130,000 articles at March 2010.) and the knowledge-rich web pages (The English Wikipedia: 3.24M articles at March 2010) using translation.

At first, we extract infoboxes both from Wikipedia dump[2] (for Korean infoboxes) and DBpedia (for English infoboxes), and then use interlanguage links to determine which infoboxes are equivalent in both languages. Interlanguage links are links from an article in one Wikipedia language to the same subject in other Wikipedia languages [1]. For example, the article *Banana* in the English language Wikipedia may have interlanguage links to the article 바나나 in the Korean language Wikipedia, *Banane* in the German language Wikipedia, and so on.

After extracting infobox template in both languages, we retrieve the template-pairs between two languages. At the same time, we also extract schema of infoboxes. This schema reflects which properties is the most relevant properties for each infobox. There are some articles have more than one infobox template, we did not deal with this case in this paper.

There are several types of imbalances of infobox information between two different languages. According to the presence of infobox, we classified interlanguages linked pairs of articles into three groups: Short Infobox (S), Distant Infobox (D), and Missing Infobox (M):

[2]http://download.wikimedia.org/

- The ***S-group*** contains pairs of articles which use the same infobox template but have a different amount of information. For example, an English-written article and a non-English-written article, which have an interlanguage link and use the same infobox template, but have a different amount of template attributes.

- The ***D-group*** contains pairs of articles which use different infobox templates. D-group emerges due to the different degrees of each Wikipedia communities' activities. In communities where many editors lively participate, template categories and formats are more well-defined and more fine-grained. For example, *philosopher, politician, officeholder* and *military person* templates in English are matched just *person* template in Korean. It appears not only a Korean Wikipedia but also non-English Wikipedias.

- The ***M-group*** contains pairs of articles where an infobox exists on only one side.

The forms of each types are shown in Figure 2.

We also refined template schema that is necessary for several reasons. The first reason is that many templates use the abbreviation for its name, for example, `SW books`, `NRHP` and `MLB player`. In the cases of `SW books` and `NRHP`, these stand for `Star Wars books` and `National Register of Historic Places` respectively. However, a `SW` is widely used for an abbreviation of a `software` and `NRHP` is used only in Wikipedia. Thus it is difficult to understand.

The second reason is that the template name does not have same meaning as the common sense in a dictionary. For example, according to the Wiktionary[11] `television` means 'An electronic communication medium that allows the transmission of real-time visual images, and often sound'. However, for example, the `Template:Infobox television` represents television program. Its attributes are shown in below:

- **Infobox_television**={show_name, image, caption, genre, format, creator, developer, writer, director, producer, country, opentheme, ...}

Moreover different format of properties with same meaning should be refined. In order to overcome the problems of multiple property names being used for the same type, for example, 'birth place', 'birthplace and age' and 'place birth' are mapped to 'birthPlace'. In addition, the 'img' property is replaced into the 'image', because the later form is better to translate. If this is solved, it will be a great help for organizing templates and the efficiency of our synchronization framework will be higher.

We extracted English triples in infoboxes are simply translated into Korean triples. For the dictionary-based term translation [8], we use bilingual dictionary which is originally created for Englsih-to-Korean translations from Wikipedia interlanguage links [1], with our pre-constructed bilingual resources from SWRC [3]. The dictionary based translation is easy to set up. It just requires access to a bilingual dictionary. However, using multiple translation entries from a dictionary can generate large amounts of ambiguity which is a classic problem of translation. Another problem comes from the uncollected names. Such as proper names, uncollected names are still not correctly translated using the dictionary-based translation. After we constructed the translation resource, we added several translation patterns for multi-terms. Multi-terms are set of single terms such as *Computer Science*. We extracted 11 translation patterns(9 patterns for syntactic solving and 2 patterns for Date) using bilingual alignment of collocations (See Table 1). It uses the simple method of word co-occurrence in two or more aligned sentences across languages. We used the English-Korean multilingual corpus from SWRC. It consists of 60,000 pairs of sentences. We can choose from any of those aligned parallel sentences and determine the relative word order in the collocation [3].

[3]Semantic Web Research Center http://swrc.kaist.ac.kr

Table 1: Syntactic Translation Patterns between English and Korean multi-terms

English	Korean
A B	A B
A and B	A B
A or B	A B
A, B and C	A B C
the A	A
A of B	B A
A in B	B 에서 A
A from B	B 에서 A
A on B	B 의 A

After the translation step, we try to align the bilingual metadata. For the alignment, we manually constructed the mapping table between Korean and English template schemas based on translation results.

The details of this method are described in the following.

4. EXPERIMENTATION

We introduce the details of experimental dataset and the processing of the data.

4.1 Metadata Extraction

We need two types of data sets for experimentation. One type is the Wikipedia dump and the other type of data is the DBpedia dataset. In this paper, we only use Korean and English data, but our approach can be applied on any language of data in Wikipedia. We extracted 37,576 Korean articles contain infobox using Wikipedia dumps at March, 2010. Once the data is ready, all articles are parsed to extract attribute-value pairs of infobox. We got 1,042 infobox templates in Korean. We got 2,792 infobox templates in English using DBpedia dataset, and 1,642 template-pairs. It is noted lots of templates are duplicate uses. Thus, the number of template-pairs is much bigger than the number of templates in Korean. However, there are many articles having infobox without the name of template, it is a hurdle in the progress and development of the infobox extraction.

4.2 Metadata Translation

We executed the translation from English triples in infoboxes to Korean triples. In our experiments, we used DBpedia 3.4 as resource for the translation, a comparison of datasets is 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, its quality is still quite poor. Because lots of triples include sentences or phrases such as below:

TRIPLE 1. *"%21%21%21", "currentMembers", "**Nic Offer Allan Wilson Mario Andreoni Tyler Pope Dan Gorman Sean McGahan Shannon Funchess Paul Quattrone**"*

TRIPLE 2. *"14226_Hamura", "discoverer", "**Lincoln Laboratory Near-Earth Asteroid Research Team**"*

Also, a large amount of translated triples consists of only numbers. Our translation approach is compared to a statistical machine translation system by Google Translate API. The Google Translate API is an initial prototype used a statistical MT system based on Moses and trained on Europarl. Overall, the Google Translate API performed better and executed faster than our dictionary based translation. However, in case of proper nouns and abbreviations, our approach yields slightly higher accuracy than Google Translate API.

5. DISCUSSIONS

Today, we created the metadata using translated results. However, we did not check the consistency between existing things and new things. To solve this problem, we try to utilize this consistency management to construct a large and fine-grained ontology by using infobox template. Thus we have built a template ontology, OntoCloud[4], from DBpedia and Wikipedia to efficiently build the template structure.

The construction of OntoCloud 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*. For the ontology building, similar types of templates are mapped to one concept. For example, the *Template:infobox baseball player* and *Template:infobox asian baseball player* describe *baseball player*. Using the OntoCloud, we could be check the consistency of templates.

6. CONCLUSIONS

As the web grows in number of pages and amount of information, there is an increasing interest towards supporting tasks such as organizing, and enriching. We have proposed a novel idea on using term translation not only to synchronize but also to enrich information of multilingual web resources, and presented an effective approach to implement this idea in Wikipedia. Our work is ongoing for technical improvements, such as better alignment between bilingual metadata, and more precise translating. After the verification of template consistency, the Korean Wikipedia and DBpedia can be updated automatically. This will be helpful to guarantee that the same information can be recognized in different languages. Moreover, will be helpful to edit articles and to create infoboxes when a new article is created. It can support the authors by suggesting the right template. As future work, it is planned to support more standardization for the Korean language and improve the quality of the translated datasets.

[4]http://swrc.kaist.ac.kr/ontocloud/

7. REFERENCES

[1] E. Adar, M. Skinner, and D. S. Weld. Information arbitrage across multi-lingual wikipedia. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 94–103, New York, NY, USA, 2009. ACM.

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007)*, pages 722–735. November 2008.

[3] R. Basili, D. De Cao, D. Croce, B. Coppola, and A. Moschitti. Cross-language frame semantics transfer in bilingual corpora. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, chapter 27, pages 332–345. 2009.

[4] G. Bouma, S. Duarte, and Z. Islam. Cross-lingual alignment and completion of wikipedia templates. In *CLIAWS3 '09: Proceedings of the Third International Workshop on Cross Lingual Information Access*, pages 21–29, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[5] S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou. Extracting semantic relationships between wikipedia categories. In *In 1st International Workshop: "SemWiki2006 - From Wiki to Semantics" (SemWiki 2006), co-located with the ESWC2006 in Budva*, 2006.

[6] C. B. Eugenio Tacchini, Andreas Schultz. Experiments with wikipedia cross-language data fusion. In *5th Workshop on Scripting and Development for the 5th Workshop on Scripting and Development for the Semantic Web (SFSW2009)*, 2009.

[7] S. Hellmann, C. Stadler, J. Lehmann, and S. Auer. Dbpedia live extraction. In *Proc. of 8th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE)*, volume 5871 of *Lecture Notes in Computer Science*, pages 1209–1223, 2009.

[8] O. Levow. Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management*, 41:523–547, 2005.

[9] O. Medelyan, D. Milne, C. Legg, and I. H. Witten. Mining meaning from wikipedia. May 2009.

[10] F. Wu and D. S. Weld. Automatically refining the wikipedia infobox ontology. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 635–644, New York, NY, USA, 2008. ACM.

[11] T. Zesch, C. Müller, and I. Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proc. of the 6th Conference on Language Resources and Evaluation (LREC)*, 2008.