Cross-lingual ontology matching as a challenge for the Multilingual Semantic Web[1]

Jorge Gracia
Ontology Engineering Group
Universidad Politécnica de Madrid

Motivation

Recently, the Semantic Web has experienced significant advancements in standards and techniques, as well as in the amount of semantic information available online. Nevertheless, mechanisms are still needed to automatically reconcile information when it is expressed in different natural languages on the Web of Data, in order to improve the access to semantic information across language barriers. In this context several challenges arise [1], such as: (i) ontology translation/localization, (ii) cross-lingual ontology mappings, (iii) representation of multilingual lexical information, and (iv) cross-lingual access and querying of linked data.

In the following we will focus on the second challenge, which is *the necessity of establishing, representing and storing cross-lingual links among semantic information on the Web*. In fact, in a "truly" multilingual Semantic Web, semantic data with lexical representations in one natural language would be mapped to equivalent or related information in other languages, thus making navigation across multilingual information possible for software agents.

Dimensions of the problem

The issue of cross-lingual ontology matching can be explored across several dimensions

1. Cross-lingual mappings can be established at *different knowledge representation levels*, each of them requiring their own mapping discovery/representation methods and techniques:
   i. conceptual level (links are established between ontology entities at the schema level),
   ii. instance level (links are established between data underlying ontologies), and
   iii. linguistic level (links are established between lexical representations of ontology concepts or instances).
2. Cross-lingual mappings can be discovered *runtime/offline*. Owing to the growing size and dynamic nature of the Web, it is unrealistic to conceive a Semantic Web in which all possible cross-lingual mappings are established beforehand. Thus, scalable techniques to dynamically discover cross-lingual mappings on demand of semantic applications have to be investigated. Nevertheless, one can imagine some application scenarios (in restricted domains for a restricted number of languages) in which computation and storage of mappings for later reuse is a viable option. In that case, suitable ways of storing and representing cross-lingual mappings become crucial. Also mappings computed runtime could be stored and made available online, thus configuring a sort of pool of cross-lingual mappings that grows with time. Such online

[1] This work was supported by the Spanish Ministry of Economy and Competitiveness within the Juan de la Cierva program and the Spanish Project BabelData (TIN2010-17550).

mappings should follow the linked data principles to favour their later access and reuse by other applications.

3. Cross-lingual links can be discovered either by *projecting* the lexical content of the mapped ontologies into a *common language* (either one of the languages of the aligned ontologies or a pivot language) e.g., using machine translation, or by *comparing the different languages directly* by means of cross-lingual semantic measures (e.g., cross-lingual explicit semantic analysis [2]). Both avenues have to be explored, compared, and possibly combined.

What is needed?

In summary, research has to be done in different aspects:

- *Cross-lingual ontology matching*. Current ontology matching techniques could be extended with multilingual capabilities, and novel techniques should be investigated as well.
- *Multilingual semantic measures*. Such novel cross-lingual ontology matching techniques above mentioned have to be grounded on measures capable of evaluating similarity or relatedness between (ontology) entities documented in different natural languages.
- *Scalability of matching techniques*. Although the scalability requirement is not inherent to the multilingual dimension in ontology matching, multilingualism exacerbates the problem due to the introduction of a higher heterogeneity degree and the possible explosion of compared language pairs.
- *Cross-lingual mapping representation*. Do current techniques for representing lexical content and ontology alignments suffice to cover multilingualism? Novel ontology lexica representation models [3] have to be explored for this task.

References

[1] J. Gracia, E. M. Ponsoda, P. Cimiano, A. G. Pérez, P. Buitelaar, and J. McCrae, "Challenges for the multilingual Web of Data," Journal of Web Semantics, vol. 11, pp. 63-71, Mar. 2012. Available at http://oa.upm.es/8848/1/Multiling.pdf

[2] P. Sorg and P. Cimiano, "Exploiting wikipedia for cross-lingual and multilingual information retrieval," Data & Knowledge Engineering, vol. 74, pp. 26-45, Apr. 2012. Available at http://dx.doi.org/10.1016/j.datak.2012.02.003

[3] J. McCrae, G. A. de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner, "Interchanging lexical resources on the semantic web," Language Resources and Evaluation, vol. 46, 2012. Available at http://dx.doi.org/10.1007/s10579-012-9182-3