Cross-lingual Linking on the Multilingual Web of Data (position statement)

Jorge Gracia, Elena Montiel-Ponsoda, and Asunción Gómez-Pérez

Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
{jgracia, emontiel, asun}@fi.upm.es

Abstract. Recently, the Semantic Web has experienced significant advancements in standards and techniques, as well as in the amount of semantic information available online. Even so, mechanisms are still needed to automatically reconcile semantic information when it is expressed in different natural languages, so that access to Web information across language barriers can be improved. That requires developing techniques for discovering and representing cross-lingual links on the Web of Data. In this paper we explore the different dimensions of such a problem and reflect on possible avenues of research on that topic.

Keywords: multilingualism, ontology matching, multilingual linked data, multilingual mappings

1 Motivation

The large and growing amount of semantic data available on the Web, mainly in the form of Linked Data [2], online ontologies, and annotated Web pages, has resulted in the emergence of the so-called Web of Data. This fact has been accompanied by significant advancements in standards and techniques, contributing to the realization of the Semantic Web vision [1]. Some issues, however, need to be solved before a fully realised Semantic Web can be achieved, as for instance, language barriers, amongst others. In this sense, mechanisms are still needed to automatically reconcile semantic data (ontologies and data underlying ontologies) when they are expressed in different natural languages on the Web, in order to enable access to semantic information across language barriers. To this respect, several challenges arise [5], specifically: (i) ontology translation/localization, (ii) cross-lingual ontology linking, (iii) representation of multilingual lexical information, and (iv) cross-lingual access and querying of linked data.

In this paper we focus on the second challenge, namely, the need of establishing, representing, and storing cross-lingual links among semantic information on the Web. In fact, in the multilingual Web of Data that we envision, 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. In the following we will refer to "cross-lingual ontology linking" in a broad sense, including (semi-) automatic ontology and instance matching methods and techniques applied to the linking of semantic data documented in several natural languages.

The problem of cross-lingual linking is a fundamental one, since more and more legacy data sources available in different natural languages are being transformed into linked data, and have to be linked to be exploited at its full potential. In fact, the establishment of links between or among multilingual data sources would also contribute to the localisation issue, since it would transform monolingual, isolated, data resources into "multilingual resources" just thanks to the links. However, the linking of resources documented in different languages is not so immediate. Several issues that arise in the localization of semantic web resources [4] would be also involved in the liking task, namely, a) conceptualization mismatches due to language and cultural discrepancies; b) conceptualization mismatches due to the perspectives from which the same domain is approached; or even c) different levels of granularity in the conceptualization.

The main purpose of this position paper is to give an insight into the problem of cross-lingual linking on the Web of Data and identify some research topics that will allow us to advance towards a truly multilingual Web of Data. In the rest of the paper (Section 2) we refer to the different knowledge representation levels in which cross-lingual links can be established. Then, we explore the problem and identify possible research lines grouped in three aspects: cross-lingual link discovery, representation, and reuse. Finally, the main conclusions of the paper are summed up in Section 3.

2 Dimensions of the problem and research lines

Cross-lingual links between ontologies and data sources can be established at different knowledge representation levels:

1. Conceptual level: links between ontology entities at the schema level.
2. Instance level: links between data underlying ontologies.
3. Linguistic level: links between lexical representations associated with ontology concepts and/or instances.

The last one is particularly important if certain lexical relations have to be represented across ontologies (e.g., translations or term variations). Each of these levels will require its own link discovery/representation methods and techniques.

In the following we propose some enhancements of available methods and techniques and suggest new avenues of research that could help overcome the problem.

2.1 Cross-lingual Link Discovery

Current ontology matching techniques have to be extended with multilingual capabilities, and novel techniques need to be investigated as well. Cross-lingual links can be discovered by means of some of these techniques:

1. 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.

2. Comparing the ontology entities directly by means of cross-lingual semantic measures, that is, measures capable of evaluating similarity or relatedness between (ontology) entities documented in different natural languages (e.g., cross-lingual explicit semantic analysis [9]).

Both avenues have to be further explored, compared, and possibly combined. There are a number of early cross-lingual ontology alignment tools that already implement the first technique[1], while the second one remains unexplored yet. Notice that such preliminary systems are intended to discover cross-lingual links at the conceptual level and that cross-lingual alignment systems operating at the instance and linguistic levels are still to come.

An alternate way to discover cross-lingual links is by using the Web of Data as a source of background knowledge. The idea is to infer links from other links already existent among online ontology entities (that are similar to the entities I intend to link). Such an approach was explored in a monolingual context by the Scarlet system [8] and could be extrapolated to a multilingual landscape.

2.2 Cross-lingual Link Representation

In principle, existing constructs of ontology languages can be utilised for representing cross-lingual mappings at the conceptual and instance levels (e.g., owl:sameAs or owl:equivalentClass), whenever the two concepts or instances can be considered cross-lingual equivalents.

Other commonly used vocabularies (e.g. rdfs:subclassOf, skos:narrower or skos:broarder) could also be re-used in case of granularity discrepancies, i.e., when one conceptualization regards a certain concept with a granularity level different from the other conceptualization. In this case, we would suggest an adaptation or enhancement of such relations for a multilingual scenario, so that finer language distinctions are captured.

In the case no equivalence exists (the one language does not conceptualize a certain phenomenon of the world, whereas the other has a concept for it), we could still provide a lexical description for the "inexistent concept" in the target language, provide a link to its closest concept, and signalize it as a specific cross-lingual case. We believe this kind of links should also be accounted for in the Web of Data.

Regarding cross-lingual mappings at the linguistic level, mappings could be established between the natural language descriptions of their concepts. At this level, lexical-semantic relations could be used (hypernym-hyponym, synonym, antonym, translation, etc.). In the simplest case in a cross-lingual scenario, a property labelled "translation" or "cultural equivalent" (for instance) might be established between the lexical realizations of the concepts [7]. Novel ontology lexica representation models [6] have to be explored for this task.

We argue that specific representation models have to be able to define specific relations between natural language descriptions in different languages, what

[1] See for instance the systems that participated in OAEI2011.5 http://oaei.ontologymatching.org/2011.5/multifarm/index.html

we term translation relations or cross-lingual relations. Highly related with this issue is the representation of term variation at a monolingual or multilingual level. A term variant has been defined as "an utterance which is semantically and conceptually related to an original term" [3]. To put it in simple words, we could define them as synonymous terms that refer to the same concept but that highlight a different aspect. We believe that the accounting for and representing term variants would also contribute to the automatic linking of the lexical descriptions associated to concepts (within or across languages).

Further, to facilitate processing and interchange of alignments, specific formats has been proposed in the literature such as the Alignment Format [2] or the EDOAL language [3]. They should be explored and, if needed, extended to accommodate the representation of cross-lingual and multilingual alignments.

2.3 Cross-lingual Link Storage and Reuse

Cross-lingual links can be discovered runtime/offline. However, 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 links are established beforehand. Thus, scalable techniques to dynamically discover cross-lingual links on demand of semantic applications have to be investigated. 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.

On the other hand, one can imagine some application scenarios (in restricted domains for a restricted number of languages) in which computation and storage of links for later reuse is a viable option. In that case, suitable ways of storing and representing cross-lingual links become crucial. Also links computed runtime could be stored and made available online, thus configuring a sort of pool of cross-lingual links that grows with time. Such online links should follow the Linked Data principles to favour their later access and reuse by other applications.

3 Conclusions

In this paper we have motivated the study of cross-lingual ontology links as one of the fundamental challenges to solve in order to attain the goals of a truly multilingual Web of Data. There are, in particular, three subproblems to treat, namely cross-lingual link discovery, representation, and reuse. We have given an overview of the characteristics of each of them, as well as identified some relevant research topics that have to be further explored to be part of the solution. For instance, representation of cross-lingual links at the linguistic level, as well as the study of cross-lingual semantic measures and cross-lingual ontology alignment techniques. In our view such topics require more atention by the community and

[2] http://alignapi.gforge.inria.fr/format.html

[3] http://alignapi.gforge.inria.fr/edoal.html

will be crucial to enable the multilingual capabilities on the Web of Data.

Acknowledgments. This work is supported by the EU project Monnet (FP7-248458), the Spanish national project BabeLData (TIN2010-17550), and the Spanish Ministry of Economy and Competitiveness within the Juan de la Cierva program.

References

1. T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.
2. C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, MarMar 2009.
3. B. Daille, B. Habert, C. Jacquemin, and J. Royauté. Empirical observation of term variations and principles for their description. *Terminology*, 3(2):197–258, 1996.
4. M. Espinoza, E. Montiel-Ponsoda, and A. Gmez-Prez. Ontology Localization. In *Proceedings of the 5th International Conference on Knowledge Capture (KCAP09)*, pages 33–40, 2009.
5. 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*, 11:63–71, Mar. 2012.
6. 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*, 46, 2012.
7. E. Montiel-Ponsoda, J. Gracia, G. A. de Cea, and A. Gómez-Pérez. Representing translations on the semantic web. In *Proc. of 2nd Workshop on the Multilingual Semantic Web, at ISWC'11, Bonn, Germany, ISSN 1613-0073*, volume 775, pages 25–37. CEUR-WS, Oct. 2011.
8. M. Sabou, M. d'Aquin, and E. Motta. Exploring the semantic web as background knowledge for ontology matching. *J. Data Semantics*, 11:156–190, 2008.
9. P. Sorg and P. Cimiano. Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data & Knowledge Engineering*, 74:26–45, Apr. 2012.