PROVIDING MULTILINGUALITY TO ONTOLOGIES: AN OVERVIEW

GUADALUPE AGUADO DE CEA
ELENA MONTIEL-PONSODA
Ontology Engineering Group (OEG)
Universidad Politécnica de Madrid

RESUMEN. *Ontologies play a decisive role in the development of the Semantic Web, since they are able to model the knowledge of a specific domain in a machine readable way. However, the need to provide multilinguality to ontologies poses new challenges in the Ontology Engineering research. In this paper we attempt to offer an overview of available strategies for the localizing process of lexical resources and ontologies. Detailed steps in the localizing process of the multilingual lexicon EuroWordNet, the multilingual ontology GENOMA-KB, and the ontology translation software LabelTranslator are presented with the aim of illustrating three different localization approaches, their main characteristics and limitations.*
PALABRAS CLAVE: *multilingualidad, ontologías multilingües, localización de ontologías*

Las ontologías desempeñan un papel esencial en el desarrollo de la Web Semántica gracias a su capacidad de modelar el conocimiento de un dominio específico para que sea entendible por las máquinas. Sin embargo, la necesidad de dotar de multilingualidad a las ontologías plantea nuevos retos a la investigación en el campo de la Ingeniería ontológica. En este trabajo pretendemos ofrecer un panorama detallado de las estrategias empleadas actualmente en la localización de recursos léxicos y ontologías. Presentamos una descripción detallada del proceso de localización del lexicón multilingüe EuroWordNet, de la ontología multilingüe GENOMA-KB, y del software de traducción de ontologías LabelTranslator, con la finalidad de ilustrar tres enfoques de localización distintos y representativos, así como sus características relevantes y principales limitaciones.
KEY WORDS: *multilinguality, multilingual ontologies, ontology localization*

1. INTRODUCTION

According to Berners-Lee (1999), the Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. Ontologies are an important pillar in the construction of the Semantic Web, basically because they "represent static domain knowledge" (Gómez-Pérez *et al*. 2003: 2). Regarding the importance conferred to ontologies within the Semantic Web, researchers in the field of Artificial Intelligence are working continuously in the improvement of this form of knowledge representation. According to Gruber (1993) an ontology is defined as "an explicit specification of a conceptualization". Later on, Studer and colleagues (Studer *et al*. 1998: 185) stretched and enriched it by stating that:

> Ontologies are defined as a formal specification of a shared conceptualization. *Conceptualization* refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. *Explicit* means that the type of concepts used, and the constraints on their use are explicitly defined. *Formal* refers to the fact that the ontology should be machine-readable. *Shared* reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted group.

If ontologies are able to model the knowledge of a specific domain in a way that can be understood by computers, the challenge now is to be able to express that knowledge so that people from diverse cultures and speaking different languages can understand it. In order to

achieve that, and integrate the new information in their knowledge structures and cultural universes, those pieces of knowledge have to undergo a process of *adaptation* or *localization*.

The process of ontology localization had not received much attention until now, since most of the ontologies available on the Web were monolingual. However, because of the incremental use of intelligent systems the need of multilingual ontologies has emerged and it is one of the main priorities in the Knowledge Engineering research. For that reason, the aim of our study, developed within the framework of the European project NeOn[1], was to analyze existent localizing strategies of lexical resources and ontologies in order to obtain an overview of methods, tools or techniques used in the localization task. What follows in the present paper is a description of different localization approaches and their suitability depending on the characteristics of each resource. In the first section, we try to clarify some basic terms that are relevant in this paper. Then, we present an overview of some multilingual ontologies. The core of the paper consists of a detailed description of the strategies followed in the localization process of three resources, and their main implications.

2. TRANSLATING VS. LOCALIZING

To *localize* means literally "to make local" or "to orient locally" (Merriam-Webster Online Dictionary[2]). In the Free Encyclopaedia Wikipedia3 we find it generally defined as "the adaptation of an object to a locality". Localization can be applied to many domains. In economics, for example, localization is the way of "adapting products for non-native environments". In web design and software, localization refers to "the adaptation of language, content and design to reflect local cultural sensitivities".

The concept of *translation* has received much more attention throughout history as the activity of translating has been carried out since different language communities exist and communicate with each other. Following the functionalist approach to translation, it can be described as "a type of transfer where communicative verbal and non-verbal signs are transferred from one language into another […]. Translation is thus an intentional, purposeful action that takes place in a given situation…" (Vermeer 1983 in Nord 1997: 11). Functionalists put emphasis on the fact that every translation is intended to fulfil a specific *function* on a specific target culture, hence the name of their approach. Translation cannot be reduced to a one-to-one-word translation, but in every translation process there are many aspects that have to be taken into account. These are:

- Intention of the text – to inform, to convince...
- Target-text addressee(s) – children, experts, scientists...
- Time and place of the text reception – a company, a country, for one year…
- Medium over which the text will be transmitted – monolingual or bilingual web pages, brochures…
- Motive for the production or reception of the text – presentation of a new product, celebration of an anniversary…

However, the most important factor to be borne in mind is the *function of the translation*, i.e., the role of the translation in the target culture.

  - If the aim of the translation is to *document* the target reader about a situation in the original language and culture, reproducing the same intention, it may result in a text with a *foreign flair* for the target reader, so that he or she is conscious of the character of a translated text.
  - If the translation aims at producing in the target reader the same effect the original text produced in the original reader, the translator may have to adapt

many aspects of the text, or even change or omit facts, so that the target reader feels the text as original of his or her culture.

Many practitioners and translator theorists agree about this difference and talk about *overt vs covert translation* (House 1977: 188), and *documentary vs instrumental translation* (Nord 1989 cited in 1997: 47).

Notwithstanding, after having defined both concepts, we have to admit, that localization and translation are equivalent, when by translating we understand the second option considered, i.e., to "produce in the target reader the same effect the original text produced in the original reader". Therefore, *localization of lexical resources will be understood as involving all steps carried out in the process of adapting a lexical resource to a concrete language and culture community.*

3. TOWARD MULTILINGUAL ONTOLOGIES

In order to systematize a possible process of ontology localization, we analyzed how existent multilingual ontologies had been localized. For this purpose we used available ontology libraries, as OntoSelect[4], which collect, analyze and organize ontologies published on the Web, with the aim of finding those ontologies that are currently available in more than one natural language. After an intensive search we realized that only very few ontologies were currently multilingual (less than a 3% of all ontologies in the OntoSelect library, for example). We could also confirm that from the existent multilingual ontologies, most of them showed important inconsistencies in the corresponding versions in each natural language. The great majority were only complete in one natural language, and presented important gaps in the other languages.

This fruitless search directed our efforts to look for other multilingual resources that had a larger tradition and that could give us hints about a possible localization process which could be adapted for the task of ontology localization.

In the first part of our research, we analysed the translation process of the following lexical resources[5]:

- Glossary localization approach: FAOTERM
- Database localization approach: FishBase
- Dictionary localization approach: Eurodicautom
- Thesauri localization approaches: Agrovoc, Eurovoc

The main conclusions drawn from that survey can be summarized in three points:

1) The translation process followed in the localization of those lexical resources was mainly *manual*, i.e., carried out by translators, terminologists and experts working together in a specific field, or *semi-automatic*, i.e., with the help of translation supporting tools. Therefore, the *process for localizing was created ad hoc* in each case.
2) *Translation supporting tools* were introduced in the localization process in the recent years and included: translation memories, machine translation programs, text alignment tools, or term extractors, among others. Those tools are *domain independent* and, therefore, can be *reused* for the localization process of other resources.
3) *Available lexical resources* (online or on paper) and *text repositories* of each domain were the basis for the translation task. Authoritative multilingual databases, glossaries, dictionaries, taxonomies or encyclopedias were systematically consulted.

4. ONTOLOGY LOCALIZATION APPROACHES

After this initial analysis of multilingual lexical resources, we centered our research on the localization process of ontologies. For this purpose we identified three representative resources:

- Lexicon localization approach: EuroWordNet (EWN)[6]
- Ontology localization approach: GENOMA-KB[7]
- Ontology localizaing software: LabelTranslator[8]

4.1. *Lexicon localization approach: EWN*

EWN is a general-purpose multilingual lexical database first created in eight languages: English, Dutch, Italian, Spanish, French, German, Czech and Estonian. The wordnets in EWN are considered "autonomous language specific ontologies", and are interconnected through an Inter-Lingual-Index (ILI), a list of unstructured meanings mainly from Princenton WordNet[9], that provide the mappings across the wordnets. Each wordnet was created independently following one of these approaches:

- Merge model: concepts and relations are defined separately in each language, and afterwards, equivalent relations to the ILI concepts are generated.
- Expand model: ILI concepts from WordNet are translated, and then adapted or extended if necessary.

Both models define the core wordnets *manually* or by using *semi-automatic techniques*, and rely strongly on *available lexical resources* in each language. Main resources were: monolingual dictionaries, taxonomies or databases; and bilingual dictionaries (English/target language). The result was a set of independent wordnets linked to each other by means of a core of concepts or ILI.

4.2. *Ontology localization approach: GENOMA-KB*

GENOMA-KB is a biomedical knowledge base for the human genome. The GENOMA-KB is built upon four independent modules: ontology module, term base module, corpus module, and entities module.

The first step was the development of the *Ontological module*, by experts in the field, based on ontology concepts and its relations. Ontology concepts were then represented by natural language labels.

The second step consisted in the compilation of the *Corpus module* with genomic domain documents selected and validated by experts.

The third step was the development of the *Term base module* in the different natural languages, which consists of specialized knowledge units extracted from the specialized corpora (*Corpus module*) and from on-line dictionaries (or other lexical resources). The extracted terms were then mapped onto the ontology.

Finally, contexts and definitions were included in the *Term base module*, and the full bibliographical data was located in the *Entities module***.**

4.3. *Ontology localizing software: LabelTranslator*

LabelTranslator was developed in order to support "the supervised translation of ontology labels" (Declerck *et al*. 2006). By *supervised translation* is meant that this approach

foresees the intervention of the domain expert or translator in case no results outcome, or they need validation. Therefore, LabelTranslator offers a *semi-automatic strategy.*

For the development of LabelTranslator already available multilingual semantic resources and basic natural language processing tools were reused for providing a semi-automatic translation of ontology labels. In the current version of the LabelTranslator platform three types of multilingual resources were included:

- EuroWordNet
- Wikipedia, the encyclopaedia on the Web
- BabelFish[10], an on-line translation service used as "fallback position" (Declerck *et al*. 2006).

The steps followed for localizing ontologies are the following:

First, an ontology has to be uploaded in the LabelTranslator platform, and the *ontology labels* to be translated have to be selected.

In the second step, the system accesses the EWN database to find the selected term (or part of a term). In the following phase result(s) are displayed, if the matching is successful. Users can then validate the suggestions, modify the translation and save it in the database. If the matching in EWN is not successful, the system checks in Wikipedia, which also uses a mechanism for relating entries in the various available languages. If the previous steps do not provide any results, the system turns to BabelFish. If the translation is not satisfactory yet, the user can enter a translation, together with part-of-speech information and a definition.

5. IDENTIFIED LOCALIZATION APPROACHES

After this survey we can identify three general localization strategies represented by different localization approaches, each of them focusing on one aspect of the localization process.

1) *Localization approach based on multilingual ontologies*. This localization process is represented by the EWN lexicon. In the development of each wordnet (ontology) the process is carried out by translators, terminologists or experts in the field, who base their decisions mainly on already available lexical resources. Each wordnet is developed independently, and it is then related to the core wordnet. The process requires a high degree of human intervention, and the establishment of equivalences can be laborious, since each wordnet represents a different language conceptualization. On the other hand, wordnets are expected to guarantee language specific properties. This localization approach is adequate for general purpose multilingual ontologies.

2) *Localization approach based on a language independent ontology linked to a multilingual lexical resource*. This localization approach is followed by the GENOMA-KB. The process starts with the definition, by experts in the field, of ontology concepts and relations, which are supposed to be independent of any language. This can be true for highly specific domain areas. The establishment of equivalences is feasible because of the agreed ontology concepts and multilingual domain corpus. Term extraction and translation can be supported by automatic tools, reducing in this way human intervention.

3) *Localization approach based on the in situ translation process of monolingual ontologies*. LabelTranslator is a software developed for supporting the translation of ontologies which are available in a natural language. This localization process is strongly based on already available lexical resources. By adding more specific resources, it could be adequate for translating specific ontologies. Human intervention is very high, since the expert or translator has to take decisions *in situ*. Results would need a further validation by other experts in the field. This tool is very useful if we think of the great amount of monolingual ontologies existent nowadays on the Web.

6. CONCLUSIONS

The main conclusions we can draw after the analysis of the localization process of lexical resources and ontologies can be summarized as follows:

1) Both lexical resources localization and ontologies localization are characterized by the intervention of translators, terminologists or experts in the field, with the help of translation supporting tools, available lexical resources, and text repositories.
2) The ontology localization process presents additional strategies depending on
    - The purpose of the resource: general vs. specific
    - The existence of the ontology in a natural language

    Then, if the ontology represents general knowledge, specificities of each language and culture universe will be better captured by independent ontologies in each language, following the so-called *localization approach based on multilingual ontologies*. However, if the degree of specificity of the ontology is high, the *localization approach based on a language independent ontology linked to a multilingual lexical resource* will represent the respective equivalences in each language in a more suitable way.

    On the other hand, if the ontology already exists in a natural language, localizing software tools will come to solve the problem of ontology localization, without forgetting the need of later revision and agreement by experts of the domain, thus following the *localization approach based on the in situ translation process of monolingual ontologies*.

ACKNOWLEDGEMENTS

The research described in this paper is supported by the European Commission's Sixth Framework Programme (IST-2005-027595) under the project name: *Lifecycle support for networked ontologies (NeOn)*. We would also like to thank Inmaculada Álvarez de Mon y Rego, José Ángel Ramos and Mari Carmen Suárez-Figueroa for their help with some ontological aspects.

NOTAS

1. NeOn is a project involving 14 European partners and co-funded by the European Commission's Sixth Framework Programme under grant number IST-2005-027595. More information in http://www.neon-project.org/web-content/
2. http://www.m-w.com/
3. http://en.wikipedia.org/wiki/Main_Page
4. http://sioc-project.org/node/192

5. Fully described in Deliverable 2.4.1 of the NeOn project
6. http://www.illc.uva.nl/EuroWordNet/
7.http://genoma.iula.upf.edu:8080/genoma/corpSearch.do;jsessionid=C5F6DA7C2954A5084D48F35666F8B0DE?operation=init
8. (Cf. Gantner 2004; Declerck *et al*. 2006)
9. http://wordnet.princeton.edu/
10. http://babelfish.altavista.com/

BIBLIOGRAPHY

Berners-Lee, T. 1999. *Weaving the web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. New York: HarperCollins Publishers.

Cabré, M. T., C. Bach, R. Estopà, J. Feliu, G. Martínez and J. Vivaldi. 2004. "The GENOMA-KB project: towards the integration of concepts, terms, textual corpora and entities". Lisboa: *Proceedings of LREC 2004*. 87-90.

Declerck, T., A. Gómez-Pérez, O. Vela, Z. Gantner, and D. Manzano-Macho. 2006. "Multilingual Lexical Semantic Resources for Ontology Translation". Genoa, Italy: *Proceedings of LREC 2006*. 1492-1495.

Gantner, Z. 2004. *TermTranslation – A Tool for the Semiautomatic Translation of Ontologies*. Technical report written at the OEG of the UPM, Spain [unpublished].

Gómez-Pérez, A., M. Fernández-López, and O. Corcho. 2003. *Ontological Engineering*. London: Springer-Verlag.

Gruber, T.R. (1993). *A translation approach to portable ontology specification.* Knowledge Acquisition 5(2):199-220

House, J. 1977. *A Model for Translation Quality Assessment*. Tübingen: Gunter Narr.

Nord, Ch. 1997. *Translating as a Purposeful Activity*. Manchester: St. Jerome.

Studer, R., V.R. Benjamins, and D. Fensel. 1998. *Knowledge engineering: principles and methods*. IEEE Transactions on Data and Knowledge Engineering 25(1-2): 161-197.

Vermeer, H. 1983. *Aufsätze zur Translationstheorie*, Heidelberg.

Vossen, P. 2004. "EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index". *Semi-special issue on multilingual databases* (IJL 17/2, June 2004).