LabelTranslator: Multilingualism in Ontologies

Thierry Declerck, Ovidiu Vela
DFKI GmbH
Language Technology Lab
Stuhlsatzenhausweg, 3 66123 Saarbruecken, Germany
declerck@dfki.de

Asunción Gómez Pérez, Zeno Gantner, David Manzano-Macho
UPM
Laboratorio de Inteligencia Artificial, Campus de Montegancedo
28660 Boadilla del Monte, Spain
asun@fi.upm.es

Abstract

We describe the integration of some multilingual semantic resources and basic natural language processing steps that are helpful in providing ontologies, which are normally using concept labels in just one (natural) language, with multilingual facility in their design and use in the context of Semantic Web applications, supporting both multilingual semantic annotation and ontology extraction based on multilingual text sources. The system can be demonstrated.

1 Introduction

Ontologies in Semantic Web applications are used, among others, for providing semantic and content annotations of multilingual web pages. Therefore we dedicated work in the Esperonto project[1] for providing a strategy and a platform that supports the multilingual extension of ontologies existing in just one natural language, and in doing so to allow the semantic annotation of multilingual web documents using multilingual labels of ontologies.

In Esperonto we investigated the use of available multilingual language resources and basic natural language processing tools for providing for a supervised automated translation of labels in ontologies.

In this poster we present the multilingual semantic resources we used for the implementation of the platform and the general strategy offered for supporting the supervised translation of the labels of ontologies.

[1] Esperonto was a project of the Information Society Tecnologies (IST) Program for Research, Technology Development & Demonstration under the 5th Framework Program of the European Commission, with number IST-2001-34373. The project ran from 2002 to 2005. See[2].

2 The multilingual semantic resources

Two main types of multilingual lexical resources have been considered in the Esperonto project: the lexical semantic approach (mainly the WordNet initiatives, like EuroWordNet (EWN), see [4], and the lexical approaches more properly speaking, being the morpho-syntactic encoding of lexical entries or terms (the Parole/Simple framework, see [6]). In the actual version of the platform, only EWN has been included.

Since EWN comes along with part-of-speech information associated with the terms encoded in the *synsets*[2], this information is also displayed to the user, who can decide which reading to select for the translation. So for example the term "book" in the source ontology can be translated either by the verb "reservar" (*to book*) or by the noun "libro" (*the book*).

Some EWN resources include also so-called "glosses" offering for a short definition of the term under consideration. Those glosses are also displayed to the user in order to support her decision for a term in the target language. But the glosses are also used by the system itself for disambiguating the list of proposals the system is extracting from the EWN resources. We used EWN for Spanish, English and German.

Another type of multilingual information has been considered for being able to translate labels of ontologies: the Wikipedia resource on the Web (see [10]), which we use additionally to EuroWordNet. Wikipedia is not based on a lexical perspective but on a dictionary perspective that encodes knowledge of the world instead of knowledge of the words. We see thus in Wikipedia a real complementary multilingual resource to EWN and similar lexical semantic resources. Wikipedia article names in one language are also

[2] "Synsets" are the organising strcutures for lexical semantic realtion between terms, like synonymy and hyperonomy.

linked to a multilingual database of corresponding terms. We used here the resources for German, English, Spanish and Catalan. Wikipedia terms are linked, when available, to entries of the "Wiktionary" (see [8]), which lists also synonyms and available term translations that are displayed to the user for easing his selection of possible target labels in the ontology to be generated by the translation.

3 The supervised translation strategy

The user can select within a GUI an ontology to be translated (from English to German and to Spanish). Once the ontology is uploaded, the user can select a label of the ontology to be translated in a selected target natural language. The GUI is then displaying candidate labels in the target language. This set of possible translations is strongly reduced by the system, which provides for part-speech disambiguation, comparison of EWN glosses, and contextual constraints given by the terms of the ontology already translated, before displaying the candidates for the term translation.

The processing chain can be roughly summarized like this:

1) If the concept label in the ontology is already available in the target language in our database, then just display it, with all relevant available information (linguistics and world knowledge). The user can modify the translation if wished.
2) If this is not the case, then use first EuroWordNet (EWN) and check if the label is present in the WordNet of the source language (English in our case).
If this is so, 2 cases are possible:
The label in the ontology is a multiple word unit (MWU): check if the multilingual index associated with the WordNet entry in the source language is pointing to an existing entry of the target language. Display the EuroWordNet entry of the target language if the matching is successful.
If this is not the case, check if the main words of the multi word unit are present in the EuroWordNet of the source and target languages (using again the multilingual index of EuroWordNet, which relates entries in the various languages). Display the results if the matching is successful. With "main words" we understand the words that are not to be considered as the so-called "stop words" (Determiners like 'the', prepositions like 'on' etc.). Main words belong in our case mostly to the class of nouns, but also to the class of adjectives.
3) If the EuroWordNet approach is not successful, use the same strategy described in 2) to the multilingual term resources of Wikipedia, which uses also an interlinking mechanism for relating entries in Wikipedia in the various languages available.
If 2) and 3) are not successful, use a fallback solution and access free accessible translation engines on the web and display their results, if any.
If no (satisfactory) result is displayed by the platform, the user can enter his/her own translation. In case satisfactory results are shown, the user can validate them, whereas the results can be edited for some improvement.

Evaluation

We have been thinking about a first evaluation scenario that allows statements about the added value of the Esperonto platform for supporting multilingualism in ontologies. We will have to show that the use of a combination of language resources, as proposed in Esperonto, allows a higher degree of automation in the translation process of ontologies and a better quality of proposed translations submitted to the domain expert, as for example using only online translation services. The first evaluation will be something like defining a continuous line of using only:
EWN,
EWN+Wikipedia,
EWN+Wikipedia+Ling.Analysis (for the analysis of Glosses and Definitions)
…
We should then be able to say how many words/terms can be translated without an active intervention of the domain expert, so that he/she cam just validate results of the translation process.
We will also compare the results of the Esperonto platform with the output of the online translation services, whereas we will have to take in consideration the cases where either EWN/Wikipedia or the online translation services are not providing any results.

Conclusions

The actual state of the platform is offering choices for the translation of ontologies that is based on various type of information: lexical semantic (EWN), encyclopaedic (Wikipedia) and usual translation services.
As the implementation of certain features that includes some linguistic processing and information is progressing, as well as the analysis of the whole ontology to be translated, we expect a higher degree of automation dealing with EWN and Wikipedia data that makes the platform a real alternative to sole translation services, since the platform is offering to a certain degree a knowledge driven translation that is supported by natural language analysis. The knowledge is the one accessed in EWN, Wikipedia and within the structure of the ontology being translated.

References

[1] http://www.globalwordnet.org

[2] http://www.esperonto.net

[3] http://www.cogsci.princeton.edu/~wn/w3wn.html

[4] http://www.hum.uva.nl/~ewn

[5] http://www.icsi.berkeley.edu/framenet

[6] http://www.ub.es/gilcub/SIMPLE/simple.html

[8] http://en.wiktionary.org/wiki/Main_Page

[9] http://www.lsi.upc.es/~nlp/projectes/ewn.html

[10] http://en.wikipedia.org/wiki/Main_Page