Multilingual Lexical Semantic Resources for Ontology Translation

Thierry Declerck[1], Asunción Gómez Pérez[2], Ovidiu Vela[1], Zeno Gantner[2], David Manzano-Macho[2]

[1] DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg 3, D-66123 Saarbrücken
declerck@dfki.de
[2] UPM, Laboratorio de Inteligencia Artificial,
28660 Boadilla del Monte, Spain
asun@fi.upm.es

Abstract

We describe the integration of some multilingual language resources in ontological descriptions, with the purpose of 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 the semantic annotation of textual documents with multilingual ontology labels and ontology extraction from multilingual text sources.

.

1. Introduction

In this paper we present in some details the multilingual semantic resources and the strategy we used for the implementation of a platform supporting the supervised translation of ontology labels.

(Domain) ontologies can be defined as a (possibly) complex data structure that introduces formal concepts and describes the relations existing between those concepts. The main goal of ontologies is to formalize (domain) knowledge for ensuring a more compact description of it and a more efficient access to it. The concepts described by the (domain) ontologies are in general not relying on the words or terms in use in a particular natural language, but the praxis has been very often to label the concepts in using English terms. The levels of description in ontology are not necessarily corresponding to the various levels present in the distinct natural languages. Concepts in ontologies might have no correspondence to any lexicalised form in a specific natural language.

Ontologies in Semantic Web applications are used, among others, for providing semantic and content annotations of multilingual web pages. In the Esperonto project (see Esperonto) a strategy and a platform have been implemented for supporting 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.

We currently continue part of this work within the eContent LIRICS project (see Lirics), where within standardisation efforts for natural language resources, we also investigate the relation between lexicon, syntax and semantic, also at the level of domain ontologies.

2. The multilingual semantic resources

Two main types of multilingual lexical resources have been considered in the Esperonto project: the lexical semantic approach of EuroWordNet (see EuroWorNet), and the lexical approach of the Parole/Simple framework (see Simple). In the actual version of the platform, only EWN has been fully included. 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 Wikipedia), which we use additionally to EuroWordNet. Wikipedia is based on an encyclopaedic perspective that encodes knowledge of the world instead of knowledge of the words. In this, Wikipedia is a real complementary multilingual resource to EWN and similar lexical semantic resources for the translation of ontology labels. Wikipedia article names in one language are also linked to a multilingual database of corresponding terms.

As a part of the service proposed, we used as a fall-back position classical on-line translation services, like BabelFish.

3. The Platform

We implemented a platform that allows the user to upload a specific ontology, to select labels of the ontology and the language in which this label should be translated. Once the user has made her selections, the systems accesses the EWN and Wikipedia databases for finding if (parts of the) selected term are encoded in the resources and displays the results of the search to the user, who can then decide if the suggestions made by EWN or Wikipedia are appropriate. Since EWN comes along with part-of-speech information associated with the terms encoded in the *synsets*, 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.

A fallback position is given by accessing on-line translation systems. The user can also enter his/her own translation

4. The supervised translation strategy

It is important to point here that we implemented some heuristics for disambiguating the possible set of results generated by EWN. This is done on the base of part-of-speech disambiguation, the comparison of EWN glosses associated to EWN entries in the source and target languages, but also on the base of contextual constraints given by the terms of the ontology already translated. In doing so we can reduce considerably the number of answers provided by EWN.

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 it this is the case, 2 things are possible:

a) 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.

b) 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 1), applying it to the multilingual term resources of Wikipedia, which uses also an interlinking mechanism for relating entries in Wikipedia in the various languages available.

If 1), 2) and 3) are not providing results, use a fallback solution and access free accessible translation engines on the web and display their results, if any. In fact, the system displays for the time being always the results of the feely available online translation service (BabelFish). The user can also enter his/her own translation (together with POS Info and a definition).

5. Some general considerations about the processing chain

As the reader can see, we give priority to the EuroWordNet resource. This is due to the fact that the EuroWordNet resources are organised in such a way that we expect a high quality in the resulting "translation" of a concept, since the multilingual index associated with a term in EuroWordNet has been built following semantic considerations and validated by language and/or domain experts.

EWN also offer glosses (in English) that give a definition to the terms listed in EWN. Those glosses can provide help when mapping a label in the ontology.

But it turns out to be difficult when there is more than one possible entry in EWN that can be referred to from the label in the ontology (ambiguity problem). We are investigating here two approaches for using the glosses, a rule-based one and a statistical approach.

The rule-based strategy is twofold: 1) if in the glosses of the EWN terms of the target language, terms are occurring that are also present in the ontology to be translated, then the EWN entry having this gloss is a better candidate for the translation as the EWN entry in which gloss no such terms are occurring, and only the preferred translation will be displayed; 2) if the source and target EWN entries share the same or similar glosses (string matching), then the corresponding entry of the target language will be selected, discarding entries of the target languages that have distinct glosses as the entry of the source language. Dry exercises have shown that when the rule-based approach provides results at all, those results seem to be correct.

The statistic approach is based on two gloss-based similarity measures in the Perl package WordNet::Similarity. This package implements two algorithms, called "The adapted Lesk" (see Banerjee, 2002) and the "Vector" algorithm (see Patwardhan, 2003). We provided for a first evaluation of those algorithms, and can report that the Lesk algorithm performs better than the Vector one. But even for the Lesk algorithm we suggest a hybrid approach combining the Lesk algorithm with the rule-based approach. But we did find time yet to implement this hybrid approach.

But in any case, one has also to be aware that the EWN resource is far for being exhaustive and having an equal coverage for the different languages involved. Also not all the language specific WordNets do make use of the glosses with the same strength. So in our case, compared to the English WordNet, the German WordNet has not a large coverage, whereas the Spanish WordNet is poorly "decorated" with glosses.

In the second place of the processing chain, we search the Wikipedia domain. Wikipedia is a Web-based multilingual dictionary resource developing quite fast and being currently extended to many languages. Wikipedia gives us an encyclopaedic view on the terms used in the ontology rather than the lexical semantic view of EuroWordNet. The definition article associated with the terms in Wikipedia can be considered as similar to the glosses in EWN, but are larger and more difficult to be processed for supporting the translation task. An advantage in using Wikipedia for supporting ontology translation is that the user can go to the Wikipedia articles and really check that the content associated with a term is the one he/she wants to have in the target ontology label.

In the actual implementation already some use is made of the structural organisation of the ontology. So the translation of terms is passed down in the taxonomy. Another use of the structural hierarchy consists in using it for guiding the translation process. Here an example for clarifying: consider the label "book" as a subtype of the label "publication". Knowing that the word "publication" is a substantive (it is encoded like this in the English

EWN), the system can then filter out the verbal readings of the word "book" (in the case of booking a travel for example), and so not display to the user the Spanish verb "reservar" but only the nominal Spanish entries, like "libro"[1].

5.1. Some linguistic issues with EuroWordNet

There are some problems related to EuroWordNet (and partially to Wikipedia): all the terms are listed using the ground form of the words. So translating for example the English sequence "technical documentation" into Spanish, the following will be actually delivered by our system (using EWN) to the user: "tecnico" and "documentacion". Two words are given, since the multi word unit "technical documentation" is not in EWN, but each word alone is covered by EWN. We have two linguistic problems here, due to the word-by-word EWN based translation:

1) The word "tecnico" is the masculine form of this adjective. But the substantive "documentacion" bears feminine gender in Spanish. So the system has to generate the form "tecnica". This has been implemented in our platform, adding to the EWN data for Spanish (and for German as well) a (morphological) rule that generates the feminine gender of the adjectives in the case it is associated with a noun bearing the feminine gender (in German we also have to consider the neutral gender). Alternatively we can augment the EWN database with all the morphological forms that can occur in German and Spanish. We think that the rule-based approach is to be preferred, since it does not modify the EWN data.

2) The second problem concerns the word order: the word-by-word translation of "technical documentation" is "tecnico documentacion". Once we have generated the right feminine form for the word "tecnico", we still have to provide for the right word order in Spanish, which is "documentacion tecnica". Here again a rule-based approach has been defined, applying to the proposed translation by EWN. In case this approach is failing, the user has still the possibility in the GUI to re-arrange the order of the translated words.

So at least two linguistic "interventions" are needed for solving this problem: provide for the right morphological forms of the translated words, and for the right word order. More formally the rules look like (whereas we subsume both Adjectives and Articles under the category "Modifier"):

a) If Gender(Head-Noun of EWN translated term) eq FEM => generate FEM-Form(Modifier of EWN translated term)
b) If Gender(Head-Noun of EWN translated term) eq NEUT => generate NEUT-Form(Modifier of EWN translated term)

These rules are meant to deal with the morphological properties of the terms (for Spanish and German). They do not apply to all Spanish adjectives, and we have a list of the adjectives for which the rules do not apply. Dealing with the word order problem (relevant only for Spanish, since German and English have the same word order within nominal phrases):

a) If Sequence(translated terms) eq Adj-Noun => generate_sequence(Noun-Adj)
b) If Sequence(translated terms) eq Noun(1)-Noun(2) => generate_sequence(Noun-Prep-Noun)

The case a) is dealing with the improvement of the word-by-word translation of "technical documentation" -> documentacion tecnica". The b) case is dealing with the word-by-word translation of "message receiver" -> "recipiente del messaje", where a prepostion has to be added in the target language (Spanish).

Another linguistic "intervention" might also be very useful: parsing the glosses (in EWN) and definitions (in Wikipedia), in order to give to those a linguistic structure, which is more appropriate for detecting relevant expressions that can help the translation process of the ontology. So the platform for multilingual ontologies will be extended in order to search into linguistically annotated glosses and definitions, instead of pure text.

5.2. Some Linguistic issues with Wikipedia

Wikipedia is using only full form words. But in the Wikipedia "family" there is also now an open dictionary, which displays the ground forms of the word. An example is given in the following URL: http://open-dictionary.com/Arts, where the groundforms of the word "arts" is given in many languages. And quite interesting: the Wiki dictionary also links to the WordNet definition! So that we can close here a circle between the word based semantic net (WordNet) and the encyclopaedic based semantic network. Here we still have implementation work to extract the morphological forms from the Wiki Dictionary and the links between Wikipedia terms and EuroWordNet terms.

6. Evaluation

We have been thinking about a first evaluation scenario that allows statements about the added value of the platform for supporting multilingualism in ontologies. We will have to show that the use of a combination of language resources, as proposed in Esperonto and Lirics, 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

[1] Here we have to mention that the version of EWN we use lists three types of word categories: Verbs, Nouns and Adjectives. An EWN entry can be part of more than one category, so the example of "book" that can be a verb and a noun. The ambiguity problem here is of purely syntactic nature. There are also semantic ambiguities, which are more difficult to cope with in our case.

domain expert, so that he/she cam just validate results of the translation process.

We will also compare the results of our 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.

7. 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 on-line 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 resources and analysis. The knowledge is the one accessed in EWN, Wikipedia and within the structure of the ontology being translated.

In the next future, we will have to look for a real formal integration of multilingual information within ontologies, a topic that will be addressed as well in the LREC Workshp Ontolex 2006 “Interfacing Ontologies and Lexical Resources for Semantic Web Technologies”.

8. Acknowledgements

The work presented in this paper has been partially funded by the IST project “Esperonto” which was supported by the Information Society Technologies (IST) Program for Research, Technology Development and Demonstration under the 5th Framework Program of the European Commission, with the number IST-2001-34373.

Current extensions are done within the context of the eContent project LIRICS, under the contract e-Content-22236-LIRICS.

9. References

Banerjee S. (2002) Adapting the Lesk algorithm for word sense disambiguation to WordNet. Master Thesis, University of Minnesota, Duluth, 2002.

Patwardhan S. (2003) Incorporating dictionary and corpus information into a vector measure of semantic relatedness. Master Thesis, University of Minnesota, Duluth, 2003.

Esperonto: http://www.esperonto.net
EuroWordNet:http://www.hum.uva.nl/~ewn
FramNet: http://www.icsi.berkeley.edu/framenet
GlobalWordNet: http://www.globalwordnet.org
Lirics: http://lirics.loria.fr
Simple: http://www.ub.es/gilcub/SIMPLE/simple.html
Wiktionary: http://en.wiktionary.org/wiki/Main_Page
Wikipedia: http://en.wikipedia.org/wiki/Main_Page
WordNet:
http://www.cogsci.princeton.edu/~wn/w3wn.html