Aligning unequal multilingual thesauri

Introduction

This paper describes a method used for solving the practical problem of aligning the English and Romanian versions of the Eurovoc thesaurus and reports on the ongoing work for the Romanian version. The problem is not trivial due to the missing of the unique identifiers of the Eurovoc in the Romanian version, neither simple, for the two versions of the thesaurus have different sizes. The nature of this task is very similar to the one of aligning ontologies, a major concern for the Semantic Web community.
Eurovoc is principally used to index the Acquis Communautaire (the EU legislation and international treaties) as it is a multilingual, polythematic thesaurus (Steinberger et al., 2002). Its last release (4.2) is available in 21 versions, out of which 16 being official, each version corresponding to an EU language. The thesaurus is comprised of 6645 terms or descriptors (519 top terms), covers 21 fields (from environment or industry, to politics and international relations) and is structured into 127 microthesauri. Every field and microthesaurus has an unique identifier in all languages allowing multilingual navigation (as mentioned above, the Romanian thesaurus is lacking these identifiers). Each term is a node in one of the 519 trees rooted by the top terms. There are five types of Semantic Relationships in Eurovoc, out of which only the *hierarchical relationships* (6669 reciprocal BT (broad term) and NT (narrow term)) were relevant for our task. When a descriptor has no broader terms, it is called *Top Term*.

The Task

The Romanian version of the Eurovoc thesaurus was developed by the General Secretariat of the Chamber of Deputies of the Romanian Parliament and covers about 70 percent of the English descriptors together with the semantic relationships between them. Strangely, the translators did not preserve the unique IDs of the term descriptors making the Romanian version useless for multilingual intended purposes. Our aim was to align the Romanian partial version of the thesaurus to the English complete version in order to automatically recover the terms IDs. Unfortunately, the uncompleted Romanian version of the thesaurus raised another problem. Obviously, hierarchical relationships and the top terms ensure the existence of as many trees as the number of the top terms. If the Romanian thesaurus had been entirely translated, we would have expected the same number of trees with the same tree structure. In our case, we definitely had to expect that not all of the considered top terms in the Romanian version corresponded exactly to the top terms of the English version and that we would have to align incomplete tree structures, too.

The Method

The method proposed has two phases. In the first one, all descriptors of the Romanian version are aligned to those in the English version, while in the second one, all the missing terms are identified and suggestive raw translations are being given by the application.

The first phase of the alignment has also several stages. In the beginning, we have to produce a translation equivalence dictionary suitable for the words from which the descriptors in the thesaurus are made of. We constructed this dictionary using the EM algorithm on the Romanian-English sub-corpus of the Acquis Communautaire 21-languages parallel corpus, in a Giza Model 1 fashion. The next step is aligning the top terms of the two languages using the translation equivalence dictionary previously made. This is done as it follows. For each word of a Romanian

term, each English translation equivalent is introduced into a hash table (available for the term) along with its estimated probability. If the equivalent is already in the hash table, than its estimated probability is updated with the greatest value between the old and the new one. In this same way, for all the Romanian terms, hash tables are constructed. After this, a multi-iterative process starts. For each Romanian term, and for each English term, we compute a translation score as the sum of the estimated probabilities (only if above a threshold), of the words which form the English term, in the Romanian term hash table created above. If an English word composing a term can not be found in the hash table, than the score is nil. The maximum translation score wins. This score is only used in the process of finding the best translation for a term. The Romanian top terms are 508 while the English, 519. These are small numbers, so the entire process is not time or memory consuming. After this, we notice that different Romanian terms are translated with the same English term. This is not acceptable so, only the term with the greatest score is kept or, in case of equality, the shorter term is kept (this is due to the fact that the English translations of the Romanian terms tend to be shorter in number of words than these terms). The process above reiterates without the resolved terms until the number of these terms remains unchanged. Every pair of top terms aligned means a pair of tree roots aligned. Once the roots are successfully aligned, their respective sub-trees should also get aligned using both a tree structure breadth-first partial matching algorithm and the translation equivalence dictionary; otherwise the roots alignment should be reconsidered.
The not aligned Romanian terms are to be looked after between the terms of the sub-trees of not aligned English top terms, layer after layer, in order not to increase the need for resources. In our case, about 30% of the top terms remained unidentified. This figure keeps the same ratio as the completeness of the Romanian version of Eurovoc but no connection can be made between the facts.

The final phase, which is a completion step, takes care of the English descriptors that does not have a corresponding Romanian descriptor. The mapping tree-structure algorithm passes through the hierarchies of the two parallel thesauri and inserts dummy nodes in the Romanian thesaurus for the missing terms (not yet translated). The translation equivalence dictionary is used to make suggestive rough translations for an expert who usually is expected to edit it and to validate the proposed terms. The expert is also helped by an easy-to-use GUI, which allows him/her to visualize the alignment, to correct translations, to change the alignments or to add new information (such as multiple non-descriptors, not necessarily paralleled in the English version).

Conclusions

Aligning unequal multilingual thesauri is a resource-demanding task when is manually done. Even though a specific requirement motivated the work presented in this paper, the system developed is applicable to any other similarly structured thesaurus and is easy to extend/adapt for working with more elaborated hierarchical knowledge structures such as ontologies of the Semantic Web. The time and memory costs can be kept significantly low if the work is gradually focused on hierarchical levels.

References

R. Steinberger, B. Pouliquen, J. Hagman, *Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus Eurovoc*, Springer-Verlag, 2002;
YYY, ZZZ, WWW, XXX, ---, Ann Arbor, June 2005;