# EUROTERM
# Extending EWN using both the expand and merge model

*Stamou Sofia* *Ntoulas Alexandros* *Hoppenbrouwers Jeroen* *Saiz-Noeda Maximiliano* *Christodoulakis Dimitris*

## *Abstract*

EuroTerm aims at expanding EuroWordNet with domain specific terminology for a set of European languages. EuroWordNet is a lexical database representing semantic relations among basic concepts for West European languages, which are combined with a so-called Inter-Lingual-Index. EuroTerm's main purpose is to combine effectively multilingual domain specific terminology into a common lexical database through a Terminology Alignment System, in order to expand EuroWordNet and the Inter-Lingual-Index with terms restricted to the conceptual domain of environment.

## *1. Introduction*

EuroTerm[1] is an EC funded project (EDC-2214) that aims to extend EuroWordNet (EWN) with environmental terminology for the following languages: Greek, Dutch and Spanish. EWN is a multilingual database with generic WordNets for eight European languages (Vossen, 1996). Individual WordNets incorporated in the central database form autonomous semantic networks linked to each other through an Inter-Lingual-Index (ILI). The aim of EuroTerm is to enrich EWN with domain specific terminology for the above languages. Each of the monolingual WordNets will comprise of ~1,000 synsets and will be stored in a common database, which will be linked to the central EWN database under the domain label "Environment". There are two approaches to build a semantic network, namely the expand and the merge model approach. The former concerns the translation of English concepts in the respective languages and then the actual development of synsets whereas the latter implies the independent development of monolingual synsets and then their linking to the most equivalent synset in the ILI (Vossen 1996). For the implementation of EuroTerm a combination of the two models has been adopted aiming at maximal conceptual coverage across languages and maintenance of the language-dependent differences. Thus, unlike the extension of EWN with computer terminology where the expand model was followed (Vossen 1999) our approach slightly differentiates in the sense that both the merge and expand model approaches were followed. For the implementation of the project common English lexical resources were used as the starting point for the terminology extraction, thus the expand model approach was firstly adopted. Once the first set of terms was extracted the merge model was adopted in order to check these terms against monolingual lexical resources and enrich them with missing terms. The final set of terms will be incorporated in the database through a Terminology Alignment System that will enable linking across languages. In the following section (2) we present our approach for the terminology acquisition and we continue with a brief description of the Terminology Alignment System (3). Finally, some applications of the EuroTerm project are discussed (4) and some early conclusions are drawn (5).

## *2. Our Approach: A Combination of the Expand and the Merge Model*

The main objective of EuroTerm is to enrich EWN and the ILI with domain-specific terminology for the following languages: Greek, Dutch and Spanish. To achieve sufficient overlap across languages and to overcome vocabulary completeness and coverage issues attention should be paid during the selection of terms that will be incorporated in the semantic network. Thus, after a close investigation of the two different approaches (merge and expand

[1] EuroTerm EDC-2214, "Extending the EuroWordNet with Public Sector Terminology" funded by the EC

model) already followed for building semantic networks and having in mind the application of EuroTerm we concluded that a combination of both approaches would result in more consistent and reliable results. The reason for applying both methods is twofold. First and foremost we wanted to assure sufficient overlap in the coverage of monolingual WordNets and still maintain language-specific properties and secondly we wanted to enable interpretation of the differences found across WordNets once they are incorporated in Information Retrieval (IR) systems, which is the envisaged application of the project. The basic idea is that the selection of the candidate terms is performed in two phases. During the first phase the expand model is applied according to which common English lexical resources are used for the terminology extraction. Once the first set of terms had been determined the merge model was applied during which monolingual lexical resources were used for the terminology acquisition. We started off with a common English environmental corpus of 429 manually collected documents and English glossaries comprising in total of 4,972 terms along with their glosses. We applied a Part-Of-Speech (POS)-tagger (Daelemans 1999) and a lemmatizer to the corpus Then the TF*IDT metric was applied to count frequencies for each lemma, which were then sorted by their frequency. So far the extraction process has been conducted automatically. Afterwards the 4,500[2] most frequent terms were semi-automatically[3] checked against English environmental glossaries and those found in them were candidates for the environmental WordNet. The candidate terms were again semi-automatically checked against ILI to trace which of them were already present with an environmental gloss attached. Environmental terms missing from the ILI had also to be checked against monolingual resources prior to their incorporation in the database. At this stage the merge model was applied. In particular, the missing terms were manually checked against monolingual domain specific lexica and corpora and their glosses found in explanatory dictionaries were investigated in order to conclude on their importance for the underlying languages. In addition, environmental terms found in monolingual resources but not in the list of English candidates were also checked against monolingual glossaries or thesauri and if they were important they were included in the candidates' list. Importance at this stage was determined by the occurrence frequency of a term in the corpus and by its presence in the glossaries. Once the selection process is finished English terms missing from the ILI were manually translated to the respective languages and receive an environmental tag. Moreover, terms found in monolingual resources were also manually translated in English and were incorporated into the ILI. Afterwards the actual development of synsets will follow based on the language internal relations used in the EWN project. Since the project is still running and due to the fact that the testing phase has not started yet we cannot report on concrete results about terminology mismatches and alignment across languages. However, it is estimated that the problems that might appear will mostly concentrate on language particularities and will not be due to the methodology we followed for their acquisition.

## *3. Terminology Alignment System*

The Terminology Alignment System (TAS) (Hoppenbrouwers, 2001) is a key part of the infrastructure underlying the EuroTerm project. Through the TAS, partners can communicate and coordinate their work on the individual WordNets in Spanish, Greek, and Dutch. We call the TAS an alignment system, since it helps terminologists to align their work on the local WordNets. The TAS is not a unified central database in which local WordNets are merged. Instead, it is a simple link database, designed to facilitate cooperative work on local WordNets. Although the project itself is based on a common starting fund of 1,000 terms, future

[2] 4,500 terms was a sufficient amount of terms to process in order to end up with 1,000 environmental terms, which was our target.

[3] They were firstly checked automatically and then the selection of the candidate terms was verified manually.

applications will require a much looser connection between partners. Therefore, the concept of federation, where all partners cooperate but there is as little mandatory standardization as possible on the local level, is exploited to the maximum. Certain standards must always be followed, but we assume that the existing practice in EWN provides this necessary minimum. The plethora of existing WordNet tools, such as Polaris and the various open source systems, can and will still be used for local WordNet maintenance. The TAS is a federated, networked tool that links up the local tools and has been built as a database-driven Web.

## *4. Applications of the EuroTerm Project*

The most immediate application of the multilingual domain specific WordNet concerns the incorporation of the environmental domain into an Information Retrieval (IR) system so that documents are semantically and not lexically represented in the index of such systems. Thus, query terms will be compared against documents not only by weighting term co-occurrence but also by measuring the semantic similarity of query and document sets of indexes. Towards this direction some modifications will need to be done in the search engine(s) in which EuroTerm network will be incorporated. In particular, a directory named Environment will have to be incorporated in the interface of the engine so that users have the ability to specify whether they are interested in the domain of environment or not. In addition, the engine should keep two separate indexes. In the first one environmental documents will be clustered whereas in the second one the rest documents will be stored. Our objective focuses on achieving better precision scores of the obtained results if the search is performed against the environmental index. In addition, it is expected that by mapping query terms against the correct environmental synsets would significantly improve retrieval results in terms of both precision and recall, making the obtained results more meaningful to the end user.

## *Conclusions*

We have described a combination of the expand and merge model approach to extend EWN with environmental terminology through a Terminology Alignment System. The use of EuroTerm will be demonstrated in an IR environment with the expectation that it will improve recall and precision of the obtained results in a meaningful way.

## *Acknowledgements*

We wish to thank all partners of the project for their contribution and Dr. Piek Vossen and Prof. Christiane Fellbaum for their support.

## *References*

Daelemans W., Zavrel J. (1999) “*Recent Advances in Memory-Based Part-Of-Speech Tagging*” In Actas del VI Simposio Internacional de Comunicacion Social, Santiago de Cuba pp.590-597, pub:ILK-9903

Hoppenbrouwers J. (2001) “*Requirements of the Terminology Alignment System”*, Technical Report EuroTerm EDC-2214, D.3.2

Vossen P. (1996) *Right or Wrong: Combining Lexical Resources in the EuroWordNet Project*. In Proceedings of the Euralex Conference, pp. 715-728.

Vossen P., Bloksma L., Peters W., Kunze C., Wagner A., Pala K., Vider K., Bertagna F. (1999) “*Extending the Inter-Lingual-Index with new Concepts”*. Deliverable 2D010, Euro WordNet, LE2-4003

## *Affiliation of the authors*

S. Stamou *{stamou@cti.gr}*, A. Ntoulas *{ntoulas@cti.gr}* and D. Christodoulakis *{dxri@cti.gr}*, can be reached at Databases Laboratory, of Computer Engineering & Informatics Department, Patras University, Greece. J. Hoppenbrouwers *{hoppie@kub.nl}* can be reached at CentER Applied Research, Tilburg University, The Netherlands and M. Saiz-Noeda *{max@dlsi.ua.es}* can be reached at the Department of Software & Computing Systems, Alicante University, Spain