Expanding EWN with Domain-Specific Terminology Using Common Lexical Resources: Vocabulary Completeness and Coverage Issues

Stamou Sofia   Ntoulas Alexandros   Kyriakopoulou Maria   Christodoulakis Dimitris

Abstract

EuroTerm is a multilingual semantic network comprising domain-specific terminology for Greek, Dutch and Spanish, which will be linked to the EuroWordNet lexical database. Two approaches have been widely adopted for the development of WordNets, namely the merge and the expand model. The former is considered as the one that ensures a better representation of language particularities in a lexical database whereas the latter assures sufficient overlap in the coverage of WordNets. For the development of EuroTerm a combination of both models was followed in order to ensure vocabulary completeness and coverage across concepts.

1. Introduction

EuroTerm[1] is an EC funded project (EDC-2214) that aims at developing a multilingual domain-specific lexical database, consisting of individual WordNets in three European languages (Greek, Dutch and Spanish), which will be incorporated into the EuroWordNet (EWN) semantic network. EWN (Vossen, 1996) is a lexical database representing semantic relations among basic linguistic concepts for eight West European languages, which are efficiently linked through the usage of an unstructured set of English concepts, namely the Inter-Lingual-Index (ILI). The main goal of EuroTerm is to expand EWN and the ILI with domain specific terminology by incorporating in it an “Environmental“ domain and an “Environmental Health” sub-domain. The domain-specific WordNets will be stored in a common lexical database, which will be linked to the central EWN database. Deviations from EWN might be due to different structure of the lexical resources and quality of the tools used for the terminology acquisition. Two different models have been used for developing WordNets, the merge and the expand model. The first one implies the independent development of each monolingual WordNet and then their linking to the most equivalent synset in the ILI (Vossen, 1996) whereas the expand model implies the translation of English concepts in other languages. For the development of EuroTerm, a combination of both models has been followed in order to assure compatibility and maximize control over the data across the individual WordNets while maintaining language-dependent differences. More specifically, the actual development of WordNets took place in two phases. During the first phase common English lexical resources were used for the terminology acquisition trying to ensure a common starting point for all languages in terms of quality and quantity of the resources. Consequently, the expand model was followed during the first phase of WordNets’ development. During the second phase the merge model has been adopted in the sense that monolingual lexical resources were used for the extraction of the environmental terminology to be incorporated in the monolingual WordNets. Since compatibility with the EWN database is desirable, the Language Independent Module, the Top-Concept Ontology and the ILI of the EWN are maintained. The only expansion concerns the incorporation of an “Environmental” domain label to the already existing ones. In the following section (2) we briefly present previous work conducted in this area and we continue with a description of our approach towards terminology acquisition (3) with emphasis given on the expand model. Following on from this, a discussion of the obtained results is provided (4) and we pinpoint some coverage and completeness issues. In the remaining sections we present some applications of the EuroTerm domain specific network (5) and we conclude with an overall assessment (6) of our approach.

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

2. Related Work

A feasibility study on the incorporation of domain specific terminology into EWN has been conducted while considering at the same time other kinds of extensions of the ILI records. In order to test how EWN can be extended for a conceptual domain, the ILI was adapted for the domain of computing (Vossen, 1999), which can be accessed via the domain labeling. For achieving that, common English lexical resources were used while terminology in other languages relied, as much as possible, on individual resources containing translational equivalencies for the selected computer terms. The final set of the selected terms consists of 444 terms most of which were assigned the computing domain label, and the remaining were spread under six sub-domains of computing (Vossen, 1999), which were then added under the EWN "Computer_Terminology" label.

3. Applying the Expand Model for Terminology Acquisition: the selection phase

The basic idea during the first phase of terminology acquisition is that terms to be added into the EWN database under the "Environmental" domain and "Environmental Health" sub-domain labels would be selected and verified from common English resources (both corpora and lexica) so that there is a common starting point in terms of quantity and quality. By starting off with a common set of terms we ensure that the core of the individual WordNets are richly encoded and comparable, since they have the same conceptual coverage. The corpora used for the terminology extraction, consist of 429 environmental documents comprising a total of 1,733,869 terms, and were manually collected from the following URLs:

- EEA, European Environment Agency (http://eea.eu.int)
- EPA, United States Environmental Protection Agency (http://www.epa.gov)
- Greenpeace (http://www.greenpeace.org)
- NOAH, New York Online Access to Health (http://www.noah-health.org)
- NRDC, Natural Resources Defense Council (http://www.nrdc.org)

The need for domain specific glossaries is more apparent however, when dealing with large amounts of relatively unstructured information, stored in various formats. Thus, various English glossaries comprising 4,972 environmental terms in total with their glosses have been collected in order to facilitate checking and verification of the quality and coverage of the terminology extracted from the corpora. For the terminology extraction the selected documents were converted to plain ASCII format, were POS tagged[2] in accordance with the Penn Treebank POS tags (Santorini, 1990) and lemmatized in order to ensure that all individual lemmas are detected even if they share common wordforms with each other. Once the previous process was completed we obtained a list of terms found in the corpora followed by their lexical category and lemma. The list was in the following format: *<word>* *<tag>* *<lemma>*. Lemmatization process did not deal with single terms only, but with compounds as well. As compounds, we considered two or more consecutive nouns and disregarded other compound categories (e.g. noun-adjective) due to time constraints and due to the fact that the environmental WordNet had to contain mainly nouns (~85% of the terms had to be nouns). Lemmatized lists were processed and the occurrence frequency of each term was obtained in the following form for single terms: *<lemma>* *<lemmaCount>* *<lemmaFrequency>* and for compounds: *<word1>*, *<word2>*, *<word3>* *<count>* *<compoundFrequency>*. Examples of the lemmatized and frequency lists are illustrated in ***Table 1*** & ***2*** respectively whereas examples of compound terms along with their frequency weights are illustrated in ***Table 3.*** Extracted terms were sorted in terms of frequency, calculated on the basis of TF*IDF values. TF*IDF metrics include a component based on the frequency of the term in a document (TF)

[2] The POS-tagger and lemmatizer were provided by Center Applied Research (Tilburg University)

and a component based on the inverse of the frequency within the document collection (IDF). Usually, the TF*IDF values are applied on diverse document collections in order to measure relevance of a term for a single document. However, we applied IF*IDF metrics for more than one document collection since we dealt with a quite homogeneous collection in terms of conceptual and vocabulary coverage. The basis for measuring term's importance is based on its frequency weight within the document. Lemmas with high IDF value tend to appear in fewer documents but hold a more precise meaning, whereas lemmas with low IDF scores (e.g. *thing*, *state*) tend to occur in most of the documents. Frequency lists, consisting of ~18,000 terms were tokenized and stop words were eliminated. For the purpose of EuroTerm[3] terms having a POS tag other than noun (NN, NNS), verb (VB, VBD, VBG, VBN, VPZ) or adjective (JJ, JJR, JJS) were considered as stop words and thus excluded from the list of candidate terms.

<word>	<tag>	<lemma>
toxics	NNS	toxic
wastes	NNP	waste
cultivated	VBV	cultivate
forests	NNP	forest
ozone	NN	ozone

Table 1: Lemmatized list example

lemma	LemmaCount	LemmaFrequency	% lemmas[4]
water	6844	0.004286	0.4286%
waste	5679	0.003556	0.3556%
ozone	3613	0.002262	0.2262%
forest	716	0.000448	0.0448%
thing	81	0.000050	0.0050%

Table 2:Frequency list example

First Word	Second Word	Third Word	Count	Compound Frequency
lung	cancer	—	81	0.004036
solid	waste	disposal	60	0,005670
drinking	water	—	56	0,002790
carbon	dioxide	emissions	27	0,002151

Table 3: List of Compounds with frequency weights attached example

One of the criteria applied for the selection of the terminology to be incorporated into the EuroTerm database is the occurrence frequency of terms within the corpus. In addition, the presence of a wordform in the ILI was examined in order to ensure that terms to be included are not already present. In case an environmental term already existed in the ILI an environmental label was attached to it indicating that the underlying term is domain-specific. Otherwise, if a term was present but without an environmental sense then an environmental gloss was attached to it, which was extracted form the English environmental glossaries described above. Incorporation of new terms in the ILI was performed through a Terminology

[3] Only nouns, verbs and adjectives will be incorporated into the EuroTerm domain specific network.

[4] The fourth column gives the percentage of the term-frequencies and it is provided for a better understanding of the figures

Alignment System (TAS) developed within the framework of the project[5]. New terms were added to the ILI through the TAS with term identifiers to the local WordNets. The final set of candidates was translated in each of the three languages with the use of bilingual dictionaries. Then, monolingual synsets were developed for each of the translated terms and were added under the "Environmental" domain label of each WordNet. At the end of the first phase the core monolingual WordNets for the three languages had been developed according to the expand model. The expand model reassured a reasonable level of overlap across monolingual WordNets but there was a risk that terminology might be biased by English lexicalizations since experiments have shown that there is a considerable variation in the way semantic information for equivalent words is coded across languages. In order to overcome such problems we decided to follow the merge model during the second phase of synset development and use monolingual lexical resources for the expansion of the core WordNets, thus achieving a higher degree of consistency. This model is followed in the ongoing work, and some preliminary results[6] show that the quality of the outcome is going to meet our expectations. It should be noted, finally, that we did not follow the merge model from the very beginning due to the hypothesis that specialized concepts limited to a specific domain tend to share a single meaning overcoming thus lexical ambiguity problems. In addition, due to time considerations and due to the fact that domain specific lexical resources were not widely available for the participating languages we decided to start with a common set of English resources and then enrich the extracted terminology with concepts derived from monolingual resources. An assessment of the results obtained during the first selection process is discussed in the following section with emphasis given on vocabulary coverage and completeness.

4. Discussion of the Obtained Results: Coverage and Completeness Issues

According to Hearst surprisingly useful lexical information can be obtained by applying simple analysis techniques on unrestricted texts (Hearst, 1998). However, the set of concepts that need to be covered by a semantic network cannot be readily obtained from a (domain-specific) corpus (Buitelaar, 2001). Thus, our basic hypothesis is that a domain-specific WordNet should comprise of corpus-representative terms in conjunction with terms found in domain-specific lexica. During the first phase of the project, domain-specific sense assignment was semi-automatically performed using a manually constructed environmental corpus. In this section we report on the methodology followed for determining domain-specific relevance of terms extracted from the corpus.

Occurrence frequency is not by itself a sufficient indicator of a term's importance in a document since terms of high frequency tend to hold many senses. As reported by Krovetz and Croft (Krovetz, 1992) word senses have skewed frequency distribution and an anomalous frequency distribution can be useful for determining domain-specific senses of general vocabulary terms. In addition, they found that general vocabulary terms holding also a domain-specific meaning appear to have low frequency but they might also have high semantic ambiguity. In order to overcome the first problem and reassure that the extracted terminology would be representative of the domain of environment we applied the standard TF*IDF values to measure the importance of terms based on the hypothesis that weighting words in inverse proportion to their number of senses should give similar effectiveness to weighting based on inverse collection frequency (Krovetz, 1992). Moreover, in order to deal with semantic ambiguity we extensively used environmental glossaries in order to check which of the extracted terms had an environmental sense. Checking terms against

[5] TAS has been developed by CentER AR with overall responsibility of Dr. Jeroen Hoppenbrouwers

[6] Unfortunately, since results lacked a consistent state at the time of this contribution, they are not shown here.

environmental glossaries was conducted semi-automatically and took place in two phases. During the first phase all terms extracted from the corpus were stemmed,[7] converted to lowercase and then automatically checked (string matching) against glossaries in order to detect which of them were already present in the glossaries. Terms present both in the corpus and the glossaries were considered as candidate terms. During the second phase terminologists manually checked the glosses of the candidate terms as given by glossaries in order to decide which of them held an environmental sense. The candidate terms that had also an environmental sense were the ones that formed the core environmental WordNet. Some preliminary results show that most of the terms extracted from the corpus do hold an environmental sense even if they belong to the general vocabulary. This is justified partly due to the uniformity of the conceptual domain of our corpus. On the other hand lexical ambiguity was dealt on the basis of the domain-specific glossaries against which terms were checked. As pointed out by Pirkola (Pirkola, 1998) terms of special dictionaries are often unambiguous and thus specialized semantic classes limited to a specific context are non-polysemous ones. By using domain specific glossaries we reassured that all terms included in our WordNet would be related to the conceptual domain of environment. In addition, each term included in the ILI had also a gloss attached in order to reassure that the correct sense of the term was present. Another way of measuring conceptual relatedness of a term to a domain could be through term collocations extracted from the corpus. However, we decided to use domain-specific glossaries instead of collocations in order to overcome lexical ambiguity problems and thus maximize conceptual coverage and completeness of the environmental domain. Through the proposed approach, vocabulary completeness and coverage can be assured since a combination of term frequencies extracted from corpora and information from domain-specific lexica were used for selecting terminology representative of a conceptual domain.

5. NLP Application of EuroTerm

EuroTerm can be used as a resource for semantic information in many NLP applications varying form Information Retrieval (IR) to dictionary publishing as a means to separate generic from domain-specific vocabulary. One envisaged application concerns the incorporation of the domain specific network in IR systems in order to test its performance against domain-dependent text retrieval. Our objective focuses on using EuroTerm database to index documents and queries not in terms of wordforms but in terms of their conceptual meaning. The main idea we adopt against conceptual indexing is targeted towards a semantic representation of documents in the index and their mapping to the *correct* synsets of the environmental domain. Consequently, we aim at conceptual text retrieval as opposed to exact keyword matching, since a document might be relevant to a search request even if it does not use the same words with the query. EWN has shown a potential for IR but the lack of word-sense disambiguation still handicaps the development of concept-based text retrieval (Gilarranz, 1997). A possible explanation for the limited performance of EWN in IR tasks might be the differentiation of conceptual equivalencies across languages, which in some cases account for diverging mappings from local WordNets to the ILI concepts, meaning that conceptual equivalencies are sometimes linked to distinct ILI concepts reflecting different senses of the same word (Peters, 1998). However, in EuroTerm such problems are rarely faced firstly due to the semantically restricted domain of concepts it contains and secondly due to the fact that the core individual WordNets are based on a common set of environmental terms extracted from common resources. Thus, the environmental network can be directly used in IR applications as a way of clustering semantically related concepts, resulting in better precision scores of the obtained results when it comes to domain-specific text retrieval.

[7] Stemming included only suffix-stripping, i.e. we simply eliminated common suffixes of terms (e.g. –s, -es).

Nevertheless, EuroTerm will not solve all problems related to IR but at least an integrated WordNet that anchors specialized terminology in generic vocabulary opens wider possibilities to develop applications for non-expert users.

6. Conclusions and Future Plans

We have presented a combination of the merge and expand models followed for enriching EWN with domain specific terminology achieving maximal overlap and compatibility across languages. Also, we discussed how vocabulary coverage and completeness can be assured when using common resources as a starting point for building semantic networks and how the latter can be incorporated in an IR environment. Future work includes evaluation of the two models in order to conclude on their contribution to domain-specific terminology acquisition.

Acknowledgements

We wish to thank all partners of the project, namely Center for Applied Research (Tilburg University) and Department of Software & Computing Systems (Alicante University) for their valuable contribution. We would also like to thank the reviewers for their comments and Dr. Piek Vossen and Prof. Christiane Fellbaum for their support.

References

Buitelaar P., Sacaleanu B. (2001) *Ranking and Selecting Synsets by Domain Relevance* In Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations, NAACL 2001 Workshop, Carnegie Mellon University, Pittsburg, 3-4 June 2001

Gilarranz J., Gonzalo J., Verdejo F., Stanford CA (1997) *An Approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database*. In “Working Notes of AAAI Spring Symposium of Cross-Language Text and Speech Retrieval”

Hearst M. (1998) *Automated Discovery of WordNet Relations.* In “WordNet: An Electronic Lexical Database” pp.131-151, ed. Christiane Fellbaum, MIT Press.

Krovetz R. Croft B. (1992) *Lexical Ambiguity and Information Retrieval*. In “Proceedings of the CAN Transactions on Information Systems”. Vol. 10(2) pp.115-141.

Peters W., Peters I., Vossen P., (1998) *Automatic Sense Clustering in EuroWordNet*. In “Proceedings of the LREC Conference”

Pirkola A. (1998) *The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval*. In “Proceedings of the 21st Annual International ACM SIGIR Conference”, Melbourne Australia, pp.55-63.

Santorini B. (1990) *Part-Of-Speech Tagging Guidelines for the Penn Treebank Project.* Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

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

Sofia Stamou *{stamou@cti.gr}*, Alexandros Ntoulas *{ntoulas@cti.gr}*, Maria Kyriakopoulou *{kyriakop@cti.gr}* and Dimitris Christodoulakis *{dxri@cti.gr}* can be reached at the Databases Laboratory, of Computer Engineering & Informatics Department, Patras University, GR26500, Greece.