Language-independent text retrieval with the EuroWordNet Multilingual Semantic Database

Julio Gilarranz, Julio Gonzalo, Felisa Verdejo

DIEEC, UNED
Ciudad Universitaria, s.n.
28040 Madrid - Spain

[jtejada, julio, felisa]@ieec.uned.es

Abstract

The EC-funded EuroWordNet (EWN) project aims to build a multilingual database coding basic semantic relations between words for some European Languages (Dutch, Italian, Spanish and English) resembling Princeton WordNet 1.5. In addition to the relations in WordNet 1.5, EWN includes domain labels, cross-category and cross-language relations, which are directly useful for multilingual information retrieval.

The multilingual structure of the database is given by an InterLingual Index that can be used for language-independent indexing of documents and queries in an information retrieval system. We explore a possible application to cross-language text retrieval based on such conceptual indexing, as an interesting alternative to knowledge-based query expansions and corpus-based approaches such as Cross-Language Latent Semantic Indexing.

1 Introduction

The recent explosion of - largely unstructured - online information from the World-Wide Web, online databases, text archives, etc., gives a prominent role to information retrieval techniques. In particular, such flow of information -especially on the Internet - is of a multilingual nature and, thus, cannot be managed assuming a monolingual perspective or any keyword matching approach, regardless of its sophistication. Specific Cross-Language Retrieval (CLTR) techniques are required.

The predominant approaches to CLTR are (Oard, 1997):

- Dictionary-based expansion of queries. Each term in the query is replaced with an appropriate set of terms in the target language. The challenge here is to restrict the ambiguity of translations, to avoid the loss of precision that a naive expansion produces - about 50% with respect to its monolingual counterpart (Davis, 1996)-.
- Approaches based on Parallel Corpora; perhaps the most successful one being Cross-Language Latent Semantic Indexing (Dumais et. al. 1997), where a vectorial representation of queries and docuemtns is arranged according to semantic correlations automatically extracted from corpora. Two limitations usually arise related to these techniques: a) the limited availability of multilingual parallel corpora and b) the worse performance with document collections not directly related to the training corpora.

A large-scale multilingual ontology offers an interesting alternative; namely, the semantic indexing of queries and documents independently from any training corpora.

Our research group at the UNED is involved in the EC-funded EuroWordNet (EWN) project (Vossen 1996), whose purpose is to develop a multilingual database resembling WordNet that stores semantic relations between words in four different languages of the European Community: Dutch, Italian, Spanish and English. The EWN database is generated semi-automatically by using tools and techniques previously developed by the partners to extract information from Machine Readable Dictionaries and other sources.

There is a special interest within EWN project in the utility of such a database for Text Retrieval. The WordNet structure has been extended to include information, such as domain labels, that is of specific interest to TR tasks. In the last stages of the project, *Novell Linguistic Development*, as industrial partner of the project, will test the quality of the final database in their TR software environment.

We are also involved in the Spanish funded ITEM project, that has as a primary goal the creation of a multilingual text retrieval environment featuring Spanish, Catalan and Basque languages. This environment will combine natural language processing approaches with statistical text retrieval techniques.

Both projects started up on 1996 and will last for three years. Some preliminary results of EWN are reported in (Bloksma, Diez-Orzas and Vossen 1996,

Climent, Rodríguez and Gonzalo 1996). We are focused now in the development of our large-scale, multilingual semantic resources.

We first describe in this paper the features of the EuroWordNet database, from the perspective of its potential usage for text retrieval. The we discuss its potential for CLTR purposes, proposing a language-neutral representation of queries and documents in terms of the concepts in the EWN *InterLingual Index*.

2 EuroWordNet : a Multilingual Lexical Knowledge Base

The aim of the EuroWordNet project is to develop (semi-automatically) a multilingual database resembling WordNet that stores semantic relations between words in four different languages of the European Community: Dutch, Italian, Spanish and English. The project began in March 1996 and has a duration of 36 months. Partners involved in the project are the University of Amsterdam (coordinator), the Universities of Pisa and Sheffield, the Universidad Politcnica de Catalunya, the UNED (Spanish National Distance University) and Novell Linguistic Development.

WordNet is a freely available lexical database for English. It consists of semantic relations between English words, which can be accessed as a kind of thesaurus, in which words with similar meanings are grouped together into so-called *synsets* (synonym sets). Besides synonymy (implicit in the definition of synsets), other relations are established between synsets or word forms: hyponymy/hyperonymy (ISA relation), which gives the network a hierarchical structure); meronymy/holonymy (HASA relation) in its *part*, *member* and *substance* variants; and *antonymy* (between opposite word forms). With these relations, the WordNet lexical database is configured as a web of 168,000 synsets (concepts) that contain 126,000 different word forms.

The structure of the monolingual wordnets in EWN follows closely the design of WordNet 1.5, but contains some specific features due to:

- The multilingual nature of the database
- Its potential application for Cross-Language Text Retrieval.
- The Machine-Readable Dictionaries (MRDs) from which the data will be semi-automatically extracted.
- the purpose to achieve maximum compatibility across the different resources.

The main features of the EuroWordNet database, from a text retrieval point of view, are:

- It will contain about 50,000 senses correlating the 20,000 most frequent words (only for nouns and verbs) in each language. This size is reasonable to experiment with generic, domain-independent text retrieval in a multilingual setting. It is planned to expand the database to a higher level of detail for some concrete domain, in order to test its adequacy to incorporate domain-specific thesauri.
- Each monolingual wordnet will reflect semantic relations as a language-internal system, maintaining cultural and linguistic differences in the wordnets. Although the four languages involve similar linguistic conceptualizations of world-knowledge, they do not match completely. In Dutch, for instance, the water to make coffee and the water to extinguish a fire have different names (koffiewater and bluswater) and are understood as different concepts, whereas for Italian, English and Spanish such differentiation does not exist.
- All wordnets will share a common top-ontology. Whereas the wordnets will be extracted semi-automatically from different resources, the common top-ontology have been manually derived, and it is the result of a deep discussion among all partners of the consortium. The criteria to reach this common ontology are of a practical nature; no theoretical claims are made about it.
- Precise criteria have already been defined (Climent, Rodríguez and Gonzalo 1996, Alonge 1996) for the relations and structures, with linguistic tests for every relation in every language. Together with the common top-ontology, these criteria should guarantee compatibility and uniformity between individual wordnets.
- Synsets will have domain labels. In WordNet, concepts as "tennis shoes" and "tennis racquet" are not related. Such associations will be possible in EuroWordNet through domain labels, and they should improve recall for a wide range of queries.
- Nouns and verbs will not be separate networks. EWN includes the cross-part-of-speech relations:
  - noun-to-verb-hypernym: angling → catch (from angling: sport of catching fish with a hook and line)
  - verb-to-noun-hyponym: catch → angling
  - noun-to-verb-synonym: adornment → adorn (from adornment: the act of adorning)
  - verb-to-noun-synonym: adorn → adornment
- Again, these relations establish links that are significant from the point of view of text retrieval. In particular, adorn and adornment are equivalent for retrieval purposes, regardless of their different categories.
- It will contain multilingual relations from each individual wordnet to English (WordNet 1.5) meanings. Such relations will form an Interlingual Index (ILI) whose structure is still open to discussion.

2.1 Multilingual structure of the EuroWordnet database

The building of the EuroWordnet database consist on an iterative process, each of its steps taking place in two phases:

a) Production of the correct language-internal structures, mainly from a monolingual perspective.
b) Comparison and, if necessary, modification of the monolingual wordnets or the equivalence relations. Connection of wordnets via an InterLingual Index (ILI).

Each monolingual wordnet is produced semi automatically from different resources depending on the site. The strategies to build the wordnets vary from site to site, depending on the different resources available. For instance, the Dutch site makes use of a database of Van Dale in which some of the relations for Dutch are already explicitly coded between senses, while the Spanish site first translates WordNet 1.5 into Spanish (using bilingual dictionaries, taxonomies and corpora) and then post-process the result.

Base Concepts and Top-Ontology Clustering

Nevertheless, a global methodology guarantees that the different wordnets will be compatible and a that a coherent multilingual structure for the final database will be achieved. The first step in the process of creating coherent wordnets across the sites involved the selection of a set of Base Concepts that represent the most important concepts in each resource (the ones having more relations and prominent places in the hierarchy). This set is being used as a common starting point for the individual wordnets. The specific selections by each site have been manually compared and merged to get the set of common Base Concepts.

From a total of 4257 noun senses and 1873 verb senses (summing up selections from all sites), we have decided to select all senses that occur in at least two selections: 497 noun synsets and 131 verb synsets (471 senses) of WordNet 1.5.

The set of Base Concepts has been clustered by creating a top-ontology of 87 concepts to which each one of the Base Concepts have been related. This ontology was initially based on the top-classifications in WordNet 1.5, Acquilex, Sift and Aktions-Art models. Next, the base concepts have been classified by linking them to the closest and most relevant ontology class. This last step led to some restructuring of the ontology to represent the diversity of concepts. This ontology has been cross-checked again with the base concepts for each site.

This way, the set of base concepts have been done manually with a maximum of consensus and overlap. These meanings occupy major hierarchical positions and has a large number of relations. Therefore, they play a major role in the database, and it is crucial that coverage and interpretation are the same across sites. Once this compatibility is guaranteed, the risks of divergence of the monolingual wordnets for the vertical extensions of the database are considerably lower.

Interlingual Index

Equivalence relations between the synsets in different languages have made explicit through and Interlingual Index. The starting point for such index has been WordNet 1.5. The Interlingual Index links to WordNet 1.5:

- Synsets for every individual wordnet.
- The top-concept ontology.
- The hierarchy of domain labels which relate concepts on the basis of scripts or topics (e.g. “sports”, “winter sports”, “computers”, etc.).

The consortium has considered two ways of building the ILI in order to enhance language-neutrality (see Figure 1):

- The first option is to introduce new synsets in the WordNet 1.5 structure when a language-specific sense does not have a corresponding WordNet 1.5 synset. Then every language-specific synset could be linked to such modified version of WN 1.5 by simple cross-language synonym links. For instance, a new concept *“finger 1 or toe 1”* is introduced to reflect Italian *dito 1* and Spanish *dedo 1* . Another concept *head 2 (head of a human)* is introduced to reflect dutch lexicalization *hoofd 1*. Then dito 1, dedo 1 and hoofd 1 can be linked to the ILI via simple cross-language equivalence links.
- The second option is to keep the ILI as an unstructured set of links to WordNet 1.5, expressing the different conceptualizations via complex cross-language translation links. For instance, *dito 1* and *dedo 1* would both be linked to *toe 1* and *finger 1* by means of cross-language hyperonym links.

Actually the second option wapreferred, as each monolingual wordnet can be linked to WN 1.5 independently; no consensus is needed to introduce new concepts in WN 1.5 as an interlingual index. The drawback is that the result is not a true interlingua, the cross-relations between languages are more difficult to establish and, in general, the resulting database is less informative.

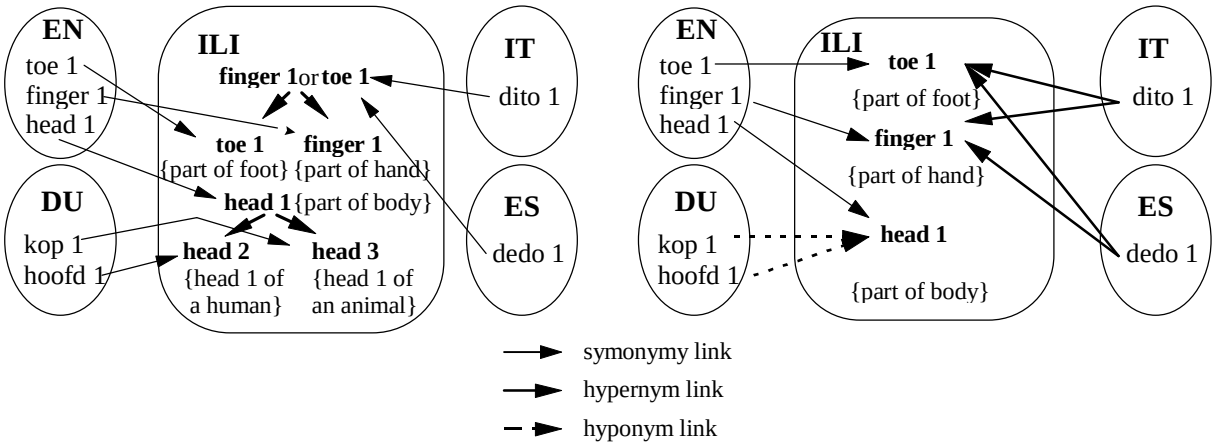

Figure 1: Alternative Structures for the EuroWordNet Interlingual Index

In any case, the Interlingual Index makes the EWN database an excellent resource to perform cross-language, conceptual text retrieval.

The project includes a final phase in which Novell Linguistic Development will make a demonstration of the database within their Information Retrieval System. The main focus of the project, though, is the development of the database itself, which may serve as well as backbone of any semantic database, as a starting point for large lexical knowledge bases, as a source of semantic information to improve grammar and spelling checkers, etc.

3 Conceptual indexing in terms of EuroWordNet

3.1 WordNet and monolingual text retrieval

The semantic content and the large coverage of WordNet makes it a promising tool to perform conceptual text retrieval (as opposed to exact keyword matching). Two different kinds of applications have been attempted in monolingual text retrieval:

- Expansion of the query to include synonyms and other semantically related words. It is thus possible to retrieve documents conceptually related to the query, even in the absence of the particular terms of the query.
- Comparison of queries and documents not only by weighting co-ocurring words, but measuring the semantic similarity of query and document sets of indexes.

Novell experiments within the EuroWordNet project (Bloksma, Diez-Orzas & Vossen 1996) show that query expansion with WordNet can significantly increase recall, but also decreases precision. The results improve when queries are manually disambiguated, but it is unfeasible to perform manual disambiguation of indexed documents.

The reasons for such decrease in precision are obvious: if the wrong meaning for a polysemous word is chosen, the query will be expanded with words totally unrelated to the original meaning. Actually, the major problem to get good results is the absence of a reliable method for word sense disambiguation. However, (Smeaton, Kelledy & O'Donnell 1995) experimented with expanded queries where the original words in the query had been removed in the context of TREC-4, finding 347 out of 6501 relevant documents that contained no query terms, showing that the role of semantically related words to enhance recall might be crucial.

On the other hand, retrieval based on measuring semantic distances has not given good results, mainly because of the absence of a good definition of semantic distance over wordnet-like structures. Besides that, WordNet 1.5 structure itself is quite unbalanced, introducing additional noise in the measures.

In spite of the absence of good word-sense disambiguation methods and accurate conceptual distance measures, (Smeaton & Quigley 1996) showed that WordNet-based techniques make significant improvements to traditional approaches when working with short documents (image captions in an image database in their experiment). The system was able to relate queries such as "children running on a beach" with image captions such as "boys playing in the sand". The small amount of indexing words made statistical approaches much less reliable.

3.2 WordNet-like structures and cross-language text retrieval

An interesting issue for cross-language text retrieval is that the effect of expanding queries with synonyms should not affect dramatically to precision - in comparison to other CLTR approaches - as, in fact, translating terms from the original language into the target language is already a way of expanding a word into (cross-language) synonyms.

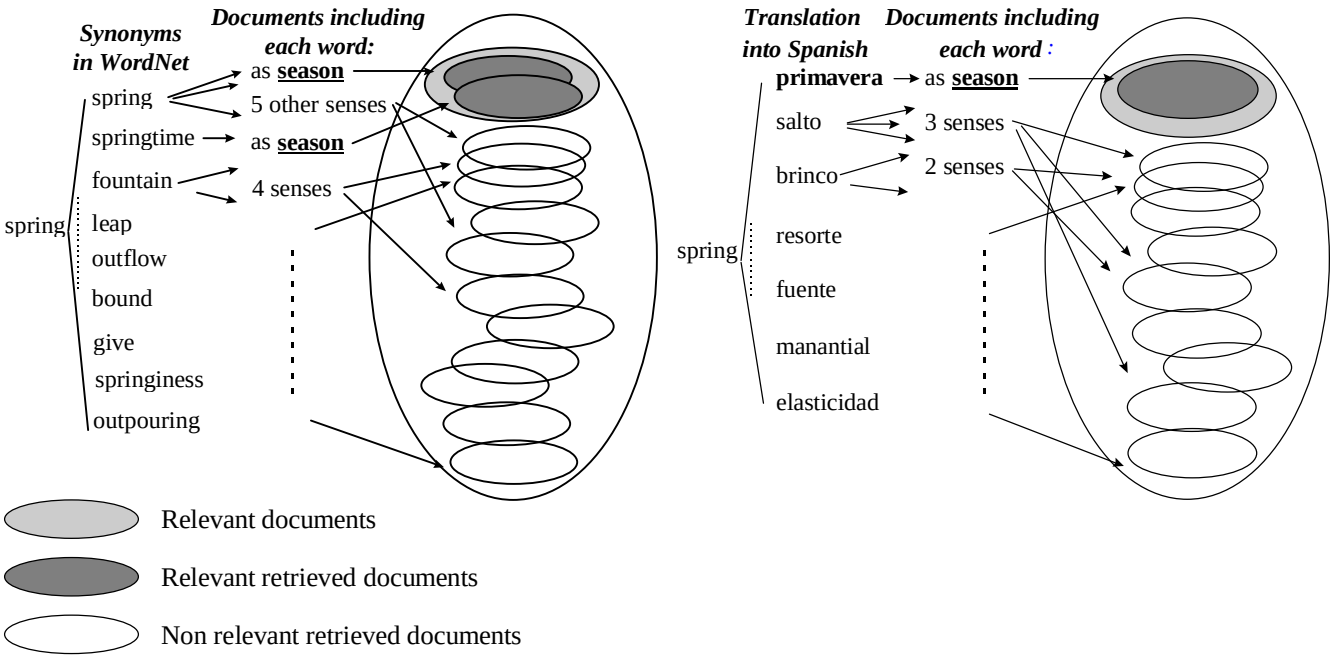

Figure 2: Monolingual and multilingual synonymy expansion

This is illustrated in Figure 1. The left picture shows the effect of expanding *spring* with WordNet 1.5 in a monolingual setting. If the intended meaning for spring *is season of growth*, we can retrieve documents that contain the word *springtime*, even if *spring* is not present, and thus we can potentially increase recall. But a more accused effect is that we are retrieving also documents containing *fountain, leap, outflow, bound, give, springiness, outpouring* ... all synonyms for inadequate senses of *spring*. The traditional problem of polysemy in text retrieval grows enormously if we also take synonyms for incorrect acceptions of the word form. In the overall, precision decreases significantly.

However, in the picture of the right side we see that the same effect is obtained in any knowledge-based multilingual setting: to go from one language to another, seeking for translations of *spring* leads us to many -say- spanish words from which only one corresponds to the intended meaning - *primavera* - while *salto*, *brinco*, *resorte*, *fuente*, etc.. correspond to unrelated concepts (*leap*, *fountain*, etc).

Therefore, CLTR based on multilingual versions of WordNet should not have worse precision, a priori, than any other knowledge-based approach to CLTR. This makes wordnet-based text retrieval even more interesting in a multilingual environment.

3.3 Our approach: conceptual indexing in terms of the Interlingual Index

Our aim is to use the EWN database to index documents and queries, not in terms of word forms or language-specific synsets, but in terms of the EWN Interlingual Index. As we explained above, the Interlingual Index is essentially WordNet 1.5 plus a set of complex translation links that relate language-specific synsets and WordNet synsets. We plan to assign a vectorial representation of every query and document in the space of the Interlingual Index synsets, by means of the translation relations. The core system could perform comparisons with traditional vectorial techniques; the difference is that the indexing space is a language-neutral set of concepts.

Such setting would be a truly multilingual one, rather than a set of cross-language techniques. Language-specific techniques (stemmers, etc) would extract the relevant terms for documents and queries. Word-sense disambiguation would match terms against EWN language-specific synsets, and the ILI would provide a - to some extent - language independent vectorial representation. Every comparison of queries and documents can be made in this representation space. Such approach has the potential of combining multilingual, concept-based retrieval with well-known vectorial techniques.

It is also a good framework to experiment with more sophisticated forms of measuring semantic similarity between documents and queries.

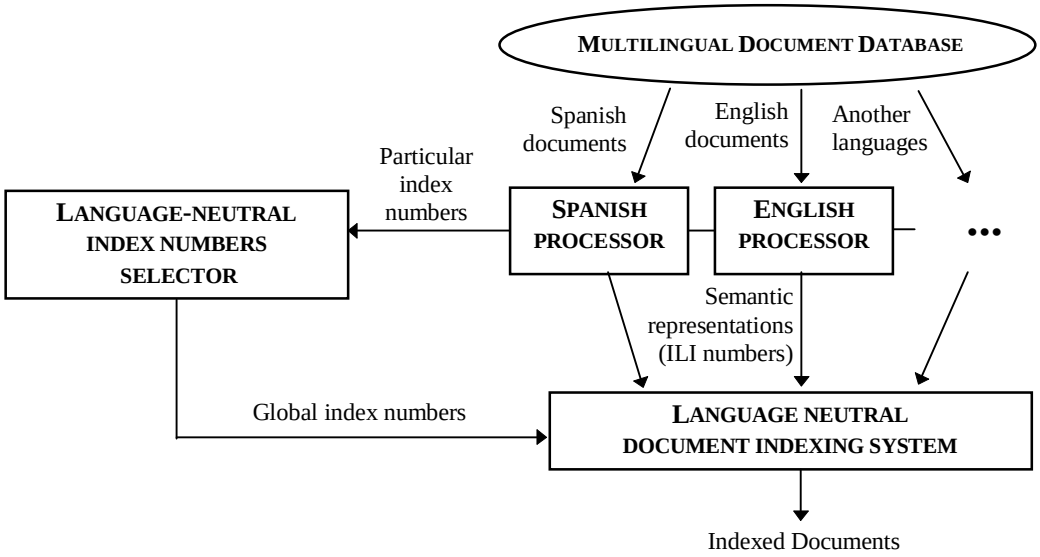

Figure 3: The indexing process

The indexing process for every document is sketched in Figure 2 and has a language-dependent stage (shown in Figure 3) and a language-independent one:

A: Language-dependent processing

1. POS tagging (we will only consider nouns and verbs). We will also consider the possibility of shallow, statistically driven parsing with tools from the ITEM project. The ITEM and EuroWordNet projects offer a good environment to experiment and test the usage of NLP techniques for Information Retrieval, though we are aware of the difficulty of getting good results in this area.
2. Search for canonical forms of words (stemming and reconstruction). We plan to use morphological analysis tools from EuroWordNet and ITEM projects.
3. Mapping of word forms into EuroWordNet (into the particular wordnet for that specific language). This process implies high quality word-sense disambiguation, which remains to be an open research question. We plan to use and adapt for our purposes the notion of *Conceptual Distance* as it is presented in (Aguirre & Rigau 1995) for word-sense disambiguation using WordNet. The *conceptual distance* measure proposed there is sensitive to a) depth in the hierarchy, b) density of the related hierarchy and c) lenght of the shortest paths between concepts. It is defined for sets of concepts and it is independent of the size of that sets, and it gives a promising accuracy on unrestricted texts.
4. Mapping into the language-neutral InterLingual Index. According to the EWN arquitecture, this is just a matter of following the cross-language translation links between each individual wordnet and WordNet 1.5 (as ILI).

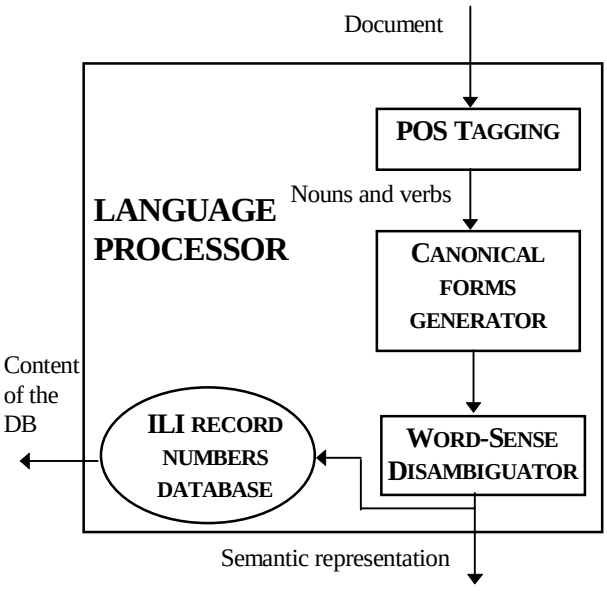

Figure 4: The language processor

B: Language-independent processing

1. Selection of relevant synsets for indexing. Besides removal of stop words and those having a high interdocument frequency, the EuroWordNet hierarchy may provide additional criteria to discard irrelevant synsets. For instance, being too high in the hierarchy or being unrelated to the rest of synsets could contribute to discard a synset. An interesting possibility is to develop a general stop list of synsets that would be applicable regardless of the language. Such list would contain too general meanings.

2. Vectorial representation and weighting. Occurrences of synset numbers will be weighted by
    - frequency of the synset in the document.
    - overall frequency of the synset in the collection of documents.
    - position of the synset within the EuroWordNet hierarchy.

The result is a language-independent, conceptual representation of the document.

Though the representation achieved should hopefully be a conceptual one, it is formally just a traditional vectorial representation. We will, at least in a first stage, compare queries and documents with a traditional vector comparison approach. We find some reasons to do so:

- This permits considering just closest synonymy between word forms. Experiences with WordNet in monolingual text retrieval indicate that considering other semantic relations introduces too much noise in the representation and may affect drastically to precision.
- It will permit a more accurate comparison to other knowledge-based approaches to CLTR. If we combine conceptual representations (with the noise associated to wrong word-sense disambiguations) and conceptual proximity comparisons (an ellusive concept that is difficult to tune for text retrieval) it will be difficult to evaluate the results. A separate evaluation of both issues seems more reliable.
- It will also permit a direct comparison to Cross-Language Latent Semantic Indexing (Dumais *et al*. 1997), a corpus-based approach that uses a vectorial representation arranged according to semantic correlations automatically extracted from corpora. The positive results of this technique challenges conceptual retrieval based on a large-scale multilingual thesaurus. An interesting question is whether the two approaches are incompatible, or if they can be combined somehow.

The process to extract the representation of a query, in contrast to the representation of documents, does not have to be fully automatic. The short lenght of queries permit more sophisticated natural language processing and interactions with the user to refine it. We will experiment with shallow parsing techniques and interactive disambiguation of the queries, balancing the accuracy of the representation with the effort demanded to the user in order to refine his query. A possible mechanism would be:

- Query expressed in natural language by the user.
- Preprocessing of the query to extract relevant synsets.
- Presentation of polysemous words not disambiguated with their associated meanings, to get refinement from the user.
- Expansion of the query. As we already have a semantic representation of it, the need for expansion has to be carefully balanced with the risk of losing precision. A first - reasonable - approach is to include synonym sets of different categories. Other options, which need a careful evaluation, include expansion to hyponyms, hypernims, etc.

After building up the vectorial representation for the query, it can be passed to a conventional information retrieval machine. Documents retrieved should match the user necessities, regardless of the language they are expressed in.

4 Expected Results

The approach that we have proposed here relies on two open issues. The first one is a precise mechanism to perform word-sense disambiguation to get an accurate representation of documents in terms of concepts. This is still an open research question, though the notion of Conceptual Distance that is being used within the EWN project is giving promising results.

The second one is the quality of the EWN database itself: its coverage, generality and homogeneity. And, in particular, the quality of the cross-language relations. Potentially, the structure of the EWN database is more adequate for text retrieval purposes than WordNet itself. However, it will be constructed in only three years with semi-automatic methods, and thus its quality can be estimated but not guaranteed in advance.

Even with these restrictions, our framework has some benefits over previous approaches.

Compared to WordNet-based monolingual text retrieval, Cross-Language retrieval based on the EuroWordNet database benefits from:

a) Cross-POS relations and domain labels, which should improve performance based on conceptual proximity.
b) The expansion to synonyms that caused a decrease of precision in a monolingual setting is common to every knowledge-based approach in cross-language settings; therefore, precision should not decrease when compared with other cross-language approaches.
c) Although the size of WordNet 1.5 is considerably higher, the scope and coverage of EWN will be more balanced and thus, it will have a more predictable behavior for text retrieval, regardless of the domain.

Compared to dictionary-based approaches to CLTR, our proposal has also some additional benefits:

a) It is a truly language-independent setting for text retrieval, rather than a method to expand queries.
b) It provides the structure to perform word-sense disambiguation when indexing documents and queries.
c) It provides language-independent criteria to identify stop words.

Finally, the EuroWordNet database offers some advantages over Cross-Latent Semantic Indexing:

a) It does not require parallel corpora to be trained, which is a strong requirement for more than two languages.
b) Adding languages can be done incrementally.
c) It seems more promising for unrestricted multilingual retrieval as World-Wide Web searches, as it is independent from any training corpora.

Based on these facts, we are convinced that the building of the EuroWordNet database offers an excellent opportunity to experiment with truly multilingual text retrieval.

Acknowledgments

This research is being supported by the European Community, project LE #4003, and the Spanish government, project TIC-96-1243-CO3-O1. Julio Gilarranz is supported with a grant from the Spanish Ministerio de Educación y Cultura.

References

Aguirre E.; and Rigau G. 1995. A Proposal for Word Sense Disambiguation Using Conceptual Distance. International Conference oon Recent Advances in Natural Language Processing. Tzigov Chark, Bulgaria.

Alonge, A. 1996. Definition of the links and subsets for verbs in the EuroWordNet Project. Deliverable d006, EC funded project LE #4003.

Bloksma, L.; Diez-Orzas; P. and Vossen, P. User requirements and Functional Specification of the EuroWordNet Project. Deliverable d001, EC funded project LE #4003.

Climent, S.; Rodríguez, H.; and Gonzalo, J 1996. Definitions of the links and subsets or nouns in the EuroWordNet Project. Deliverable d005, EC funded project LE #4003.

Davis, M. 1996. New experiments in cross-language text retrieval at NMSU's Computing Research Lab. In Harman, D. K., ed., *The Fifth Text REtrieval Conference (TREC-5).* NIST. To appear.

Miller, G., R. Beckwith, C. Fellbaum, D. Gross and K. Miller. 1990. Five Papers on WordNet. CSL Report 43, Cognitive Science Laboratory, Princeton University.

Oard, D. 1997. Alternative Approaches for Cross-Language Text Retrieval. 1997. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence. To appear.

Richardson, R.; and Smeaton, A.F. 1995. Using WordNet in a Knowledge-Based Approach to Information Retrieval. In Proceedings of the BCS-IRSG Colloquium, Crewe.

Vossen, P. 1996. Eurowordnet: building a multilingual wordnet database with semantic relations between words, Technical and Financial Annex, EC funded project LE #4003.

Smeaton, A.F.; Kelledy, F.; and O' Donnell, R. 1995. TREC-4 Experiments at Dublin City University: Thresholding Posting Lists, Query Expansion with WordNet and POS tagging of Spanish.

Smeaton, A.F. and Quigley, A. 1996. Experiments on Using Semantic Distances between Words in Image Caption Retrieval, in Proceedings of the 19th International Conference on Research and Development in IR.