An Approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database

Julio Gilarranz, Julio Gonzalo and Felisa Verdejo
DIEEC, UNED
Ciudad Universitaria, s.n.
28040 Madrid - Spain
[jtejada, julio, felisa]@ieec.uned.es

Abstract

This paper explores the application of a multilingual, WordNet-like lexical knowledge base, *EuroWordNet*, to Cross-Language Text Retrieval. EuroWordNet (EWN) is a multilingual database with basic semantic relations between words for some European languages (Dutch, Italian, Spanish and English). In addition to the relations in WordNet 1.5, EWN includes domain labels, cross-language and cross-category relations, which are directly useful for Multilingual Information Retrieval. We propose conceptual, language-independent indexing on the basis of EWN Interlingual Index (ILI) as a promising approach to cross-language text retrieval.

Introduction

There is a general consensus in the domain of Cross-Language Text Retrieval (CLTR) about the necessity to develop large-scale resources in order to improve the capacities and performance of current CLTR systems (Fluhr 1996; Oard & Dorr 1996). In particular, large-scale databases storing basic semantic relations between words can be used for reformulation of queries, automatic indexing and other central issues of the text retrieval task. Such a database is already available for American English (Miller *et al.* 1990), and similar monolingual databases are being constructed for other languages.

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 using tools and techniques previously developed by the partners to extract information from Machine Readable Dictionaries and other sources. Other statistical techniques on corpora will be applied as well.

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 CLTR 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 information retrieval 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, Díez-Orzas, & Vossen 1996; Climent, Rodríguez, & Gonzalo 1996). We are focused now in the development of our large-scale, multilingual semantic resources, and this symposium is a perfect occasion for us to contact the TR community and exchange ideas about a) potential usage of our resources and b) how to optimize the design of that resources to fit text retrieval needs.

In this paper, we first review previous applications of WordNet to monolingual information retrieval. We show that the decrease of precision that the usage of WordNet seems to imply is not higher than the loss of precision that any knowledge-based CLTR system implies. Thus, wordnet-based approaches should perform comparatively better in a multilingual setting than in a monolingual one. Then we make a brief presentation of EWN focused on its potential for CLTR purposes. Finally, we discuss the possibility of doing conceptual, language-neutral text information retrieval using the EWN database.

WordNet and Information Retrieval

WordNet 1.5 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 related meanings are grouped together.

WordNet classifies word meanings into four lexical categories: nouns, verbs, adjectives and adverbs. Essentially, only relations between meanings in the same category are considered. These are the main ones:

Synonymy : similarity of meaning. The basic units in the WordNet database are synsets - or synonym sets - i.e., groups of word forms (single words or collocations) with the same meaning. Synonymy is, thus, an implicit relationship in WordNet, while most of the remaining relations are defined between senses, i.e., between synsets. It is the most important relation for us, because semantically similar words usually can be interchanged in most contexts.

Hyponymy/hypernymy : ISA relation. A hyponym has all the features of its hypernym and adds at least one distinguing feature. This relationship produces a hierarchical organization of synsets for every category.

Meronymy/holonymy : HASA relation. A meronym is a part, a member or a substance of its holonym in WordNet 1.5.

Antonymy : WordNet 1.5 also considers opposition of meanings, although this relationship - which is defined between word forms, not between synsets - is not a fundamental organizing relation for nouns and verbs.

With these relations, the WordNet lexical database is configured as a web of word meanings. It contains 168,000 synsets with 126,000 different word forms.

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).

Perhaps the most immediate application to information retrieval is expanding the query to include synonyms of the relevant words in the query and other semantically related words. Besides that, WordNet also offers the possibility of comparing queries against documents not only by weighting co-occurring words, but measuring the semantic similarity of query and document sets of indexes.

Actually, the major problem to get good results is the absence of a reliable method for word sense disambiguation. If we choose the wrong meaning for a word in a query, we will expand the query with other words that can be totally unrelated (for instance, if the original query includes "spring" and it refers to springtime, it would only add noise to include the set of synonyms fountain, outflow, outpouring, natural_spring). This problem equally applies to similarity measures: if the proper meanings are not identified, the semantic similarity between query and index may lead to meaningless results.

With respect to query expansion, Novell experiments within the EuroWordNet project (Bloksma, Díez-Orzas, & Vossen 1996) show that query expansion with WorNet can significantly increase recall, but also decreases the precision. The results improve when the queries are manually disambiguated, but it is unfeasible to perform manually disambiguation of indexed documents. In the context of TREC-4, (Smeaton, Kelledy, & ODonnell 1995) experimented with expanded queries where the original words in the query had been removed. The system discovered 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.

The Information Retrieval group at Dublin City University (DCU) is one of the groups that has carried out extensive research on using WordNet for indexing and expansion of queries. Their long-term approach (Smeaton *et al.* 1995) is to index queries and documents by the words which occur within them, but when computing the degree of similarity between query and document they incorporate a quantitative measure of the semantic similarity between index terms. In order to determine word-word similarity, they use Hierarchical Concept Graphs derived from WordNet and augmented with a statistical analysis of word frequency occurrences in a corpus. Their earlier results were discouraging: the system performed worse that traditional approaches, due to the lack of a reliable word-sense disambiguation and to the deficiencies of their measure of semantic similarity. However, they shown in (Smeaton & Quigley 1996) that the system made significant improvements to traditional approaches when working with short documents; in that particular case, image captions in an image database. The system was able to relate queries such as "children running on a beach" with image captions as "boys playing in the sand". The small amount of indexing words made statistical approaches much less reliable.

In brief, WordNet has shown a potential for information retrieval, though the lack of good word-sense disambiguation methods and measures of semantic similarity still handicaps the development of concept-based 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 com-

parison to other CLTR approaches - as, in fact, translating terms from the original language to the target language is already a way of expanding a word into (cross-language) synonyms. This is illustrated in Figure 1. The left picture shows the effect of expanding *spring* with WordNet 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.* 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* lead us to many -say- spanish words, from which only one corresponds to the intended meaning, while *salto, brinco, resorte, fuente*, etc. correspond to different 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.

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 University of Sheffield, the Istituto di Linguistica Computazionale del CNR, the Universitat Politècnica de Catalunya, the UNED (Spanish Distance Learning University) and Novell Linguistic Development.

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 has been manually derived, and it is the result of an in-depth discussion between 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, & 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. Any

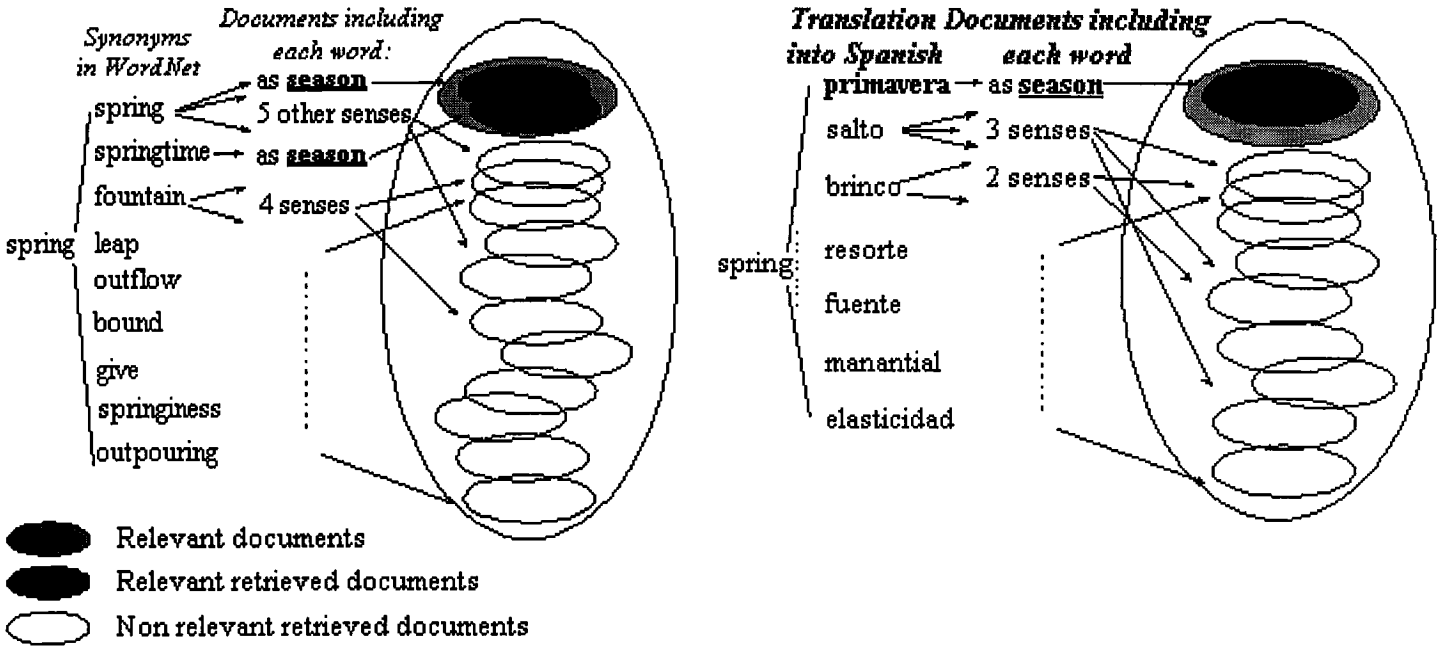

Figure 1: Comparison between expanding a query with WordNet in a monolingual setting and expanding a query with *any* knowledge-based CLTR approach in a multilingual setting (including a multilingual wordnet)

Cross-Language task, including text retrieval, will rely on such Interlingual Index.

The consortium has considered two ways of building the ILI in order to enhance language-neutrality. 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. The second option is to keep WN 1.5 untouched, and permit more complex links between language-specific synsets and WN 1.5, such as hyponym or hyperonym equivalencies.

Actually the second option is preferred, 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. In any case, the Interlingual Index makes the EWN database an excellent resource to perform cross-language, conceptual text retrieval. We see this symposium as an excellent occasion to discuss which setting for multilingual information would better suite multilingual retrieval tasks.

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.

A Proposal for Conceptual Indexing

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 pro-

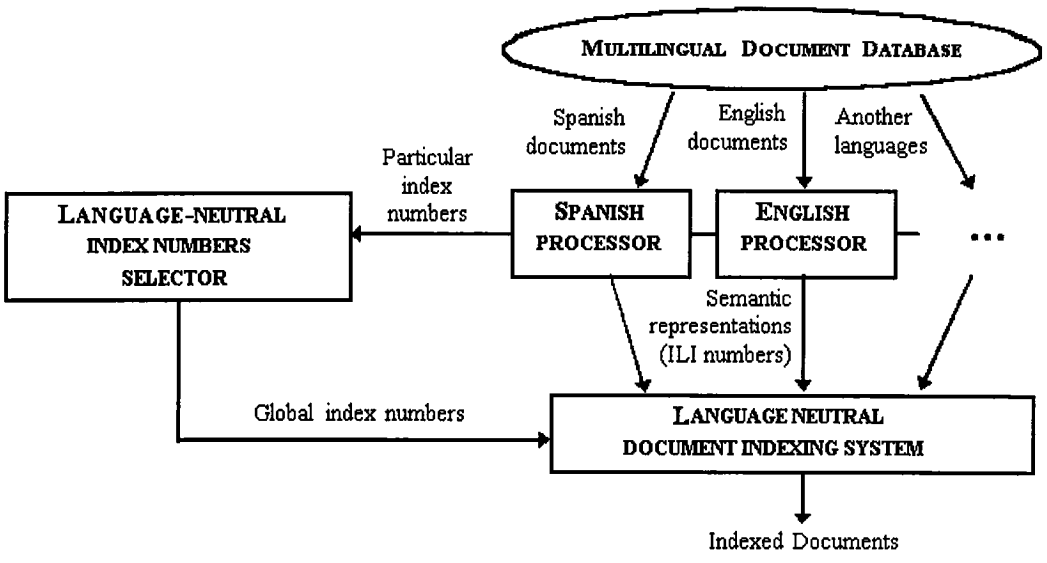

Figure 2: The indexing process

vide 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.

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

- POS tagging (we will consider nouns and verbs only). 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 to get positive results in this area.
- Search for canonical forms of words (stemming and reconstruction). We plan to use morphological analysis tools from EuroWordNet and ITEM projects.
- Mapping of word forms into EuroWordNet (into the 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 propsed 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.
- Mapping into the language-neutral InterLingual Index. According to EWN architecture, this is just a matter of following the cross-language translation links between each individual wordnet and WordNet 1.5 (as ILI).

B: Language-independent processing

- 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.
- 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 a 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 anyhow.

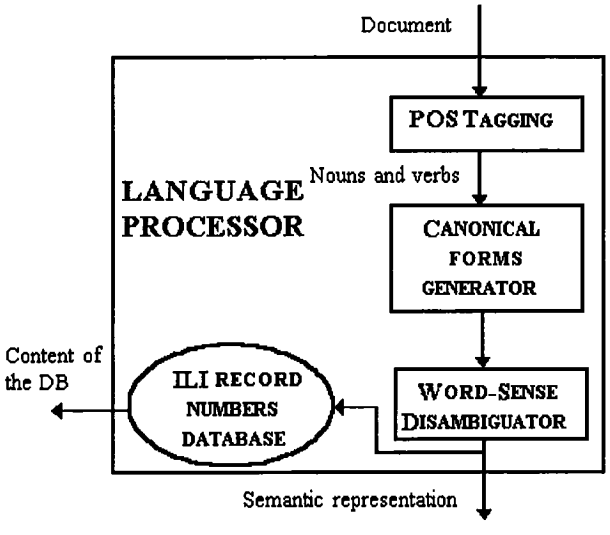

Figure 3: Language-dependent part of the indexing process.

The process to extract a representation of the query, in contrast to the representation of documents, does not have to be fully automatic. The short lenght of queries permits more sophisticated natural language processing and interactions with the user to refine it. We will experiment with shallow parsing 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 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, hypernyms, 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.

Expected Results

As we have explained previously, CLTR based on EuroWordNet should perform better than its monolingual counterpart, as the decrease in precision that synonymy expansion produces affects to every knowledge-based approach in a cross-language environment. The conceptual indexing that the EWN database facilitates offers, thus, a promising way of increasing recall while keeping, at least, standard precision rates.

The approach that we have proposed here relies on two main issues. The first one is a precise mechanism to perform word-sense disambiguation to get an accurate representation of documents in terms of concepts. The second one is the quality of the EWN

database itself: its coverage and 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 only in three years with semi-automatic methods, and thus its quality can be estimated but not guaranteed in advance.

In comparison with Cross-language Latent Semantic Indexing, which also performs some kind of conceptual retrieval, EWN offers the advantage that there is no need for a multilingual parallel corpus that covers simultaneously every language considered (this is actually a strong requirement for more than two languages). In EWN, every language can be linked independently to the structure of the database. It is, in fact, independent from any training corpus, so it seems more promising in unrestricted multilingual retrieval as, for instance, World-Wide-Web searches. As a drawback, its performance relies on accurate word-sense disambiguation and on the quality of the final database, which are still open questions.

Regardless of future results, however, we are convinced that the design 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. In *International Conference on Recent Advances in Natural Language Processing.*

Alonge, A. 1996. Definition of the links and subsets for verbs in the eurowordnet project. Technical report, Deliverable D006, EC-funded project LE # 4003.

Bloksma, L.; Díez-Orzas, P.; and Vossen, P. 1996. User requirements and functional specification of the eurowordnet project. Technical report, Deliverable D001, EC-funded project LE # 4003.

Climent, S.; Rodríguez, H.; and Gonzalo, J. 1996. Definition of the links and subsets for nouns in the eurowordnet project. Technical report, Deliverable D005, EC-funded project LE # 4003.

Dumais, S.; Letsche, T.; Littman, M. L.; and Landauer, T. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Symposium on Cross-Language Text and Speech Retrieval.* American Association for Artificial Intelligence.

Fluhr, C. 1996. *Survey of the State of the Art in Human Language Technology.* Center for Spoken Language Understanding, Oregon Graduate Institute. chapter Multilingual Information Retrieval.

Miller, G.; Beckwith, C.; Fellbaum, D.; Gross, D.; and Miller, K. 1990. Five papers on wordnet, csl report 43. Technical report, Cognitive Science Laboratory, Princeton University.

Oard, D. W., and Dorr, B. 1996. A survey of multilingual text retrieval. Technical report, UMIACS-TR-96-19 CS-TR-3615.

Smeaton, A., 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.*

Smeaton, A.; Kelledy, F.; ODonnell, R.; Quigley, I.; and Townsend, E. 1995. Using linguistic resources or language processing. In *Proceedings of IA95, Montpellier.*

Smeaton, A.; Kelledy, F.; and ODonnell, R. 1995. Trec-4 experiments at dublin city university: Thresholding posting lists, query expansion with wordnet and pos tagging of spanish. In *Proceedings of TREC-4.*

Vossen, P. 1996. Eurowordnet: building a multilingual wordnet database with semantic relations between words. technical and financial annex. Technical report, EC-funded project LE # 4003.