Applying EuroWordNet to Multilingual Text Retrieval

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

Carol Peters
Istituto di Elaborazione della Informazione
Consiglio Nazionale delle Ricerche
Via S. María, 46
I-56126 Pisa - Italy
carol@iei.pi.cnr.it

Nicoletta Calzolari
Istituto di Linguistica Computazionale del CNR
Via della Faggiola 32
56100 Pisa - Italy
glottolo@ilc.pi.cnr.it

(Received ..... ; Accepted in final form .....)

Abstract.
We discuss ways in which EuroWordNet (EWN) can be used in multilingual information retrieval activities, focusing on two approaches to Cross-Language Text Retrieval that use the EWN database as a large-scale multilingual semantic resource. The first one indexes documents and queries in terms of the EuroWordNet Inter-Lingual Index, thus turning term weighting and query/document matching into language-independent tasks. The second one integrates the information in the EWN database with corpus-based techniques, thus allowing retrieval for domain-specific terms that may not be present in our multilingual database. They show the potential of the EWN database as a promising alternative to existing approaches (dictionary-based and corpus-based) to Cross-Language and Multilingual Text Retrieval.

Key words: Cross-Language Text Retrieval, Multilingual lexical resources, Large-Scale Ontologies

Abbreviations: EWN – EuroWordNet; CLTR – Cross-Language Text Retrieval; ILI – InterLingual Index

1. Introduction

The current challenge for Natural Language Processing (NLP) technology is to demonstrate its cost-benefit for real-world tasks. Although it is widely recognised that reusable linguistic resources and tools are needed to address the complexity of building large-scale NLP applications, our task now is to reduce the distance between the research lab and the marketplace. Experience tells us that already available resources can be exploited by software developers to improve their products by including, for instance, some language component. A good example of this is the growing number of applications using WordNet mentioned in the literature, since it first became freely available.

Multilinguality has been a central concern for the very beginning of NLP research. Substantial efforts were dedicated to machine translation, perceived as a major priority in the building and exploiting of applications. The international spread of computers in both the working place and at home, together with the dramatic increase of computer network users, are opening opportunities for many other applications involving multilingual resources. To mention just a few, the improvement of access to international multimedia information services, the enhancement of language learning aids, the assisted production of multilingual manuals, technical or legal documents, the automatic generation of reports in highly structured domains, such as traffic bulletins or weather information. For all of them, a generic multilingual database with wordnets for different languages will be a key resource.

One of the most interesting applications for the EuroWordNet (EWN) database is probably in the area of Cross-Language Text Retrieval (CLTR). The explosive growth of universally accessible information over the international networks - information that is unstructured, heterogeneous and multilingual by nature - has made CLTR one of the most compelling challenges that information technology has to face nowadays. For this reason, we have decided to focus the last article in this collection on a description of how it will be possible to apply the EuroWordNet multilingual database in CLTR activities.

The main approaches to CLTR being experimented today use either knowledge-based or corpus-based techniques:

Knowledge-based approaches apply bilingual or multilingual dictionaries, thesauri or general-purpose ontologies to cross-language text retrieval.

USING THESAURI: So far, the best known and tested approaches to CLTR are thesaurus-based, although they are generally used in controlled-text retrieval, where each document is indexed (mainly by hand) with keywords from the thesaurus. A thesaurus is an ontology specialized in organizing terminology; a multilingual thesaurus organizes terminology for more than one language. ISO 5964 gives specifications for the incorporation of domain knowledge in multilingual thesauri and identifies alternative techniques. There are now a number of multilingual thesaurus-based systems available commercially. However, although the use of such thesauri has been shown to give good results for CLTR (Salton, 1970) thesaurus construction and maintenance is expensive, and training is required for optimum usage. In addition, domain-specific thesauri are not very useful outside of the particular domain for which they have been designed.

USING DICTIONARIES: Some of the first methods attempting to match the query to the document for free-text (as opposed to controlled-text) retrieval have used bilingual dictionaries. It has been shown that dictionary-based query translation, where each term or phrase in the query is replaced by a list of all its possible translations, represents an acceptable first pass at cross-language information retrieval although such – relatively simple – methods clearly show performance below that of monolingual retrieval. Automatic machine readable dictionary (MRD) query translation has been found to lead to a drop in effectiveness of 40-60% of monolingual retrieval (Hull and Grefenstette, 1996; Ballesteros and Croft, 1996). There are three main reasons for this: general purpose dictionaries do not normally contain specialized vocabulary; failure to translate multiword terms; and the presence of spurious translations (a problem derived from polysemy which is further discussed in section 2).

Corpus-based approaches : The above considerations have encouraged an interest in corpus-based techniques in which information about the relationship between terms over languages is obtained from observed statistics of term usage. Corpus-based approaches analyze large collections of texts in multiple languages and automatically extract the information needed to construct application-specific translation techniques. The collections analyzed may consist of parallel (translation equivalent) or comparable (domain-specific) sets of documents. The main approaches that have been experimented using corpora are vector space and probabilistic techniques. A recent, comparative evaluation of some representative approaches to free-text retrieval (Carbonell et al., 1997) has shown that corpus-based approaches - and in particular some applications of example-based machine translation - significantly outperform simple dictionary-based term translation.

The first tests with parallel corpora were on statistical methods for the extraction of multilingual term equivalence data which could be used as input for the lexical component of MT systems. Some of the most interesting recent experiments, however, are those using a matrix reduction technique known as Latent Semantic Indexing (LSI) to extract language independent terms and document representations from parallel corpora (Dumais et al., 1996). LSI applies a singular value decomposition to a large, sparse term document co-occurrence matrix (including terms from all parallel versions of the documents) and extracts a subset of the singular vectors to form a new vector space. Thus queries in one language can retrieve documents in the other (as well as in the original language).

The problem with using parallel texts as training corpora is that test corpora are costly to acquire – it is difficult to find already existing translations of the right kind of documents and translated versions are expensive to create. For this reason, there has been a lot of interest recently in the potential of comparable corpora. A comparable document collection is one in which documents are aligned on the basis of the similarity between the topics they address rather than because they are translation equivalent. Methods have been studied to extract information from such corpora on cross-language equivalences in order to translate and expand a query formulated in one language with useful terms in another (Sheridan and Ballerini, 1996; Picchi and Peters, 1996). Again, as with the parallel corpus method reported above, it appears that this method is very application dependent. A new reference corpus would have to be built to perform retrieval on a new topic.

From this discussion, we can conclude that any single method currently being tried presents limitations. Existing resources – such as electronic bilingual dictionaries – are normally inadequate or insufficient for the purpose; the building of resources such as domain-specific thesauri and training corpora is expensive and such resources are generally not fully reusable; a new multilingual application will require the construction of new resources or considerable work on the adaptation of previously built ones.

It should also be noted that most of the systems and methods in use so far concentrate on pairs rather than multiples of languages. This is hardly surprising. The situation is far more complex when an attempt is made to achieve effective retrieval over a number of languages than over a single pair; it is necessary to study some kind of interlingual mechanism – at a more or less conceptual level – in order to permit multiple cross-language transfer. This is why we feel that a large-scale, multilingual semantic database such as EWN offers an interesting alternative: performing conceptual, language-neutral retrieval without training and without parallel corpora. This paper thus discusses ways in which EWN can be used in order to achieve this goal, and discusses why it should work and what can be learned from it.

First, we consider previous WordNet-based monolingual text retrieval experiments. Although they have not been a complete success, we show that a cross-language scenario can keep the advantages of a Word-

Net approach without its major drawback (an unnecessary excerbation of problems of polysemy). Section 3 review the major features of the EWN database from the point of view of Text Retrieval, highlighting its advantages over WordNet 1.5.

In sections 4 and 5, we discuss two proposals that use EWN for CLTR in different retrieval contexts. The first approach is to index documents in terms of the Interlingual Index of the database, turning term weighting and query/document comparison into language-independent tasks and taking full advantage of EWN features for text retrieval. This approach seems specially well-suited for searches over unrestricted, heterogeneous collections, such as World-Wide-Web searches, as it does not need for training or parallel corpora. The second one suggests how knowledge from the EWN database could be integrated with a corpus-based approach in a multilingual interface for a digital library system. This integration with corpus-based techniques seems best-suited for searches over homogeneous corpora with domain-specific vocabulary that is not likely to be found in the EWN database. Finally, in section 6, we conclude by mentioning ways that EWN can be used in information retrieval tasks such as acquisition of semantic information or word-sense disambiguation.

2. Lessons from WordNet-Based Monolingual Text Retrieval

WordNet is a large-scale lexical database configured as a semantic net of 91,587 *synsets* (sets of synonymous word-senses) containing 119,217 words and basic semantic relations (synonymy, hyponymy, meronymy, etc.) between them. The semantic content and the large coverage of WordNet means that it is a promising tool for conceptual text retrieval, as opposed to exact keyword matching. The obvious advantage is that we are able to retrieve relevant documents that do not contain query terms but are semantically related to the query.

Two main kinds of applications have been attempted in monolingual text retrieval using WordNet:

- Expansion of the query to include synonyms and other semantically related words (thus enhancing recall). (Smeaton et al., 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.

- Comparison of queries and documents via some measure of conceptual distance. Rather than scoring only query/document coocurrence of terms, this kind of comparison takes into account semantic relatedness of terms. If the measure is accurate, both recall and precision can be enhanced.

Experiments by Novell within the EWN project (Bloksma et al., 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. In general, experiments with WordNet have produced similar results (Richardson and Smeaton, 1995). The best results are obtained when working with short documents. (Smeaton and Quigley, 1996) used image captions in an image database, being able to retrieve image captions such as *boys playing in the sand* from a query as *children running on a beach*, while standard IR techniques do not perform as well. The reason is that with very short documents the chance of finding the original terms in the query are much lower than for average-size documents (that typically include many phrasings for the same concepts). An expansion of the query into semantically related words solves this problem.

Despite the fact that these indications are encouraging, in general the impression has been that there is a relative lack of good results when experimenting WordNet in information retrieval tasks. Explanations for this include:

The WordNet database has some weaknesses for text retrieval. In particular:

- It lacks cross-part-of-speech relations. This means, for instance, that *paint* and *painting* or *adornment* and *adorn* are not related in WordNet, whereas the connection between them is of great relevance in text retrieval.

- It lacks topic or domain information. For instance, *tennis shoes* and *tennis racquet* have no semantic connection in WordNet; they belong to different parts of the hierarchy. This is another dimension of relatedness that is crucial for text retrieval, as the topic determines the context of a word within a document to a large extent.

- The WordNet hierarchy is unbalanced; the coverage of the subdomains is not homogeneous. This causes a distortion of most measures of semantic distance used in the literature.

– Sense distinctions in WordNet are too much fine-grained. For instance, *people* has four different senses in WordNet:

  1. people – any group of persons (men or women or children) collectively.
  2. citizenry, people – (the body of citizens of a state or country)
  3. people – (members of a family line)
  4. multitude, masses, mass, hoi polloi, people – (the common people generally)

  Such fine-grainedness does not favor retrieval, as queries do not usually imply such distinctions from a human-judgment point of view (which is the only golden standard!).

In addition, we can consider:

– Lack of accurate word-sense disambiguation (WSD) methods. This is still an open research question. Experiments with unrestricted WSD on large corpora show that the best algorithms do not outperform the *'most frequent'* heuristic that consists simply in always choosing the most frequent sense and has a success rate of around 50% (with variations depending on the texts being disambiguated).

– Lack of a good measure of conceptual distance. This is something difficult to evaluate, and the measures actually used in the experiments did not fit the criteria (see section 4).

But above and beyond these reasons, the most important one, in our opinion, is that WordNet-based approaches augment exponentially the problem of polysemy. Let us suppose that the query contains the word *spring* in the sense of *spring, springtime – Season of growth.* The standard problem of polysemy when retrieving by exact term matching is that, besides pertinent documents, we will also retrieve documents talking about metal devices, fountains, etc., that correspond to other meanings of *spring.* Now, what happens if we expand the query with WordNet? We will obtain a set of words *spring, springtime, fountain, leap, outflow, bound, give, springiness, outpouring* of which only the first two correspond to the intended meaning; unfortunately, the other words are probably polysemous as well, and none of their senses will be related to the original idea of the query. So the problem of ambiguity is expanded, by adding terms for the wrong meanings of the query term. Recall may be slightly improved by the addition of *springtime*, but precision will be very much affected by the large set of irrelevant words that have also been added.

The interesting point is that this is exactly what happens - either explicitly or implicitly - within any approach to cross-language text retrieval. When translating a query from the source to the target language, the task of finding equivalent terms in the target language is identical to finding synonyms; in fact we are looking for synonyms, but in another language. When we translate *spring*, we obtain a set of words *primavera(season), salto(leap), brinco(leap), elasticidad(springiness), manantial(source of water), fuente(source), resorte(metal device)* where only one meaning of one of the words is related to the intended meaning of the query. This expansion of ambiguity makes cross-language text retrieval much harder than its monolingual counterpart.

Thus a WordNet-based (or better, a EuroWordNet based) approach to cross-language text retrieval expands ambiguities just as any other cross-language approach, *but not more*. From the term *spring* we may obtain its synonyms in any of the project languages - eg. for Spanish the above set of words -. While WordNet introduces one more level of expansion in monolingual retrieval, EuroWordNet expands ambiguity as any other cross-language approach. Thus, the advantages of using a WordNet-like structure should be enhanced in a multilingual setting. This makes WordNet-based retrieval probably more interesting in a multilingual setting than in a monolingual one. We will further discuss this question in the next section.

3. EWN features for Text Retrieval

The major feature of the EWN database, comparing to WordNet 1.5, is obviously its multilingual nature. But besides this, there are a number of features that are relevant from the point of view of text retrieval. The database is described in detail in this volume (**?**; **?**; **?**), but we will review here the relevant issues for text retrieval.

- It will contain about 50,000 word meanings correlating the 20,000 most frequent words (only nouns and verbs in the first stage) in each language. This size should be sufficient to experiment with generic, domain-independent text retrieval in a multilingual setting without the need for training with bilingual parallel corpora. The individual monolingual databases will be considerably smaller than WordNet 1.5, but the different coverage is only on specific subdomains; the coverage of most frequent words and more generic terms will be similar in both databases. The EWN database will be expanded to a higher level of detail for a particular domain, in order to test its adequacy to incorporate domain-specific thesauri.

– Synsets have domain labels that relate concepts on the basis of topics or scripts rather than classification. This means that *tennis shoes* and *tennis racquet* will be related through a common domain label *tennis*. As we have mentioned above, topic relations are essential to word-sense disambiguation and retrieval tasks.

– Nouns and verbs are not separate networks. EWN includes the cross-part-of-speech relations:

  - noun-to-verb-hyperonym: *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 nearly equivalent for retrieval purposes, regardless of their different Part-Of-Speech.

The multilingual architecture is described in detail in (**?**) in this volume; we summarize here its components:

Monolingual Wordnets. Each language has its individual wordnet with internal relations that reflects specific properties of that language. However, each monolingual wordnet is being built from a common set of 1085 *base concepts* (concepts which are relatively high in the semantic hierarchies and which have many relations with other concepts) verified manually to fit all monolingual wordnets. This is one of the measures that guarantees overlap and compatibility between wordnets, reducing spurious mismatches in the hierarchy.

Interlingual Index(ILI). a superset of all concepts occurring in the monolingual wordnets. The ILI began as a collection of records that matches WordNet 1.5 synsets, and it grows as new concepts are added. It will also be modified with respect to WordNet 1.5, as too fine-grained sense distinctions will be collapsed. (see section 2). **??** describes this process in detail.All interlingual relations and language-independent information is linked to the ILI, as explained below.

Cross-language relations. Each wordnet is linked to the ILI via cross-language equivalence relations, namely:

- cross-language synonymy: *It:anitra EQ-NEAR-SYNONYM duck*
- cross-language hyperonym: *Dutch:hoofd (human head) HAS-EQ-HYPERONYM head*
- cross-language hyponym: *Sp:dedo (finger or toe) HAS-EQ-HYPONYM finger*, *Sp:dedo HAS-EQ-HYPONYM toe*

Cross-language complex relations (hypernyms and hyponyms) indicate potential new ILI records. After each building stage all complex relations are collected and compared across languages and new ILI records will be added if appropriate. These relations facilitate cross-language retrieval.

Top-concept Ontology. hierarchy of about 63 language-independent concepts reflecting explicit opposition relations (e.g. *object* vs. *substance*) This ontology is linked to the base concepts through the ILI -see (**?**)-.

Hierarchy of domain labels. Also linked to the ILI and thus inherited by every monolingual wordnet.

4. Language-Independent Text Retrieval with EWN

Rather than expanding queries with terms for all languages, our first proposal (Gilarranz et al., 1997) is to index documents in terms of the Interlingual Index records (which, in practice, serves as a language-independent ontology). We use the vector space model, but instead of a vector of weighted terms, each document is represented as a vector of weighted Interlingual Index records. This means that we can consider language-independent criteria for term weighting and comparison query/document. Two major processes have to be considered: document indexing and query/document matching.

4.1. Indexing of Documents

Indexing of documents has two stages: a language-dependent one that maps terms to ILI records, and a language-independent one that assigns weights to the representation:

Language-dependent stage

1. POS Tagging. In principle, only nouns and verbs will be considered.

2. **Term Identification.** This step includes stemming and reconstruction, and identification of multiwords. It is interesting to note that the detection of multiwords is of great importance to text retrieval tasks, and WordNet is rich in multiword information, offering a potential for retrieval refinement that should be exploited. However, an appropriate treatment of multiwords from a multilingual perspective is not at all simple.

   On one hand, the detection of lexicalized multiwords greatly enhances precision. For instance, *hot spring* can be identified in a document as a lexicalized multiword simply inspecting WordNet 1.5 entries. We can thus assign a single meaning to *hot spring*, avoiding a separate inclusion of meanings for *hot* and *spring*, which would not reflect the content of the document. Even when WordNet 1.5 includes non-lexicalized phrases such as *a great distance* or *fasten with a screw*, it would seem helpful to use these in order to refine term identification and matching.

   On the other hand, non-lexicalized phrases cannot be used naively for text retrieval. Consider, for instance, lexical items in one language that do not have equivalents in another. An example is *toe*, which does not have an equivalent in Spanish. The closest lexical item is *dedo*, which means *finger or toe.* Thus it appears that going from one language to the other we would lose information, so it seems better to have a Spanish synset with a multiword, even if it is not lexicalized, to describe the concept in Spanish (such non-lexicalized multiwords are very common in WordNet 1.5, which oscillates between lexical and conceptual criteria to include synsets). The appropriate multiword in Spanish would be *dedo del pie* (*del pie* = *of the foot*). But actually when a Spanish document is talking about toes, it will probably use just *dedo.* A retrieval system looking for *dedo del pie* as a single bound item could miss relevant information; the best solution is probably to search for both *dedo* and *pie* in close coocurrence and also just *dedo. pie* can be used as a weight for document ranking.

   The methodology adopted in EWN - based on this and many other facts - has been to include in the database only lexicalized expressions (such as *hot spring* above), to avoid making wrong predictions about expressibility (**?**).

3. **Word-Sense Disambiguation.** This is a crucial step within our approach; the quality of the representation depends mainly of accurate word-sense disambiguation.

Our intention is to disambiguate by combining different algorithms that exploit different kinds of contextual information. Semantic relatedness of words in the context can be exploited via the *conceptual density* measure defined in (Agirre and Rigau, 1996). Unlike previous measures of conceptual distance used in WordNet-based text retrieval, this one is sensitive to both a) length of the path connecting the concepts, b) depth in the hierarchy (concepts deeper in the hierarchy are closer than concepts higher in the hierarchy) , and c) density of the hierarchy (concepts in a dense region are closer than those in a more sparse region). It is interesting to note that the word-sense disambiguation algorithm using this measure has only been tested on English - that is what WordNet 1.5 provides -. With the EWN database it can, in principle, be applied for all the languages in the database.

Besides this measure, the domain labels offer complementary information about topic relations that is also very helpful to disambiguate. Finally, we may consider not only the context of the words in the query, but also a context for the actual meanings to be disambiguated: the gloss and the synonym sets are obvious candidates.

It is usually assumed that information retrieval performs an implicit disambiguation when comparing query and document, because the adequate senses for a term are reinforced by the terms in the context (Krovetz and Croft, 1992). It is very interesting to check this point explicitly: what is the impact of explicit word-sense disambiguation on the overall performance of the system? We do not know of previous approaches that explicitly disambiguate the document before comparing to queries. Our approach will be to optimize our word-sense disambiguation algorithm balancing retrieval accuracy with computational cost.

4. **Mapping into Interlingual Index.** Once the words are disambiguated in terms of the monolingual wordnet, the synsets are mapped into the InterLingual Index via cross-language equivalence relations.

Language-independent stage

5. **Weighting.** Now weighting can be done with language-independent criteria. In particular:

   – frequency in the document (as usual)

   – inverse of the frequency in the database (as usual)

- depth and position in the conceptual hierarchy. This is a very interesting possibility. Synsets very high in the hierarchy are not sufficiently specific and can be removed (the ILI thus provides a list of *stop synsets*). The rest can be weighted increasingly with depth in the hierarchy.
- disambiguation confidence. Reliable data should be weighted higher. If a term cannot be disambiguated reliably, all candidate meanings can be included with low weights.

The same process will be applied on queries, although performing disambiguation is much more difficult because queries are very short compared with documents and thus offer little context information. Interactive user refinement of the query will be tested within the Novell CLTR environment.

4.2. Query/Document Matching

We will experiment with three approaches to query/document comparison. Each approach adds some kind of information to the previous one:

a) Cosine comparison As formally we have a classical vector model, we could use classical cosine comparison as a baseline. Thus we can evaluate separately the impact of the indexing process and the methods for comparison.

b) Weighted expansion The vector can be expanded - still in a language-independent manner - by including related ILI records. The first candidates are cross-POS synonyms, which usually have strongly related meanings (see previous sections). Hyponyms are also good candidates. Other relations should be included very carefully to avoid loss of precision.

c) Measure of semantic relatedness Instead of simply matching identical concepts, we can measure the semantic relatedness of query and document indexing concepts. The first candidate here is the measure of conceptual distance mentioned above, but also topic relatedness should score.

4.3. Development Steps

The stages to develop such a system are:

1. Monolingual experiments with WordNet.
   - In a monolingual (English) environment, we are setting up a word-sense disambiguation algorithm that should balance accuracy and efficiency.
   - Experiment with different synset weightings against cosine comparison.
   - Experiment with different comparison methods.
   - Evaluate the impact of accurate WSD for each different comparison method.

   As a monolingual test set to setup WSD methods, probably the best corpora available are Semcor (Miller et al., 1994) and the corpus described in (Ng and Lee, 1996), both annotated with WordNet senses.

2. Bilingual experiments (English-Spanish). Once WSD is tuned, one of the best corpora for cross-language retrieval evaluation would be the one described in (Carbonell et al., 1997), where most current approaches have been tested in identical conditions. This would produce a reliable evaluation, although experiments with different corpora and query lengths should also be performed.

3. Multilingual experiments. Our plan is to find or construct a corpus that could resemble WWW searches (possibly by crawling on the web); such corpus is a real challenge for a text retrieval system (variety of topics, heterogeneity and multilinguality are difficult tasks for text retrieval), but will provide a better evaluation of the system for unrestricted searches.

4.4. Summing Up

This proposal for cross-language text retrieval has attractive advantages over other techniques:

- It performs language independent indexing, providing
  - a semantic structure to perform explicit WSD for indexing.
  - language-independent weighting criteria.
- It permits language-independent retrieval, by
  - Concept comparison rather than term comparison

- Topic comparison

– It does not require training or availability of parallel corpora (a great advantage when thinking of more than two languages or when performing retrieval on unrestricted texts, such as World-Wide-Web searches).

– It has most of the advantages of WordNet, but not its disadvantages:

  - Words can be conceptually related even if they have different POS.
  - Besides classification relations, synsets have also topic information (domain labels), which are especially useful for text retrieval.

5. Using EWN in a Corpus-Based Cross-Language Query System

The previous section described a way in which EWN could be used for multilingual text retrieval which is well-suited for retrieval over sparse, heterogeneous document collections where training is unfeasible, the best example being a World-Wide-Web search. This approach is not feasible for technical corpora, because the relevant domain-specific vocabulary is not likely to be found in the EWN database. In this section we will discuss how the EWN database could be integrated with a corpus-based approach to overcome this limitation.

We feel that a promising and cost-effective solution for CLTR will probably be an integration of a multilingual ontology such as EWN with corpus-based techniques which extract information on cross-language lexical equivalents from large collections of multilingual text data. In this way, it should be possible to both exploit the wide-range of cross-language equivalence relations contained in the semantic database (retrieving not just cross-language synonyms but also other kinds of related words) while at the same time being able to find possible target language query terms even when the source language search item is not included as an entry in the database.

5.1. A corpus-based system for domain-specific document bases

At Pisa, we are studying the design of a cross-language query system for a text collection containing documents in more than one language. Our intention is to implement a system which will integrate a dictionary/thesaurus-type search with a corpus-based strategy. The aim is to be able to match a query formulated in one language against documents stored in other languages, even when the query terms themselves are not included in the multilingual lexicon. However, the problem with most corpus-based cross-language systems is that the acquisition of a suitable set of relevant documents on which to train the retrieval system is extremely resource consuming. To overcome this problem, we have focused our attention on developing a system for a multilingual digital library; in this case, the training corpus will be extracted from the collection of documents contained in the digital library itself. Digital library applications are becoming of crucial importance in today's information-based society. As more and more information is made available in a multiplicity of languages on the World Wide Web, frequently organized in distributed domain-specific document bases, there is growing interest in the development of methods which allow users to formulate queries in one language and retrieve matching documents which have been written and stored in other languages.

Our corpus-based strategy is based on the concept of comparable corpora. Comparable corpora are sets of texts in pairs (or multiples) of languages with the same communicative function, i.e. generally on the same topic or domain. Such corpora are sources of natural language lexical equivalences across languages and as such can provide much useful data for contrastive language studies. For this reason, language scholars are now beginning to acquire such collections of texts for particular linguistic or terminological studies. However, a digital library system which contains document archives on the same domain but in different languages is actually a real world implementation of the comparable corpora principle.

The corpus query system is based on the assumptions that (i) words acquire sense from their context, and (ii) words used in a similar way throughout a sub-language or special domain corpus will be semantically similar. It follows that, if it is possible to establish equivalences between several items contained in two different contexts, there is a high probability that the two contexts themselves are to some extent similar. We thus use lexical and linguistic knowledge extracted from a domain-specific corpus in one language and project it onto a comparable corpus in the other; i.e. given a particular term or set of terms in the texts in one language (L_1), the aim is to be able to identify contexts which contain equivalent or related expressions in the texts of the other (L_2). To do this, we attempt to isolate the vocabulary related to that term in the L_1 corpus – hypothesising that lexically equivalent terms will be associated with a similar vocabulary in L_2. So far, we have tested this system with success for just two languages: using a bilingual Italian/English dictionary and an Italian/English comparable corpus (Picchi and Peters, 1996). The multilinguality of EWN gives us the opportunity to extend this methodology and apply it to document collections in more than two languages.

5.2. Methodology

If we have a collection of documents on a given topic in two or more of the EWN languages, for any term of interest, T, searched in one language (L_1) the system will automatically construct a context window containing T and up to n lexically significant words (nouns and verbs only) to the right and left of T for the set of documents in our collection in L_1; The value for n can be varied. For each of these co-occurrences of T, morphological procedures identify the source lemma(s). The significance of the correlation between these items and T is then calculated using a statistical procedure. We are currently using Church and Hanks' Mutual Information Index (Church and Hanks, 1990) in our procedures, although we are also testing a different measure based on the likelihood ratio as formulated by (Dunning, 1993).

The set of most significant collocates derived makes up the vocabulary, V_1, that is considered to characterize our term T in the particular sub-domain corpus being analyzed. Next, using morphological procedures, and the EWN database, we can construct an equivalent vocabulary of translation equivalents (V_2) for each L_2 set of documents to be searched. Words or expressions that can be considered as lexically equivalent to our selected term in the L_1 texts will then be searched in the L_2 document sets, i.e. we do this by searching for those contexts in the second language texts in which there is a significant presence of the L_2 vocabulary for T. The significance is determined on the basis of a statistical procedure that assesses the probability for different sets of L_2 cooccurrences to represent lexically equivalent contexts for T. The L_2 contexts retrieved are written in a file and listed in descending order of relevance to our L_1 term.

In this way, we can create a multilingual search tool which combines the features of a knowledge-based tool with that of a corpus-based procedure and thus searches for both terms for which multilingual equivalences can be identified in EWN but also for cross-language lexical equivalences which are not included in our multilingual database.

5.3. Search Term Disambiguation

Although the problem of polysemy is greatly reduced in a domain specific corpus, it is still present – to a varying degree depending on the type of texts being treated. The construction of the L_1 vocabulary which characterizes our term T permits us to obtain a clustering of the most relevant terms connected to T. If the corpus contains a predominant sense for the term then the vocabulary should represent this sense – secondary senses that appear rarely will not cause a representative vocabulary of collocates to be constructed. If, in the corpus, there is more than one relevant sense for T then we would expect two or more distinct clusterings of significant collocates. For example, the Italian noun *accordo* has two distinct senses in the Italian component of EuroWordNet: the general sense which is translated by *agreement, understanding*, and the very specific musical sense translated by *chord.* Testing our system on a corpus of parliamentary debates, we found no examples of the second sense. However, if there had been, we would expect to obtain two distinct clusterings of significant collocates with little or no overlap. Thus, using an automatic statistical clustering technique, it should be possible to distinguish between common technical terms which are used with different meanings in different scientific areas. Think, for example, of the different usages of *protocol* in the medical and software engineering domains. Very different sets of collocates would be constructed for the different acceptions of this term and thus searching for the appropriate sense would be facilitated.

5.4. Target Term Disambiguation

When constructing the L_2 vocabularies of significant collocates for the L_1 term being searched, our procedure will take as input all the cross-language equivalent synonyms provided by EWN, regardless of sense distinctions. Spurious or inappropriate translations are eliminated by the fact that we normally do not find them together with a significant number of items from the L_2 vocabulary for the term being searched. This makes it possible for us to perform a sense disambiguation on the target terms proposed. For example, if we examine all the occurrences of the Italian noun *sicurezza* in our test corpus, we find that the sense is that of *safety*, or *security.* This is confirmed by the set of significant collocates for this term; the top ten are the Italian equivalents of toy, hygiene, reactor, health, nuclear, maritime, council, road, provisions, Euratom.

The EWN entries for *safety*, and *security* include the following possible synonyms in the different senses: for the six senses of *safety*, we have *refuge, guard, base hit, condom, rubber, prophylactic*; for the seven senses of *security*, we find *protection, certificate, surety, guarantee, warranty*. Of these potential synonyms, only *protection* is actually relevant, as a cross-language equivalence of the Italian term *sicurezza* that we are searching. Most of the irrelevant word senses are not found in our test corpus; but even if they should appear, the contexts containing them are ranked very low as they will not contain significant collocates from the target language vocabulary of related cooccurrences constructed for our source language search term: *sicurezza*. Thus, our approach helps us to identify the correct sense of the target terms offered by EWN and to provide a ranking of the best L_2 matches for the L_1 term searched.

5.5. Summing Up

In our proposal for integrating EWN in a corpus-based approach, queries will be translated using the multilingual database but will also be expanded by applying the comparable-corpus based strategy in order to associate with each query term, not only its direct translations but also a vocabulary which defines its probable immediate context, in the query language and in the target language(s).

In this way, we search for both pre-identified translation equivalents and also cross-language lexical equivalences. When EWN offers no translation equivalent, the search for cross-language equivalent contexts is still possible. Documents retrieved are ranked with respect to (i) translation equivalents of query terms, (ii) statistical value assigned to associated significant collocates.

One of the limitations of this type of statistically-based querying over domain-specific archives is that it is only feasible when the text collection is sufficiently large and sufficiently homogeneous to be able to derive a statistically meaningful set of collocates for the terms queried. These conditions should be satisfied by the average digital library.

Considering the two approaches that we have discussed here, we are convinced that the building of the EuroWordNet database offers an excellent opportunity to experiment with truly multilingual text retrieval. And viceversa: CLTR is an excellent testbed to measure the quality of the EWN database.

6. Other Related Applications of EuroWordNet

Before concluding, it is important to observe that a semantic net can be used in a number of techniques and tasks related to disambiguation and acquisition of semantic information, i.e. in tasks which are clearly complementary and/or subsidiary to the broad application goal of Information Retrieval. It is thus clear that a semantic net of the EWN type can be applied to similar activities in the multilingual and/or multiple language context.

In our first approach we have seen one example: semantic indexing involving word-sense disambiguation (WSD), and the EWN database is an excellent source of information to perform WSD. We review here applications where WordNet has been tested in such tasks for English, as we can safely infer from this the usability and usefulness of similar semantic nets for other languages. The purpose is, in fact, to highlight the potentialities of the EWN database in a number of challenging and related tasks, which can be seen as components of different application systems, while - at the same time - showing that this database should be considered as just *one* important component for reaching better results. We are well aware that it is far from being *the* solution. It will always need to be integrated by other methods and/or data.

As is known, WordNet, or other thesauri, are being used extensively in word similarity measurement techniques (Kurohashi and Nagao, 1994; Li and Abe, 1995), which are at the basis of different NLP applications and tasks. The intuitive underlying assumption is that words closer to each other in a hierarchical structure can be considered to have a higher degree of semantic similarity. (Fujii et al., 1997), after evaluation of the results of statistics-based and human-crafted thesaurus-based approaches, claim that better results can be obtained by an integration of the two approaches. This combined technique is evaluated also in its application to a word sense disambiguation task, where it is also shown to be effective. The use of WordNet helps to solve problems of data sparseness suffered by semantic disambiguation tasks based on probabilistic models.

Selectional preferences on the arguments are important information to be encoded for a predicate, to be used e.g. in resolving many structural and lexical ambiguities. The advantages of using WordNet in the task of automatic acquisition of lexical information (e.g. selectional preferences) are reported in (McCarthy, 1997), as part of the LE-SPARKLE project, but go back to (Grishman et al., 1992), (Ribas, 1995) and (Li and Abe, 1995). In these approaches, WordNet is used, in different ways, to obtain classes at an appropriate level of generalization, thus reducing again the sparseness problem. However, in lexical acquisition tasks there is also the problem of having to acquire information from an ambiguous input, and therefore the problem of semantic disambiguation. McCarthy tries to tackle the lack of semantic disambiguation in the input, which causes, for the presence of various erroneous senses, a sort of over-generalization or dispersion in non-appropriate semantic areas. Also sense disambiguation is handled, although in a crude manner, using WordNet as a classification tool to disambiguate heads in parsed texts.

It is clear that these two tasks of word-sense disambiguation and acquisition of semantic preferences are strictly, and somehow circularly, interconnected, but a bootstrapping and iterative approach can be adopted, where even crude disambiguation can help acquisition.

One interesting observation - common to many experiments using WordNet so far - is that better results can be obtained with less granular word-senses than those found in WordNet. Within EuroWordNet (as stated in section 3) a decision in this sense has already been made, i.e. to allow for less granular sense distinctions.

Semantic similarity, assessed with reference to WordNet (Resnik, 1995), is used also by (Sanfilippo, 1997), again within LE-SPARKLE, to disambiguate syntactic collocates, and preliminary results are shown to be encouraging, also considering the reduction of input data requirements with respect to statistical techniques. Here too the task at stake is the automatic acquisition of semantic co-occurrence restrictions from text corpora.

Semantic tagging using WordNet associated to the Hidden Markov Model disambiguation method (as used in POS-tagging), is described in (Segond et al., 1997), which improves the baseline from 81% to 89%.

WordNet is also used to enhance information extraction systems. (Chai and Bierman, 1997) use WordNet hierarchies to automatically generalize patterns in the training of one information extraction system in new domains.

To sum up, it appears clear that the EWN database, which consists of separate but linked monolingual wordnets, besides its cross-lingual capacities, offers the chance to apply several information extraction and word-sense disambiguation techniques for many languages other than English, and that many of these techniques may in addittion be improved by the available cross-language information.

7. Conclusions

The existence of large-scale semantic databases such as EuroWordNet opens new and challenging possibilities for cross-language and multilingual text retrieval, among other applications.

We have presented two approaches: the first one indexes documents and queries in terms of the EuroWordNet InterLingual Index, thus turning term weighting and query/document matching into language-independent tasks. The second one integrates the information in the EWN database with corpus-based techniques, thus allowing retrieval for domain-specific terms that may not be present in our multilingual database.

Each methodology addresses a different aspect of text retrieval: Our first approach is well-suited for searches over heterogeneous document collections such as World-Wide-Web searches, as it is language-independent and does not require training with multilingual parallel corpora. The second approach, on the other hand, is best suited for retrieval over digital libraries, where the document collections are homogeneous and technical vocabularies play an essential role for retrieval. Finally, we also review some applications of WordNet to information extraction and word-sense disambiguation tasks: the EWN database, besides its cross-lingual capacities, offers the chance to apply such techniques to many languages other than English.

We hope that we have shown the potential of the EuroWordNet database for truly multilingual (rather than just bilingual or cross-language) text retrieval and information extraction.

Acknowledgments

This research is being supported by the European Community, project LE #4003 and also partially by the Spanish government, project TIC-96-1243-CO3-O1.

References

Agirre, E. and Rigau, G. (1996). Word sense disambiguation using conceptual density. In *Proceedings of COLING'96*.

Ballesteros, L. and Croft, W. (1996). Dictionary-based methods for cross-lingual information retrieval. In *Proc. of the* 7^{th} *International DEXA Conference on Database and Expert Systems Applications*, pages 791–801.

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.

Carbonell, J., Yang, Y., Frederking, R., Brown, R., Geng, Y., and Lee, D. (1997). Translingual information retrieval. In *Proceedings of IJCAI'97*.

Chai, J. and Bierman, A. (1997). The use of lexical semantics in information extraction. In Vossen, P., Adriaens, G., Calzolari, N., Sanfilippo, A., and Wilks, Y., editors, *Proceedings of the ACL/EACL'97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources*.

Church, K. and Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.

Dumais, S., Landauer, T., and Littman, M. (1996). Automatic cross-linguistic information retrieval using latent semantic indexing. In *Working Notes of the Workshop on Cross-Linguistic Information Retrieval, ACM SIGIR'96*, pages 16–23.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1).

Fujii, A., Hasegawa, T., Tokunaga, T., and Tanaka, H. (1997). Integration of hand-crafted and statistical resources in measuring word similarity. In Vossen, P., Adriaens, G., Calzolari, N., Sanfilippo, A., and Wilks, Y., editors, *Proceedings of the ACL/EACL'97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources.*

Gilarranz, J., Gonzalo, J., and Verdejo, M. (1997). An approach to cross-language text retrieval with the eurowordnet semantic database. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval.* AAAI Spring Symposium Series, to appear.

Grishman, R., Macleod, C., and Sterling, J. (1992). New york university description of the proteus system as used for muc-4. In *Proceedings of the Fourth Message Understanding Conference*, pages 223–241.

Hull, D. and Grefenstette, G. (1996). Querying across languages. a dictionary-based approach to multilingual information retrieval. In *Proc. of the* 19^{th} *ACM SIGIR Conference*, pages 49–57.

Krovetz, R. and Croft, W. (1992). Lexical ambiguity and information retrieval. *ACM Transactions on Infromation Systems*, 10(2):115–141.

Kurohashi, S. and Nagao, M. (1994). A method of case structure analysis for japanese sentences based on examples in case frame dictionary. *IEICE TRANSACTIONS on Information and Systems*, E77-D(2):227–239.

Li, H. and Abe, N. (1995). Generalizing case frames using a thesaurus and the mdl principle. In *Proceedings of Recent Advances in Natural Language Processing*, pages 239–248.

McCarthy, D. (1997). Word sense disambiguation for acquisition of selectional preferences. In Vossen, P., Adriaens, G., Calzolari, N., Sanfilippo, A., and Wilks, Y., editors, *Proceedings of the ACL/EACL'97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources.*

Miller, G., Chodorow, M., Landes, S., Leacock, C., and Thomas, R. (1994). Using a semantic concordance for sense identification. In *Proceedings of the ARPA Human Language Technology Workshop.*

Ng, H. and Lee, H. (1996). Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. In *Proceedings of ACL'96.*

Picchi, E. and Peters, C. (1996). Cross language information retrieval: A system for comparable corpus querying. In Grefenstette, G., editor, *Working Notes of the Workshop on Cross-Linguistic Information Retrieval, ACM SIGIR'96*, pages 24–33.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI.*

Ribas, F. (1995). On learning more appropriate selectional restrictions. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, pages 112–118.

Richardson, R. and Smeaton, A. (1995). Using wordnet in a knowledge-based approach to information retrieval. In *Proceedings of the BCS-IRSG Colloquium, Crewe.*

Salton, G. (1970). Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21:187–194.

Sanfilippo, A. (1997). Using semantic similarity to acquire co-occurrence restrictions from corpora. In Vossen, P., Adriaens, G., Calzolari, N., Sanfilippo, A., and Wilks, Y., editors, *Proceedings of the ACL/EACL'97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources.*

Segond, F., Schiller, A., Grefenstette, G., and Chanod, J. (1997). An experiment in semantic tagging using hidden markov model tagging. In Vossen, P., Adriaens, G., Calzolari, N., Sanfilippo, A., and Wilks, Y., editors, *Proceedings of the*

ACL/EACL'97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources.

Sheridan, P. and Ballerini, J. (1996). Experiments in multilingual information retrieval using the spider system. In *Proc. of the* 19^{th} *ACM SIGIR Conference*, pages 58–65.

Smeaton, A., Kelledy, F., and O'Donnell, R. (1995). Trec-4 experiments at dublin city university: Thresolding posting lists, query expansion with wordnet and pos tagging of spanish. In *Proceedings of TREC-4.*

Smeaton, A. and Quigley, A. (1996). Experiments on using semantic distances between words in image caption retrieval. In *Proceedings of the* 19^{th} *International Conference on Research and Development in IR.*