Evaluating wordnets in Cross-Language Information Retrieval: the ITEM search engine

Felisa Verdejo, Julio Gonzalo, Anselmo Peñas, Fernando López and David Fernández

Depto. de Ingeniería Eléctrica, Electrónica y de Control
UNED, Spain
{felisa,julio,anselmo,flopez,david}@ieec.uned.es

Abstract

This paper presents the ITEM multilingual search engine. This search engine performs full lexical processing (morphological analysis, tagging and Word Sense Disambiguation) on documents and queries in order to provide language-neutral indexes for querying and retrieval. The indexing terms are the EuroWordNet/ITEM InterLingual Index records that link wordnets in 10 languages of the European Community (the search engine currently supports Spanish, English and Catalan). The goal of this application is to provide a way of comparing in context the behavior of different Natural Language Processing strategies for Cross-Language Information Retrieval (CLIR) and, in particular, different Word Sense Disambiguation strategies for query translation and conceptual indexing.

1. Introduction

The aim of this paper is to present the ITEM multilingual search engine. The ITEM project (`http:sensei.ieec.uned.es/item/principal.htm`), financed by the Spanish government, started in 1996 and finished in 1999, and had two main goals: 1) integrating a variety of language resources and tools for Natural Language Processing in Spanish, Catalan, Basque and English, and 2) demonstrating the application of such resources and tools in a multilingual information retrieval system.

The multilingual search engine is one of the results of the ITEM project and is publicly available for online testing at `http://terral.ieec.uned.es/clir`. The search engine permits querying and retrieving documents in three languages (Spanish, Catalan and English); the user is allowed to select from a number of Natural Language Processing options, and to refine the results of such lexical processing.

In its current state, the search engine integrates lexical databases and NLP modules (morphological analyzers, lemmatizers, Part-Of-Speech taggers and Word Sense Disambiguators) for Spanish, Catalan and English (Rodríguez et al., 1998; Màrquez and Padró, 1997; Rigau et al., 1997) Basque is also expected to be integrated in a near future.

The language resources developed within the project - and integrated in the search engine- have a close relation to the EuroWordNet project (Vossen, 1998), and include a lexical database with semantic relations for words in English, Spanish, Catalan and Basque that follows the EuroWordNet design (henceforth EWN/ITEM database)(Benítez et al., 1998; Farreres et al., 1998). The main feature of the EWN/ITEM multilingual semantic network is an InterLingual Index where all monolingual wordnets are connected. Such index permits finding equivalent concepts between any pair of languages in the database. The InterLingual Index is the superset of all concepts occurring in all the languages.

In the search engine, the documents in the text collection are fully processed to obtain the lexical information that permits a conceptual indexing of each document in terms of the EuroWordNet/ITEM InterLingual Index. The collection used for the web interface covers around 10000 newspaper articles from the *International* section of the "Washington Post" (English), "El Pais" (Spanish) and "El Periódico" (Catalan) from April 1998 to May 1999. The language resources and tools, however, are not tuned to any particular domain.

The goal of this application is to provide a way of comparing in context the behavior of different Natural Language Processing strategies for Cross-Language Information Retrieval (CLIR) and, in particular, different Word Sense Disambiguation strategies for query translation and conceptual indexing.

Next two sections describe two different approaches to CLIR using the EWN/ITEM database: the first one is translating the query from the source language to the target languages via the InterLingual Index. The second one is mapping both queries and documents into the InterLingual Index as a language-neutral indexing space. Section 4 enumerates briefly the lexical tools integrated in the search engine. Section 5 describes the web interface to the search engine and, finally, last section draws some conclusions and further developments on the multilingual search engine.

2. Query translation via EuroWordNet

The ITEM search engine implements two alternative approaches to Cross-Language Retrieval. The first one is translating the query from the source language into the other two target languages, and then performing three different monolingual searches with the standard search engine INQUERY (Callan et al., 1992).

This approach is close to dictionary-based CLIR, where a source word is substituted by the translations offered by a bilingual dictionary after some statistical filtering (especially to translate phrases). However, using the EuroWordNet/ITEM semantic network offers a number of advantages over a set of bilingual dictionaries:

- English, Spanish and Catalan wordnets play the role of six bilingual dictionaries. The appeal of having an Interlingual Index grows quickly with the number of

languages involved, and the potential number of languages for the search engine are currently 10 (English, Spanish, Catalan, Basque and the additional EWN languages: Dutch, Italian, French, German, Estonian and Czech).

- Word Sense Disambiguation can be performed explicitly at a language-independent stage (the InterLingual Index representation). Disambiguation gives appropriate ILI records, and ILI records are linked to sets of synonym terms in every target language.

- The semantic relations in the EWN/ITEM lexical database permit a controlled expansion with semantically related terms: hyponyms, meronyms, etc.

- The hypernymy/hyponymy relations on the InterLingual Index permit obtaining approximate translations for source terms that do not have equivalents in the target language(s). For instance, "governor's race" does not have an equivalent in Spain, and therefore there is no Spanish term with the same meaning. However, "governor's race" can be linked to "elecciones" via "elections", which is the direct hypernym of "governor's race". Other example is "grand jury", that has no equivalent in Spanish but can have as approximate translation "jurado" as a Spanish equivalent for the "jury" concept, which is the direct hypernym for "grand jury".

The query translation process is illustrated in Section 5.

3. Indexing by concepts

Translating the query is the most popular approach to Cross-Language Text Retrieval, because it demands much less computation effort than translating or processing the documents in any way.

However, the availability of the EuroWordNet/ITEM database and its InterLingual Index permits exploring an attractive alternative to query translation: using InterLingual Index records to index both queries and documents, getting closer to concept retrieval rather than keyword retrieval. The major advantages over the previous approach (translating the queries) are:

- The comparison between documents and queries is done at a conceptual level, getting rid of the polysemy of words as indexing terms and identifying synonym terms as single indexing units.

- The comparison is done in a language-neutral space, simplyfing the problem of merging results from three monolingual queries in three different databases. All texts can be indexed with the same indexing terms regardless of the source language of the texts.

- In an interactive CLIR system the refinement of the query could be done largely at a conceptual level, avoiding a refinement process for every target language involved.

4. Lexical processing

In order to perform any of the two approaches to CLIR described above, there is a need for full lexical processing software connected with the EWN/ITEM database.

In the ITEM search engine, documents (in the conceptual indexing approach) and queries (in both approaches) are processed by a cascade of lexical analyzers, where only lemmatizers and taggers are language dependent:

1. **Lemmatization and Part-Of-Speech disambiguation**. Spanish and Catalan are processed with the MACO+ morphological analyzer and the RELAX tagger [1,4]. English is processed with the publicly available version of the Brill tagger (Brill, 1992) and the WordNet lemmatizer (Miller et al., 1990).

2. **Detection of multiword expressions**. The detection of multiwords is, in this approach, a language-independent task that considers only expressions included in the EuroWordNet/ITEM lexical database.

3. **Word Sense Disambiguation**. All nouns in the document (or query) are disambiguated, assigning probabilities to every possible sense of each noun. We are currently using a fast implementation of an unsupervised algorithm inspired in (Agirre and Rigau, 1996) that only uses hierarchical information and conceptual distance measures to perform disambiguation. Being unsupervised was a hard constraint on our system, as there are -to our knowledge- no corpora with hand-annotated senses for Catalan or Spanish. Our current WSD system performs (for the Semcor test collection) large below the "First sense" heuristic in picking up the correct sense, but its probability distribution seems to perform slightly better than just picking the first sense in a Cross-Language Information Retrieval task (Vossen et al., 1999).

   In the present state of the search engine, two additional options are available for WSD (both in documents and queries): a "First sense" option that always takes the first sense in the database, and an "all senses" option that takes all possible senses for a noun as equally good indexes. It is necessary to remark, however, that in the Spanish and Catalan wordnets the first sense does not necessarily implies that this is the most frequent sense.

It is worth mentioning that the EWN/ITEM database has not been enriched or adapted to the particular domain for the Internet demo (namely the International newspaper section). While this is an obvious drawback for certain types of queries, our intention was precisely to measure what one can get, and what cannot be expected, from a large scale lexical resource in such an application. An evaluation of the cost of semi-automatically adapting the lexical database to particular retrieval domains will be undertaken in forthcoming extensions of the project.

Figure 1: A snapshot of the ITEM search engine.

5. The search interface

The off-line evaluation of our approach to CLIR in terms of test collections and precision/recall measures has been reported elsewhere (Vossen et al., 1999; Gonzalo et al., 1999b; Gonzalo et al., 1999a). However, precision/recall measures, taken in isolation, tend to disguise both the benefits and the problems of language processing depending on the type of query.

The ITEM search engine, in contrast, provides a direct experience with the use of the EWN/ITEM multilingual semantic network and with the peculiarities of using full lexical processing. The web interface to the search engine allows the user to set a number of parameters regarding the NL processing of the query (and the documents), refine the results of the query processing and compare results with different parameter setups. A snapshot of the web interface (`http://terral.ieec.uned.es/clir`) can be seen in Figure 1. The upper box in the interface (see Figure 1) is used to state a query; the user picks up the processing options in the buttons immediately below. The options are:

- Language of the query. The query may be stated in Spanish, English or Catalan.
- Target language. The target language depends on the newspaper collection chosen: Spanish for "El Pais", English for "Washington Post" and Catalan for "El Periódico".
- Document Representation. The options are "Textual" or "Conceptual". In the textual option, the query is translated into the target language via the InterLingual Index, and then a standard monolingual retrieval is performed against the original text collection. In the "Conceptual" option, query processing stops at the conceptual level representation, and it is compared against the conceptual representation of the documents in terms of the InterLingual Index.
- Word Sense Disambiguation. The user may choose to keep all possible senses of each word ("All Senses"), take always the most common sense ("First Sense"), or use the WSD algorithm described above ("Conceptual Density"). If the document representation selected is "textual", the WSD choice only affects to query translation, restricting the number of concepts translated into the target language. If the document representation selected is "conceptual", then the WSD choice affects also to how the documents are indexed.

  The probability distribution given by the WSD program to the candidate concepts for every noun is used in this way: for document representation, the senses

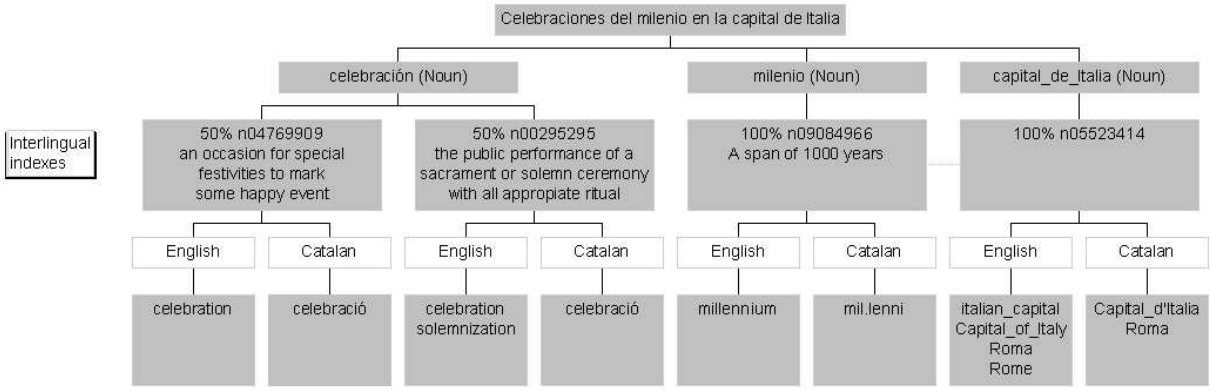

Figure 2: An example of lexical processing of a query.

that score at least 80% of the highest scored sense are chosen as valid indexes, and the rest is discarded. For queries (where context is usually too short to disambiguate accurately), all senses are kept in the expanded query, but weighting them according to their probabilities.

- Multi-word detection. When activated, multiword expressions in documents and queries are used as single-indexing units. In a near future, a more sophisticated treatment of multi-word expression will be incorporated, distinguishing different types of multi-words. A phrase will be considered a single indexing unit only for exocentric compounds, such as "fisher cat", where the meaning of the components is unrelated to the meaning of the multi word expression. Other multiword expressions, such as "abstract art" will be treated giving credit to the meanings of the component words.

Once the query is run, the system provides:

- In the box below the buttons, the expanded query with the target language terms (for text retrieval, as in the figure) or the InterLingual Index records (for concept retrieval). The user may refine the processed query adding/deleting terms (or concepts) in this box and performing a direct search (without further lexical processing) with this refined query in the target language (or in the conceptual representation).

- In the lower box, a ranked list of relevant documents in the newspaper selected. The user may click on the title to see the complete text.

The processed query (which is passed to the standard retrieval engine INQUERY) is obtained from lexical processing (as described in Section 4) and some post-processing to encode the lexical information according to the INQUERY syntax (using, for instance, INQUERY phrase and synonym operators #phrase and #syn). For instance, the original Spanish query

celebraciones del milenio en la capital de Italia

gives, after lexical processing, the result in Figure 2. The main processing steps represented are: 1) Identification of multiword expressions, lemmas and adequate Part-Of-Speech, 2) Representation in terms of the InterLingual Index, with the probabilities assigned by the WSD algorithm, and 3) Expansion into the target languages. The information in steps 2 and 3 is used to build the final query according to the options selected by the user.

For instance, when the document representation is "textual" and the WSD option is "All senses", the result is:

#sum(#sum(celebration #sum(celebration solemnization)) millennium #sum(#phrase(capital of italy) #phrase(italian capital) roma rome))

When the WSD option is "First sense" the result would be:

#sum(celebration millennium #sum(#phrase(capital of italy) #phrase(italian capital) roma rome))

and when the WSD option is "Conceptual density", the weights are used for the query syntax:

#sum(#wsum(100 50 celebration 50 #sum(celebration solemnization)) #wsum(100 100 millennium) #wsum(100 100 #sum(#phrase(capital of italy) #phrase(italian capital) roma rome)))

Now, if the document representation is "Conceptual" and the WSD strategy is, for instance, "First Sense", the query turns into:

#sum(n04769909 n09084966 n05523414)

where, for instance, `n05523414` stands for

```
n05523414
English: Rome, Roma, Italian capital, capi-
tal of Italy
Spanish: capital de Italia, Roma
Catalan: capital d'Itàlia, Roma

    => hypernym: n05483778
          English: national capital
```

```
Spanish: capital de nació
Catalan: capital de nación
```

The user can then refine the query, either by restating the original query or, what is more interesting, directly adding/deleting terms from the expanded query. For instance, the user may choose an "All senses" expansion into the target language, and then manually delete the translation terms that are not appropriate, and then directly interrogate the database with the result. In the next months we expect to provide also a graphical description of the concepts involved and a suggestion of semantically related concepts/terms. Furthermore, we also expect to introduce new WSD possibilities exploiting other sources of contextual information beyond conceptual density.

Note that all the processing options are kept for evaluation and experimentation reasons only, because they demand a combinatory number of different indexations of the whole text collection. The multilingual news collection used in the demo takes around 450Mb of different indexing versions from around 50Mb of original text. All these combinations would not be feasible on larger text collections, where the processing parameters should be fixed (transparent to the user) and non-trivially optimized for space and processing time.

6. Conclusions and future work

A number of experiments regarding the use of concepts in CLIR (Gonzalo et al., 1999a; Gonzalo et al., 1999b; Vossen et al., 1999), together with the direct experience using the search interface, permit us giving first conclusions on the quality of the resources and tools employed and on the utility of language resources and tools for NLP.

For query translation, the EuroWordNet and EWN/ITEM databases offer interesting features compared to bilingual Machine-Readable dictionaries. The semantic relations in the InterLingual Index permit finding approximate translations when a direct equivalence is not available (or does not exist in the target language), and is able to suggest other semantically related terms. However, as for Machine Readable Dictionaries, it is necessary to tune the system to the domain in order to provide adequate translations for domain specific terms and meanings, specially with multiword expressions.

Conceptual indexing is an attractive option, a priori, to perform Cross-Language Retrieval. But it faces two major challenges:

- the sense distinctions for a given word in the lexical database should reflect differences in context usage; otherwise, such distinctions are harmful for retrieval performance. This requisite means that we must find ways of clustering EWN senses into coarser meanings more appropriate for IR purposes. Our experience with the ITEM search engine confirms that the appropriate granularity of senses depends on the application, and for Information Retrieval it is crucial to have the required granularity.

- Word Sense Disambiguation is still an open research question, especially when the task is performing semantic annotation on all nouns and verbs in a text collection in three different languages. Our algorithm fulfills the coverage requirements, as it is unsupervised and language independent (for languages with a wordnet database). However, it is not precise enough, as every other unsupervised WSD system known to us. However it seems that the weighting produced by our system works better than a first sense heuristic in Information Retrieval, even if it detects the most suitable sense worse that the first sense heuristic.

We believe that the optimal way to take advantage of Language Engineering software in CLIR is integrated within interactive search interfaces able to suggest terms and concepts and guiding the user to obtain an optimal combination of terms for his information needs. An advantage of a concept retrieval approach in an interactive retrieval setting is that picking up appropriate concepts is done only once for all target languages, while picking up appropriate translations must be done once for every target language.

To conclude, we believe that the evaluation of Language Resources and Tools can hardly be conducted without measuring their impact on final applications such as Machine Translation systems, Information Extraction tools or Information Retrieval engines. Cross-Language Retrieval is one of the challenges for NLP researchers in the so-called Information Society, and one of the reasons why NLP and IR communities are getting closer to each other. The ITEM search engine is a contribution to allow for qualitative and quantitative tests on the impact of the lexical databases and lexical processing tools in Information Retrieval systems.

7. Acknowledgements

This work has been supported by the Spanish Government, Comisión Interministerial de Ciencia y Tecnología (CICyT), project ITEM (TIC96-1243-C03-01), and also partially by the the European Commission, EuroWordNet project (LE #4003).

8. References

Eneko Agirre and German Rigau. 1996. Word sense disambiguation using conceptual density. In *Proceedings of COLING'96*.

L. Benítez, S. Cervell, G. Escudero, M. López, G. Rigau, and M. Taulé. 1998. Methods and tools for building the catalan wordnet. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98)*.

E. Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*.

J. Callan, B. Croft, and S. Harding. 1992. The INQUERY retrieval system. In *Proceedings of the 3rd Int. Conference on Database and Expert Systems applications*.

X. Farreres, G. Rigau, and H. Rodríguez. 1.998. Using wordnet for building wordnets. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*.

J. Gonzalo, A. Peñas, and F. Verdejo. 1999a. Lexical ambiguity and information retrieval revisited. In *Proceedings of EMNLP/VLC'99 Conference*.

J. Gonzalo, F. Verdejo, and I. Chugur. 1999b. Using EuroWordNet in a concept-based approach to cross-language text retrieval. *Applied Artificial Intelligence*, 13(7):647–678.

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

L. Màrquez and L. Padró. 1997. A flexible POS tagger using an automatically acquired language model. In *Proceedings of ACL/EACL'97*.

G. Rigau, J. Atserias, and E. Agirre. 1997. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of ACL/EACL'97*.

H. Rodríguez, M. Taulé, and J. Turmo. 1998. An environment for morphosyntactic processing of unrestricted spanish text. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98)*.

P. Vossen, W. Peters, and J. Gonzalo. 1999. Towards a universal index of meaning. In *Proceedings of SIGLEX'99*.

P. Vossen. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers.