Browsing by Phrases: Terminological Information in Interactive Multilingual Text Retrieval

Anselmo Peñas
Dpto. Lenguajes y Sistemas Informáticos
UNED,Spain
anselmo@lsi.uned.es

Julio Gonzalo
Dpto. Lenguajes y Sistemas Informáticos
UNED,Spain
julio@lsi.uned.es

Felisa Verdejo
Dpto. Lenguajes y Sistemas Informáticos
UNED,Spain
felisa@lsi.uned.es

ABSTRACT
This paper present an interactive search engine (*Website Term Browser*) which makes use of phrasal information to process queries and suggest relevant topics in a fully multilingual setting.

Categories and Subject Descriptors
Retrieval Issues: *Cross-lingual retrieval, Text Retrieval, Browsing.* Social Issues: *Multilingual access.*

Keywords
Multilingual Information Access, Interaction, Natural Language Processing, Terminology Extraction.

1. INTRODUCTION
In an interactive setting, phrasal information has been used to suggest the user ways of enhancing and refining queries or browsing/classifying search results:

- Handcraft hierarchies based on thesauri (e.g. ERIC) or topic hierarchies (e.g. Yahoo) to browse the document space.
- Automatic building of terminological hierarchies. For instance, automatic clustering of documents into nested classes [3] or subsumption relations between terms [7].
- Extraction of links between documents with similar keywords [4].
- Query expansion with phrases suggested by the system [1].

Most or all of this work has been done only for monolingual retrieval. It is, however, in a multilingual environment where phrasal information is most likely to enhance retrieval, as shown e.g. in [2]: the ambiguity produced by translating separately each term in the query can be greatly reduced by considering possible translations for larger indexing units.

This paper proposes a way of extracting and using phrasal information in an Interactive Multilingual Retrieval environment. The system, "Website Term Browser" (WTB[1]), applies NLP techniques to perform the following tasks:

1. Terminology Extraction and Indexing.
2. Query Processing and Translation
3. Browsing by phrases.

The next sections explain each part of the system in greater detail.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

2. TERMINOLOGY-BASED INDEXING
The collection is processed to obtain a large list of terminological phrases. The detection of the phrases in the collection is based on syntactic patterns (figure 1) applied over the tagged documents. The selection of phrases is based on document frequency and term subsumption. Such processing is performed separately for each language (Spanish, English and Catalan in the current version).

1. N N	1. A N [N]
2. N A	2. N N [N]
3. N [A] Prep N [A]	3. A A N
4. N [A] Prep Art N [A]	4. N A N
5. N [A] Prep V [N [A]]	5. N Prep N

Figure 1. Syntactic patterns for Spanish.and English

Rather than relying on lexical dispersion, as in [1], we reuse, in a relaxed way, a terminology extraction procedure [6] originally meant to produce a terminological list to be used by documentalists in a thesaurus construction process. For our purposes, such a list is more useful than the final thesaurus items, which are more conceptual and less related to language usage.

3. QUERY PROCESSING
In query translation for Cross-Language Retrieval, term translation ambiguity can be drastically mitigated by restricting the translation of the components of a phrase into terms that are highly associated as phrases in the target language [2]. This process is generalized in the Website Term Browser as follows:

1. Lemmatized query terms are expanded with semantically related terms in the query language and all source languages using the EuroWordNet lexical database [8].
2. Phrases containing some of the expanded terms are extracted. The number of expansion terms is usually high, and the use of semantically related terms (such as synonyms or meronyms) produce a lot of noise terms. However, the ranking via phrasal information discards most inappropriate combinations, both in the source and in the target languages.
3. Unlike batch cross-language retrieval, where phrasal information is used only to select the best translation for individual terms according to their context, in this process all salient phrases are retained for the interactive selection process.

4. In a first pass, documents are ranked primarily according to the frequency and salience of the relevant phrases that they contain.

4. BROWSE BY PHRASES INTERFACE

Figure 2 shows the WTB interface. The query process produces a ranking of documents and a ranking of phrasal expressions that are salient in the collection and relevant to the user's query. Both kinds of information are presented to the user, who may directly click on a document or browse the ranking of phrases.

Phrases in different languages are shown to users ranked and hierarchised, according to:

1. Number of expanded terms contained in the phrase. The higher the number of terms within the phrase, the higher the ranking. Original query terms are ranked higher than expanded terms.
2. Salience of the phrase according to their weight as terminological expressions. This weight is reduced to within-collection document frequency if there is no cross-domain corpus to compare with.
3. Subsumption of phrases. For presentation purposes, a group of phrases containing a sub-phrase are presented as subsumed by the most frequent sub-phrase in the collection. That helps browsing the space of phrases similarly to a topic hierarchy.

5. CONCLUSIONS

The Web Term Browser is, to our knowledge, the first interactive search engine that makes use of phrasal information to process queries and suggest relevant topics in a fully multilingual setting. This work should help bridging the gap between research in CLIR algorithms (that use phrasal information to restrict the set of candidate translations) and interactive CLIR, where the focus has been on interactive selection of translation terms and foreign-language document selection [5].

6. REFERENCES

[1] Anick, P. G. and Tipirneni S. The Paraphrase Search Assistant: Terminological Feedback for Iterative Information Seeking. Proceedings of ACM. 1999.

[2] Ballesteros, L. and Croft W. B. Resolving Ambiguity for Cross-Language Information Retrieval. Proceedings of ACM SIGIR'98. 1998.

[3] Hearst, M. A. and Pedersen J. O. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. Proceedings of ACM SIGIR'96. 1996.

[4] Jones, S. and Staveley M. S. Phrasier: a System for Interactive Document Retrieval Using Keyphrases. Proceedings of ACM SIGIR'99. 1999.

[5] Oard, D. Evaluating Cross-Language IR: Document selection. Proc. CLEF'2000: Springer-Verlag; 2001.

[6] Peñas, A., Verdejo, F. and Gonzalo, J. 2001. Corpus-based Terminology Extraction applied to Information Access. In Proceedings of Corpus Linguistics 2001, Lancaster University, UK.

[7] Sanderson, M. and Croft B. Deriving concept hierarchies from text. Proceedings of ACM-SIGIR'99. 1999; 206-212.

[8] Vossen, P. Introduction to EuroWordNet. Computers and the Humanities, Special Issue on EuroWordNet. 1998.

1. The system is available for testing at http://rayuela.lsi.uned.es/wtb

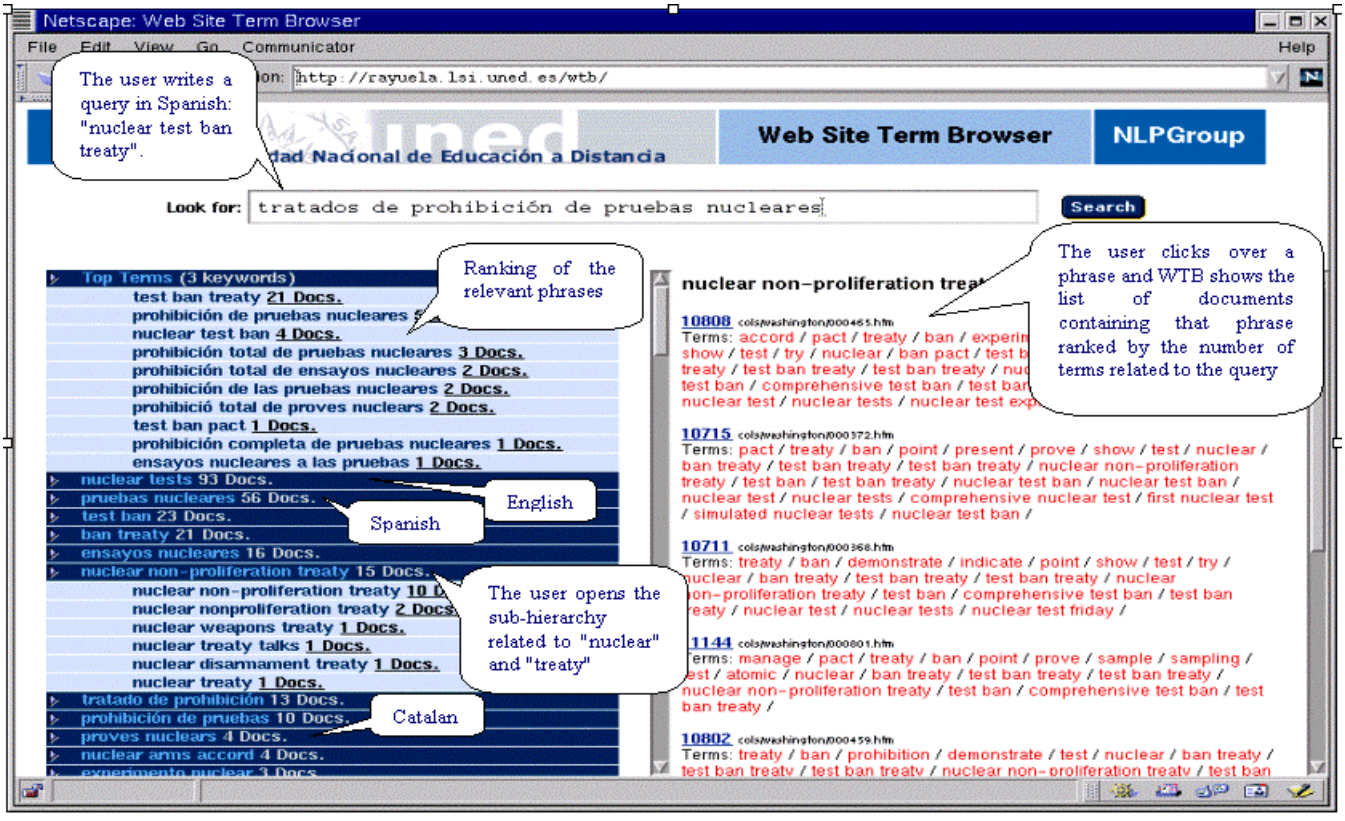

Figure 2. Website Term Browser interface