

Cross-Language Information Access through Phrase Browsing

Anselmo Peñas, Julio Gonzalo and Felisa Verdejo

Dpto. Lenguajes y Sistemas Informáticos

UNED

Ciudad Universitaria, s/n

28040 Madrid, Spain

{anselmo,julio,felisa}@lsi.uned.es

Abstract: This paper presents a cross-language retrieval system which integrates shallow parsing and lexical semantic databases in an interactive approach to information access. At indexing time, the system extracts a list of phrases for every language in the collection. At search time, the system bridges the gap between the user's query and the relevant phrases in the collection in any language, expanding and translating individual terms and retaining the phrases that are actually relevant in the collection. The user can access information via a standard ranked list of documents or via a hierarchy of phrasal information, in which the selection of a phrase modifies the ranked list and provides access to the documents related to the phrase. This interactive setting, to our belief, optimises the use of simple and robust Natural Language resources and techniques to facilitate cross-language information access.

1 Introduction

Phrasal information has been explored by many researchers as a way to improve text retrieval, in general with a moderate degree of success.

Used in a non-interactive mode (to select indexing terms or to expand/modify queries), phrases have proved to be useful in some contexts (e.g. Internet search engines in [LP99]), but in general the cost of acquiring phrasal information does not correspond with a guarantee of improving results. An interesting example is monolingual retrieval in German as evaluated in the first CLEF (Cross-Language Evaluation Forum) campaign. To some extent, German provides single indexes for many multi-word units, as compounding is largely a morphological process comparing to, e.g., English or Spanish. However, the best results in monolingual German Retrieval were obtained by the systems that performed morphological analysis to break noun compounds into the stems forming the compound.

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

- Handcraft hierarchies based on thesauri (e.g. ERIC¹) or topic hierarchies (e.g. Yahoo²) to browse the document space.
- Automatic building of terminological hierarchies. For instance, automatic clustering of documents into nested classes [IIP96] or subsumption relations between terms [SC99].
- Extraction of links between documents with similar keywords (e.g. [JS99]).
- Query expansion with phrases suggested by the system (e.g. [AT99]).

The work described in [IIP96] is purely monolingual and based on automatic clustering, and thus the access is not clearly based on phrases or complex keywords. The possibility of extending this work by performing a multilingual clustering to obtain a (language-neutral) set of nested classes is very attractive, but little work has been done, to our knowledge, in the field of multilingual document clustering.

[AT99] exploits the tendency of key domain concepts to participate in families of semantically related lexical compounds. The lexical dispersion hypothesis is that "key concepts within a document collection are more likely than other terms to participate in a wide variety of semantically related lexical compounds". While lexical dispersion seems a fruitful idea, a problem is that many key concepts are multiwords and the lexical dispersion of their components does not allow to identify them.

Drawing upon the ideas of [FR86], [SC99] defines subsumption between two terms x and y as "term x subsumes term y if the documents which y occurs in are a subset of the documents which x occurs in". Document frequency provides an ordering from general to more specific terms. Subsumption permits building hierarchies of terms, although the meaning of such relationship is not completely clear. As the aim of Sanderson is building concept hierarchies, one of the main problems is word ambiguity: subsumption should be applied to concepts instead of words. Precise word sense disambiguation tools are not available yet, and thus subsumption cannot be applied to word senses. For this reason, subsumption is applied to the top documents of a search, where terms are expected to be used with similar senses. This local analysis is forced, too, by the computational cost of comparing all pairs of words for subsumption detection.

[JS99] uses a tool to extract keyphrases from a collection. About 10 key phrases are assigned to each document. Through these key phrases it is possible to access the documents that share some of them. Documents are then interlinked and can be navigated using that links. A vector space with all the key phrases in the collection is built, where similarity between documents is used to rank related documents.

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 [BC98]: the ambiguity produced by translating separately each term in

¹ ERIC: <http://ericae.net>

² Yahoo: <http://www.yahoo.com>

the query can be greatly reduced by considering statistically plausible translations for larger indexing units.

This paper proposes a way of extracting and using phrasal information in an Interactive Cross-language Retrieval environment. The system, "Website Term Browser" (WTB³), applies known NLP techniques to perform the following tasks:

1. *Phrase extraction.* The collection is processed to obtain a large list of phrases. For specific-domain collections, statistics are compared with a different domain in order to assign higher weights to domain-specific phrases [PVG01]. For broader collections, the weighting is based in document frequency. Such processing is performed separately for each language (Spanish, English and Catalan in the present version). Rather than relying on lexical dispersion, as in [AT99], we reuse a terminology extraction procedure originally meant to produce a terminological list to be used by documentalists in a thesaurus construction process [PVG01]. For our purposes, such a list shows to be more useful than the final thesaurus items, which are more conceptual and less related to language usage.
2. *Query processing.* The system looks for syntactic/semantic/translation variants of the terms in the query, and locates the most salient phrases in every target language. These phrases are used to produce a ranking of the documents. Such process is related to query translation in Cross-Language retrieval, where candidate translations are disambiguated implicitly by choosing translations that form salient phrases in the target languages. In our interactive system, however, there is no need to retain only the most probable phrase translation; all the salient phrases in the target language are presented as potentially relevant information in the interactive refinement process.
3. *Interactive browsing environment.* The relevant phrases in every language are organised and presented to the user hierarchically, together with the ranked documents. The user selects the phrase that addresses or refines his query, and the documents are re-ranked according to this feedback information. The ambiguity problem described in [SC99] is largely mitigated by the fact that phrases tend to be much less ambiguous than single terms. Unlike [JS99], the system retains all phrases appearing in a document, and it is only at query time that such phrases are ranked and organised for presentation to the user.

This approach seems particularly well suited for specific domains where the phrasal vocabulary is rather specialised, but the searcher is not necessarily familiarised with the preferred vocabulary used in the collection (specially in a foreign language collection). The system, to our belief, is unique in covering the distance between the query vocabulary and the collection terminology, considering variations in document language, syntax and semantics to match relevant phrases in the collection given a user's query.

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

³ Website Term Browser is available for testing at <http://rayucla.lsi.uned.es/wtb>

2 Phrase extraction and indexing

Since Terminology Extraction (TE) deals with the identification of terms which are frequently used to refer to concepts in a specific domain, our first approach to phrase extraction followed an automatic TE methodology which is divided in three steps [Bo92] [FA99]:

1. *Term extraction* through morphological analysis, part of speech tagging and shallow parsing.
2. *Term weighting* with statistical information, measuring the relevance of each term in the domain.
3. *Term selection*, through the ranking and truncation of terminological lists by weight thresholds.

We implemented our particular version of these steps within the EC-funded project "European Schools Treasury Browser" (ETB⁴) [PGV01]. One of the subtasks of the project is to create a multilingual thesaurus in the domain of multimedia educative resources for primary and secondary school. We designed an automatic procedure to extract a ranked list of terminological expressions from a corpus in the domain, which has been used by documentalists as a preliminary resource to build such a thesaurus.

The collection is processed to obtain a large list of terminological phrases. The detection of phrases in the collection is based on syntactic patterns (figure 1) applied over the documents tagged on their part of speech.

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

However, while the goal in the Terminology Extraction task is to decide which terms are relevant in a particular domain, in the Information Retrieval task it is the user who decides which are the relevant terms according to his information needs. As the query will give the relevant terms, the indexing task is concerned with recall rather than precision of the extracted phrases. This implies:

1. Terminology list truncation is not convenient.
2. Relaxing of terminological phrase patterns is possible⁵.
3. The focus of the TE procedure shifts to a new task: *phrase retrieval*, which requires a new indexing level: phrases are retrieved from the lemmas of the query terms, and documents are retrieved from phrases (figure 2).

⁴ European Schools Treasury Browser, IST-1999-11781 (Information Society Technologies) <http://etb.eun.org>

⁵ For example, the pattern (N [N/A]* [Prep [Art] [N/A]⁺]*) is being used for English, Spanish, Catalan, French and Italian.

3. *Phrase retrieval*: phrases containing some of the expansion terms are retrieved. The number of expansion terms is usually high, and the use of semantically related terms (such as synonyms or meronyms) produces a lot of noise. However, the ranking via phrasal information discards most inappropriate combinations, both in the source and in the target languages.
4. *Term ranking*: unlike batch cross-language retrieval, where phrasal information is used only to select the best translations for individual terms according to their context, in this process all salient phrases are retained for the interactive selection process. The phrases are ranked according to
 - the number of expanded query terms that they contain,
 - their weight as lexicalised expressions in terms of document frequency.
5. *Document ranking*: documents are ranked according to the frequency and salience of the relevant phrases that they contain.

4. Interaction through phrase browsing

The query process produces both a ranking of documents and a ranking of phrasal expressions which 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 a hierarchy of phrases.

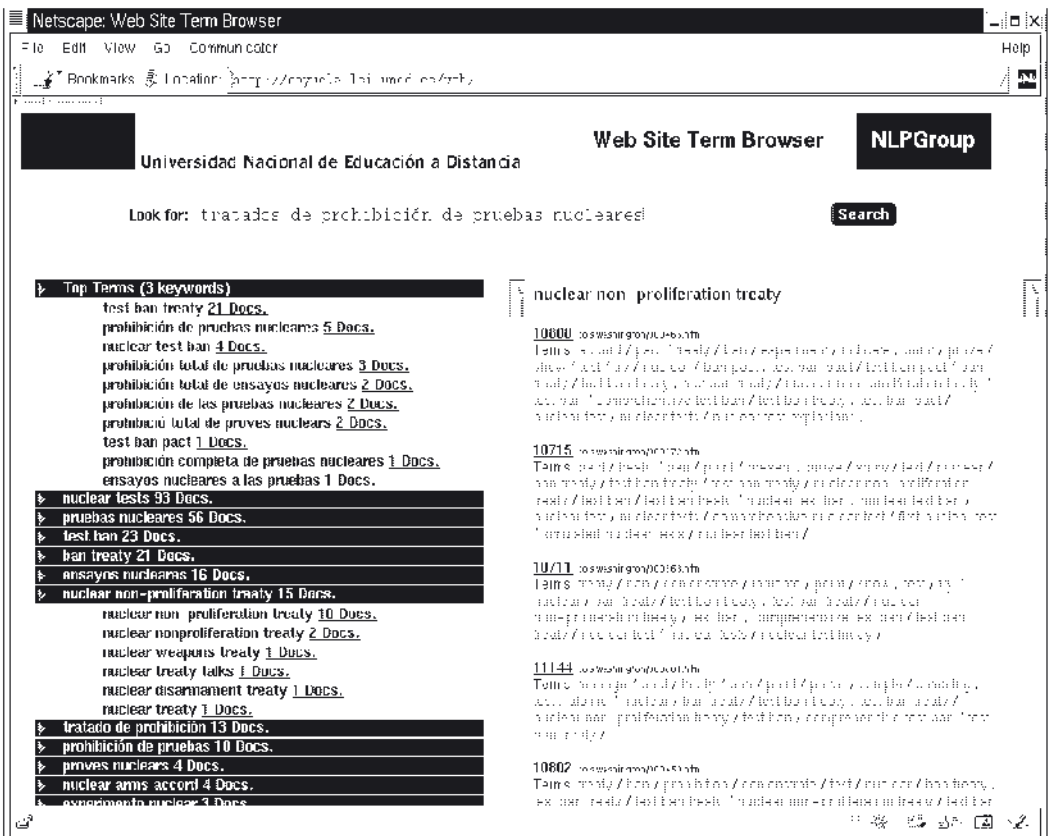


Figure 4. Website Term Browser interface

The figure 4 is a snapshot of the search interface on a collection of international news in English, Spanish and Catalan. The user has written a Spanish query ("tratados de prohibición de pruebas nucleares"). After the query expansion and translation, relevant phrases have been retrieved in the three languages. For instance, "test ban treaty" is an English phrase, "prohibición total de ensayos nucleares" is a Spanish phrase, and "prohibició total de proves nuclears" is a Catalan phrase. The user has selected the English phrase "nuclear non-proliferation treaty" and WTB has unfolded the corresponding sub-hierarchy of terms, altering the document ranking in order to give the user the list of documents which contain the phrase.

5 Evaluation

Our current effort is put on evaluation of the system. Evaluation of interactive retrieval systems can be an elusive task, as proved by previous TRIEC experiences. Previous designs to evaluate interactive cross-language systems do not suit the Web Term Browser, as they are devoted to evaluate a) how the systems helps choosing adequate target language terms when the user has no familiarity with that language [Oa01] [Og99], and b) how the system shows documents written in a foreign language so that the user can judge about their relevance without knowing the language [Oa01]. The Website Term Browser, however, is intended for users that a) have a reasonable passive vocabulary in the target languages, but b) are not necessarily familiarised with the domain-specific terminology used in the collection of documents.

Our first evaluation has been conducted indexing the public documents in UNED⁶ domain (*uned.es*). The collection contains 39,324 web pages written in Spanish (predominantly) and English (mostly containing research information). We have prepared a web interface for searching pages in this domain, that has been made available to teachers and students.

Given a query, the search interface presents two kinds of information to users: the relevant phrases (in both languages) in the left side and, in the right hand side, the ranked documents found by the Google⁷ search engine in the *uned.es* domain. At any time, the user can:

- a) EXPLORE DOCUMENT, select a document to view its contents,
- b) EXPLORE PHRASE, select a phrase to view the related documents in the right hand side, and
- c) RECONSULT WITH PHRASE, query google again with any of the phrases displayed.

The hypothesis is that users will only use phrasal information when Google does not fulfil the information need directly, and one or more phrases seem useful suggestions to the user. To verify this hypothesis, all interactions with the system are logged. The interactions are grouped in sessions, where a session begins with a query to the system, con-

⁶ Distance Learning University of Spain, <http://www.uned.es>

⁷ Google: <http://www.google.com>

tinues with any combination of actions in a), b) or c), and ends when the user leaves the system or poses a new query.

Table 1 shows some statistical information about the first 1000 search sessions made in the system.

	Actions				Total
	QUERY	EXPLORE PHRASE	EXPLORE DOCUMENT	RECONSULT WITH PHRASE	
% of sessions with ACTION	100%	68.8%	71.7%	18.4%	-
Average number of actions per session	1	1.94	1.88	0.36	5.18
First action after QUERY	-	54.6%	37.2%	8.2%	100%
% of sessions finishing with ACTION	-	34.4%	57.5/1000	8.1%	100%
Last action before finishing the session with EXPLORE DOC.	255/575 44.4%	268/575 46.6%	-	52/575 9%	100%

Table 1. Statistics of the first 1000 sessions over WTB

The results indicate that phrasal information is helpful in the searching process. After posing a query, EXPLORE PHRASE is the first action in 54% of the sessions, whereas EXPLORE DOCUMENT (thus preferring Google outcome) is the first action in 37% of the sessions. This means that terms give better expectations of relevance than Google's ranking.

The last row in the table provides evidence on how useful phrases are compared to Google's ranking. It considers sessions where the last action is a document exploration, which include successful sessions ending in an appropriate document. Note that only 575 sessions finish with a document exploration. In a 46.6% of that sessions, the previous action was a phrase selection, while in a 53.4% of them, the previous action was a Google's ranking. This is a strong indication that WTB phrasal information can substantially complement the document rankings provided by the standard search engines.

6 Conclusions

The Website Term Browser is, to our knowledge, the first interactive search engine that makes use of phrasal information to process queries and suggest relevant terms in a fully cross-language setting. This work should help to bridge 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.

An interesting feature of the system is that integrates NLP with a low computational cost: morphosyntactic information (including lexical databases, lemmatisation, part of speech tagging and shallow parsing), multilingual semantic knowledge (via the EuroWordNet database) and implicit disambiguation of translation candidates.

Our approach uses a simple phrase extraction procedure, leaving the tasks of term selection, organisation and presentation as a query-dependent process. This perspective reduces the cost of phrase indexing and copes easily with irrelevant or incorrect phrases: user queries will simply not retrieve them in most cases. The system allows the access to documents from the (most descriptive) phrases they contain, permitting an identification of relevant concepts as an (optional) intermediate step between query formulation and document selection.

The first evaluation of the system has been conducted with 1000 interactive retrieval sessions in the *uned.es* domain. The results show that phrasal information is chosen by users as the best indicator for relevant contents in a significant percentage of searching sessions.

Acknowledgments

This material is based on work supported by ETB project IST-1999-11781.

References

- [AT99] Anick, P. G. and Tipimani S. The Paraphrase Search Assistant: Terminological Feedback for Iterative Information Seeking. Proceedings of 22nd ACM SIGIR Conference Research and Development in Information Retrieval. 1999; 153-159.
- [BC98] Ballesteros, L. and Croft W. B. Resolving Ambiguity for Cross-Language Information Retrieval. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1998; 64-71.
- [Bo92] Bourigault, D. Surface grammatical analysis for the extraction of terminological noun phrases. Proceedings of 14th International Conference on Computational Linguistics, COLING'92. 1992; 977-981.
- [Ch01] Chugur, I. Peñas A. Gonzalo J. and Verdejo F. Monolingual and bilingual dictionary approaches to the enrichment of the Spanish WordNet with adjectives. Proceedings of NAACL Workshop on WordNet and other lexical resources: applications, extensions and customizations.; 2001; Carnegie Mellon University, Pittsburgh. 2001.
- [FA99] Frantzi, K. T. and S. Ananiadou. The C-value/NC-value domain independent method for multiword term extraction. Journal of Natural Language Processing. 1999; 6(3):145-180.
- [FR86] Forsyth R., Rada R. Adding an edge in Machine Learning: applications in Expert Systems and Information Retrieval. Ellis Horwood Ltd. 1986; 198-212.
- [HP96] Hearst, M. A. and Pedersen J. O. Reexamining the Cluster Hypothesis: Scat-

ter/Gather on Retrieval Results. Proceedings of 19th ACM SIGIR Conference on Research and Development in Information Retrieval. 1996.

- [JS99] Jones, S. and Staveley M. S. Phrasier: a System for Interactive Document Retrieval Using Keyphrases. Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval. 1999; 160-167.
- [LP99] Lima, E. F. and Pedersen J. O. Phrase Recognition and Expansion for Short, Precision-biased Queries based on a Query log. Proceedings of 22nd ACM SIGIR Conference on Research and Development in Information Retrieval: 1999.
- [Oa01] Oard, D. Evaluating Cross-Language Information Retrieval: Document selection. Cross-Language Information Retrieval and Evaluation: Proceedings of CLEF2000: Springer-Verlag; 2001.
- [Og99] Ogden, W. Cowie J. Davis M. Ludovik E. Molina-Salgado H. and Shin H. Getting information from documents you cannot read: an interactive cross-language text retrieval and summarisation system. Joint ACM DL/SIGIR Workshop on Multilingual Information Discovery and Access: 1999.
- [PVG01]Peñas, A. Verdejo F. and Gonzalo J. Corpus-based Terminology Extraction applied to Information Access. Corpus Linguistics 2001; 2001; Lancaster University, UK.
- [SC99] Sanderson, M. and Croft B. Deriving concept hierarchies from text. Proceedings of 22nd ACM-SIGIR International Conference on Research and Development in Information Retrieval. 1999; 206-212.
- [Vo98] Vossen, P. Introduction to EuroWordNet. Computers and the Humanities, Special Issue on EuroWordNet. 1998.