Corpus-based terminology extraction applied to information access

Anselmo Peñas, Felisa Verdejo and Julio Gonzalo

{anselmo,felisa,julio}@lsi.uned.es
Dpto. Lenguajes y Sistemas Informáticos,
UNED, Spain

Abstract

This paper presents an application of corpus-based terminology extraction in interactive information retrieval. In this approach, the terminology obtained in an automatic extraction procedure is used, without any manual revision, to provide retrieval indexes and a "browsing by phrases" facility for document accessing in an interactive retrieval search interface. We argue that the combination of automatic terminology extraction and interactive search provides an optimal balance between controlled-vocabulary document retrieval (where thesauri are costly to acquire and maintain) and free text retrieval (where complex terms associated to domain specific concepts are largely overseen).

1 Introduction

Although thesauri are widely used in Information Retrieval, their development requires labor-intensive processes with a high manual cost. On the other hand, new domains with specific conventions and new terminology are continuously appearing. The development of terminology lists is a previous step in thesaurus building that allows the use of automatic techniques in order to facilitate documentalists' labor.

Terminology lists can be seen as an intermediate point between free and controlled access to information. They are used as indexes for document and resources access. In the context of education, they are used in schools, libraries and documentation centers for database access (ERIC). Although there are several available educational thesauri, the domain of new technologies in education is more specific and requires the addition of new terms.

This work has been developed inside the European Treasury Browser (ETB) project which is aimed to build the needed structures to organize and retrieve educational resources in a centralized web site server. One of the main resources which is being developed inside ETB, is a multilingual thesaurus whose terms will be used for describing the educational resources.

We describe the corpus-based method for the terminology extraction procedure used to extract and suggest to documentalists the Spanish terms in the domain of new technologies, primary and secondary education. The method is based on the comparison of two corpora extracted from the web: the first one, an appropriate corpus in the domain and, the second, a corpus in a different and more general domain (international news in our case). The comparison of terms in both corpora facilitates the detection of specific terms of our domain. Section 2 will describe the methodology followed for the Terminology Extraction (TE) procedure based on corpora and morphosyntactic analysis. Exploration and evaluation of the results will be given in section 3.

Thesauri and controlled vocabularies are widely used in Information Retrieval (IR). We argue that the cost of thesaurus construction can be skipped for IR purposes if constrains over terminology are relaxed. This is possible in a framework of interactive text retrieval. TE then, becomes an appropriate tool for providing richer indexing terms to IR indexes. Section 4 analyses the differences between IR and TE which allows the relaxing of TE process, and shows a first prototype where Corpus-Based TE is applied to interactive IR.

2 Corpus-based Terminology Extraction

Terminology Extraction (TE) tasks deal with the identification of terms which are frequently used to refer to the concepts in a specific domain. Typically, automatic terminology extraction (TE, Term Extraction) (ATR, Automatic Terminology Recognition) is divided in three steps (Bourigault, 1992) (Frantzi et al., 1999):

1. Term extraction via morphological analysis, part of speech tagging and shallow parsing.
2. Term weighting with statistical information. The weight is a measure of the term relevance in the domain.
3. Term selection, ranking and truncation of terminological lists by thresholds of weight.

These steps need a previous one in which corpora are obtained and prepared for the TE task. We will distinguish between one word terms (mono-lexical terms) and multi-word terms (poly-lexical terms), extracted with different techniques. The following subsections explain all the TE process performed for the domain of interest: multimedia educative resources for primary and secondary school in Spanish.

2.1 Construction of the corpora

The corpora have been constructed from web pages harvested with crawlers. These web pages need preprocessing because they contain information which can disturb the term extraction process:

1. Treatment of html tags.
2. Deletion of pages in other languages than Spanish. For this task a language recognizer has been used.
3. Deletion of repeated pages and chunks. Due to the continuous update of web site contents, pages with different names but the same content are very frequent. This becomes a problem since identical sequences of words in different documents gives a positive evidence of terminological expressions. Repeated pages and chunks produce noise in statistical measures.

As mentioned above, the automatic terminology Extraction method (Manning and Schütze, 1999) used in this work is based in the use of two corpora:

Educative Resources Corpus

The first corpus is related to the domain and contains useful terminology in order to classify, organize and retrieve multimedia resources for secondary school. The pages of the two following web sites have been collected with a crawler:

- Programa de Nuevas Tecnologías:
  http://www.pntic.mec.es/main_recursos.html
- Aldea Global:
  http://sauce.pntic.mec.es/~alglobal

This corpus has 1,075 documents and 670,646 words.

International News Corpus

With the aim of discarding frequent terms which are not domain specific, a second corpus has been collected from the web. This corpus is composed by 7,364 international news from an electronic newspaper (http://www.elpais.com), and has a size of 2.9 million words.

As explained below, the comparison of term frequencies in both corpora gives a relevance measure for domain terminology.

2.2 Term detection

Texts were first tokenized. Abbreviations, erroneous strings and words from other languages than Spanish were ignored. In order to obtain mono-lexical terms, texts were tagged on their part of speech using (Márquez et al., 1997; Rodríguez et al., 1998) and only nouns [N], verbs [V] and adjectives [A] were extracted. Different forms of the same word are counted by considering their lemma.

In order to detect poly-lexical terms, syntactic pattern recognition has been applied to the collection. Complex noun phrases can be splitted into simpler ones, but all of them must be considered for terminology extraction. The selection of the appropriate, even correct ones will be decided in subsequent steps. For example, "distance education teachers" is a noun phrase which contains a simpler one: "distance education". Term detection phase requires the extraction of all candidate phrases.

The use of syntactic patterns is adequate for this task. Patterns enable to find all the phrases that match them in all the documents of the collection. Patterns are defined as morphosyntactic tag sequences. If the text contains a word sequence whose tags match the pattern, then a new phrase has been recognized. The patterns do not attempt to cover all the possible constructions of noun phrases, but only the ones that appear more frequently in terminological expressions. They were obtained empirically after an iterative refinement of prototypes. The patterns used are listed in figure 1. Along the pattern recognition process, a record of phrase occurrences and the documents in which they appear is built.

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

Figure 1. Syntactic patterns for Spanish.

Patterns recognized 72.453 candidate phrases in the "Educational Resources Corpus". 75% of them appear only once in the whole corpus and largely consist of wrong expressions or irrelevant terms for the domain. As discussed later, it is not necessary to discard wrong expressions for Information Retrieval purposes, but for term extraction, the correctness of the identified expressions is preferable even if some relevant expressions are lost. In other words, precision is more important than recall. Therefore, for the Terminology Extraction task a threshold for the number of phrase occurrences has been defined, and all the expressions that appear only once in the educational corpus have been discarded.

The results of the term detection phase are two lists:

1. a list of lemmas (mono-lexical terms), and
2. a list of terminological phrases (poly-lexical terms).

Every term is associated with its frequency and the number of different documents in which it appears. Such statistics are obtained both for the educational corpus and for the newspaper corpus.

2.3 Term weighting

Term weighting gives a relevance value to every detected term in order to select the most relevant terms in the domain. We have defined a weighting formula that satisfies the following constraints:

1. Less frequent terms in the domain corpus should have less relevance.
2. Highly frequent terms in the domain corpus should have higher relevance, unless they are also very frequent in the comparison corpus or they appear in a very small fraction of the documents in the domain corpus.

The formula considers:

1. Term frequency in the collection.
2. Document frequency of terms in the collection.
3. Term frequency in a more general domain.

$$\text{Relevance}(t, sc, gc) = 1 - \frac{1}{\log_2\left(2 + \frac{F_{t,sc} \cdot D_{t,sc}}{F_{t,gc}}\right)}$$

where

$F_{t,sc}$: relative frequency of the term *t* in the specific corpus *sc*
$F_{t,gc}$: relative frequency of the term *t* in the generic corpus *gc*
$D_{t,sc}$: relative number of documents in sc where *t* appears.

Although a majority of documents in the educational corpus are related to the domain under study, some documents contain very specific terms belonging to different domains (e.g. tales for children about witches). If this kind of documents are long enough they can give high frequencies for non relevant terms. To solve this problem, the measure considers document frequency. One term in the domain must appear in several documents to be considered relevant.

2.4 Term selection

Three criteria have been used in order to reduce the number of candidate terms:

1. Removal of unfrequent terms in the educational corpus (threshold=10). Terms not frequent in the corpus have a low probability of being representative in the domain.
2. Removal of very frequent terms in the newspaper Corpus (threshold=1000). Terms which appear very frequently in other domains have a low probability of being specific to our domain.
3. Selection of the first n (n=2000) terms ranked according to the relevance measure.

The thresholds used for the previous list truncation depend on the number of terms that will be handled in the following phases. As we want to evaluate manually the precision of the automatic term extraction process, the thresholds were adjusted in order to obtain between 2000 and 3000 candidate terms.

Poly-lexical term frequencies do not behave as mono-lexical frequencies. Poly-lexical terms are much less frequent than mono-lexical terms, but a couple of occurrences of a poly-lexical term give high evidence of lexicalised expressions. For this reason, further criteria were needed to add poly-lexical terms to the selection list attending to the relevance of their components. Through the exploration of poly-lexical terms we corroborate that very few compounds without relevant components were indeed relevant. Those terms were ignored and all the poly-lexical terms with relevant components were added to the term list for manual revision.

3 Evaluation of the term extraction procedure

3.1 Visual exploration of results

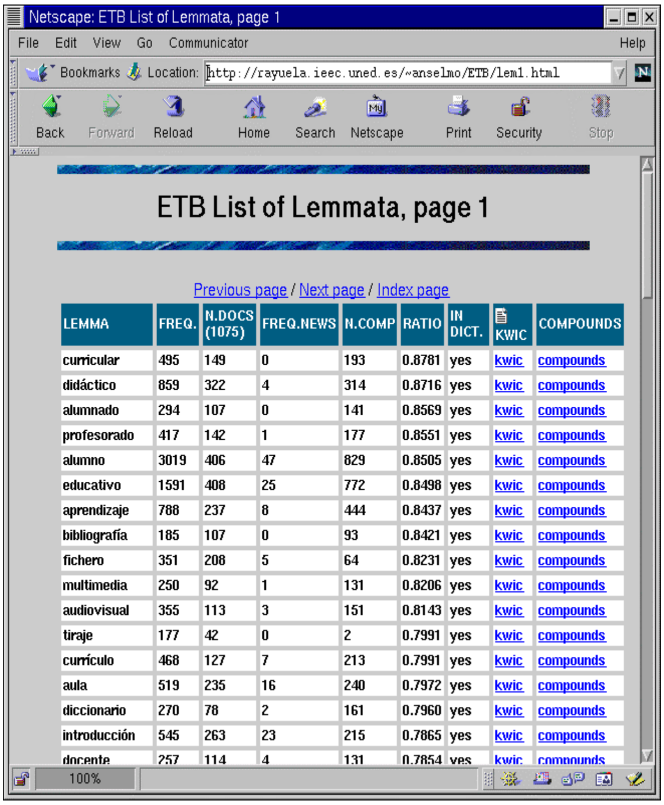
Visual exploration of results is needed during the whole process:

- to help in the decisions of prototype development and refinement,
- to evaluate measures and techniques, and suggest modifications and improvements,
- to give documentalists the possibility of exploring data in order to assist final decisions in thesaurus construction.

We needed a simple, intuitive and comfortable way for data exploration. Thanks to hyperlinking, the use of html pages has resulted appropriate for this task. The html pages containing the extracted data were automatically generated in each iteration of the prototype. Figure 2 shows the pages with the statistical data for mono-lexical relevance measure computation. In this case, terms are ordered by relevance weight. The columns contain the following data:

1. Term frequency in the educational corpus
2. Document frequency in the educational corpus.
3. Term frequency in the newspaper corpus.
4. Number of compounds containing the term.
5. Relevance weight.
6. Whether the term is contained in a electronic dictionary or not.
7. Hyperlink to the page with all the contexts where the term appears in any inflected form.
8. Hyperlink to the page with all the compounds which contain the term in any inflected form.

The pages for poly-lexical terms contain, again, statistical information for each term, and hyperlinks to keyword in context (KWIC) exploration pages. The KWIC pages have hyperlinks to the documents the context belongs to in two versions, text and part of speech tagged files. These links allow a deeper analysis of term contexts and better discrimination of the term senses in the collection.



Location: http://rayuela.ieec.uned.es/~anselmo/ETB/lem1.html

ETB List of Lemmata, page 1

Previous page / Next page / Index page

LEMMA	FREQ.	N.DOCS (1075)	FREQ.NEWS	N.COMP	RATIO	IN DICT.	KWIC	COMPOUNDS
curricular	495	149	0	193	0.8781	yes	kwic	compounds
didáctico	859	322	4	314	0.8716	yes	kwic	compounds
alumnado	294	107	0	141	0.8569	yes	kwic	compounds
profesorado	417	142	1	177	0.8551	yes	kwic	compounds
alumno	3019	406	47	829	0.8505	yes	kwic	compounds
educativo	1591	408	25	772	0.8498	yes	kwic	compounds
aprendizaje	788	237	8	444	0.8437	yes	kwic	compounds
bibliografía	185	107	0	93	0.8421	yes	kwic	compounds
fichero	351	208	5	64	0.8231	yes	kwic	compounds
multimedia	250	92	1	131	0.8206	yes	kwic	compounds
audiovisual	355	113	3	151	0.8143	yes	kwic	compounds
tiraje	177	42	0	2	0.7991	yes	kwic	compounds
currículo	468	127	7	213	0.7991	yes	kwic	compounds
aula	519	235	16	240	0.7972	yes	kwic	compounds
diccionario	270	78	2	161	0.7960	yes	kwic	compounds
introducción	545	263	23	215	0.7865	yes	kwic	compounds
docente	257	114	4	131	0.7854	yes	kwic	compounds

Figure 2. Visual exploration of terms

3.2 Evaluation

The final list of candidate terms contained 2,856 mono and poly-lexical. The terms were manually revised and classified to test the accuracy of the extraction process. Terms were classified as:

- *Incorrect*, when the term is not acceptable in the language (Spanish in the example).
- *Non lexicalised*, when it is correct but it does not have a specific meaning further than the combination of meanings of its components.
- *Not in the domain*, when the term is lexicalised but does not belong to the domain.
- *Adequate*.
- *Specific domain*, when it should be part of a microthesaurus inside the domain.

- *Computers domain.*
- *Variant*, when the term has already been considered in some other flexive form.

Tables 1 and 2 show the classification of the candidate terms.

Adequate	Specific domain	Computers	Variants	Total of terms
1235 43.24%	513 17.96%	59 2.07%	78 2.73%	2856 100%

Table 1. Correct terms.

Incorrect	Not lexicalised	Not domain	Total of terms
151 5.29%	515 18.03%	305 10.68%	2856 100%

Table 2.Incorrect and not adequate terms.

The appropriate terms with the manual classification were used as the input to documentalists to produce the thesauri. They were 66% of the terms automatically selected, an indication that the automatic procedure can be useful in detecting phrases for Information Retrieval purposes, as it is discussed in the next sections.

4 Terminology based Information Retrieval

Traditional Information Retrieval only uses mono-lexical terms for collection indexing. Arbitrary consideration of N word sequences (n-grams) generate indexes too large to be useful. An intermediate approach is to add only terminological phrases to the collection index. In this way, the term extraction procedure described above has been applied to indexing tasks for Information Retrieval.

However, besides the addition of useful terms to document indexes, the use of terminology gives another possibility: instead of navigating through the collection documents, it's possible to navigate through the collection terminology and access the documents from the relevant terms (Anick et al., 1999; Jones et al., 1999). Furthermore, the consideration of two areas, one for document ranking and a second for term browsing, opens an interesting way for interactive information retrieval.

Along these lines, a first monolingual prototype has been developed in order to explore term browsing possibilities for accessing information. Indexing and retrieval are described in the following subsections.

4.1 Indexing

In the terminology extraction task, the goal is to decide which terms are relevant in a particular domain. In the Information Retrieval task, on the other hand, users decide the relevant terms according to their information needs. This implies that, for IR purposes, precision at indexing time should be sacrificed in favor of a higher recall, delaying the determination of which phrases are relevant until the user poses a particular query. For this reason, the terminology extraction procedure described above has been used here ignoring one step: term list truncation.

The term extraction process has been adapted to keep all indexable phrases, ranked but without a cutting threshold. The ranking is used to structure and organize the relevant phrases for a given user's query.

In the indexing phase, the lemmas of nouns and adjectives are linked to the phrases in which they participate. The phrases, in turn, are linked to the documents in which they appear. Indexing levels are shown in figure 3.

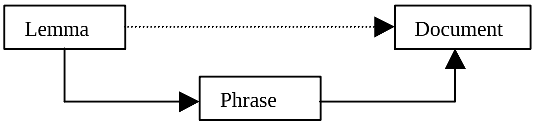


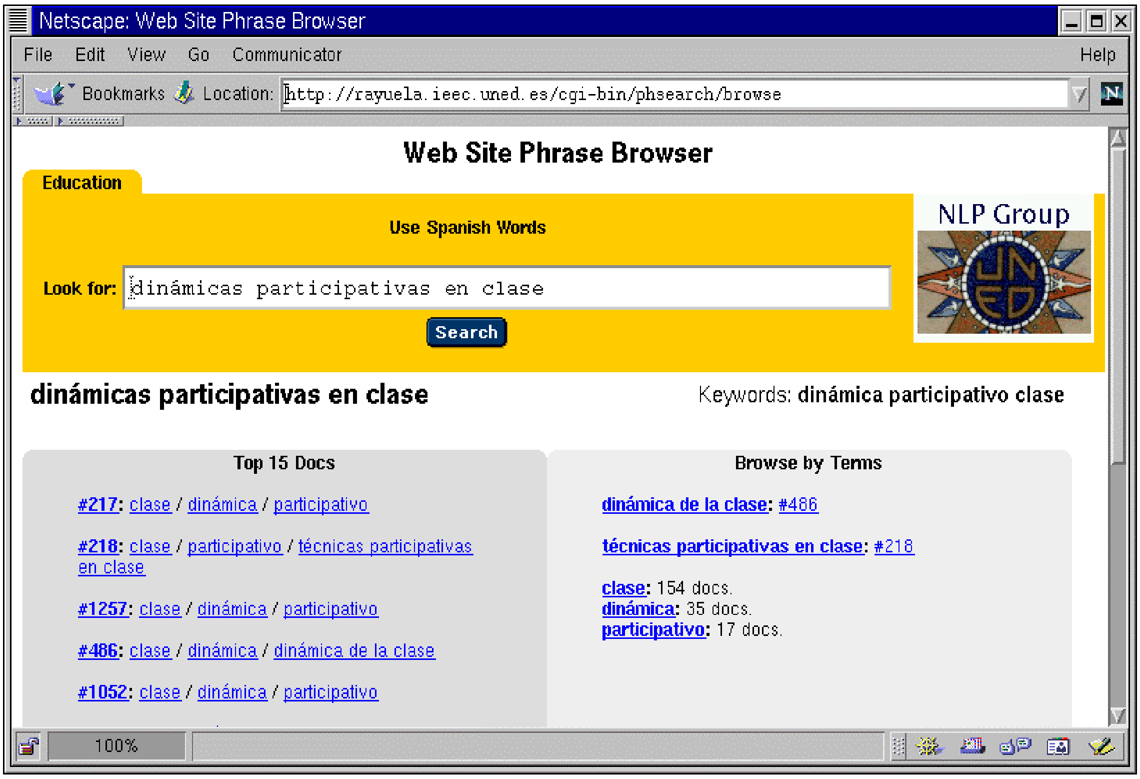
Figure 3. Indexing Levels

The first level of indexing provides access to phrases from the query's isolated terms. The second level supplies access to documents from the selected phrases.

4.2 Retrieval

From the above indexing schema, the retrieval process follows this steps:

1. Through the interface shown in Figure 4, the user provides to the system the searching terms ('*Look for*' text area) . Terms do not need to have any syntactic structure as phrases or sentences.
2. The system obtains the mono-lexical terms related to the query through their lemmatization and categorization.
3. Poly-lexical terms are retrieved from mono-lexical ones through the corresponding index.
4. From poly-lexical terms, documents which contain them are retrieved.
5. According to their relevance weight, terms are organized and shown to users (rightmost area in Figure 4) to allow document access directly from terms.
6. Documents are also ranked and shown in a different area (leftmost area in Figure 4). Ranking criteria are based on the number of identified terms contained in documents.


Netscape: Web Site Phrase Browser
File Edit View Go Communicator Help
Bookmarks Location:
NLP Group
UNED
Search
100%

http://rayuela.ieec.uned.es/cgi-bin/phsearch/browse

Web Site Phrase Browser

Education

Use Spanish Words

Look for: dinámicas participativas en clase

dinámicas participativas en clase

Keywords: **dinámica participativo clase**

Top 15 Docs

#217: clase / dinámica / participativo

#218: clase / participativo / técnicas participativas en clase

#1257: clase / dinámica / participativo

#486: clase / dinámica / dinámica de la clase

#1052: clase / dinámica / participativo

Browse by Terms

dinámica de la clase: #486

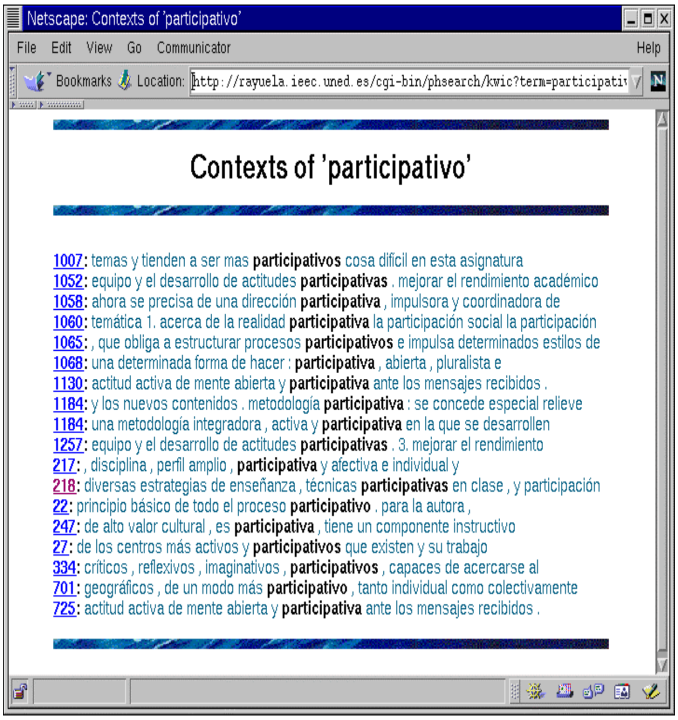
técnicas participativas en clase: #218

clase: 154 docs.
dinámica: 35 docs.
participativo: 17 docs.

Figure 4. Website Term Browser interface.

Both areas, term area and document area have two kind of links:

1. Links for exploring the selected documents.
2. Links for exploring the term contexts in the collection (Figure 5). From these contexts user can select and access the relevant documents.


Netscape: Contexts of 'participativo'
File Edit View Go Communicator Help
Bookmarks Location: http://rayuela.ieec.uned.es/cgi-bin/phsearch/kwic?term=participati

Contexts of 'participativo'

1007: temas y tienden a ser mas **participativos** cosa dificil en esta asignatura
1052: equipo y el desarrollo de actitudes **participativas** . mejorar el rendimiento académico
1058: ahora se precisa de una dirección **participativa** , impulsora y coordinadora de
1060: temática 1. acerca de la realidad **participativa** la participación social la participación
1065: , que obliga a estructurar procesos **participativos** e impulsa determinados estilos de
1068: una determinada forma de hacer : **participativa** , abierta , pluralista e
1130: actitud activa de mente abierta y **participativa** ante los mensajes recibidos .
1184: y los nuevos contenidos . metodología **participativa** : se concede especial relieve
1184: una metodología integradora , activa y **participativa** en la que se desarrollen
1257: equipo y el desarrollo de actitudes **participativas** . 3. mejorar el rendimiento
217: , disciplina , perfil amplio , **participativa** y afectiva e individual y
218: diversas estrategias de enseñanza , técnicas **participativas** en clase , y participación
22: principio básico de todo el proceso **participativo** . para la autora ,
247: de alto valor cultural , es **participativa** , tiene un componente instructivo
27: de los centros más activos y **participativos** que existen y su trabajo
334: críticos , reflexivos , imaginativos , **participativos** , capaces de acercarse al
701: geográficos , de un modo más **participativo** , tanto individual como colectivamente
725: actitud activa de mente abierta y **participativa** ante los mensajes recibidos .

Figure 5. Term contexts with links to documents.

5 Conclusions and future work

The work has been developed in the context of ETB project with two objectives. First, to provide to documentalists the Spanish terminology for a thesaurus building in the domain of new technologies, primary and secondary education. This thesaurus will provide multilingual structure for resources organization and retrieval. This paper has shown the methodology used for the automatic extraction of Spanish terms.

Second, the extracted terminology has been used in a first prototype for information access. The developed search engine gives an intermediate way for information retrieval between free searching and thesaurus-guided searching in an interactive framework. In this prototype, documents are accessible from the terminological phrases suggested by the system after user's query. As users usually don't make use of the same phrases contained in the collection, this approach bridges the distance between the terms used in queries and the terminology used in the collection.

Our present interest is focused in extending this work to cross-language information access, where the use of phrases provides not only a way for document accessing but also an excellent way for reducing ambiguity in query translation (Ballesteros et al., 1998). We plan to extend this way of disambiguation not only for query translation but also for query expansion and term variation. For query expansion and translation we plan to use the synonymy, hyper/hyponymy and meronymy relations of EuroWordNet (Vossen 1998), developing a complete multilingual framework for interactive information access.

Acknowledgments

This work has been partially supported by the European Commission, ETB project IST-1999-11781.

References

ERIC: http://ericae.net

Anick, P. G. and Tipirneni 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.

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.

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.

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.

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.

Manning C. and Schütze H. Foundations of Statistical Natural Language Processing. MIT Press. Cambridge, MA. 1999.

Márquez, L. an Padró L. A flexible POS tagger using an automatically acquired language model. Proceedings of ACL/EACL'97. 1997.

Rodríguez, H. Taulé M. and Turmo J. An environment for morphosyntactic processing of unrestricted Spanish text. Proceedings of LREC'98. 1998.

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