Integrating Spanish Linguistic Resources in a Web Site Assistant

Paloma Martínez*, Ana García-Serrano†, Alberto Ruiz-Cristina†

* Universidad Carlos III de Madrid
Avd. Universidad 30, 28911 Leganés, Madrid, Spain
pmf@inf.uc3m.es

†Universidad Politécnica de Madrid
Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid, Spain
{agarcia, aruiz}@isys.dia.fi.upm.es

Abstract

This work describes a proposal to improve web document retrieval by facing the main problems in document searching: first, traditional web search engines miss documents that are relevant to the user query and retrieve many that are not. Second, the query formulation is not as accessible as it could be, and some users have difficulties in expressing boolean queries. To improve the quality of Internet search engines, two main approaches have typically been adopted: One is the creation of a metasearch engine that makes use of multiple search engines by unifying both the query language and the type of results returned by the different search engines; the other one involves applying NLP techniques for query extensions in order to handle morphological, lexical, semantic and syntactic variations. Focusing on the second approach, we present the research project MESIA (project CAM 07T/0017/1998) for the Madrid Local Government web site (www.comadrid.es). Its main goal is to exploit general purpose linguistic resources to extend user queries in order to enhance the answers provided by AltaVista search engine.

1. Introduction

When we think about the problems posed by the Web, two kinds of problems are distinguished (Baeza-Yates and Ribeiro-Neto, 1999); first, those problems related to the *data* itself (heterogeneous, unstructured and redundant data, distributed resources, high percentage of volatile data, etc.) and second, problems regarding the *user* and his interaction with the retrieval system, especially, the issues of how to specify a query and how to interpret the answer provided by the system. Focusing on the second type of problems, the main topics in the current search engines are:

- Many of the documents retrieved for general queries are irrelevant to the subject of interest and other documents are missing because the query does not include the exact keywords (simple pattern-matching does not deal with morphological, lexical, semantic and syntactical variation).
- This produces inadequate levels of recall and precision.
- Searches are based on the existence of terms in documents (although the META tag could be of great interest) and not in the semantic content of documents.
- Boolean expressions are not so intuitive (novice users have serious problems in formulating boolean queries).
- Furthermore, a given sequence of words does not represent the same query in all search engines: some make stemming, others consider stop words, etc.

So, in late years to improve the quality of the search on Internet, two main approaches have emerged: first, the creation of metasearch engines that make use of multiple search engines performing unification of both the query language and the type of results returned by the different search engines and second, to integrate natural language processing (NLP) techniques such as using query extensions or improving the quality of retrieved information using NLP-based systems.

Focusing on the second issue, we present the MESIA research project.

2. The MESIA Project

2.1. General description

MESIA is an assistant search engine for the Madrid Local Government web site (www.comadrid.es) whose main goal is to enhance the answers provided by AltaVista search engine. This is achieved by extending user queries and answers, as well as by sorting and filtering the results.

The query extension is achieved with the use of linguistic resources. The challenge of integrating Spanish general-purpose linguistic resources consists of managing different information domains; the application is not oriented to a specific domain, as the Madrid Local Government web site contains documents regarding different subjects (culture, education, sports, research…).

The sorting and extension of results are supported by an ontology which stores domain information associated to keywords. This allows to make a semantic interpretation of the extended query words, leading to the extension of results. The sorting is also made following semantic criteria.

2.2. Linguistic resources for the Spanish language

As any system that incorporates analysis and understanding of language, MESIA makes use of lexical resources.

- ARIES (www.mat.upm.es/~aries/), (Goñi et al, 1997) is a Spanish lexical platform developed by the Universidad Politécnica de Madrid and the Universidad Autónoma de Madrid. ARIES is composed of a Spanish lexicon including around

38,000 lemma entries, with 21,000 nouns, 7,300 verbs, 10,000 adjectives and around 500 entries for prepositions, conjunctions, articles, adverbs and pronouns; some access utilities and a morphological analyzer/generator are also included.

The morphological analyzer assigns part-of-speech tags to words. Moreover, a DCG morphological generator for deriving word variants is being incorporated in the MESIA system. This generator allows, for instance, obtaining number and gender variants from a given nominal lemma. Two ARIES lexical entries for the word *becas* ("grants") are shown in Figure 2.

```
"becas"

cat        =        n        /* noun */
concat     =        wl       /* word that accepts a
number morpheme */
agr gen    =        fem      /* fem gender */
agr num    =        plu      /* plural number */
nut        =        plu1     /* plural derivation */
lex        =        beca     /* lemma */

cat        =        v        /* verb*/
concat     =        vl       /* verbal lexeme */
agr pers   =        fem      /* fem gender */
agr num    =        plu      /* plural number */
vinfo tense=        pres     /* present tense*/
vinfo mood=         ind      /* indicative mood */
lex        =        bec      /* lemma */
```

Figure 2: Example of an ARIES entry

- EuroWordNet is a multilingual lexical database for European languages. The Spanish EuroWordNet (EWN), www.hum.uva.nl/~ewn (Vossen, 1998), contains 24.000 nouns and 4.100 verbs. They are structured in "synsets", which are sets of synonymous words related to a single concept. The synset structure of EuroWordNet is based on the American Wordnet for English.

  There are basic semantic relationships between synsets, including hyponymy and hyperonymy. Wordnets are linked to an Inter-Lingual-Index which allows to connect synsets that are related to the same concept but contain words in different languages.
- Finally, the syntactic component of the system is a set of finite cascade automata, (Martínez and García-Serrano, 2000), which solves the ambiguity produced by ARIES POS analyzer and also extracts the relevant query terms (heads and modifiers in a first approach).

2.3. The MESIA approach to query extension

A layered Natural Language Processing (NLP) component has been installed (Figure 1) on a proprietary search engine, Altavista, which is used without further modifications. The current prototype is running in Ciao Prolog (Bueno et al, 1999), at http://tornado.dia.fi.upm.es/mesia/mesia_demo.html (Figure 2 shows the MESIA interface).

The aim is to provide linguistic mechanisms that transform and extend the user questions by integrating use of ARIES morphological database (for handling morphological variation), shallow parsing (for disambiguating POS tags) and use of Spanish EuroWordNet semantic database (for dealing with lexical and semantic variations). Additionally, an ontology for storing document metadata that defines a set of terms and relationships that characterize a domain area is being used.

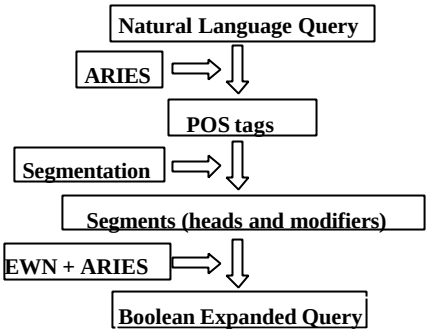

Figure 1: Layered architecture of the Formal Query Generator

In a first step the natural language query is tokenized, and ARIES morphological analyzer assigns the possible part-of-speech (POS) tags to each query word. Then, a segmentation process is performed in order to solve the syntactic ambiguity produced by ARIES part-of-speech analyzer. It also detects the query segments; this task is carried out by a simple phrase segmenter based on cascade finite automata that identify simple noun, prepositional and verbal phrases in the query. This shallow parsing also extracts the significant query terms or keywords (phrase heads and modifiers) that have to be extended.

The use of the segmenter is not always possible: users of search engines do not usually submit properly written queries, but parts of sentences or just separate words. Although one of the MESIA goals is to allow the understanding of natural language sentences, in these cases the system must be flexible enough to extract the significant words from the query without being helped by the syntactic component.

Later in the process, every significant word is expanded using the EWN lexical resource: from the lemma extracted for each keyword, EWN obtains semantically related terms; in a first approach, synonymy and hyperonymy relations are considered. In order to minimize ambiguity, only synonyms with the same POS tag of the word can be considered, as stated in section 2.4. This category is obtained in the segmentation step.

In the next step, the ARIES resource is used again. This time, the morphological generator provides the morphological variants for original query keywords and also for those proceeding from the EWN extension. Nominal and adjectival keywords are then expanded with gender and number variations, while verbal keywords are expanded with all their corresponding number/tense variations. This step is necessary, as the search engine does not perform any kind of stemming/morphological inflection for Spanish language.

Finally, keywords and their variants are converted into a conjunct of disjuncts (conjunctive normal form) in order to obtain the boolean expanded query (the lexical and morphological variants are joined by OR to the original terms and all these sets are linked by AND operator). The EWN query enlargement adds all the synonyms of a keyword in a first approach. The basic hypothesis was that the retrieval process itself performs a disambiguation (the conjunction of terms, as a restriction, eliminates many of the spurious forms).

As an example, the result for the user query "Háblame de las resoluciones de becas postdoctorales" ("*tell me about resolutions of postdoctoral grants*") is:

(resoluciones OR resolución OR soluciones OR solución OR resultados OR resultado)
AND
(becas OR beca)
AND
(postdoctorales OR postdoctoral)

In this example there are three keywords (*resolución*, *beca* and *postdoctoral*) that have been expanded with their corresponding number variations (singular and plural) and one of them (*resolución*) has two synonyms (*solución* and *resultado*).

Figure 2: The MESIA interface

2.4. Preliminary experimental work

In order to evaluate how the linguistic knowledge affects the retrieval results, a first study of seven short (2-12 words) queries (García-Serrano and Martínez, 2001) has been carried out in four types of experiments in order to measure the precision values. Only the top 20 documents obtained in each search process are taken into account, given that common users do not usually explore more results. Queries have been entered in *Altavista Advanced Search* mode. The main features of these experiments are:

- Experiment 1 (baseline): boolean query with relevant terms combined by AND operator.
- Experiment 2 (expansion with ARIES): each relevant term is expanded with gender, number, etc. variations, all of them linked by OR operator (these sets combined with the AND operator)
- Experiment 3 (expansion with EWN): relevant terms are enriched with synonyms/hyperonyms connected by OR operator.
- Experiment 4 (expansion with ARIES and EWN): it includes expansions from experiments 2 and 3.

The result of this first study reveals that ARIES morphology generally enhances the retrieval results. However, extensions with synonyms/hyperonyms, although contribute to obtain as a result documents that are not retrieved in the baseline experiment, affect to precision values (see Table 1).

Baseline	ARIES	EWN	ARIES+EWN
0,558	0.683	0.377	0.454

Table 1: Average precision of first set of queries

In a second study, 20 user queries (3-10 words/query) have been tried using the same four previous defined experiments, but making the following modifications: when a relevant term is a verb, only the infinitive lemma is added as its morphological variation; concerning the lexical variations, only the synonyms extracted from EWN have been considered taking into account the keyword POS tag (noun, verb, adjective, etc.); finally, the *Order by* field provided by Altavista has been filled with the original query relevant terms, in order to obtain the most relevant links at the top of the result list. Table 2 presents the results of this second study, showing that a combination of morphological and semantic features enhances information retrieval. Nevertheless, a separately deep analysis of several queries shows that EWN expansion affects the precision value.

Baseline	ARIES	EWN	ARIES+EWN
0,551	0.667	0.685	0.779

Table 2: Average precision of second set of queries

2.5. Use of domain ontology to improve retrieval results

The integration of the domain ontology is an attempt to increase the semantic component of the system, as there are extensions that cannot be handled by linguistic resources. A user who asks for a topic is probably also interested in related subjects; the trouble is that these subjects are not necessarily described by the query words. For example, if the user makes a query about the Spanish poet Luis Cernuda stating only his name, the previously described linguistic extensions will not lead to obtaining poetry related results, as there are no synonyms or morphological variations that apply.

The domain ontology for the MESIA system is a hierarchical tree structure in which nodes represent issues that are developed in the domain web pages. Nodes are described and related among themselves by a set of keywords. For example, the "theater_festival" node is associated with the words "theater", "festival", "representation", "show" and "play".

In a first approach, the ontology management works with a weight system: if one of the keywords associated to a node appears in the user query, the node gets a

numerical weight, and so do the nodes near it in the tree structure, which get a lower increment. In each access, this process is repeated for every keyword and every node in the ontology. A keyword can be associated to different nodes (for example, keyword "festival" is also associated to the "music_festival" node), and this node will also get a weight. When the weighting process is finished, the nodes which received some weight are sorted according to their numerical value. The input keywords for the ontology access are the query words extended with linguistic resources.

The problem with domain ontologies is the maintenance effort that they require; however, the benefits in the retrieval are important. If the domain is not extremely big, each node can have associated not only keywords but also a list of related links. These links will be displayed as a result, having a 100% accuracy. Every node stores a describing title, so that when their associated links are displayed, the title appears at the top, improving the presentation and making easier for the user to decide whether a link will be interesting or not, before actually entering its associated web page. This is currently working in the MESIA system.

However, if it is impossible to maintain this link list, another approach could take the semantically expanded keywords and perform another search having them as a new query.

In any case, the ontology will still be useful to sort the results: this sorting is made according to the weight obtained in the ontology access. Figure 3 represents part of the domain ontology that is already running in the MESIA system, which is related to cultural issues in Madrid.

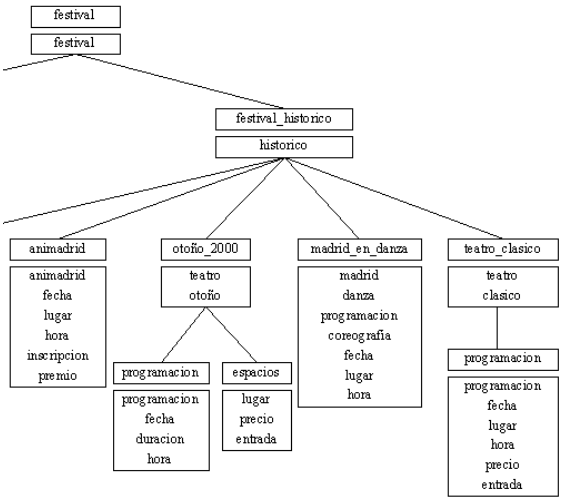

Figure 3: Partial view of the domain ontology

3. Future work: OMNIPAPER Project

As future enhancements for the MESIA system, the use of weighting information (depending on the acceptance of IR system) is proposed; besides, new lexical operators could be considered, such as paragraph or sentence (Mihalcea, 1999).

In some cases, expansion with EWN/WordNet must be constrained; otherwise precision will suffer drastically. However, EWN/WordNet poses some problems for IR (Gonzalo et al, 1998; Mihalcea, 1999): too much fine-grained sense distinctions (the degree of granularity required is task-dependent); lack of domain information (it would be very interesting to manage different domains that allow to store semantic preferences depending on the activated domain); and EWN/WordNet does not include cross-part-of-speech semantic relations.

The large amount of terms included in the ARIES database slows down the analysis and generation processes, specially with complex sentences. In order to minimize this access time, the ARIES database is being redesigned as a "trie" structure.

Regarding the domain ontology, it was stated before that a strong maintenance effort is required; this could be minimized if the web pages of the domain could be classified and integrated in the ontology automatically or semiautomatically; this process would make use of metadata fields. In the simple approach considered in MESIA, if the web designer includes a *meta* HTML field in every document header with a few words describing the page content, a basic ontology access like the formerly described would classify the page in the node which gets the highest weight in the access process.

These subjects, as well as all the other techniques applied in the MESIA project, will be studied and developed in a new project called OMNIPAPER (IST-2001-32174). The main goal of OMNIPAPER project is to create an intelligent and uniform gate to a large number of European digital newspapers, increasing reader's satisfaction in retrieval considering their topics of interest in a self-learning environment. To achieve this, a multilingual navigation and linking layer on top of distributed information resources will be developed. OMNIPAPER project will provide:

- A prototype to have simultaneous and structured accesses to articles of digital European newspapers.
- Multilingual access to different types of distributed information resources.
- A common multilingual thesaurus over distributed digital collections linking them to each other.
- Automatic metadata and keyword extraction as well as document classification techniques that will speed up the process of including new information in the system and will improve information retrieval precision and recall.
- A reference guide (Blueprint) for knowledge retrieval combining emerging and powerful standards (XML, RDF meta-data, Topic Maps).
- A high quality retrieval through learning from user behaviour.

4. Conclusions

The main objective of this work is to show how linguistic knowledge enhances information retrieval as well as user interaction with the Web, specially working on Spanish queries. MESIA metasearcher enlarges usual search (natural language queries) with new morphological and semantic capabilities.

There can be established important contributions for interactive systems intended for casual users that tend to formulate short queries; for instance, in order to solve ambiguity problems in short queries, an adequate proposal is to give the user the possibility of specifying the query related terms from those derived from EWN (not only synonyms) and, in this way, to prepare different query extensions to be displayed to the user.

5. References

Baeza-Yates, Ribeiro-Neto (1999), Ricardo Baeza-Yates, Berthier Ribeiro-Neto: Modern Information Retrieval. Chapter 3. Addison Wesley, 1999.

Bueno, F., Cabeza, D. Carro, M. Hermenegildo, M. López, P. and Puebla, G. The Ciao Prolog System: A Next Generation Logic Programming Environment, TR CLIP 3/97.1(www.clip.dia.fi.upm.es/Software/Ciao/), 1999.

García-Serrano, A. and Martínez, P. An interface Agent with linguistic skills. Applications of Natural Language to Information Systems. Moreno and van de Riet (Eds.), Lectures Notes in Informatics, pp. 45-54, 2001.

García-Serrano, A., Martínez, P., Ruiz-Cristina, A. Linguistic Engineering Approach To The Enhancement Of Web Searching. Proceeding of the 2001 IEEE Systems, Man and Cybernetics Conference.

Gonzalo, J. Verdejo, F. Chugur, I. and Cigarran, J. Indexing with WordNet synsets can improve Text Retrieval, Proc. COLING/ACL '98 Workshop on Usage of WordNet for NLP, Montreal, 1998.

Goñi, J. M. González, J. C. and Moreno, A.ARIES: A lexical platform for engineering Spanish processing tools. Natural Language Engineering, vol 3 no 4, pp. 317-345, 1997.

Martínez, P. and García-Serrano, A. The role of knowledge-based technology in language applications development. Expert Systems with Applications, vol. 19, no 2, pp. 155-160, 2000.

Mihalcea, R. Word Sense Disambiguation and its application to Internet search, Master Thesis School of Engineering and applied Science, Southern Methodist University, 1999.

Vossen, P. The EuroWordNet Base Concepts and Top Ontology. Version 2. EuroWordNet (LE 4003) Deliverable, 1998.