Research in Linguistic Engineering: Resources and Tools[•]

Ana García-Serrano, José M. Goñi-Menoyo ([1])

NLP&IR Group, ETSI Informática, Universidad Nacional de Educación a Distancia (UNED)
C/ Juan del Rosal 16, 28040 Madrid
agarcia@lsi.uned.es

([1]) ETSI Telecomunicación, Universidad Politécnica de Madrid (UPM)
Avda. Complutense, Nº 30, 28040 Madrid
josemiguel.goni@upm.es

Abstract

In this paper we are revisiting some of the resources and tools developed by the members of the Intelligent Systems Research Group (GSI) at UPM as well as from the Information Retrieval and Natural Language Processing Research Group (IR&NLP) at UNED. Details about developed resources (corpus, software) and current interests and projects are given for the two groups. It is also included a brief summary and links into open source resources and tools developed by other groups of the MAVIR consortium (www.mavir.net).

1. GSI - UPM Research interests

One of the GSI main research interest is the analysis of the current available methods and techniques in the field of Multilingual Information Retrieval and their application, especially to the case in which one of the involved languages is Spanish. In addition to that, contributing to the application of new techniques, the improvement of existing ones, and the hybridation or combination of all them are also our objectives. One important aspect in most of the group research is the applied research we use to perform. In the case of linguistic resources, systems and application we use to evaluate all in international evaluation whenever possible.

The availability of linguistic resources is a prerequisite for the development of linguistic engineering systems, whether they implement symbolic or statistical models. Their compilation is a costly task, and special abilities are needed, as well as a deep knowledge of languages. So, linguistic resources should be developed by linguists with expertise and computer engineers. Multilingual and Multimedia Information Retrieval (MMIR) has new challenges as to extend the number of supported languages or to include written and spoken language, images (captions) or audio or video (transcriptions). One of our objectives is the extension of existing systems to languages such as Arabic, at least to a basic level (some of the goals of our Bravo project).

One of our main goals is to explore the usefulness of these systems in real exploitation scenarios, as the web pages of a Local government (Mesia project, TIC-07T/0017/1998 Computational Model for Selective Information Extraction from Short Texts) Technical Support Services (Rimmel, TIN2004-07588-C03-01 Multilingual & Multimedia Information Retrieval and its Evaluation), Recommended Systems (Advice, IST-1999-11305 Virtual Sales Assistant for the Complete Customer Service in Digital Markets), News retrieval and search (Omnipaper IST-2001-32174 Smart Access to European Newspapers, and Nedine, IST-2003-224, V FP) Intelligent News Distribution Network for Multinational Business News Exchange and Dissemination). More details can be found at http://nlp.uned.es/~agarcia or UPM group web pages (http://www2.upm.es/observatorio/vi/index.jsp?pageac=grupo.jsp&idGrupo=265.

In these environments, a huge amount of information is available, both textual –whether structured or not- and images –whether annotated or not-, that must be accessed in a simple, easy and quick way for maintenance technicians. In addition to that, the application domain need semantic techniques and domain modelling (ontologies) to complement standard techniques (statistical), in order to improve the retrieval results. So, metadata and Semantic Web modelling techniques are also tackled in our prototypes.

[•] This work has been partially supported by the Spanish R+D National Plan, by means of the project BRAVO (Multilingual and Multimodal Answers Advanced Search – Information Retrieval), TIN2007-67407-C03-03; and by the Madrid's R+D Regional Plan, by means of the project MAVIR (Enhancing the Access and the Visibility of Networked Multilingual Information for the Community of Madrid), S-0505/TIC/000267.

These general objectives can be implemented via GSI current projects (Bravo and MAVIR), by the development of a platform that integrates linguistic components and tools and resources (existing ones or new-developed) in order:

- To apply/enhance this platform to Multilingual and Multimedia Information Retrieval.
- To apply this platform for Information Extraction and the semi-automatic generation of ontologies/ repositories, using techniques from Knowledge Engineering for restricted domains..
- To evaluate the tools in existing international evaluation fora.
- To incorporate linguistic resources to obtain a better results.

1.1 Available Linguistic resources

The available linguistic resources and software tools developed in the research group (some of them as outcomes of research projects) are:

- **Lexical resources:** The Lexical Database ARIES was developed in former research projects by part of the research team. This resource is under continuous enhancement, partly during the development of our projects. The database consists of the lexical database, declarative rules for inflectional morphology and for allomorph expansion, and related documentation ARIES was developed due to the dramatic lack of lexical resources for Spanish language in 1995.

Authors: José M. Goñi-Menoyo, José C. González-Cristóbal (Universidad Politécnica de Madrid), Antonio Moreno Sandoval (Universidad Autónoma de Madrid). The "Grupo de Sistemas Inteligentes" (GSI-UPM) of Universidad Politécnica de Madrid, and the "Laboratorio de Lingüística Informática" of Universidad Autónoma de Madrid (LLI-UAM), are well-known research groups, with ample activity in several Natural Language Processing issues. Licence agreement for "BASE LEXICA DE LA PLATAFORMA ARIES" granted to "DAEDALUS". May, 2006 (http://www.daedalus.es). This resource was inscribed in the Spanish Intellectual Property Registry in 1996.

Reference: (Goñi-Menoyo et al, 1995)

- **TRIELIB library:** As a joint outcome of the work of some researchers of the UPM team and from the company DAEDALUS-Data, Decisions and Language, S.A., a university spin-off and EPO of this project. TRIELIB is a trie-based software library for indexing textual entries and its associated information. Its main characteristic is that access time for a lexical entry is independent of the size of the lexical database size. An indexing and retrieval system for information retrieval has been built on top of TRIELIB. It is implemented in standard C++ suitable for installing in Linux or Windows platforms.

Inscription requested in the Spanish Intellectual Property Registry on December 2006. It was developed jointly by Universidad Politécnica de Madrid and the company DAEDALUS-Data, Decisions, and Language, S.A. Authors: José Miguel Goñi Menoyo, José Carlos González Cristóbal (UPM), Jorge Fombella Mourelle, and Julio Villena Román DAEDALUS. The exploitation rights of TRIELIB were granted to DAEDALUS in September 2006. A Licence for using TRIELIB has been granted to Mundinteractivos, S.A. (propietary company of elmundo.es), June 2006.

Reference: (Goñi-Menoyo et al 1996)

- **Lexical resources management tools:** A library of tools for lexical resources management based on the TRIELIB library as the **Morphological processing module:** The morphological model used in the ARIES lexical database is well adapted to the trie data structure used in the TRIELIB library. It allows us a very efficient processing of the morphemes grouped in continuation classes.

- **Specific tools:** Initial versions of specific tools for the different linguistic processes are available, such as stemmers, text segmenters, stopwords lists, and others. Some of them are obtained from other research groups or companies, being open-source software. **IDRA for InDexing and Retrieving Automatically** collection of texts, is distributed with licence GPL 3.0 http://sourceforge.net/projects/idraproject/.

1.2 Evaluation in international fora

The research team has participated in the CLEF international Cross-Language Evaluation Forum, (http://www.clef-campaign.org), with other participants from UC3M, Daedalus and UAM organized in the so-called Spanish project MIRACLE (TIC-07T/0055/2003 Multilingual Information Retrieval and its CLEF Evaluation). We have been in seven evaluation campaigns as many other international research groups, whose research targets related to multilingual IR. As an example, in 2006, the groups were 90 (74 in 2005): 60 from Europe, 14 from USA, 10 from Asia, 4 from South-America and 2 from Australia.

The MIRACLE contribution to CLEF has been in several tracks:

- **AdHoc:** General experiments for Monolingual, Bilingual and Multilingual IR from a multilingual document collection. This is the CLEF original track, which attracts most participation.
- **ImageCLEF:** Information Retrieval from image collections, using IR techniques over textual annotations to images as well as content-based retrieval.
- **GeoCLEF:** Multilingual IR on the multilingual document collection, using geographical information to support the queries.
- **WebCLEF:** Multilingual IR over a collection of pages extracted from official sites from several European countries (not in 2006).
- **iCLEF:** Interactive experiments for IR, targeted to evaluate the efficiency and facility of use of the systems (not in 2006).
- **QA-CLEF:** Question Answering experiments over multilingual document collections.

These campaigns do not permit to sort the different participating systems and approaches in an absolute manner, since the evaluation results compare only a particular type of experiments but not the others, the texts are of a very specific category (news, caption of images etc) and sometimes they have not changed significatively for years. In spite of that, we can say that they have been very satisfactory for us, and we had submitted experiment runs in a significative number of languages. In fact, almost all possible languages in CLEF.

In 2005, we started an initial participation in an IR evaluation forum, oriented towards Asiatic languages (Chinese, Japanese and Korean), the NTCIR. Our results obtained are limited to have been able to submit results for such different languages.

For next years we can organize a new participation in:

- The new CLEF (the original one ends 2009), maybe limited to a lesser number of tracks.
- NTCIR (*NII-NACSIS Test Collection for IR Systems)* (http://research.nii.ac.jp/ntcir/)
- TC-STAR (TC-Star Evaluation Workshop on Speech-to-Speech Translation, http://www.elda.org/tcstar-workshop/),
- TREC (Text Retrieval Evaluation Conferences, http://trec.nist.gov/),
- ACE (Automatic Content Extraction, http://www.nist.gov/speech/tests/ace/),
- INEX (Initiative for the Evaluation of XML Retrieval, http://inex.is.informatik.uni-duisburg.de/), and others.

1.3 IDRA: An on-going research system

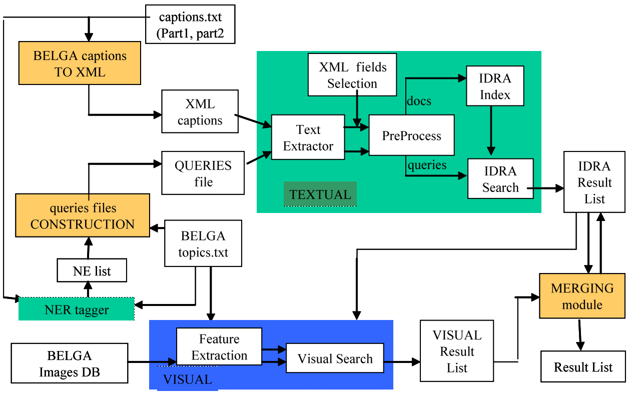

Figure 1: UPM global IR tool

The global Information Retrieval Tool (shown at Fig. 1) includes our own implemented system IDRA (InDexing and Retrieving Automatically), and the Valencia University CBIR system. This year we participate at CLEF 2009 campaign with this tool[1] (ImageCLEF Photo Retrieval Task, http://www.imageclef.org/2009/photo). The main goal, using IDRA with such a large collection (BELGA photo collection, Belga News Agency, http://www.belga.be.), was to analyze how the obtained results from the textual module could be improved using information from the content-based module. A global strategy for all experiments has been that the Content-Based module always starts working with a selected textual results list as part of his input data (different from our participation at ImageCLEF 2008 (Garcia-Serrano et al, 2008), (Granados, 2009a)).

The so-called "BELGA Collection" which contains 498,920 images from Belga News Agency. Each photograph is accompanied by a caption composed of English text up to a few sentences in length. One example of a query for the collection is:

```
<top>
<num> Number: 0 </num>
<title> soccer </title>
<clusterTitle> soccer belgium </clusterTitle>
<clusterDesc> Relevant images contain photographs of the Belgium team in a soccer match. </clusterDesc>
<image> belga38/00704995.jpg </image>
<clusterTitle> spain soccer </clusterTitle>
<clusterDesc> Relevant images contain photographs of the Spain team in a soccer match. </clusterDesc>
<image> belga6/00110574.jpg </image>
<clusterTitle> beach soccer </clusterTitle>
<clusterDesc> Relevant images contain photographs of a soccer beach match. </clusterDesc>
<image> belga33/06278068.jpg </image>
<clusterTitle> italy soccer </clusterTitle>
<clusterDesc> Relevant images contain photographs of the Italy team in a soccer match. </clusterDesc>
<image> belga20/1027435.jpg </image>
<clusterTitle> soccer netherlands </clusterTitle>
<clusterDesc> Relevant images contain photographs of the Netherlands team in a soccer match or the teams in Netherlands' league. </clusterDesc>
<image> belga10/01214810.jpg </image>
<clusterTitle> soccer -belgium -spain -beach -italy –Netherlands </clusterTitle>
<clusterDesc> Relevant images contain photographs of any aspects or subtopics of soccer which are not related to the above clusters. </clusterDesc>
<image> belga20/01404831.jpg </image>
```

IDRA textual retrieval is based on the VSM approach using weighted vectors based on the TF-IDF weight. Applying this approach, a representing vector will be calculated for each one of the image captions in the collection. The components of the vectors will be the weight values for the different words in the collection. When a query is launched, a vector for that query is also calculated and compared with all the vectors stored during the index process. This comparison will generate the ranked results list for the launched query.

The textual retrieval task architecture can be seen in the Fig. 1. Each one of the components takes care of a specific task. These tasks will be sequentially executed:

[1] Section extracted from reference (Granados, 09b)

- **Text Extractor.** Is in charge of extracting the text from the different files. It uses the JDOM Java API to identify the content of each of the tags of the captions XML files. This API has problems with some special characters, so it is needed to carry out a pre-process of the text to eliminate them.

- **Preprocess.** This component process the text in two ways: special characters deletion (characters with no statistical meaning, like punctuation marks, are eliminated) and stopwords detection (exclusion of semantic empty words from a new constructed list, different from last year one).

- **XML Fields Selection.** With this component, it is possible to select the desired XML tags of the captions files, which will compound the associated text describing each image. In the captions XML files there are eight different tags (DOCNO, TITLE, DESCRIPTION, NOTES, LOCATION, DATE, IMAGE and THUMBNAIL). In the index process, the selected tags from the captions XML files had been three: TITLE, DESCRIPTION, and LOCATION.

- **IDRA Index.** This module indexes the selected text associated with each image (its XML caption). The approach consists in calculate the weights vectors for each one of the images selected texts. Each vector is compounded by the TF-IDF weights values [16] of the different words in the collection. TF-IDF weight is a statistical measure used to evaluate how important a word is to a text in a concrete collection. All the weights values of each vector will be normalized using the Euclidean distance between the elements of the vector. Therefore, the IDRA Index process update the next values for each one of the words appearing in the XML captions collection.

- **IDRA Search.** For the query text is also calculated his weights vector in the same way as above. Now, the similarity between the query and an image caption will depend on the proximity of their associated vectors. To measure the proximity between two vectors we use the cosine. This value of similarity will be calculated between the query and all the images captions indexed, and the images will be ranked in descending order as the IDRA result list.

To index the BELGA collection, the system needs approximately 2 days to index each one of the 5 parts in which the collection was divided to be indexed. These 5 indexations processes can be executed concurrently. Queries file response time depends on the concrete queries file launched (on the large of the queries texts), but it takes over 10 hours to obtain a results file for 119 queries (119 queries at cluster level).

The general aim of using a Named Entity Recognition (NER) module in our approach was to perform retrieval using the named entities extracted from both the documents and the queries. The C&C taggers (Clark, Curran 2007) were used off-the-self. After that, our task was reduced to extract those unique linguistic expressions that were tagged as Location, Person or Organization.

However, we soon realized that the resulting annotation was not correctly picking up information which seems to be crucial to determine the topic of the images. More specifically, "Time" would refer to an Organization whereas the description "Time magazine correspondent" would refer to a person and, as such, the modifier correspondent seems relevant to describe the situation captured by a given image. Most of these modifiers would not be picked-up by a NER tagger because they are not in uppercase.

Our proposal aims (Agerri et al 2009) to exploit the interaction between the various levels of annotations (POS, NER and Chunks) provided by the C&C taggers in order to obtain a better bracketing of named entities. The general idea is to create foci consisting of those words or expressions marked-up as named entities. Whenever the C&C tagger annotates a word as a named entity, a chunk/phrase is built around it by attaching those surrounding/satellites terms that act as modifiers of the named entity according to their POS and membership to a particular chunk. We are currently able to deal with periods and abbreviations, with prepositions, adjectives and nouns. This approach allows us to extract entities such as Paris-Roubaix race, princess Mathilde, Leonardo da Vinci international airport (instead of Leonardo da Vinci), District of Columbia, Royal Palace of Brussels, etc.

The following caption (id 1470132) is illustrative of our procedure:
"American photojournalist James Nachtwey in a file photograph from May 18 2003 as he is awarded the Dan David prize in Tel Aviv for his out standing contribution to photography. It was announced by Time magazine on Thurs day, 11 December 2003 that Nachtwey was injured in Baghdad along with Time magazine senior correspondent Michael Weisskopf when a hand grenade was thrown into a Humvee they were traveling in with the US Army. Both journalists are reported in stable condition and are being evacuated to a US military hospital in Germany."
On the one hand, the named entities annotated by the tagger for this text are: American James Nachtwey, Dan David, Tel Aviv, Baghdad, Michael Weisskopf, US Army, US, Germany, and Nachtwey. On the other, the descriptions extracted by our system are: American photojournalist James Nachtwey, Dan David prize, Tel Aviv, Baghdad, Time magazine senior correspondent Michael Weisskopf, US Army, US military hospital, Germany and Natchtwey. It is particularly noticeable that our system was able to recognize Time magazine, and that the topic of the caption is about the Dan David prize and a US military hospital in Germany.

The total process of annotation, analysis, extraction of entities/descriptions for the BELGA collection and queries took 87 machine hours (on a standard Pentium 4 PC).

After the evaluation by the task organizers (Paramita et al, 2009), obtained results for each of the submitted experiments are: the mean average precision (MAP), the R-Precision, the precision at 10 and 20 first results, the number of relevant images retrieved (out of a total of 34887 relevant images in the collection), and the cluster recall at 10 and 20. Average values from all the experiments presented to the task for these metrics are also used, as well as the best value obtained for each of the metrics.

We have sent five runs. MirFI1, is our best run for precision metrics (very similar to MirFI2 and MirFI3), and appears in the 16th position in R-Precision classification and in the 19th in MAP one (from a total of 84 submitted experiments). 19 groups participate in the task and only 6 of them obtain better precision results than our best experiment.

Regarding the diversity metrics (cluster recall at 10 an 20, CR@10 and CR@20), MirFI4 and MirFI5 obtain our best diversity values, appearing in position 11th (over 84) in cluster recall classification, and being the 5th best group from all the 19 participating ones.

Comparing obtained results from experiments MirFI1 and MirFI2, we can see that not using CD (cluster description) tag from the topics is a quite better for precision results and very similar in diversity ones. So we can say that the addition of this field in the queries construction step was not very useful. Obtained results for experiments MirFI2 and MirFI3 are almost the same. So we can conclude that the use of the ENRICH merging algorithm with the visual re-ranked results list, does not affect the results in a significant way.

MirFI3 and MirFI4 are differentiated in the way of constructing the second half of the queries (topics from 26 to 50). As explained in sections 2.2 and 3, MirFI4 extracts the text from the captions corresponding to the example images included in the second part of the topics. The evaluation of the results shows that MirFI3 obtains better precision results than MirFI4, but worse diversity ones. One reason is that the use of the captions text adds more information to the queries, which is useful for the diversity aim, is noise for the precision one.

The goal of experiment MirFI5 was to analyze if results could be improved with the use of NER techniques for the construction of the queries. The obtained precision results for this experiment was our worse ones (a lot of noise was introduced) but the addition of entities information to the queries, improves the diversity results. Experiments MirFI4 and MirFI5, also show how this additional information improves the diversity results, but makes the precision ones worse.

2. NLP&IR – UNED research interests

The NLP & IR Group has been involved in research projects and scientific competitions related to design and evaluation of information retrieval systems (both on monolingual and cross-lingual environments), creation and application of large scale lexical and semantic databases, natural language interfaces and discourse modelling in educational contexts. The main research lines of the NLP & IR Group are:

1. *Intelligent Information Access:* multilingual information access; foreign-language search assistants for document retrieval, question answering and image retrieval; text entailment, question answering and information synthesis tasks; multilingual phrase browsing and terminology retrieval; shallow and efficient NLP for information access; intelligent organization, visualization and browsing of search results; evaluation (including design of evaluation metrics and organization of international evaluation campaigns) and knowledge based management using ontologies.
2. *Acquisition and Representation of Linguistic Knowledge:* lexical and semantic databases; multilingual semantic networks (EuroWordNet); multilingual thesauri; word sense disambiguation and web mining for lexical discovery.

The NLP&IR Group is currently involved in the following ongoing projects (more information about past projects can be found at http://nlp.uned.es/projects.html):

- *QEAVIs: Quantitative Evaluation of Academic Websites Visibility.* (Ministry of Science and Technology, Spanish government 2008-2010): Automated Classification of academic websites by topic and language, in order to create ranks with them. The main goal of the project is to improve the accessibility and visibility of academic information on the World Wide Web.

- *TrebleCLEF: Evaluation Best Practice and Collaboration for Multilingual Information Access (*Coordination and Support Action. European Commission, Seventh Framework programme. 2008-2009)*:* TrebleCLEF supports the development and consolidation of expertise in the multidisciplinary research area of multilingual information access (MLIA and CLEF) and disseminates this knowhow to the application communities through a set of complementary activities.

- *MAVIR: Mejorando el acceso y visibilidad de la información multilingüe en red para la Comunidad de Madrid* (2006-2010).*:* Research network co-funded by the Regional Government of Madrid, involving researchers from different fields (computer scientists, technical experts, linguists and documentalists) in order to integrate NLP, IR, Information Extraction and Semantic Web technologies with scientific communication in the Web.

2.2 Available resources

Some of the available resources at UNED research group are:

- **IR system based on Conceptual Clustering** (Authors: Juan M. Cigarrán, researcher, and Julio Gonzalo) http://bender.lsi.uned.es/ModuloWeb/jbraindead.html. It is an information retrieval system that performs clustering of results by automatically selecting descriptors of the documents, formal concept analysis and latent semantic index techniques. The IR system analyzes the results and retrieves a concept lattice (showing the he most general information on top and the most specific on the bottom) and allowing the user to browse across the relevant documents in a different fashion. The system integrates an IR technology, information extraction and formal concept analysis techniques.

Main innovation it is that the technology is a result of a research project (technological push). The added value is that we propose a new fashion of exploring the results and browsing across the relevant information, which allows discovering non explicit relations.

Technical Requirements: The system needs a IR module that retrieves a set of relevant documents from a query expressing the user's information needs. The system is currently using the Yahoo! and Google API to retrieve documents from the Web, but this module is independent.
References: (Cigarran et al, 2005), (Cigarran et al, 2004)

- **Eurowordnet**. (http://www.illc.uva.nl/EuroWordNet/) This research group is the responsible of the Spanish wordnet for Eurowordnet, developed with EU funds, 1996-1999, 4FP (Telematics, LE 4003). The project aimed at building a multilingual lexical database with semantic relations between words in 8 european languages (Spanish, English, Italian, Dutch, French, German, Estonian and Czech). Every monolingual wordnet is linked to the others via an InterLingual Index derived from Wordnet 1.5.

- **HERMES_192 dataset**: a comparable corpus for multilingual news clustering evaluation. This corpus is a compilation of news written in Spanish and English in the same time period from the news agency EFE (http://www.efe.com/), and compiled by the HERMES project (http://nlp.uned.es/hermes/index.html). Manual clustered HERMES_192 dataset is made up of 35 clusters, 2 monolingual and 33 multilingual.

The news were automatically categorized and belong to a variety of IPTC categories (http://www.iptc.org/std/IIM/4.1/specification/IIMV4.1.pdf), including: "politics", "crime law / justice", "disasters / accidents", "sports", "lifestyle / leisure", "social issues", "health", "environmental issues", "science / technology" and "unrest conflicts \ war", but without subcategories.

Some news belongs to more than one IPTC category according to the automatic categorization. Since they were interested in a multilingual document clustering which goes beyond the high level IPTC categories, making clusters of smaller granularity, we carried out a manual clustering with each subset. Three persons read the news and grouped them considering the content of each one. They judged independently and only the identical resultant clusters were selected.

The following data is currently available:

(1) *Analyzed corpus: PoS tagging and Named Entity detection and classification (http://nlp.uned.es/~vfresno/multilingual-clustering-benchmark/stats-hermes-analyzed.html).*
The linguistic analysis of each document was done by means of FreeLing tool (http://garraf.epsevg.upc.es/freeling/) (specifically: morpho-syntactic analysis, lemmatization, and recognition and classification of Named Entities).

(2) *News + Summary - XML format*
(http://nlp.uned.es/~vfresno/multilingual-clustering-benchmark/ HERMES_192.rar)

Reference of this dataset: (Montalvo et al, 2007)

- **ISCORPUS** (http://nlp.uned.es/ISCORPUS/index.html). Has been created for qualitative and quantitative studies of Information Synthesis tasks.

The use of this corpus is only allowed for research purposes. This corpus contains the next directories:

- *Traces*: It contains, for each query and user, all the monitorized actions realized by users along synthesis process, anotated in files. Each line contains three fields: time||action||index of the treated sentence or document. The action field values are: VISUALIZANDO DOCUMENTO: The user get in a new document. ANOTANDO FRAGMENTO: The user adds the sentence to his report. BORRANDO FRAGMENTO: The user deletes the sentence form SALIENDO DEL DOCUMENTO: The user get out from the document. CUESTIONARIO REALIZADO: The user has completed the form.

The form answers are registered at the end of each file: PERSONAS: Answer to the question "Who are the main people involved in the topic?" ORGANIZACIONES: Answer to the question "What are the main organizations participating in the topic?" FACTORES: Answer to the question "What are the key factors in the topic?" P1: Were you familiarized with the topic? P2: Was it hard for you to elaborate the report? P3: Did you miss the possibility of introducing annotations or rewriting parts of the report by hand? P4: Do you consider that you generated a good report? P5: Are you tired? NADA=Nothing. POCO=Little, ALGO=Something. MUCHO=A lot.

- *Contents*: This directory contains, for all queries, the index and content of sentences sets, explored by the user.

- *Reports:* The /reports/ directory contains the indexes of the sentences selected by users during synthesis process, for each query and user.

- *Query texts*: The queryText fich contains the texts showed to the users in order to lead the user task.

Reference: (Amigo et al, 2004)

- **Automatic association of Web directories to word senses**. The aim of this research is the development and application of algorithms to combine lexical information with web directories, in order to associate Wordnet word senses (http://www.cogsci.princeton.edu/~wn/) with ODP (Open Directory Project, http://dmoz.org/) directories. Such associations can be used as rich domain labels and to acquire sense-tagged corpora automatically, cluster topically-related senses and detect sense specializations.

The current algorithm has been evaluated for the 29 nouns (147 senses) used in the Senseval 2 competition, obtaining 148 word sense/ Internet directory associations covering 88% of the domain-specific word senses in the test data with 86% accuracy. The results indicate that, when the directory/word sense association is correct, the training samples acquired automatically from the Internet directories are as valid for training as the original Senseval 2 training instances.

The following data is currently available:
(1) The full set of associations between noun senses in wordnet 1.7 (http://www.cogsci.princeton.edu/cgi-bin/webwn1.7.1) and ODP directories, and a summary file with the coverage of the system (http://terral.lsi.uned.es/ODP/Coverage.gz).
(2) The full set of hyponyms (sense specializations) extracted for Wordnet 1.7 from ODP directories.
(3) The training material obtained automatically for the Senseval 2 WSD test suite, which can be compared with the corresponding hand-tagged training material provided by the Senseval 2 organization, and tested against Senseval 2 test material.

Reference: (Santamaria et al, 2003)

3. MAVIR-TIMM available resources

The Spanish MAVIR project and the Spanish Thematic Network TIMM (Tratamiento de Información Multilingüe y Multimodal (http://sinai.ujaen.es/timm), facilitates the development of the first version of a catalogue for Spanish developed resources, systems and generic software in linguistic engineering (Peinado and Garcia-Serrano 2009).

Even if the MAVIR project is a consortium of research groups in the region of Madrid collaborating in research projects in Information Retrieval, Extraction, Question Answering and related fields, the catalogue includes items from other Spanish groups. MAVIR participant groups are adscribed to Universidad Carlos III de Madrid, Universidad Autónoma de Madrid, Universidad Nacional de Educación a Distancia (coordinator), Universidad Europea de Madrid and Centro de Información y Documentación Científica del Consejo Superior de Investigaciones Científicas.

The contents are classified into:

1. Córpora, Databases and other Linguistic Resources (34 items)
2. Analysers, Taggers and Classifiers (24 items)
3. Task-driven Systems
   a. Dialogue-based systems (3)
   b. Navigators and searchers (5)
   c. Question-answering systems (2)
   d. Automatic translation systems (1)
   e. Automatic summarization systems (2)
4. General purpose tools and software libraries (12)
5. Others and Know-how (2)

The current seven Arabic related items included in this catalogue are:

1. The Arabic WordNet (AWN) is a lexical database of the Arabic language following the development process of Princeton English WordNet and Euro WordNet.
2. ANERcorp is an Arabic NER corpus which consists of 150,000 tokens (which go up to 200,000 tokens after segmentation).
3. SVM Arabic. A Named Entity Recognition model which is trained using an SVM-based approach over a 125,000 Arabic tokens training file.The model allows the user to extract the named entities with an open-domain text and classify them into 4 different categories, namely: person, location, organization and miscellaneous.
4. ANERgazet is a set of 3 Arabic gazetteers (people, locations and organizations) which might be used mainly for the Arabic NER task, but still can be used for other Arabic NLP tasks. Each gazetteer contains a list of Arabic names belonging to the concerned class. The gazetteers were extracted automatically from Arabic Wikipedia and the Web resources and then manually filtered.
5. The Arabic QA. A list of documents that consists of Arabic newswire articles collected from the Web. We have manually built a set of Arabic questions in order to ensure that each question has the correct answer in the documents. The proportions of the question type (factoid, list, ...) are similar to the ones used in the CLEF 2006. The three sets of data are meant to be used as a test-platform for an Arabic Question Answering system.
6. Arabic JIRS is a passage retrieval system for Arabic texts which return a set of relevant passages to the user's query (which is written in Arabic). The Arabic JIRS, indexes Arabic documents and then provides an interface in order to extract passages which are relevant to the user's query.
7. The Arabic WikiPedia XML Corpus. The 30 most frequent categories of the Arabic Wikipedia XML corpus were selected in order to provide a testbed for the single-label categorization task in the Arabic language. The aim of this corpus is to support experiments of supervised and unsupervised classifiers with Arabic-witten texts. The gold standard is provided, as well as the tokenized and untokenized versions of this corpus.

References

(Agerri et al, 2009) R. Agerri, R. Granados, A. García-Serrano, "Extracting descriptions for Image Photo Retrieval" Working Notes (Edition digital) 7th International Workshop in Adaptive Multimedia Retrieval (AMR09) http://nlp.uned.es/amr2009/ 24-25/09/ 2009, ETSI Industriales, UNED, Madrid, PÁGS: # 8, 2009

(Amigo et al, 2004) Enrique Amigo, Julio Gonzalo, Victor Peinado, Anselmo Peñas and Felisa Verdejo. "An Empirical Study of Information Synthesis Task", 42nd meeting of Association for Computational Linguistics (ACL), july 2004, Barcelona.

(Cigarran et al, 2005) J. Cigarrán, A. Peñas, J. Gonzalo, F. Verdejo. (2005) "Automatic selection of noun phrases as document descriptors in an FCA-based Information Retrieval system". International Conference on Formal Concept Analysis (ICFCA 2005). Lecture Notes in Computer Science. Springer-Verlag, vol. 3403.

(Cigarran et al, 2004) J. Cigarrán, J. Gonzalo, A. Peñas, F. Verdejo (2004). "Browsing search results via Formal Concept Analysis: Automatic selection of Attributes". Concept Lattices Proceedings of the Second International Conference on Formal Concept Analysis (ICFCA 2004). Lecture Notes in Computer Science. Springer-Verlag.

(Granados et al, 2009b) R. Granados, X. Benavent, R. Agerri, A. García-Serrano, J.M. Goñi, J. Gomar, E. de Ves, J. Domingo and G. Ayala Título: MIRACLE (FI) at ImageCLEFphoto09 http://clef-campaign.org/2009/working_notes/rgranados-paperCLEF2009.pdf, REF: Working Notes (Edition electronic) Cross Language evaluation Forum (CLEF 2009) 30-09 a 02-10/ 2009, Corfú, Grecia PÁGS: # 10, 2009

(Clark, Curran 2007) Clark, S., Curran, J.: Wide-coverage efficient statistical parsing with CCG and log-linear models. Computational Linguistics, vol. 33, n. 4, pp. 493--553. (2007)

(Garcia-Serrano et al, 2008) Ana García-Serrano, Xaro Benavent, Rubén Granados and José Miguel Goñi-Menoyo. *Some Results Using Different Approaches to Merge Visual and Text-Based Features in CLEF'08 Photo Collection*. Lecture Notes in Computer Science, Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, Revised Selected Papers. ISSN: 0302-9743 (Print) 1611-3349 (Online). Volume 5706/2009. ISBN: 978-3-642-04446-5. Pág.: 568-571.

(Granados, 2009a) Granados, R., García-Serrano, Ana., Goñi, J.M.: La herramienta IDRA (Indexing and Retrieving Automatically). Demostración en la XXV edición del Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural 2009 (SEPLN´09). (2009)

(Goñi-Menoyo et al, 1995) Goñi-Menoyo, J.M.; González-Cristóbal, J.C.; and Moreno-Sandoval, A. (1995). Manual de Referencia de la Plataforma Léxica ARIES, versión 5.0. Universidad Politécnica de Madrid

(Goñi-Menoyo et al 1996) Goñi-Menoyo, J.M.; Fombella-Mourelle, J.; González-Cristóbal, J.C.; and Villena-Román, J. (2006). Biblioteca "TRIELIB". Guía de uso. Informe técnico. Universidad Politécnica de Madrid y DAEDALUS-Data, Decisions, and Language

(Montalvo et al, 2007) Montalvo, S., Martínez, R., Casillas, A. and Fresno, V. *Multilingual news clustering: Feature translation vs. identification of cognate named entities*. Pattern Recognition Letters 28 (16) 2305-2311, Elsevier (2007).

(Paramita et al, 2009) Paramita, M., Sanderson, M., Clough, P.: Diversity in photo retrieval: overview of the ImageCLEFPhoto task 2009. CLEF working notes 2009, Corfu, Greece. (2009)

(Peinado and Garcia-Serrano 2009) Catalogue for spanish developed resources, systems and generic software in linguistic engineering. 2009 On-line edition at www.mavir.org.

(Santamaria et al, 2003) Santamaría, C., Gonzalo, J. and Verdejo, M. F. Automatic Association of Web directories to word senses. (2003) Computational Linguistics 29 (3), MIT Press.