

HERMES, Hemerotecas electrónicas: Recuperación multilingüe y extracción semántica, TIC2000-0335-C03

M.F.Verdejo, J.Gonzalo* LL.Màrquez, LL.Padró, H.Rodríguez** E.Agirre***
NLP Group, UNED TALP Group, UPC IXA Group, UPV

Abstract

The main goal of the HERMES project is to explore computational linguistics techniques in order to evaluate their potential for accessing information contained in multilingual collections of documents, either static or dynamic. We aim to provide fully evaluated and scalable search prototypes to help users searching both a stable news repository and a set of dynamic online news servers. For this, Human Language Technologies in English, Spanish, Catalan and Basque are being developed in order to optimize their use in such Information Access applications. Results obtained so far are given and open challenges are described.

Keywords: Human language technologies, cross-lingual information retrieval

1 Goals and work plan of the project

The main goal of the HERMES project is to explore computational linguistics techniques in order to evaluate their potential for improving multilingual information retrieval, providing accurate responses to users requests. Our approach considers the design and building of resources and techniques in four languages, namely Spanish, Catalan, Basque and English, to process multilingual textual information available in a variety of electronic repositories such as digital libraries, Internet or Intranets sites. To reach this goal, both basic and applied research are needed. Firstly, algorithms to deal with multilingual information are required; some are known for one language and need to be adapted to other languages and for this, corpora are required. Others are open problems and the work to be done will contribute to the state of the art. Secondly, from the application point of view, the tools implemented need to be efficient and reusable, in order to explore better ways of indexing, searching and accessing multilingual information in real scenarios, as well as to offer more functionalities such as summarization. Furthermore, resources and tools should comply and contribute to international efforts in standardization and evaluation. Two applications are foreseen: (1) An environment to consult a large virtual library in four languages (2) A search engine for news-on-line from a variety of Internet newspapers sites. Both applications raise at least two challenges: going beyond

*url: <http://sensei.lsi.uned.es/NLP/>

**url : <http://www.lsi.upc.es/~nlp/>

***url : <http://ixa.si.ehu.es/Ixa>

TIC2000-0335-C03

information retrieval techniques based on matching search terms, and linking documents and queries in an independent way of their source language.

The following set of intermediate, more concrete objectives, outlines the task structure of the work plan:

- a) Improve and/or propose new word sense disambiguation techniques (task 3)
- b) Develop robust techniques to process a document in any of the 4 languages to perform: Language identification, multiword and term recognition, entity recognition and co-reference resolution. (task 3)
- c) Develop new algorithms, to be used as components in the applications, to perform clustering and summarization of documents (task 4)
- d) Create robust techniques to map a query in one language in any of the other languages (task 4)
- e) Define interactive cross-lingual information retrieval tasks, including as proof of concept linguistic techniques. Propose an evaluation framework. (task 5)
- f) Design and implement two applications, to show the performance with different requirements: closed vs. open collections, on-line/off line,...(task 5)
- g) Create, refine, and compile the linguistic resources needed for the above objectives (task 2).

The project structures these objectives in four tasks, plus coordination (task 1), each one organized in subtasks, as detailed in the technical annex. The duration of the project is from January 2001 to December 2003. Relevant aspects of each task are described in the next section.

2 Approach, work done and future directions

Task 2: Document Collections, Resources and Evaluation Criteria

For any sophisticated NLP application such as this project goals, it is necessary to rely on robust, wide-coverage language resources, which includes lexical databases, corpora, and linguistic processors. Additionally, it is necessary to evaluate the coverage and performance of all these resources. Task 2 deals with the creation, updating and adaptation of new and existing resources for NLP processing. The task is organized in three subtasks: collection creation and international evaluation campaigns (this last listed in section 3), 2.2 resources adaptation and 2.3 evaluation.

Regarding **subtask 2.1** (collection creation), the building and compiling stage have been completed, and all the corpora and document collections required in the project have been acquired. Nevertheless, useful corpora will continue being compiled. The collection includes: a) a Spanish-Catalan parallel corpus from *El Periódico* online newspaper. A webbot collects every day the bilingual edition of this newspaper. The data is cleaned, converted to XML, paragraph-aligned and stored, b) a Spanish-Basque parallel corpus from the official gazette of the Basque Country. The data is cleaned, converted to XML, paragraph-aligned and stored, c) a Spanish-Catalan-English comparable corpus from *EFE* news agency. All *EFE* news from 2000 in these three languages has been standardized through XML annotation. The marks include date, author, topic, etc, d) a

Basque corpus from *Egunkaria* newspaper from 2000. It has been standardized through the same XML annotation. These collections have been used in text categorization experiments, and constitutes a valuable resource for cross-lingual Information Retrieval experiments (see task 4) e) Basque and Spanish corpora hand annotated with Named Entities (see task 3.2) Additionally, several specific actions have been undertaken: A 100,000 word Spanish corpus has been annotated manually with PoS tags, in order to train, tune and evaluate taggers and lemmatizers with high fiability. Also, a 800,000 word subset of *EFE* Spanish corpus has been annotated with Named Entities (see task 3.2), enabling the training and evaluation of NE recognisers and Classifiers for Spanish.

Task 2.2 consists of adapting existing resources to the current project, and has two main threads.

(1)WordNet enrichment- on the one hand, the WN-Domain labels developed by (Magnini and Cavaglia, 2001) for WordNet, have been ported to EuroWordNet, enabling its use in Spanish, Catalan and Basque. The use of WN Domains is a useful information source in tasks such as text categorization and word sense disambiguation. Basque EuroWordNet is being extended and revised. Nominal synsets are under thorough revision to ensure the presence of basic lexical items and senses. On the other hand, procedures for hierarchy mapping (Daudé et al., 2001) have been developed, and applied to different WN versions, providing accurate version independence to WN-based applications and resources. The mapping procedures have also been applied to multilingual hierarchy mapping between Spanish dictionary-acquired taxonomies and WN. Currently, prototypes are available for the following areas: a) Example acquisition system. Occurrences of EuroWordNet concepts are automatically acquired from online texts, b) Topic signatures. EuroWordNet synsets are enriched with characteristic term vectors (Agirre et al., 2001), c) Relation acquisition. Derivation and syntagmatic synset relations are acquired from machine-readable dictionaries (Agirre et al., 2002a), d) Selectional restrictions. English verbs are enriched with selectional restrictions (Agirre and Martínez, 2002), e) Synset clustering based on topic signatures, f) WordNet enrichment using Internet directories (Santamaria et al., 2001). Currently up to a 30% coverage, g) Automatic creation of a phrase bilingual Spanish-English dictionary from comparable corpora and alignment algorithms. Currently more than 2 million phrases of two and three words have been collected. These prototypes should be definitive and operative to enrich all EWN concepts before the end of 2002.

(2)Linguistic Processors adaptation- this is mainly an engineering task, which aims to ease the integration of analysers and resources for all languages involved in the project. The proposed architecture is a distributed system based on a CORBA layer, which makes each component independent of the others. In this sense, two XML DTDs have been defined for the corpus storage and analyser outputs. The document-oriented off-line DTD is used to annotate corpus that will be accessed through the IR/QA techniques, and contains information about document attributes (date, author, category), structure (title, paragraphs), and Named Entities appearing in the text. The query-oriented on-line DTD defines the output of the basic distributed Linguistic Processor modules, and encodes information about morphology, part-of-speech, and Named-Entity recognition. Following these DTDs, processors have been embedded into CORBA objects, which can be accessed from any remote machine, following the architecture proposed in (Carreras and Padró, 2002). In addition, a C++ recoding of the slower Perl modules is currently being undertaken in order to speed up the system, which is a critical issue, especially on the on-line processors, since they will be used from the IR/QA end-user applications. Additionally, the robustness and customisation of the Basque processors have been improved (Alegría et al., 2002).

TIC2000-0335-C03

Task 2.3 Deals with the evaluation of the project components at different levels: At the linguistic processor level, the evaluation is performed through usual precision/recall scores, according to each specific processor. At application level, a more general evaluation is required, including user viewpoint evaluation. This implies the evaluation of tasks, for which standard test sets -such as those of CLEF- may be used. In addition, a document (Amigó, 2002) has been issued stating project evaluation criteria and protocols, both to the specific resource level and to the application level.

Task 3: Advanced techniques for language processing

In this task we have joint efforts to produce state-of-the-art systems in the following areas: a) Word sense disambiguation, b) Language identification, c) Recognition and classification of Named Entities and terms and d) Co-reference resolution

Word Sense Disambiguation- in this area all UPV/EHU, UPC and UNED teams have reputable experience. There are two main classes of systems for WSD: a) *Supervised systems*: they need tagged material to train, where tagging is usually performed by hand. There is only enough training material for a handful of words in Spanish, English and Bay enosof d)ofongtheyr ids

Recognition and classification of Named Entities and terms- the UPC team has developed a framework to perform wide-coverage Spanish Named Entity Recognition, which is described in (Arévalo et al., 2002). First, a linguistic description of the typology of Named Entities is proposed. Following this definition architecture of sequential processes is defined for addressing the recognition and classification of *strong* (or simple) named entities and *weak* (or complex) named entities. On the one hand, strong entities are treated in an inference scheme combining several classifiers (based on the AdaBoost learning algorithm) learnt on the basis of very simple attributes requiring non tagged corpora complemented with external information sources (a list of trigger words and a gazetteer)(Carreras et al 2002b). This part of the system has participated in the Shared Task competition organized by the ACL's Special Interest Group on Natural Language Learning during the annual CoNLL workshop, achieving the first position both in Spanish and Dutch languages (Carreras et al., 2002a). On the other hand, the recognition of weak named entities is approached through a context free grammar for recognizing syntactic patterns, which is currently being improved and fully tested-first results in (Arévalo et al., 2002).

Co-reference resolution- the UPC group is currently developing a module for performing co-reference resolution. The architecture of the system is aimed to be language independent. The system is rule-based and its performance is controlled by three rule-sets: i) a rule set for detecting units that could be considered anaphoric descriptions (basically anaphoric pronouns), ii) a rule-set constraining the nominal phrases candidates to co-refer with such anaphoric units, and iii) a rule-set defining preferences between the candidates accepted by the second rule-set. The system follows basically (Palomar et al., 2001) for detecting co-referents of pronouns in Spanish texts. The original system has been adapted to the characteristics of the parser used within Hermes. The system is currently being implemented for Spanish. In the next future it will be tuned to Catalan and English.

Task 4: Advanced techniques for IR/IE: Document Classification, Clustering and Linking

Task 4 involves three subtasks: Document classification, clustering and linking (4.1), Automatic summarisation (4.2) and, Query expansion/translation and conceptual indexing (4.3).

Regarding **Document classification** (or categorisation) several actions have been undertaken. On the one hand text classifiers for the 4 languages involved in the project (Basque, Catalan, English, and Spanish) have been developed using a unified set of categories: the standard IPTC first level 17 categories for news agencies. In doing so, several aspects have been studied: (1) a comparison between several Machine Learning algorithms for performing the task. Specifically, we have tried Bayesian classifiers and SNoW -sparse network of linear Winnow classifiers- for categorising Basque documents, and AdaBoost and Support Vector Machines for dealing with Catalan, English, and Spanish. The overall results are over 80% (F1 measure), which can be considered state-of-the-art, and the best performing algorithm is SNoW for Basque and SVM (with a slight advantage over AdaBoost) for the other languages. (2) We also addressed the problem of collecting a labelled corpus for training the systems. For Catalan, English, and Spanish we used the *EFE* collection of news agency documents (year 2000), which are labelled with internal category codes. We developed an n to m automatic mapping between *EFE* categories to the IPTC categorization scheme in order to automatically obtain training documents. The Basque collection is extracted from the *Egunkaria* newspaper, categorized according to the newspaper sections. Mappings to the IPTC categories plus

a bootstrapping step (with human supervision) have been performed in order to obtain training examples for all categories. (3) Alternatives for the representation of documents: from the common bag of words (plus stemming and stop-words removal) to the coding of entities, and statistically based n-grams of words. For Basque, lemmatisation has proved to be very important, while the representation of documents using only nouns lead to very good results (Arregi and Fernández, 2002). (4) A knowledge-based text classification system based on the domain-labels of the nouns appearing in the documents has been developed. Linking the domains through EuroWordNet and having again a mapping between domain labels and IPTC categories, we obtained a language independent non-supervised classification system valid for the four languages of the project. This system is now under evaluation in order to see if it is competitive with the supervised machine learning systems. One issue under consideration is to test the effect of the automatic mappings between category schemes on the accuracy of the resulting systems. Another clear challenge for the future research is the combination of both approaches for text classification.

The clustering subtask deals with the problem of identifying sets of thematically related documents (e.g., a set of documents, in one or many languages, referring to the same new/event). Since these sets or categories are not defined in advance, this is a task of unsupervised learning (usually referred to as clustering). The list of research topics in which we have been working up to the present include:

- Development of a survey on similarity metrics for dealing with text (Rodríguez, 2002)
- Experimenting with several distance measures between documents, which take into account different levels of granularity between document components (paragraph, summary, document, etc.). This topic is especially relevant also for the document summarization task and for establishing the links between parts of documents. Several experiments are being carried out in order to try different clustering algorithms with several metrics using a mid-size document collection (about 16,000 documents per language). The clustering algorithms used are freely available implementations. The clustering is performed on monolingual and multilingual collections of documents.
- The problem of measuring the quality of clustering systems is very important since the non-supervised nature of the task does not allow defining “accuracy”-like measures on the resulting clusters. The quality of the clusters generated depends on many factors, which can be partially automatically measured (granularity/density/intersection degree of clusters, etc.) but always requires some manual intervention in order to assess the real utility. We have worked on defining some measures of direct and indirect utility measures of the results.

The work performed on similarity metrics and clustering is the basis for addressing the linking of documents at a level of document, paragraph, and entities involved in the documents. We plan to achieve a first prototype by the end of the year.

Automatic Summarization- the final goal of this task is to develop a multilingual summarization system. The first step performed was an extensive study of the state of the art in automated text summarization. This resulted in a document proposing three complementary classifications and a comprehensive file for analysing summarization systems (Alonso et al., 2002). Among all the existing possibilities, two complementary summarization strategies were chosen: 1) detecting lexical chains on texts, based on the work of (Barzilay, 1997), and 2) exploiting rhetorical structure of the texts, based on the work of (Marcu, 1997). The reasons to choose these two lines of research were the following: (a) They rely on general linguistic properties of texts, providing for flexibility and scalability, which are one of the main shortcomings of systems that exploit genre- or domain-dependent features of texts. (b) They can be easily adapted to other languages and to other kinds of

summarization, such as multilingual summarization. (c) They exploit properties of texts that are highly informative but that can be treated with shallow NLP techniques, namely, morphological analysis and chunking. (d) They are complementary, since they both exploit discourse properties of text, but from a different perspective: lexical and structural.

Two prototypes of indicative, extractive summarization systems for Spanish have been developed:

- A system based on lexical chains (Fuentes and Rodriguez, 2002) that selects the most informative fragments for including in the summary. The system uses lexical chains as primary source for ranking segments of the text. It also uses co-reference chains and named-entity chains. The system aims to be language independent provided we could access the required knowledge sources, basically the corresponding WordNets.
- A system based on the rhetorical structure of texts. It detects discourse markers in text and exploits the information in a discourse marker lexicon to obtain shallow discourse structure. Up to the moment, this system only performs sentence and paragraph-level summarization.

The performance of these systems has been evaluated to assess further improvements. The lexical chain system has been tested with a gold standard built from a corpus of Spanish agency news using the evaluation software MEADeval (<http://perun.si.umich.edu/clair/meadeval>). Its results were compared with two baselines: the lead method (i.e., extracting a number of paragraphs, starting on the first one, until the desired length, given the compression rate is achieved) and SweSum an available system (<http://www.nada.kth.se/~xmartin/swesum/>) allowing summarization of Spanish texts. As for the rhetorical structure system, only the discourse segmentation module has been evaluated (Alonso and Castellón, 2001). Both evaluations show competitive results and they are especially valuable to indicate future improvements. In addition, these two summarization systems have been combined to achieve an improved representation of discourse that should correspondingly improve the quality of the resulting summary. Evaluation of this combination (Alonso and Fuentes, 2002) shows that there is a qualitative (selection of text fragments) and quantitative (degree of compression) improvement on the performance of both systems when they are combined. Future improvements on the lexical chain system concern Named Entity chains, which could be improved with more accurate recognition and classification of Named Entities, for example, by using gazetteers. Additionally, Named Entities could be linked to EuroWordNet, which would relate these two kinds of information directly. Concerning co-reference chains, extended co-reference mechanisms (e.g., definite descriptions) should be taken into account to improve the performance of the system. Also developing more complex forms of chain merging could be a promising direction. The rhetorical structure system should be improved by enhancing the quantity of discourse markers in the discourse marker lexicon and the information associated to them. To address the first, bootstrapping techniques have been used. As for the second problem, we plan to obtain discourse schemata that span over high-level textual units by multiple sequence alignment techniques. A parallel line of research is on cross-language summarization for foreign-language document selection. This kind of summaries only make sense in the context of foreign-language document retrieval, and hence it is described in detail in task 5 (search applications).

Query Expansion/Translation and Conceptual Indexing- these techniques can only be evaluated in the context of Cross-Language Information Retrieval engines. For the sake of readability, we describe both the techniques and their evaluation in next section.

Task 5: Multilingual Information Access Applications

The driving force of the HERMES project is the idea that Language Engineering techniques and resources may help bridging the gap between the classic “Document Retrieval” model and the broader “Multilingual Information Access” paradigm. Traditionally, Information Retrieval has been understood as a fully automatic process that inputs a query (a statement of user needs) and a text collection, and outputs a (ranked) set of documents, which are relevant to the query. A perfect IR engine would retrieve only relevant documents (perfect precision) and all relevant documents (perfect recall). This model made many implicit assumptions: that the text collection and the user needs are static, that both query and documents are written in the same language; that the user is familiar with the terminology used in the documents; that information needs are optimally satisfied with documents (rather than information itself), etc. The advent of Internet and the so-called *Information Society* has quickly driven this paradigm into obsolescence. The term “Multilingual Information Access” refers to the broader –and now realistic- challenge of helping users to browse, search, retrieve, recognize and ultimately use information (rather than documents) from distributed, dynamic and heterogeneous sources of multimedia and multilingual hyperlinked information objects. Relevant research topics include Multilingual IR, Multimedia (video, speech, image) retrieval, Interactive Retrieval, Question & Answer systems, Digital Libraries, Internet crawlers and search engines, etc. The HERMES project focuses in the problem of accessing information contained in multilingual collections of documents, either static (news repository) or dynamic (searching online news sources). The user profile can vary along three main features: (1) Native language, and degree of familiarity with each of the other languages in the collection. (2) Degree of familiarity with the contents and the terminology of the collection. (3) Specificity of the information need, from very fuzzy (mostly navigation) to very concrete (focused search of specific information items). The ultimate goal of HERMES is providing fully evaluated and scalable search prototypes to help such users searching both a stable news repository and a set of dynamic online news servers. The companion goal, described in previous sections, is leveraging Human Language Technologies in English, Spanish, Catalan and Basque in order to optimise their use in such Information Access applications. We have currently accomplished the following tasks towards the final HERMES prototypes:

- Qualitative evaluation of Word Sense Disambiguation Strategies in Concept-based multilingual retrieval.
- Development and evaluation of a prototype, “Website Term Browser” that uses morphosyntactic information to provide a “phrase-browsing” multilingual search feature for users with reading abilities in most languages of the collection.
- Development of the first HERMES prototype, that incorporates morphosyntactic analysis, named entity recognition, summarization, paragraph-based searches and clustering in a single search interface.
- Development and comparative evaluation of search strategies for unknown languages, including a cross-language summarization technique to assist document selection and a query formulation and refinement strategy based in user assisted monolingual phrase selection combined with automatic phrase translation.

In the remainder of this Section we summarize the results obtained in each of these stages.

Word Sense Disambiguation and Concept-based multilingual retrieval- the possibility of using a language-independent inventory of concepts (the EuroWordNet InterLingual Index) as indexing space for all documents in all languages seemed a very attractive option to the HERMES consortium, for a variety of reasons: (a) Compared to other Cross-Language IR (CLIR) strategies

(essentially, query translation into the document languages), having a unique indexing space avoids the problem of merging ranked lists from individual languages, a problem that is yet unresolved in CLIR. (b) It scales better than query (or document) translation for a growing number of languages. (c) It uses WSD to solve directly some traditional problems of keyword-based retrieval (identification of synonym terms, identification of different senses of a word, etc.). (d) It permits using the conceptual relations in EuroWordNet for query expansion (broader and narrower terms, part-of relations, etc). Two main reasons, however, led to provisionally discarding this approach: (1) Word Sense Disambiguation techniques have not yet reached a sufficient degree of maturity for such an ambitious approach. On the one hand, supervised techniques are not yet scalable, in the absence of large-scale training corpora. On the other hand, unsupervised techniques are still far from being accurate enough for indexing purposes. While the HERMES unsupervised WSD systems were the best both in the lexical sample and in the all-words SENSEVAL tasks (see task 3), the recall obtained was essentially similar to a pick-first-sense strategy. (2) The qualitative evaluation of the concept indexing approach showed that there is a friction between multiword indexing units for monolingual and cross-language retrieval. Multiword expressions (either being units in the conceptual indexing or not) are not adequate for monolingual retrieval, where it is better to index all its components. This has been thoroughly tested for agglutinative languages (German and Dutch in particular) in the framework of CLEF evaluations. However, phrasal expressions are much better than individual words for translation. Conceptual indexing suffers from this problem of different granularities for different searches: if a multi-word expression is not in EuroWordNet, it will not be translated adequately into other languages; and if it is present in EuroWordNet, it will be translated properly but it will not produce good monolingual results in each individual language searched. This problem derives from understanding indexing and translating as one single task. We have, therefore, looked for alternative uses of lexical processing in multilingual search applications, with a remarkable success, as detailed below.

Development and evaluation of a Multilingual Phrase-Browsing Search Approach- as mentioned above, multiwords and, more generally, phrasal expressions are not adequate indexing units for automatic retrieval because they are too restrictive, hiding their components as partial matching units between text and query. Phrases extracted from the collection can be, however, extremely useful in interactive searches, because they provide an alternative, information-oriented perspective of the contents of the collection. This is known as the “phrase-browsing” approach to interactive retrieval. The UNED group has extended the phrase-browsing paradigm to a multilingual phrase-browsing paradigm (Peñas 2002). The main idea is to exploit the two potential benefits of phrases: as a rich source of information for interactive searching, on one hand, and for accurate lexical expansion and translation, on the other. This approach solves the problem of the conceptual indexing approach: phrases are not used for indexing (but for interactive refinement of the query), but they are used for translation (where they provide enough context for accurate translation) and expansion (again offering enough context to expand only with relevant terms). The result is the WebSite Term Browser, a multilingual phrase-browsing search engine that unifies expansion and translation of the query terms as a single process. WTB has two main components: at indexing time, efficient morphosyntactic analysis is used to recognize all noun phrases in the documents, and an index from phrases to documents is built in addition to the classical index from words into documents. An index from words to phrases completes the indexing process. At searching time, the system expands every word in the query using semantic relations in EuroWordNet, including synonyms and cross-language equivalents in all document languages. This process, initially very noisy, is used to select phrases in the collection containing a maximal number

of query-related terms; this restriction eliminates most noise. Phrases are organized conceptually and shown to the user as an alternative way of accessing the information in the collection. The evaluation of WTB was done comparing the utility of phrasal information to the ranked document lists provided by Google, which is currently one of the best Internet search engines. A search interface for the UNED domain (with approximately 50000 docs.) was designed to make this comparison. Given a user query, this interface outputs both the results of Google and the phrases retrieved by the WTB software. The user could then select a document or a phrase matching his information needs. The search interface was offered as a service of the UNED educative portal, and all searching sessions were logged for analysis. The study of the first 2000 non empty queries revealed that choosing a phrase after the initial query was even more frequent than choosing a document retrieved by Google, indicating the potential value of the approach (Peñas et al 2001).

First HERMES prototype- the success of the multilingual phrase browsing approach led to a continuation in the first HERMES prototype, which incorporates lexical resources (EuroWordNet), extends phrase recognition with detection of named entities and events, provides document and paragraph searching facilities based in all that language annotations, and provides room for summarization, and clustering. Rather than simple document retrieval, this interface provides sophisticated access to the information contained in relevant subsets of the collection being searched. In this first prototype, a search begins with an initial short query containing some keyword terms. The system filters the contents of the collection, retrieving documents that contain the query terms. From such set of documents, the systems extracts and displays information about persons, locations, events, etc., related to the query. The user can then refine his information need by selecting and combining such pieces of information. The system allows, for instance, selecting all actions taken by Arafat in relation to Israel in a particular time frame. Finally, the system displays the results of this information request. The user may select to view documents or paragraphs related to his query:

- If the documents view is selected, there is a range of options to examine them. The first prototype includes the possibility of a) clustering them according to their IPTC category, which is automatically assigned, b) viewing summaries of individual documents, and c) viewing classified entities within individual documents.
- If users prefer to view relevant paragraphs, they can chose to list them grouped by subjects, actions or objects. Once a paragraph is selected, the system shows it with the same options than documents, but also with the option of saving the paragraph in a table that plays the role of a clipboard where user can build a cross-document summary.

Hence this first prototype integrates morphological analysis, parsing, named entity recognition and summarization techniques. The next prototypes will also include the possibility of viewing documents according to their similarity (clustering), together with a multi-document summary of contents of the cluster. Multilinguality is not yet incorporated into this prototype; we believe that multilingual information access challenges are of a very specific nature, and should be thoroughly tested in isolation before incorporating them into a single search interface. According to this philosophy, we have designed and evaluated some information access strategies for unknown languages that we describe below.

Foreign-Language Search Assistance- Cross-Language Information Retrieval is stated as the problem of, given a query in some source language, being able to identify relevant documents

written in some (different) target languages. Over the last 6 years, this problem has received a growing attention from the Information Retrieval and Language Engineering research communities, and there exists now well-founded techniques to solve it and well-established comparative evaluation methodologies. As mentioned before, the UNED group is involved in the organization of an annual comparative evaluation of CLIR systems in European Languages. This problem is, however, only one of the challenges of the multilingual information access problem. For instance, if we type a Spanish query and receive a ranked list of Chinese documents, how can we recognize which of them are really relevant for our purposes? How can we refine our query taking these results into account? How can we use the information contained in documents we cannot read? These are problems that have not received enough attention from the research community yet. The default assumptions for a Cross-Language search engine are that a) commercial Machine Translations systems can be used to translate documents into the native language of the user, and b) document selection and query refinement can be done using such translations. In order to challenge such (untested) assumptions, the UNED team designed and organized, together with the University of Maryland, a common evaluation framework, called iCLEF, for the comparative study of user interaction issues in Cross-Language IR. The first evaluation campaign took place in 2001 (Oard and Gonzalo 2002), and it was devoted to document selection mechanisms: Is Machine Translation the only option available? Can it be substituted for simpler and faster methods without degradation of relevance judgements? Are there even better alternatives than Machine Translation?

At the UNED group, we devised a cross-language summarization technique based on noun-phrase alignment using comparable corpora (López-Ostenero et al. 2002a). The idea is to extract noun phrases as in previous HERMES prototypes, and align them using only bilingual dictionaries and statistical evidence from related corpora (in our case, news in different languages from the same time frame). The result is not readable as a summary in its classical sense (cf. previous section on summarization), but it can be enough for indicating whether a document is relevant for a query or not, and it can be read much faster than a full Machine Translation of the document. The evaluation of such cross-language summaries was done using the common iCLEF methodology over four topics in Spanish, the English CLEF news collection, 32 Spanish users and 128 searching sessions. The official iCLEF results stated that cross-language summaries allowed for relevance judgement with almost the same precision than Systran Professional 3.0, and with a 51% improvement in recall, because users judged documents faster with phrase-based summaries. Compared to other iCLEF results, the UNED strategy was the only one to overperform, even to match, MT results. The second evaluation campaign took place in 2002 (Gonzalo and Oard 2002), and it was devoted to study support mechanisms for interactive query formulation and refinement. While other groups studied user-assisted term translation (via inverse dictionaries, translation definitions, etc), at UNED we hypothesized that user assisted term translation would demand a high cognitive load from the user. As an alternative, we tried to assist the user to select relevant phrases from his query, and then let the system perform automatic translation of the queries using the aligned phrases resource already built the previous year. Using phrase-based summaries to explore the document contents, the users could easily add new phrases to the query, which again would be translated without user intervention. In the official iCLEF measure, our strategy performed 65% better (combined precision/recall measure) than a strategy based on user assisted term selection (López-Ostenero et al., 2002b). To complete the picture, the Maryland group checked that user-assisted translation is better than blind translation; overall, the UNED strategy is one of the most promising ones. Both results from the two iCLEF evaluation campaigns will be crucial for the design of the next HERMES prototype, where all search facilities will be enhanced

TIC2000-0335-C03

with multilingual abilities. The future HERMES prototype should offer different searching facilities according to the knowledge of the languages involved. Users with active, passive or no knowledge of some of the target languages should receive very different types of search assistance. The Website Term Browser can be a good model for high or moderately multilingual users, while the iCLEF UNED prototypes could be a model to assist monolingual users.

Multilingual Online News Meta-searcher prototype- the second HERMES application is designed to facilitate access to online news rather than stable repositories. The first prototype is already active and is able to accept queries in any of the four HERMES languages translate them into all languages and query the online searchers of *ABC*, *El País*, *El Mundo* (Spanish), *Avui* (Catalan), *Egunkaria* (Basque) and *Washington Post* (English). Most of the effort on this initial prototype has been done on the meta-search aspects of the server, and thus the cross-language mapping is still fairly simple and experimental. Three ways of cross-language expansion have been currently devised: a) just building a query with all possible translations for all possible query terms, b) all terms, but structured in a Boolean query; translations for a single term are grouped with OR operators, and translations for different terms with AND operators; c) lookup of the bilingual dictionary of phrases described above as a first translation resource, and resort to bilingual dictionary for the rest of terms.

3 HERMES results and activities in the international research community

Publications- total number of papers published: 65, where 2 are HERMES technical reports, 13 in collections indexed by ISI JCR, 17 published in major CL events (ACL, COLING, LREC), 1 in ACM collection, 13 in international and 10 in national Conferences with referees, 7 in national journals and 2 in international journals. 2 PhD Thesis have been presented, one in UPC and the other in UNED. The full list is available at the project web site. <http://terral.lsi.uned.es/hermes/>

Resources built, with public distribution: Spanish CLEF collection and evaluation suites (free for participants and for a small fee through ELDA), Basque and Spanish SENSEVAL evaluation suites (available through the official site (<http://www.itri.brighton.ac.uk/events/senseval/>)). Eurowordnet-HERMES (a lexical database including the 4 languages).

Relationships with other research groups involving PhD students -the UNED group has established collaboration with Prof. A.Kilgariff, belonging to ITRI (Brighton University) on SENSEVAL resources and evaluation. The PhD student I.Chugur has carried out a 3-month stage in 2001 carrying out a study of sense similarity (Chugur et al., 2002). The PhD student E.Amigó has participated in a summer school organized by ELSNET with a grant from the organization. The UPC group has established collaboration with the *Cognitive Computation Group* (Department of Computer Science, University of Illinois at Urbana-Champaign, <http://l2r.cs.uiuc.edu/~cogcomp>) headed by Prof. Dan Roth, on the topic of the application of automatic learning methods to natural languages processing. These techniques have been applied to entity recognition and shallow parsing. The PhD student Xavier Carreras has carried out a stage in this group at the beginning of 2002. The UPV group: the PhD Mikel Lersundi has carried out a 3 month stage in Maryland University, under the supervision of Prof. Bonnie Dorr, on the topic of analysing the semantic interpretation of prepositions and suffix in Basque, Spanish and English. This work contributes to task 2.2, the extraction of lexical semantics

relations from dictionary definitions. The PhD student David Martínez will spend three months in John Hopkins University with Prof. David Yarowsky on the topic of *bootstrapping* for disambiguation purposes.

Partnership and Collaborations

In a variety of dimensions the research groups have strong links with international research groups, through EU funded projects, the co-organization of international campaigns, such as the **SENSEVAL** and **CLEF** competitions, the participation on excellence networks, or the co-organization of workshops. Next we mention some of them. (1) **CLEF**- IST Programme of the European Union (project no. IST-2000-31002). <http://clef.iei.pi.cnr.it:2002/> Scientific coordinator: ISTI-CNR Istituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, UNED is partner and co organizer with the University of Maryland of the interactive iCLEF 2001 and 2002 interactive task (<http://terral.lsi.uned.es/iCLEF>). Task definition, guidelines, testbed development, and task coordination. See (Gonzalo and Oard, 2002). (2) **MEANING: Developing Multilingual Web-scale Language Technologies** (IST-2001-34460), coordinated by Germán Rigau, also involved as researcher in HERMES. Both projects share topics related to WordNet (T2.2) y and WSD (T3.1). (3) **ELSNET**- European Network on Excellence in Human Languages Technologies. UNED and UPC groups belong to this network. <http://www.elsnet.org/> (4) **RITOS2**, - UNED group is involved in the NLP group and is organizing a workshop on the topic *Multilingual Information Access and Natural Language Processing*, in the next Iberoamerican conference on artificial intelligence IBERAMIA 2002 to be held in Sevilla. <http://sensei.lsi.uned.es/iberamia-mlia/> (4) The ISCA (International Speech Communication Association) **Special Interest Group on Speech and Language Technology for Minority Languages**. UPV/EHU belongs to this special interest group. <http://193.2.100.60/SALTMIL/>

4 References

- 1 Agirre E., Ansa O., Martínez D., Hovy E. 2001. Enriching WordNet concepts with topic signatures. Proceedings of the NAACL workshop on WordNet and Other lexical Resources.
- 2 Agirre E., Lersundi M., Martínez D. 2002. A Multilingual Approach to Disambiguate Prepositions and Case Suffixes ACL Workshop: WSD: recent successes and future directions
- 3 Agirre E., Martínez D. 2001. Decision Lists for English and Basque. Proceedings of the SENSEVAL-2 Workshop. In conjunction with ACL'2001/EACL'2001. Toulouse
- 4 Agirre E., Martínez D. 2002. Integrating Selectional Preferences in WordNet. Proceedings of First International WordNet Conference. Mysore (India).
- 5 Alegria I., Aranzabe M., Ezeiza A., Ezeiza N., Urizar R. 2002. Robustness and customisation in an analyser/lemmatiser for Basque. LREC-2002 Workshop Customizing knowledge in NLP applications
- 6 Alonso L., Castellón I., 2001. Towards a delimitation of discursive segment for NLP applications, International Workshop on Semantics, Pragmatics and Rhetorics
- 7 Alonso L., Castellón, I., Climent, S., Fuentes, M., Padró, LL., Rodríguez H. 2002. Comparative Study of Automated Text Summarization Systems, HERMES Project.
- 8 Alonso L., Fuentes M. 2002. Collaborating discourse for Text Summarisation, Proceedings of the Seventh ESSLLI Student Session, Trento, Italy.
- 9 Amigó E. 2002 Criterios y técnicas de evaluación. HERMES Technical Report

- 10 Arévalo M., Carreras X., Márquez, L., Martí, M.A., Padró L., Simón M.J. 2002 A Proposal for Wide-Coverage Spanish Named Entity Recognition. In SEPLN Journal 28, 63-80.
- 11 Arregi O., Fernández I. 2002. Clasificación de documentos escritos en euskara: impacto de la lematización. I Jornadas de Tratamiento y Recuperación de Información, JOTRI, Valencia.
- 12 Barzilay, R. 1997. Lexical Chains for Summarization, Masterthesis. Ben-Gurion University
- 13 Carreras X., Márquez L., Padró L. 2002. Named Entity Extraction Using AdaBoost. In Proceedings of the ACL SIGDAT Workshop on CoNLL2002. Taipei, Taiwan.
- 14 Carreras X., Márquez L., Padró L. 2002 Wide-Coverage Spanish Named Entity Extraction. To appear in Proceedings of the VIII Iberoamerican Conf. on Artificial Intelligence. Sevilla, Spain.
- 15 Carreras X., Padró, L. A Flexible Distributed Architecture for Natural Language Analyzers. 2002 European Conf. on Language Resources and Evaluation (LREC'02). Spain.
- 16 Daudé J., Padró L., Rigau G. (2001), A Complete WN1.5 to WN1.6 Mapping. Proc. of NAACL Workshop WordNet and Other Lexical Resources: Applications
- 17 Escudero G., Marquez L., Rigau G.: Using LazyBoosting for Word Sense Disambiguation. Proc of the SENSEVAL-2 Workshop. In conjunction with ACL'2001/EACL'2001.
- 18 Fernández-Amorós D., Gonzalo J., Verdejo, F. The UNED system at SENSEVAL-2 Proc. of the SENSEVAL-2 Workshop. In conjunction with ACL'2001/EACL'2001. Toulouse
- 19 Fuentes M., Rodríguez, H. 2002. Using cohesive properties of text for Automatic Summarization. In JOTRI'02.
- 20 Gonzalo J., Oard D. 2002. The CLEF 2002 Interactive Track. In Proceedings of CLEF 2002, Springer-Verlag LNCS, to appear.
- 21 Chugur I., Gonzalo J., Verdejo F. 2002. A Study of Polysemy and Sense Proximity in the SENSEVAL-2 Test Suite. Proc. of the ACL 2002 work on WSD: Recent Successes and Future Directions.
- 22 López-Ostenero F., Gonzalo J., Peñas A., Verdejo F. 2002a. Noun Phrase translations for Cross-Language Document Selection. In Proc. CLEF 2001, Springer-Verlag LNCS 2406.
- 23 López-Ostenero F., Gonzalo J., Peñas A., Verdejo F. 2002b. Phrases are better than words for Cross-Language query formulation and refinement. In Proc. CLEF 2002, Springer Verlag LNCS, to appear.
- 24 Magnini B., Strapparava C., Pezzulo G., Gliozzo A.. 2002. Comparing Ontology-Based and Corpus-Based Domain Annotation in WordNet. In Proceedings of First International WordNet Conference, Mysore, India, pp. 146-154.
- 25 Marcu D. 1997. The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts. PhD Thesis, Department of Computer Science, University of Toronto
- 26 Oard D., Gonzalo J. 2002. The CLEF 2001 Interactive Track. In Proc. of CLEF 2001, Springer-Verlag LNCS 2406.
- 27 Palomar M., Ferrández A., Moreno L., Martínez-Barco P., Peral J., Saiz-Noeda M. Muñoz R. (2001) "An Algorithm for Anaphora Resolution in Spanish Text" Computational Linguistics 27.
- 28 Peñas A., 2002. *WebSite Term Browser*: un sistema interactivo y multilingüe de búsqueda textual basado en técnicas lingüísticas. Tesis Doctoral, Dep. de Lenguajes y Sistemas Informáticos, UNED.
- 29 Peñas A., Gonzalo J., Verdejo, F. 2001. Cross-Language Information Access through phrase browsing. In Proceedings NLDB, Lecture Notes in Informatics.
- 30 Rodriguez H. 2002 Similarity Measures in Computational Linguistics HERMES Technical Report.
- 31 Santamaría C., Gonzalo J., Verdejo F. (2001). Internet como fuente de información léxica: extracción de etiquetas de dominio y detección de nuevos sentidos. Procesamiento del Lenguaje Natural, 2