An Interface Agent with Linguistic Skills

Ana García-Serrano[1] and Paloma Martínez[2]

[1] Department of Artificial Intelligence
Technical University of Madrid
Campus de Montegancedo
28660 Boadilla del Monte, Madrid
agarcia@dia.fi.upm.es
[2] Computer Science Department
Carlos III University of Madrid
Avda. Universidad 30
28911 Leganés, Madrid
pmf@inf.uc3m.es

Abstract: We present MESIA[1], an interface agent that supports the web-user interaction in natural language. In the information retrieval field, three topics could be covered to enhance retrieval results: (semi)automatic query expansion, strategies that take into account features related to the documents as well mechanisms that specify search criteria from previous experience. The interface agent performs a linguistic processing to improve the quality of the web search through the modification of the user query, as well as by means of the structure and some features extracted from the documents retrieved by a conventional web search. The Interface Agent is also endowed with a knowledge-based component that uses past experience stored in an ontology and classified links to get better results from the modified user query. The prototype is implemented in Ciao-Prolog and Java, incorporating two adapted lexical resources (ARIES and Spanish EuroWordNet) and is locally running for the cultural pages of the Madrid Local Government.

1. Introduction

The growing use of Internet has motivated additional demands of new information management techniques and effective search methodologies. An adequate presentation of retrieved results and an accurate search (according to variable criteria) are two crucial aspects for the user approval in any organisation that uses an Internet based information system as business and advertising support. The goal of this paper is to show the work in progress of MESIA system, a meta-search engine with semantic capabilities for Web information retrieval (IR). By the use of natural language processing (NLP) tools, the

[1] This work was supported by MESIA project CAM 07T/0017/1998

results of existing commercial search engines could be enhanced not only in the treatment of the user queries but also through pragmatic knowledge extracted from the content of retrieved Web pages.

Currently, existing search engines (AltaVista, Yahoo and others) are based on statistical analysis that brings the discrimination and selection of Web pages related to a query. Purely statistical methods used in IR do not achieve optimal results; the exponential growing of information in Internet makes that search-based in user-specified words or patterns in their text (keyword-based search) obtains much more documents than necessary (irrelevant information); on the other hand, it is also the case that documents do no appear in the answer because they do no contain explicitly the query terms but other semantically related words. Thus, new strategies that profit from document content are required as well as new mechanisms to define search criteria from the user query taking into account the acquired experience.

In the last decade much work has been devoted to develop linguistic resources and now it is time to integrate it into the conventional web searches to improve the outcome by means of linguistic processing. Given that currently there are not complete and correct natural language understanding systems neither general purpose linguistic resources, it is only possible to work in specific knowledge domains. This allows to have dependent domain knowledge that facilitates information search. Notice that we do not propose a solution based on incorporating semantic knowledge in Web documents (such as extending HTML tags) but to extract semantic knowledge from the documents located by a traditional search engine.

At the same time, the agent-based paradigm has become increasingly popular, as there is growing evidence that agent-based architectures promotes an efficient construction of scalable software systems allowing the reusability of previous software developments. By using an agent-based architecture, we take advantages such as the concurrence in tasks processing, the distribution of the resources like knowledge and methods, and the flexible interaction among the system components.

In the information retrieval field, three topics could be covered to enhance retrieval results:

- **Modifying the original query**. The system transforms the user query that is close to natural language into a formal query by extracting the significant terms and expanding them by including morphological variants and synonyms.
- **Classification of documents** that compose the search results. An information extraction process is performed to classify the documents obtained by the Altavista and to sort them before to be displayed to the user. To this classification we need to identify structural and semantic aspects as a result of a linguistic process from parts of the document.
- **The accumulation of experience.** The system includes a knowledge manager (the librarian component) for the documents retrieved for more frequent queries. It is also foreseen to include user profiles that will allow to decide whether the query is sent to the librarian or, otherwise, a new search is launched. The currently available user model is very simple (an ontology with a description based on the foreseen use of the system for each type of user) but it allows, in some cases, to incorporate conditions into the formal query in order to delimit the answer.

In order to perform a selective search through Spanish language based on the previous topics, two types of knowledge have been identified from a manual analysis and the

foreseen utilisation of the system: (a) Knowledge about the documents structure and classification according with different criteria and (b) Linguistic knowledge about domain sub-language, specific vocabulary and expectative-based analysis considering significant expressions. A software system that incorporates and articulates this knowledge is required so, a knowledge-based architecture was designed to allow the functional distribution of knowledge.

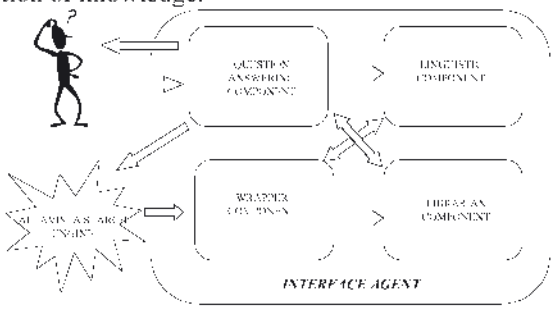

Fig. 1: Interface Agent Functional architecture

2. Interface Agent Description

Figure 1 shows the functional diagram of MESIA interface agent. It is composed of four components: linguistic, wrapper, librarian and question answering. MESIA system reflects the specialised-based structuring approach in an engineering field: (1) the organisation of knowledge matches the organisation of the different linguistic and control knowledge, (2) knowledge resides in distinct cognitive components and (3) the reasoning process is based on the centralised co-ordination of the different components tasks to integrate the different proposals into a coherent result (ranked list of links).
The prototype has been design following a knowledge-based methodology during the different phases of the development cycle, using Java for the interface and the logic programming environment CIAO-Prolog, [Bu99], which facilitated the integration of the available resources.

2.1 Linguistic Component

The linguistic agent incorporates knowledge about the sub-language and performs the expansion of user queries and the partial analysis of text units (sequences of one or more than word not necessary speech acts or sentences). In very recent years, relationships between IR and NLP have emerged; the reason is that statistical IR approaches often do not produce the desired retrieval results. NLP tools that could be incorporated in IR systems are thesauri, lexical databases, shallow parsers, etc. In order to improve the task of text retrieval, several natural language technologies could be applied, [Li98]:

- The *morphology* is the linguistic source more commonly used in IR systems; for instance, stemming algorithms are applied to queries and documents to take into account all morphological variants and to avoid the potential for obviously missed relevant documents. It should be noted that English stemming is widely used in

commercial search engines; however, for other languages with a rich morphology, such as Spanish, it is not a reality. Apart from this, a highly inflectional morphology provides more useful information for IR than it does for English.

- The *lexical* source may be used in IR both in part-of-speech tagging of query terms (nouns, adjectives, verbs an so on) or for the utilisation of lexicons from which the specific features of terms can be accessed (a common vocabulary is used in formulating adequate indexing or searching terms).
- The *syntactic* source can use the part-of-speech tagging output in order to detect phrases or significant groups both in user query and/or in document text. Thus, better indexing terms that represent the document data content as well as better searching keys in the queries (it is not the same to look for documents about "automata" that about "cellular automata").
- The *semantic* source assists in interpreting the meaning of sentences as units of understanding in opposition to individual words or phrases that compose them. Consequently, semantic disambiguation of words with various senses, identification of verb-arguments relations in a sentence or the query expansion by adding synonymous terms have to be considered. Other semantic processing approaches contemplate to use semantic vectors to represent document and queries, [Vo99].
- The *discourse* level could be used to understand what the specific role of a piece of information plays in a document (v.g. a conclusion, an opinion, a fact, etc.), that is, to take advantage of structure and organisation of documents. Additionally, the recognition and resolution of anaphora would result in an improved formulation of queries and documents.
- Finally, the *pragmatic* source concerns how the IR system understands the users in the context of their needs, history and objectives. Some basic principles of communication could be incorporated in the user interface of IR systems to facilitate the conversation between the user and the system.

The MESIA linguistic component incorporates several lexical resources as is explained below:

- **ARIES** (www.mat.upm.es/~aries/), [GGM97], is a Spanish lexical platform developed by the Universidad Politécnica de Madrid and Universidad Autónoma de Madrid. ARIES is composed of a Spanish lexicon with around 38,000 lemma entries, including 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 the query words (useful to identify the relevant terms of the query). Moreover, a DCG morphological generator for deriving word variants is being incorporated in MESIA system. This generator allows, for instance, obtaining number and gender forms from a nominal lemma. An example of an ARIES lexical entry (doctor) is shown below:

```
doctor
category              n       /* noun */
concat                wl      /* word that accepts a number morpheme */
agr gender      =     masc    /* masculine gender */
agr number      =     sing    /* singular number */
plural derivation=    plu2    /* rule for plural generation */
lex             =     doctor  /* lemma */
```

- a **simple phrase segmenter** based on cascade finite automata, [MG00], identifies simple noun, prepositional and verb phrases in the query. It is used to remove the ambiguity produced by ARIES part-of-speech analyzer. This segmentation also helps to build the formal query (selection and combination of boolean operators) to be sent to the Altavista search engine.
- **EuroWordNet** (www.hum.uva.nl/~ewn), [Vo98], [Go98], is a lexical database that is structured as an top concept ontology that reflects different explicit opposite relationships (v.g., animate, inanimate) and it can be seen as a representation of several vocabulary semantic fields. Moreover, it contains a hierarchy of domain tags that relate concepts in different subjects, for instance, sports, winter sports, water sports, etc. The most important semantic relationships in EuroWordNet are synonymy, antonymy, hyponymy, meronymy, entailment and cause. EuroWordNet is used in three different linguistic processes:
  - Expansion of query relevant terms with synonyms and other semantically related terms; The EuroWordNet database enables the user to use Domains (a hierarchy of domains labels which relate concepts on the basis of scripts or topics) to separate the generic from the domain-specific vocabularies; this is important to control the ambiguity problem in NLP.
  - Together with ARIES and phrase segmenter, it is also used by the wrapper agent and the librarian agent during its document classification task.

2.2 Wrapper Component

The wrapper component includes knowledge to identify the structure of the web-pages and the significant (clue-guided) text units extraction. This component is in charge of analysing the HTML pages returned by the Altavista search engine trying to extract the textual information that they contain The web pages are structurally classified into four classes according with the contents: (1) Database access front-end, (2) Explanatory (raw) text (3) Index-pages, that contains new links to other pages with a not predefined or common structure (4) Download pages that contains files with a short description of their contents.

2.3 Librarian Component

The librarian component responds to any other agent request finding the links related to the user request, to put in order the Web pages returned by the Altavista search engine and to update the knowledge base that contains a domain ontology with the related web-pages (links); because of the Spanish lexicon required in the specific domain we have selected is not supported by EuroWordNet, an extended thesaurus (domain ontology) containing the semantic relationships among domain terms is being developed Figure 2 shows a partial view of the thesaurus for the cultural activities of "festivals". The ontology is implemented in XML and keeps the Web links related to each concept; each node represent a concept, has a weight assigned taking into account different structural and semantic criteria gathered both from a domain analysis and also has a set of

associated keywords as well as a set of web links. The aim is to rank the links retrieved by Altavista by means of an ontology-based inference before presenting them to the user.

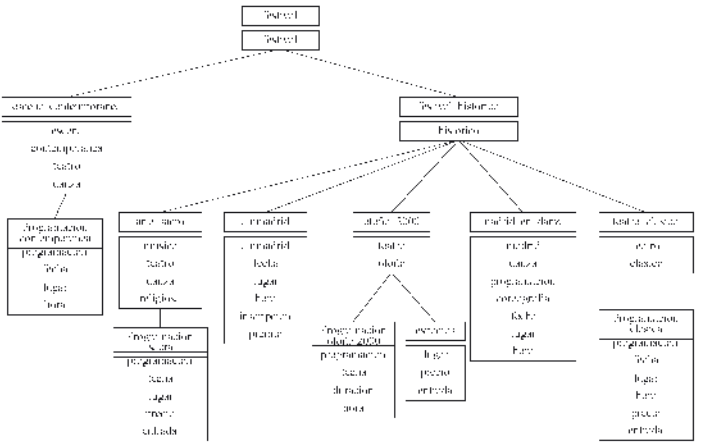

Fig. 2: Partial view of domain ontology

2.4 Question Answering Component

This module is related to the user query formulation. The approach is to improve the initial query formulation through query expansion and term rewriting by translating the query user from natural language to a formal query that can be run by the search engine. Some natural language query examples are:

- busco convocatorias abiertas de proyectos nacionales (*I'm looking for a national opened call for proposal*)
- quiero los impresos de la última convocatoria de proyectos de la CICYT (*I want the forms of the last CICYT call for proposals*)
- ¿existen becas para la movilidad de personal investigador? (*Are there any grants for research staff mobility?*)

So, it requires two functionalities provided by the linguistic component (shallow parsing and query expansion). Then a process to build the boolean query is applied. These functionalities are explained below.

Shallow parsing: ARIES morphological analyser obtains the part-of-speech tags and other features for the query words. ARIES provides the PROLOG predicate w(Lemma,Category,Gender Person,Number Tense,Word,[]) that returns the morphological features of a given word. For instance, morphological analysis applied to the word "doctores" (*doctors*) produces two types of analysis (noun and verb):

?-w(L,C,G,N,doctores,[]).

L = doctorar, C = v (verb), G = sing 2, N = pres subj;

L = doctor, C = n (noun), G = masc, N = sing.

Following, in order to extract the relevant terms of the query as well as to solve ambiguity problems, the segmentation process detects simple phrases (noun, prepositional and verbal phrases). It allows to obtain the head and modifiers of phrases that are then expanded with gender and number variations and with semantically related

terms proceeding from EuroWordNet. For instance, the user could formulate the query ¿qué becas postdoctorales hay? (Which postdoctoral grants are there?) and the shallow parser should extract the noun phrase composed of *becas* (noun) and *postdoctorales* (adjective) that will compose the formal query.

Query expansion: Once the relevant terms of the query have been extracted, they are extending in two ways:

- The relevant terms are expanded using Spanish EuroWordNet by means of synonyms and hyperonyms. For instance, for the noun "estancia" (*stay*), EuroWordNet provides:

  ?- ewn(estancia, n, Synonyms, Hyperonyms, _).

  Synonyms sin([permanencia]), Hyperonyms hyper(['actividad humana']);

  Synonyms = sin([ ]), Hyperonyms = hyper([estancia]);

  Synonyms sin([cortijo,labranza,heredad,hacienda,granja]), Hyperonyms hyper([visita, estancia]);

- Using ARIES lexical database, for each relevant term, their morphological variations are added to the set of query terms (Figure 5). This step is required because the search engine does no perform any kind of stemming for Spanish language. For example, for the noun "doctor" ARIES returns four morphological variants: *doctoras*, *doctores*, *doctora*, *doctor*.

Boolean query generation: Briefly, the strategy of boolean query generation is as follows: semantically related terms are added to each relevant term connected by ORs. For instance, the term "becas" (*grants*) is expanded as: (beca OR galardón OR apoyo OR ayuda).

Then, morphological variations are added to each term connected by ORs. For instance, the previous terms are expanded as: (becas OR beca OR galardones OR galardón OR apoyos OR apoyo OR ayuda OR ayudas).

Finally, all set of expanded terms are connected by the AND operator (in a first version, in the prototype only simple noun and prepositional phrases are handled). For instance, the user query: ¿qué becas postdoctorales hay? (Which postdoctoral grants are there?) is translated into: (becas OR beca OR galardones OR galardón OR apoyos OR apoyo OR ayuda OR ayudas) AND (postdoctorales OR postdoctoral).

This query can be sent to the Altavista search engine or to the librarian component in order to look for the documents on the ontology.

3. Preliminary Experimental Work

In order to evaluate how the linguistic knowledge affects the retrieval results, in a first study, we have run seven user queries in four types of experiment, all of them executed in the Altavista Advanced Search mode. The four types of experiment are:

1. baseline (the relevant words of the query linked by AND operator),
2. only morphological variations expansion (with ARIES),
3. only semantic expansion (with EuroWordNet),
4. expansion with ARIES and EuroWordNet

Table 1 presents the precision values (fraction of the retrieved documents which is relevant) for each user query in each type of experiment. The queries used are:

Q1: Becas para la realización de tesis doctorales
Q2: Pruebas de acceso a la universidad
Q3: Becas para estancias en el extranjero
Q4: Estancias en universidades extranjeras
Q5: Convocatorias de becas de investigación
Q6: Convalidación de estudios extranjeros
Q7: Becas postdoctorales

Query	Baseline	ARIES	EuroWordNet	ARIES and EuroWordNet
Q1	0,7	0,6875	0,47	0,48
Q2	0,375	0,3	0,075	0,075
Q3	0,33	0,85	0,375	0,65
Q4	0,25	0,73	0,4	0,53
Q5	0,92	0,84	0,24	0,26
Q6	0,33	0,37	0,08	0,18
Q7	1	1	1	1

Table 1: Precision values obtained with the set of queries

Before analyzing retrieval evaluation, it is necessary to highlight that there is no way to know what are the criteria Altavista searcher considers in order to rank the retrieved documents when the Advanced search mode is used. Moreover, in this first experiment, no ranking of query terms has been performed. After an analysis of results, we tested that the use of ARIES morphology (extension with morphological word variants) generally enhances the search results. However, the extension with synonyms/hyperonyms, although contributes to retrieve documents that are not retrieved in the other experiments, affects the precision measure because other non relevant documents are also displayed to the user; it is due to the ambiguity problem. Moreover, MESIA prototype adds hyperonyms (a more general concept) to a relevant term when there is no synonyms and, sometimes, it contributes to retrieve pages that are not relevant (but, occasionally, EuroWordNet proposes interesting hyperonyms to be added to the query).

Furthermore, in this first approach we follow a simple combination of query terms by using OR and AND operators without any order criteria and it affects the presentation of retrieved documents. We have also tested that the extension of user query also affects to the query; when user queries have many relevant terms it is more difficult to discriminate ambiguous meanings not belonging to the query context.

New experiments are being performed taking into account:

- New collocations of query terms using boolean operators giving preferences to specific terms.
- Enhancing the incorporation of synonyms and hyperonyms provided by EuroWordNet.
- Profiting from a special Altavista searcher field (*order by*) that helps to rank the documents giving priorities to some query terms.

3. Related work

Of late years, agent-based technology has been incorporated in several computer science areas: distributed AI, NLP, process control, etc. The vast amount of information in Web as well as the wide use of it makes of Internet an excellent test field for agent-based technology. Some strategies (Webmate, Amalthaea, Personal WebWatcher, etc.) consist of providing the agents with a specific functionality (specialised in specific subjects, for instance, "research papers") and also with a personalized character (agents fix their behaviour to the user preferences).
[JCS99] proposes a two-level multiagent architecture to manage Web information with learning and personal features. First level contains the personal agents that help users and second level has the specialised agents about specific topics. By a negotiation mechanism, personal agents get in touch with specialised agents that facilitate the work (using KQML language and the Speech Act theory principles).
LIETZIA system, [Li95] is a user interface agent that assists the user browsing the WWW. The user can work with a traditional browser, for instance, Netscape, and the agent tracks the user behaviour trying to anticipate the topics of interest through an autonomous and concurrent exploration of the links that can be achieved from the current user location.
From a linguistic point of view, GETESS, [St99], a system to access a Web tourist information site, uses the semantic content of documents in a restricted domain. If semantic analysis fails, then syntactic methods for retrieving are used. Particularly, the search engine is characterised by: semantic knowledge that supports information retrieval, partial but robust NL understanding, several interaction ways and combination of knowledge from structured and semi-structured documents and relational databases.
[Fe99] and [CCS00] make use of ontologies (in the sense of consensual and formal specifications of a vocabulary used to describe a specific domain) to Web searching although they are based on HTML extensions that allow to include semantic knowledge in document structure. [MHN98] also uses a domain ontology described by Description Logics to perform information retrieval as well as in [To00].

4. Conclusions and Future Work

Related linguistic resources, the use of EuroWordNet in unrestricted domains poses the ambiguity problem (synonyms that do not belong to the query context) that produces a decrease of precision values. A possible solution is to show the semantically related concepts to the user in order to perform a filtering task. Furthermore, we are working on (semi) automatically updating the domain ontology with keywords appearing in web documents. Further work on combining structure and content in the queries is required along with new visual metaphors to formulate those queries and display the answers.
Concerning the search results, IR technology is faced with several problems: relevance feedback consisting on identifying relevant and non relevant documents and ordering them according to user relevance; the inadequate specification of query terms by the user, etc. Consequently, search engines should manage user models to guide the search process and the ranking of results. In IR systems, variables studied such as indexes,

similarity functions, relevance measures, etc. do not have to do with the user. It would be convenient to use knowledge about the user state, user needs, decisions and perceptions filtering (objectivity vs. subjectivity)

Bibliography

[Bu99] Bueno, F. et. al: The Ciao Prolog System: A Next Generation Logic Programming Environment, REFERENCE MANUAL.The Ciao System Documentation Series Technical Report CLIP 3/97.1, The CLIP Group School of Computer Science Technical University of Madrid, (www.clip.dia.fi.upm.es/Software/Ciao/), 1999

[CCS00] Chiang, R.; Chua, C.; Storey V.:A Smart Web Query Engine for Semantic Retrieval of Web Data. NLDB 2000, Versailles, France, June 2000.

[Fe99] Fensel, D. t. al.: On2broker: Semantic-based access to information Sources at the WWW. Proceedings of the World Conference on the WWW and Internet (WebNet 99), Honolulu, USA. October 1999.

[Go98] Gonzalo J. et. al.: Extracción de relaciones semánticas entre nombres y verbos en EuroWordNet. Revista SEPLN, 23, 1998.

[GGM97] Goñi, J. M.; González, J. C.; Moreno, A. ARIES: A lexical platform for engineering Spanish processing tools. Natural Language Engineering, 3 (4), 1997, pp. 317-345.

[JCS99] Julian, V.; Carrascosa, C.; Soler, J. Una arquitectura de sistema multi-agente para la recuperación y presentación de la información. IV Congreso ISKO-España EOCONSID'99.

[Li95] Lieberman, H.: LETIZIA: An agent that assists web browsing.Proceedings IJCAI'95, pp.924-929, 1995.

[Li98] Liddy, E. D. Enhanced Text Retrieval Using Natural Language Processing. ASIS Bulletin, April/May, V. 24, N. 4, 1998 (www.asis.org/Bulletin/Apr-98).

[MG00] Martínez, P.; García-Serrano, A. The role of knowledge-based technology in language applications development. Expert Systems with Aplications, V. 19, N. 2., 2000, pp. 155-160

[MHN98] Möller, R.; Haarslev, V.; Neumann, B.: Semantics-based information retrieval. Proceedings on Information Technology and Knowledge Systems (IT & KNOWS). Viena and Budapest, 1998.

[St99] Staab, S. et. al.: A System for Facilitating and Enhancing Web Search. Proceedings of International Working Conference on Artificial and Neural Networks (IWANN '99). Alicante, Spain, 1999.

[Vo99] Voorhees, E.: Natural Language Processing and Information Retrieval. In Information Extraction: Towards Scalable, Adaptable Systems. Maria Teresa Pazienza (Ed.). LNAI Tutorial, Springer Verlag, 1999.

[Vo98] Vossen, P. et. al: The EuroWordNet Base Concepts and Top Ontology. Version 2. EuroWordNet (LE 4003) Deliverable, 1998

[To00] Todirascu, A. et. al.: Using Semantics for Efficient Information Retrieval. NLDB 2000, Versailles, France, June 2000.