Suggesting Named Entities for Information Access

Enrique Amigó, Anselmo Peñas, Julio Gonzalo and Felisa Verdejo

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

Abstract. In interactive searching environments, robust linguistic techniques can provide sophisticated search assistance with a reasonable tolerance to errors, because users can easily select relevant items and dismiss the noisy bits. The general idea is that the combination of Language Engineering and Information Retrieval techniques can be used to *suggest* complex terms or relevant pieces of information to the user, facilitating query formulation and refinement when the information need is not completely defined a priori or when the user is not familiar with the contents and/or the terminology used in the collection. In this paper, we describe an interactive search engine that suggests Named Entities extracted automatically from the collection, and related to the initial query terms, helping users to filter and structure relevant information according to the persons, locations or other entities involved.

1 Introduction

Current Internet search engines are quite efficient at finding information, but there are still a number of (common) search scenarios where users are not properly supported:

- *The requested information is available only in a foreign language*. Even if the user is able to read documents in some foreign language(s) (*passive vocabulary*) he might not be able to formulate adequate queries in such language(s) (*active vocabulary*), or he might just ignore in which language he will find the information he is seeking for.
- *The user is not aware of the appropriate wording for the search.* The missing piece here is a better knowledge of the document collection and the specialized terminology in the domain of the search.
- *The user need is vague or not completely defined.* Search engines are good at solving precise information needs, such as "Where can I buy soja milk online in the New York area?". But for more vague requests navigation and browsing of documents is also necessary for refining, tuning and accomplishing the information need [4].
- *The user aim is to compile or summarize pieces of information around a topic.* This kind of searching needs lots of queries and users don't receive any kind of help to cover the main concepts or entities around the topic. In a traditional Information Retrieval setting, the system retrieves a set of relevant documents, and the user has to analyse their contents and extracts the relevant information without assistance.

These challenges motivate further research on interactive search engines using NLP techniques and wide lexical resources as *CrossLexica* [2] or *EuroWordNet* [8]. TREC experiences on interactive Information Retrieval failed to establish quantitatively the benefits of interactive assistance in a classical Information Retrieval task (Interactive Track Reports of TREC/3-9[1]) but positive results are now being obtained when fuzzy information needs are considered, and when the search task is cross-lingual (the document collection is written in a language unknown to the searcher) [6][5]. The *Website Term Browser (WTB)* [6][7] is an interactive multilingual searching facility that provides, besides documents, a set of terminological expressions (mainly phrases) related to the query as an alternative way to access information. Such expressions match and refine the user needs according to the contents and the terminology in the collection. This approach, based on the automatic extraction, retrieval and browsing of terminology from the collection, was showed to be helpful for users to access information when compared to the use of Google's document ranking.

Beyond phrase suggestion, linguistic techniques permit further kinds of processing in order to improve searching facilities and overcome the limitations mentioned above. The hypothesis underlying our approach is that, within an interactive framework, robust linguistic techniques provide rich information without compromising precision, because such information is offered as suggestions where the user will make his final choices.

In this paper, we describe an interactive search engine that, along this general philosophy, suggests Named Entities which are related to the initial query terms, helping users to filter and structure relevant information according to the persons, locations or entities involved. In a hypothetical searching task where the user has to collect and summarize bits and pieces of relevant information around a topic, this kind of information may help not only finding the appropriate documents, but also finding and structuring the relevant information scattered along them.

The Named Entities are automatically extracted from the document collection using linguistic processing software. This approach has been implemented in the first Hermes[2] project prototype. The following sections describe the parts of the system aimed to extract, select and suggest Named Entities, as well as the kind of interaction that this feature introduces to help users in the searching process.

2 Linguistic processing of the document collection

The *Hermes* prototype applies NLP techniques to lemmatize documents and extract Named Entities before they are used to index the collection. The document collection currently consists of 15,000 news in Spanish (the final collection will be ten times larger and will contain also documents in English, Catalan and Basque languages). This collection has been lemmatized and POS tagged with MACO and Relax [3]. The

[1] http://trec.nist.gov

[2] This work has been supported by HERMES project under a grant (TIC2000-0335-C03-01) from the Spanish Government. http://terral.lsi.uned.es/hermes

Named Entities (NE) have been recognized and classified with an NLP package developed by the Technical University of Catalonia [1] in two steps:

1. Named Entity Recognition (NER), consisting of detecting the pieces of text that correspond to names of entities.
2. Named Entity Classification (NEC), consisting of deciding whether each detected NE refers to a person, a location, an organization, etc.

3 Search process

Figure 1 shows the *Hermes prototype* interface. Results of the querying and retrieval process are shown in four separate areas:

1. A ranking of classified Named Entities (Person, Organization, Location and Miscellaneous, on the left area) that are salient in the collection and probably relevant to the user's query.
2. A ranking of documents (right) classified by date, subject or category (document metadata fields).
3. An area where a refined query is built interactively (central part of the interface) according to the available named entities and the documents being found at each refining step.
4. An area to view individual documents (bottom part of the interface).

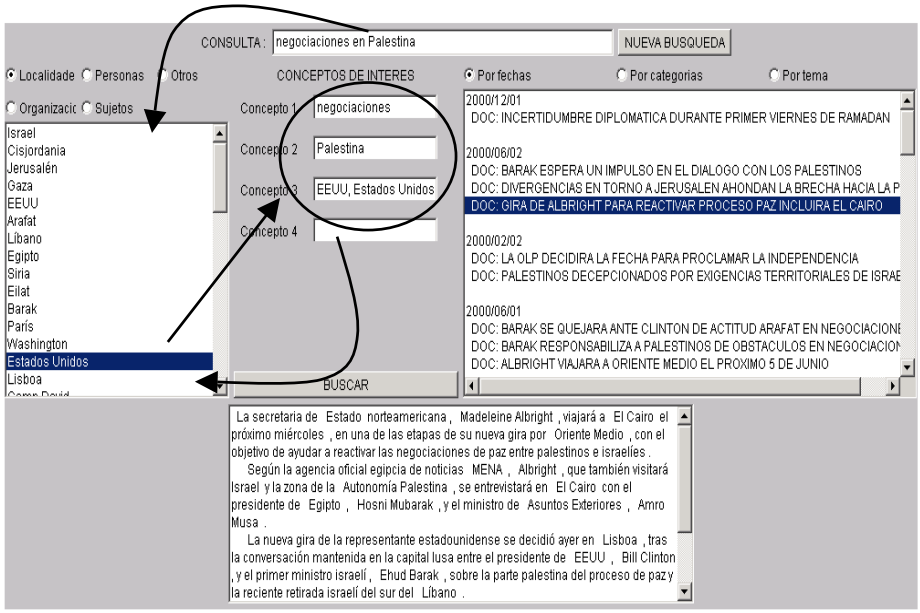

Figure 1: Hermes Search Interface (first prototype)

All this information is presented to the user, who may browse the ranking of entities, refine the query or directly click on a document to view its content. The complete search process follows four steps:

Initial querying. The user introduces some initial filtering words or a natural language expression. From this first query the system determines the subset of documents that will be explored. Then the system performs a local analysis over the document subset in order to extract relevant information. *Figure 1* shows that user has written *"Palestina"* as a query, expressing a quite vague information need.

Query refinement. The system automatically identifies and classifies the entities present in the document subset, obtaining statistics of their presence both in the subset and in the whole collection. Entities are then shown to the user, ranked and organized in a hierarchy according to:

- *Type of Entity*, i.e., Location, Person, Organization or Miscellaneous. Note that the automatic classification of Named Entities highly depends on world knowledge and context. As we don't apply deep semantic processing there can be some errors in the classification, but they are easily detected by users. For example, *Figure 1* shows that "Arafat" has been classified as a location in some of its occurrences. However, there are texts in which "Arafat" has been correctly classified as person, so it will also appear under the Person hierarchy.
- Salience of the entities weighted according to their presence (document frequency) in the pre-selected documents.
- Subsumed entities. For presentation purposes, a group of entities containing a sub-entity are presented as subsumed by the most frequent sub-phrase in the collection. For example, both *"president Bill Clinton"* and *"Bill Clinton"* are subsumed as instances of *"Clinton"* in the hierarchy. This hierarchical organization helps browsing the space of entities.

From the space of named entities suggested by the system, the user can drag and drop his choices into the query refinement area. Names can be dropped over a new field, implying a new search entity, or can be dropped over a pre-selected name, implying a synonymy for search purposes. *Figure 1* shows that the user has selected both "EEUU" and "Estados Unidos" (United States) and has dropped them into the same concept field.

The new query is submitted to the search engine as a boolean expression, producing changes both in the document area and in the Named Entities area, as illustrated by the flow arrows in *Figure 1*.

Listing of documents. Documents can be listed by date, subject or category according to their metadata. These metadata were automatically assigned in a previous classification process.

Document visualization. A selected document is shown to the user in the document visualization area, where there is an alternative feedback facility: users can click over a Named Entity in the text of the visualization area, producing the highlighting of the documents in the list that contain the selected entity.

4 Conclusions

The system described in this paper follows an interactive browse/searching paradigm to help users stating and refining their information needs. For this task, the Hermes first prototype uses automatically recognized and classified Named Entities. An initial user query determines the context of documents in which salient entities are selected automatically and presented for user selection. Entities become very significant to locate relevant pieces of information and to reduce the space of documents to be explored. This approach complements the traditional ranking of documents being helpful when users have vague or broad information needs.

Our immediate work includes incorporating multilingual aspects to the search process, scaling the system to deal with larger document sets, and designing a methodology to establish quantitative and qualitative parameters to evaluate the utility of Named Entities in interactive information access applications.

5 References

1. Arévalo, M. Carreras X. Màrquez L. Martí M. A. Padró L. and Simón M. J. A Proposal for wide-coverage Spanish Named Entity Recognition. *Revista de la Sociedad Española de Procesamiento del Lenguaje Natural*. 2002; 1(3):1-15.
2. Bolshakov, I. A. anf Gelbukh A. A very large database of collocations and semantic links. Mokrane Bouzeghoub Et Al. (Eds.) Natural Language Processing and Information Systems. Lecture Notes in Computer Science. Springer-Verlag. 2001; 1959:103-114.
3. Carmona, J. Cervell S. Màrquez L. Martí M. A. Padró L. Placer R. Rodríguez H. Taulé M. and Turmo J. An environment for morphosyntactic processing of unrestricted Spanish text. *Proceedings of LREC'98*. 1998.
4. Hearst, M. Next generation web search: setting our sites. *IEEE Data Engineering Bulleting, Issue on Next Generation Web Search*, Luis Gravano (Ed.). 2000.
5. López-Ostenero, F. Gonzalo J. Peñas A. and Verdejo F. Interactive Cross-Language Searching: phrases are better than terms for query formulation and refinement. Evaluation of Cross-Language Information Retrieval Systems, Springer-Verlag LNCS Series, to appear.
6. Peñas, A. Gonzalo J. and Verdejo F. Cross-Language Information Access through Phrase Browsing. Applications of Natural Language to Information Systems, Proceedings of 6th International Workshop NLDB 2001, Madrid, Lecture Notes in Informatics (LNI), Series of the German Informatics Society (GI-Edition). 2001; P-3:121-130.
7. Peñas, A. Verdejo F. and Gonzalo J. Terminology Retrieval: towards a synergy between thesaurus and free text searching. Proceedings of VIII Iberoamerican Conference on Artificial Intelligence, IBERAMIA 2002. Springer-Verlag Lecture Notes in Computer Science. 2002.
8. Vossen, P. Introduction to EuroWordNet. Computers and the Humanities, Special Issue on EuroWordNet. 1998.