An Open Distance Learning Web-Course for NLP in IR

Felisa Verdejo, Julio Gonzalo and Anselmo Peñas
Depto. Ingeniería Eléctrica, Electrónica y de Control
ETS Ingenieros Industriales, UNED
Ciudad Universitaria, s.n. 28040 Madrid - Spain
{felisa,julio,anselmo}@ieec.uned.es

Abstract

A computer-based course addressing the topic of applying Natural Language resources and techniques to Information Retrieval is presented. The course provides several Internet on-line resources to support a learning by doing approach in a real world context. Rationale for the design of the course is presented and a detailed description of the course structure and content is given.

1 Introduction

There are two traditional distinctions in higher education that are currently on a period of big debate. The first one concerns industrial training vs higher education objectives, the second one is related to the role of emerging technologies for distance learning, potentially blurring the until now clear separation between conventional universities and distance learning Institutions.

Changes in both Education and Training contexts are moving closer their previously separate objectives. On one hand, there is much concern about bringing educational curricula more in line with vocational needs. In a variety of disciplines, higher education courses are becoming oriented to professionally recognized qualifications, adopting approaches to integrate practice in context. On the other hand, industrial organizations are looking for a more versatile way of building personal profiles to adapt individuals to the changing needs of their organizations. Thus, acquiring more general skill is been increasingly addressed for advanced training purposes.

Distance learning Universities are operating since the seventies. Most of them were based on the industrial model characterized by the production of highly effective learning materials for independent study and the use of "one-way" media, such as print, video, radio or TV broadcasting. While in traditional education the cost of education depends on the number of students involved, this is not the case with the industrial distance education model. The cost depends on a fixed part for the preparation of materials with less investment in academic staff for tutoring tasks. A clear improvement of using new technology for this model has been the delivery of course materials through CD-ROM, and later on through the Internet. Computer-based materials can integrate presentations with simulations, problem solving tools, virtual laboratories and the like, to engage students in a process of active learning. More controversial is the embedding of human-human interactive technologies. It is an open challenge to find new ways of satisfying increasing interaction between students and tutors when staff resources remain scarce. Peer collaboration is an idea to explore, but as it is the case in many industrial organizations collaborative behavior does not happen spontaneously just because the technology is available, but rather is a shift of culture to be established.

The virtual campus [1] is a metaphor currently deserving a lot of attention. It is sometimes

presented as a bridge for conventional universities to extend their scope to reach distance learners, sometimes as an opportunity for distance learning institutions to provide an environment combining the strength of systematic teaching with elaborated materials to support more efficiently distance study. But metaphors should be carefully contrasted with the demand side perspective to foresee whether technology-based distance teaching can succeed. For example, there is, despite the quite short history of desktop computer conferencing, some failed pilot experiences in real-time multipoint teleteaching. The cost of equipment was a well-known factor, but another very practical issue was often neglected: real distance students are in fact fully reluctant to participate regularly in synchronous events, with a fixed schedule. The existing technology has demonstrated to be powerful. Paying due regard to usability issues in order to realize its potential for learning purposes remains an open question.

Over the last decade the European Commission has launched a variety of programs to encourage international partnership and cooperation in the field of education and training. Some of the programs aimed at analyzing the current situation to recommend future action. Others were projects strongly technology oriented to develop platforms and environments specially focused on distance and flexible learning. In addition, a set of pilot experiences and applications were also implemented.

Within the SOCRATES framework [2], ACO*HUM [3] is a thematic network including a working group on Computational linguistics and language engineering. A plan for developing open distance learning pilot courses was proposed in cooperation with ELSNET [4], and finally a proposal including six pilots was launched in February 98. We have developed one of them, on the topic of Information Retrieval (IR) and Natural Language Processing (NLP) [5].

This topic is especially well suited for a learning-by-doing web course. The Internet itself is the biggest IR testbed, and the web search engines are the most powerful applications of IR techniques. The students can be guided through on-line NLP software to manipulate, expand, translate queries, etc., and get first hand impressions on the utility of such processes.

Next section discusses our approach while a detailed presentation is given in section three.

2 Approach and aims

The approach for developing our project has been to capture the experience of designing materials from the distance teaching tradition, while embracing two principles of recent learning theories in the support of both education and training: (i) *learning by doing*, and (ii) *situated learning*. The first addresses how to acquire knowledge [6] and the latter highlights the ability to deploy knowledge in real world context [7]. Some of the assumptions we consider follows:

A first essential component of our approach is to provide structured access to a set of tools and learning resources, both internal and external to the site.

In order to deploy the strategy of learning-by-doing, students are required to perform tasks for each main issue. These tasks are designed to encourage lots of practical work situating the learning experience as close to the real world as possible.

A second component is to help the learner to be introduced in a professional community. In this respect we would pay special attention to establish for all actors (researchers, teachers, professionals and students) concrete ways to collaborate fruitfully in the evolution and updating of the product. This is a quite important aspect in an emerging field where changes come quite rapidly.

A third component is to enable the learner or trainee to became engaged in a learning community, breaking the feeling of being isolated. Social support and participation

improve learners' motivation: facilities for shared virtual spaces and personal communication would be included.

A fourth component relates to our strong endorsement of the role of collaboration. A culture of collaboration needs to be developed. To explore the potential of group learning we would provide means for developing collaborative activities using asynchronous technology. This would be applicable in contexts where stable small groups of learners could be organized.

Based on the above argumentation the project aims are listed below:

- To develop a computer-based course
  - addressing the topic of applying Natural Language resources and techniques to Information Retrieval or, to be more accurate, Text Retrieval, dealing with multilinguality issues.
  - providing or assembling Internet on-line resources that permit a practical experimentation of the issues considered in the course. Such facilities can be in-site software, including: Stemmers, Morphological Analysers, Part of Speech Taggers for different languages, Multilingual Lexical Databases, Cross-Language Mapping of Queries or external, public domain resources such as Internet Retrieval Engines, Machine Translation Systems, etc.
- To design a structured web-based site
  - allowing single learning on-line mode, and asynchronous collaboration mode
  - offering guidelines to support independent learners
  - providing an interface to facilitate flexible access to the content matter and the rest of the tools and resources
- To involve colleagues in contributing to current and further development of the prototype

The remaining sections of the paper describe how the background ideas have been articulated and embedded in the design of the prototype, further testing with users, as well as the internal evaluation and dissemination plan.

3 Key features of the IR-NLP web site.

Figure 1 shows the Web Site Homepage. The layout is homogeneous through the site: on the left appears the set of available options, on the right the contents for the selected function. There are two different menus: one for the Homepage, the other for the Contents of the course.

The Homepage menu (see figure 1) offers General Information and Communication Services, and provide access to the contents of the course. General Information is structured in five sections: *Using this site*, *Introduction*, *Syllabus outline*, *Requirements* and *Study Guide*. As an example, *Introduction* gives answers to the following questions: (i) What is this course about? (ii) What will you not find here? (iii) In which ways can the site be exploited for learning purposes? Communication services integrate e-mail facilities with a directory of contacts. In addition, *News* will be provided here. It also holds room for a Frequently-Asked-Questions list where answers to common problems related to the course will be build and enriched with the contribution of users.

The *Contents* button provides access to the main page of the course. The *Contents* menu (see Figure 3) contains the following functions:

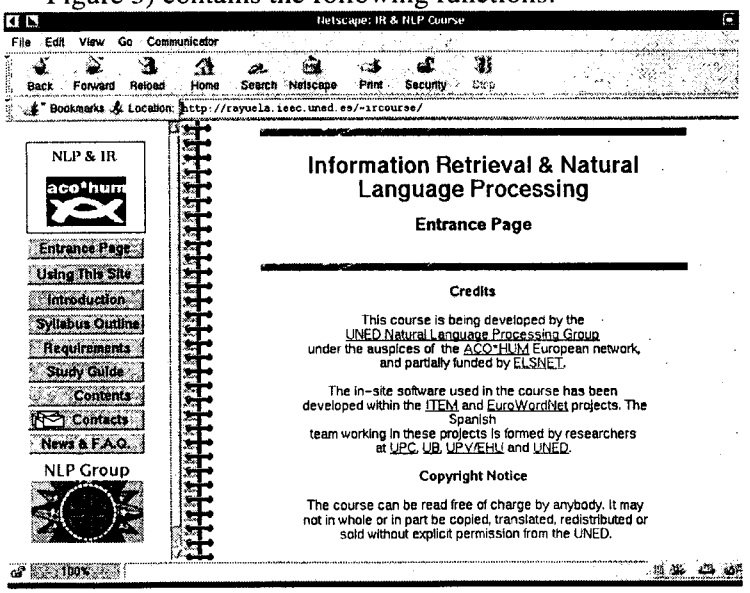

***Figure 1**: NLP & IR web-site homepage*

examples and interactive exercises. Information is now mainly hyper-textual with links to a reading list of references and the glossary. A self-test, that can be filled interactively and submitted for evaluation to obtain feedback, is also included for each chapter. Each exercise has a didactic description in terms of the following features:

- *Estimated workload*: Gives an estimation of the time needed to solve the exercise.
- *Difficulty level*: three levels are considered, easy, medium and difficult.
- *Learning Objectives*: a description of the main purpose of the exercise, either in terms of knowledge or skills.
- *Character*: either Recommended or Optional.
- *Media needed*: Paper&pencil, Software (provided by our site), or Links to external sources.
- *Solution*: One of these three possibilities: (i) The solution is directly available (ii) A method to check the solution is provided (iii) Questions to think about the results are provided.

Figure 2 shows one of the exercises for the Information Retrieval chapter.

Study Guide - offers didactic support to carry out a distance self study of the Contents. For each chapter, the following items are included: (i) A list of the main concepts/principles or techniques for the chapter. (ii) Learning Goals, a description of the concepts/ abilities the student/trainee should acquire by the end of her study. (iii) Schedule, an order of study and activities recommended for distance self-learners. The study guide provides a guided-tour through contents. It is the path recommended for the self-learner wishing to carry out the full course.

The rest of the buttons provide direct access to the collection of on-line *References*, *Links* to external sites, *Glossary*, *Examples*, *Exercises*, *Self-tests* and *Software*. In this way the site can be exploited also as a complementary resource for teachers, trainees, students or professionals to enrich in a particular way their own learning framework.

Estimated workload: **1 h.**
Difficulty level: **Easy.**
Learning objectives: **Get familiarized with IR processes and some of the WWW retrieval engines that will be used in subsequent exercises.**
Character: **Recommended.**
Media: **Link to external source.**
Solution: **Questions to think about your results are provided.**

Try the following queries on at least four search engines here, and answer the questions.

Search for 'Darbuka'.
- How many documents have you found?
- Is there a ranking of the documents retrieved?

Search for 'Buy a Darbuka'.
- Is there a big difference between the number of documents retrieved in both searches? Why?
- Is there any difference in the ranking of documents retrieved?

Try to use a boolean connective like AND (usually in "Advanced Search"): 'Buy AND Darbuka'
- What has happened?
- How many documents have been retrieved?
- Is there any relevant document for the query?
- Is the use of boolean connectives possible on all the search engines you have tried?
- Is there another way to refine the query? How?

Now try searching for 'drums' and after for 'drums from the north of Africa'.
- Which of both queries is more general?
- Which retrieves more documents?
- Can you change the expression "north of Africa" in order to get better results?
- How does the type of words used affect the retrieval? For example, Do proper nouns give better results? Why?
- What happens if we give a description of the term instead of it?

Compare the different search engines.
- Which differences have you found in the results?
- Do you detect differences in coverage? How would you check if a known document is indexed or not for a given retrieval engine?
- Which differences have you found in the way of querying? Is there any relation with the results?
- Which are the main problems you have found?

Figure 2: *Exercise 2.0. Introduction to IR Engines.*

For instance, the ***Links*** button leads to a page where the external links appearing in our site are listed under the following headings:

- On-line Search Software
  - Search Engines
  - Cross-Language Text Retrieval Search Engines Demos
- Language Resources, Tools and Services
  - On-line dictionaries
  - Lexical-Semantic Knowledge Bases
  - Morphological analyzers and taggers
  - Text Summarization and Information Access
  - Machine Translation Services
- Sites of interest

- Information Retrieval
- IR and Natural Language Processing
- Cross Language Text Retrieval

References - This is the main source of documentation to follow the course. There are three kinds of references: material from tutorials or reference books on IR and NLP, relevant papers on each of the topics covered, and reviews of the state-of-the-art and prospects for the future. We try to select only freely available on-line material, to eliminate the significant burden imposed on a distance self-learner if he has to obtain references in documentation centers.

The ***Software*** button leads to a page where a variety of tools related to the course are described. In some cases they are available for interactive use, in others they can be downloaded and installed in your own equipment. Tools are of three kinds: public domain, adapted for the course (for example the Porter algorithm) or specially licensed for the course. For the latter we have designed and build interactive web interfaces. Figure 3 shows an exercise involving the use of external (links in part 1, 2 and 6) and internal resources (part 3). Clicking on "*this interface*", a window appears where one can write a text and submit it for "word by word translation". The results of the morphological analysis and the tagger are shown on the left window, and the dictionary translation for tagged words appear on the right window.

The ***Projects*** button is for registered groups of students, following a regular course (it means under the full responsibility and organization of a tutor). It provides an environment where you can perform collaboratively learning activities.

4 Current status and Future Action

The IRL & NLP Course site is currently public accessible on the WEB, still on a working status (actually we have invested much more than the funding). Now our strategy is to carry out a personalized call to improve and enrich collaboratively the site. We foresee a first phase targeting selected experts and performing a testbed with our Ph.D. university students. Then we will carry out a dissemination plan through specialized networks such as ACO*HUM and professional events.

We will ask experts for their comments on the content and their willingness to establish different degrees of collaboration. From weak ones, such as notifying changes in their URLs, to more involved ones. For instance, contributing with practical material or tools or being included in the directory to answer potential user questions in order to create incrementally structured FAQs.

With students we will test three issues: usability factors, content evaluation as well as experimental data to contrast the current workload estimations of the study guide.

Our intention, for dissemination purposes, is to establish a cooperative framework where researchers on the field could update information or contribute with their points of view. In this way, this web site would constitute a comprehensive and updated reference for any course on the topic.

5 Acknowledgements

Special thanks to Dr. K. de Smedt for his active role in launching the ELSNET LE Training Showcase initiative, as well as to ELSNET for the financial support.

6 References

[1] Verdejo, F., and Davies, G. Editors (1998). *The Virtual Campus: Trends for Higher Education and Training*. Chapman & Hall.
[2] SOCRATES home page: http://europa.eu.int/en/comm/dg22/socrates.html
[3] ACO*HUM home page: *http://www.hd.uib.no/AcoHum*

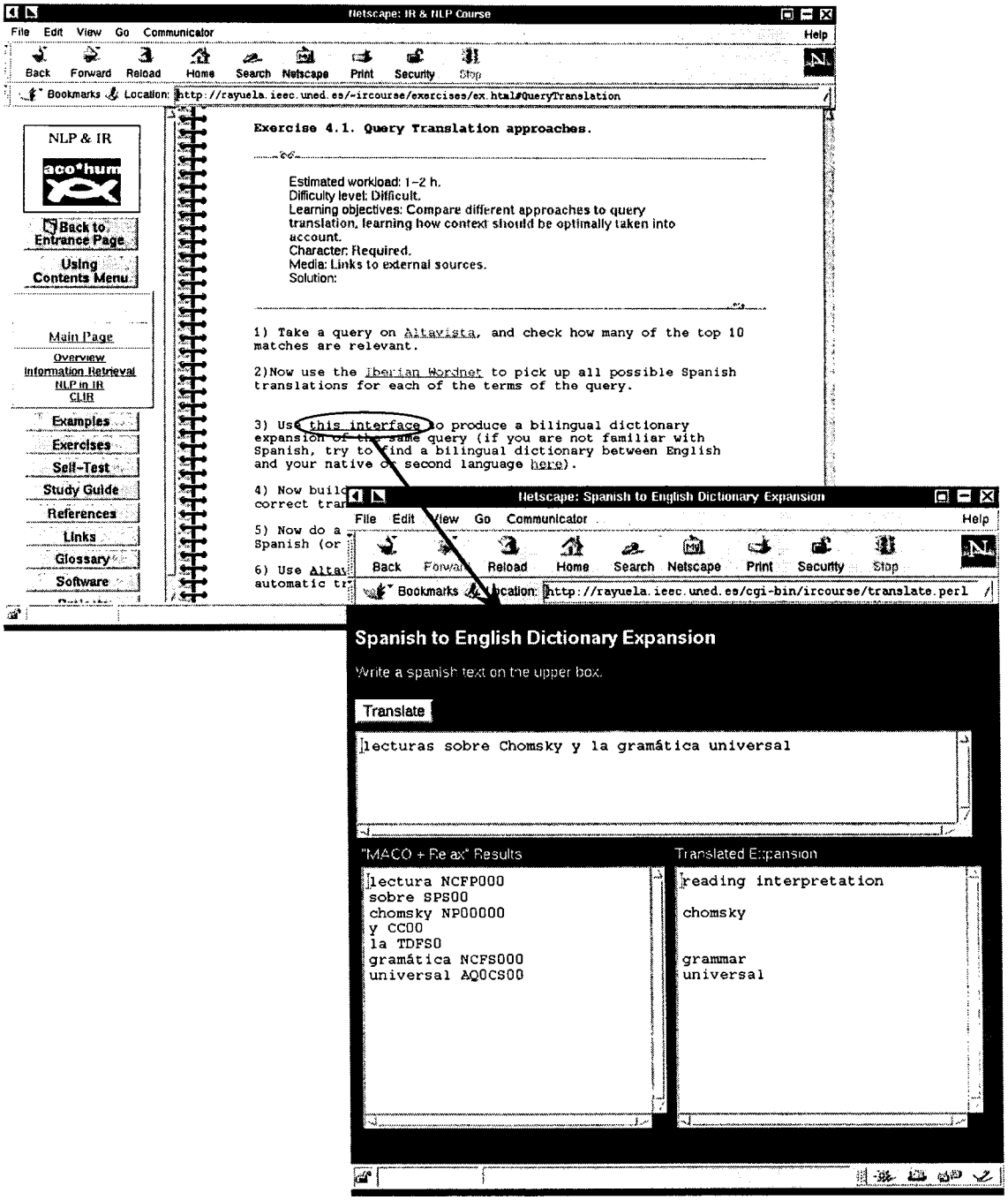

Figure 3. *Example of interaction with on-line software.*

[4] ELSNET home page: *http://www.elsnet.org*

[5] IR & NLP Course home page: *http://rayuela.ieec.uned.es/~ircourse*

[6] Schank, R. C. (1994). Active Learning through multimedia. *IEEE Multimedia*. Vol. 1(1), Spring 94, pp 69-78.

[7] Lave, J. & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.