Language Technology for eLearning

Michael Rosner

Department of Computer Science and AI
University of Malta
mike.rosner@um.edu.mt

Abstract. Given the huge amount of static and dynamic contents created for eLearning tasks, the major challenge for their wide use is to improve their retrieval and accessibility within Learning Management Systems.
This paper describes the LT4eL project which addresses this challenge by proposing Language Technology based functionalities to support semi-automatic metadata generation for the description of the learning objects, on the basis of a linguistic analysis of the content. Semantic knowledge will be integrated to enhance the management, distribution and retrieval of the learning material. We will employ ontologies, key elements in the architecture of the Semantic Web initiative, to structure the learning material within Learning Management Systems, by means of the descriptive metadata. We will also explore the use of Latent Semantic Indexing techniques for the matching of the learning objects with the user information requirements.

Acknowledgement and Disclaimer[1]

1 Introduction

In the context of enlarged Europe, eLearning (including smooth exchange, transformation of contents) will play a central role in education, information exchange and life-long training.

Given the huge amount of high-quality content - teaching materials and courses - one of the major challenges for their wide use is to improve their retrieval and accessibility. This includes:

- production and dissemination of useful and standardized metadata which describes the contents adequately;
- multilingual information retrieval and access of the learning material;
- appropriate matching of well-defined learning need with the relevant content;
- description and retrieval of dynamic contents, i.e. contributions made by learners in an interactive learning environment (e.g. Contributions to fora, chat room discussions)

[1] This article is a highly abbreviated and lightly edited version of project proposal no. 027391 LT4eL as submitted to Brussels under FP6 in March 2005. The project was accepted in July. Although the present author obviously had some role in the preparation of that proposal, the bulk of the text derives from others in the consortium listed in section 6. The present author therefore wishes to acknowledge their contribution and disclaim originality of authorship.
The project coordinator responsible for the original submission is Paola Monachesi (Paola.Monachesi@let.uu.nl) of the University of Utrecht.

Innovative solutions are needed in order to solve the problems which are beyond content production: raising wider awareness of the existence of these contents, personalising eLearning processes with adaptive Learning Management Systems (LMSs) and supporting active learning. Language Technology (LT) and the Semantic Web (SW) are areas which can offer effective solutions to these problems. They offer new technologies which open the possibility to integrate intelligent techniques in eLearning platforms and transcend the current situation in which "eLearning = e-versions of course materials +communication-tool".

Significant research has been carried out in the area of LT and SW: the aim of this project is to enhance eLearning with these technologies in order to develop innovative applications for education and training. We will improve on open source LMSs by integrating the relevant technology. In particular, we will enhance the retrieval of static and dynamic learning objects by employing LT resources and tools for the semi-automatic generation of descriptive metadata while semantic knowledge will be integrated to enhance the management, distribution and search of the learning material.

The new functionalities will be integrated and validated within the ILIAS LMS[2], developed at the University of Cologne. ILIAS is an open-source web-based learning management system that allows users to create, edit and publish learning and teaching material in an integrated system with a normal web browser. However, we plan to adopt a modular approach so that the functionalities developed within the project will be compatible with different open source platforms.

Several initiatives have been launched within LT both at national and international level aiming at the development of resources and tools in the areas of Parsing, Tagging and Corpus Linguistics. However, their integration in enhancing eLearning platforms has not yet been fully exploited. It seems feasible to draw on existing and mature language processing techniques and to integrate them into LMSs. We believe that language resources and tools can be employed to facilitate tasks which are typically performed in a LMS such as searching for learning material in a multilingual environment, summarizing discussions in fora and chatrooms and generating glossary items or definitions of unknown terms.

Several tools and techniques are under development within the Semantic Web initiative which could play a significant role also within eLearning. In particular, ontologies can be employed to query and to navigate through the learning material which can improve the learning process. The topics which constitute the object of learning are linked to ontologies allowing for the creation of individualised courses. There is thus the possibility to develop a more dynamic learning environment with better access to specific learning objects. We will investigate the best way to include ontologies to structure and to query the learning material within an LMS. In particular, we will investigate the extent to which Latent Semantic Indexing (Landauer Foltz and Laham [2]), techniques could be employed for the matching of learning objects with the requirements of the user.

The starting point of this project were the workshops on Language Resources: Integration and Development in eLearning and in teaching Computational Linguistics (http://nats-www.informatik.uni-hamburg.de/view/Main/LrecWorkshop) and on eLearning for Computational Linguistics and Computational Linguistics for eLearning (http://www.cogsci.uni-osnabrueck.de/| vreuer/ws_elearning_cl/) organized during the conference on Language Resources and Evaluation (LREC), and the conference on Computational Linguistics (COLING), respectively. They created the basis for the present collaboration which includes centres from Bulgaria, Czech Republic, Germany, Malta, Netherlands, Poland, Portugal, Romania, United Kingdom and Switzerland1.

From these workshops it transpired that not enough attention has been dedicated, within eLearning, to the commercially less attractive languages (Bulgarian, Czech, Dutch, Maltese, Polish, Portuguese, Romanian) which are represented in the consortium: our project aims at filling this gap by

[2] The ILIAS project website is at http://www.ilias.de

integrating the expertise and the results gained for more commercially attractive languages such as German and English. The new functionalities introduced will be developed for all the languages represented in the consortium.

eLearning applications are very much an emerging field, and there are no standard, general methodologies that can be used to validate effectiveness of the learning process in our specific context. We will therefore develop a suitable validation methodology that will allow us to assess the enahncement of the ILIAS Learning Management System with the added functionalities.

The remaining parts of the paper are structured as follows. Section 2 discusses the main aims and objectives of the project. Sections 3 and 4 describes the implementation philosophy and individual work packages. Section 5 attempts to show how the division of labour is organised, whilst 6 mentions the other members of the consortium.

2 Aims and Objectives

The project will focus on the following scientific and technological objectives:

1. Integration of LT resources and tools in eLearning: the project will employ open source LT resources and tools, produced in the context of other projects, for the development of new functionalities that will allow the semi-automatic generation of metadata for the description of learning objects in a LMS.
2. Integration of semantic knowledge in eLearning: the project will integrate the use of ontologoes, a key element in the SW architecture, to structure and retrieve the learning material within LMSs. Furthremore, it will investigate, through a pilot study, the use of Latent Semantic Indexing for the retrieval of the required learning object.
3. Supporting multilinguality: the project will support the multilingual character of enlarged Europe. Special attention is dedicated to the commercially less attractive languages which are represented in the consortium, such as Dutch, Polish, Portuguese, Czech, Romanian, Maltese and Bulgarian. The results already obtained for more studied languages such as German and English will be integrated. The new functionalities will be developed for the eight languages represented in the consortium.
4. Development and validation of enhanced LMS: the new functionalities will be integrated in the existing open source ILIAS Learning Management System and validated in a realistic learning environment. The developed prototype will be made available to the community and activities will be planned to stimulate its use within academia and schools.
5. Expertise dissemination: the project will stimulate information flow within and beyond the parties of the consortium by means of research and development activities, workshops and seminars, thus strengthening the integration of IST Research in Europe.
6. Awareness raising: the project will draw attention, within the eLearning community to the significant potential of Language Technology and emerging technologies such as the SW. Our aim is to bring together parties across discipline boundaries.
7. Knowledge transfer: the project will encourage the flow of knowledge from academia to industry. We wish to strengthen the cooperation with industrial partners in the area of eLearning and to encourage their participation in research activities through the organization of a user panel which will monitor the project throughout.

3 Implementation Principles

Contents for eLearning tasks are created either inside a learning management system, e.g. through built-in authoring tools, or outside a particular LMS. They will reside in a particular LMS or they will be distributed on various platforms. Modern LMSs must therefore allow for import and export of contents. This includes not only static content, which is supplied by authors and does not change frequently, but also dynamic content which is generated by the participants through the learning process, ranging from short notes, contributions to discussion fora and emails to more elaborate texts such as homework assignments.

In this project, we will address one of the major problems users of ever expanding LMSs will be confronted with: how to retrieve learning content from an LMS. We want to tackle this problem from two different, but closely connected, perspectives: content and retrieval.

On the content side, the fast growing content cannot be easily identified in the absence of systematic metadata annotation. It should thus be common practice to supply metadata along with the contents, however this is a tedious activity which is not widely accepted by authors as part of their tasks.

The solution we offer is to provide LT based functionalities that take care of the metadata annotation semi-automatically on the basis of a linguistic analysis of the (static or dynamic) content. To our knowledge no such functionalities have been integrated in LMSs yet, which makes the project quite innovative. One other thing that makes our project unique is that we want to provide these functionalities for all the nine languages represented in our project. All partners will participate in the research and development of the new functionalities for their language by providing their resources, as well as their computational and eLearning expertise.

On the retrieval side, we can observe that standard retrieval systems based on keyword matching will only look at the queries and not at systematic relationships between the concepts denoted by the queries and other concepts that might be relevant for the user. In this project, we will use ontologies as an instrument to express and exploit such relationships, which should result in better search results and more sophisticated ways to navigate through the learning objects. An ontology of at least 1000 concepts for the relevant domain will be developed as well as an English vocabulary and English annotated learning objects. All partners will contribute to the development of the language vocabularies which will be linked to the ontology as well as to the annotation of the learning objects. We will implement multilingual retrieval of learning objects, focussing on the languages of the NMS and ACC and on the language families represented in the consortium (Romance, Germanic, Slavic).

4 Work Packages

In order to achieve the goals of the project, the following work packages have been envisaged:

4.1 WP 1: Setting the scene

This involves making a survey of the state of the art in the fields which are relevant for the tasks in WP2 and WP3, in particular information extraction. The survey will cover published research as well as available systems (open source or ILIAS compatible) which perform these tasks at least partially. In addition, we will make a survey of the sort of information people are typically looking for in the domain under consideration. The main tasks are as follows:

- Survey of the state of the art for the fields relevant to WP2 and WP3 including information extraction, inventorisation and classification of existing tools and resources;

- collection and normalization of the learning material in the area of information society technologies,dealing with the use of the computer in a particular application area to be defined, possibly business or humanities. We aim at a corpus of 200,000 words (i.e. 1000 pages) for each language. We will try to look for material dealing with the same topic in the various languages. IPR issues will be considered.
- adoption of relevant standards (e.g. OLAC [1] and IMDI [3]) for linguistic annotation of learning objects;
- development of a glossary for eLearning and LT terms to be used project-wide;
- dissemination of the results through the Web portal;

4.2 WP 2: Semi-automatic metadata generation driven by LT resources

The aim of this workpackage is to improve the retrieval and accessibility of content through the identification of learning material using descriptive metadata. LT and resources for the languages addressed in the project will be employed to develop functionalities to facilitate the semi-automatic generation of metadata. Authors and managers of learning objects will be provided with a set of candidate keywords (for the keyword field). Terms and definitions will be detected and provided for glossary compilation as well as input to the ontology construction which is the main target of WP3. Here the tasks are:

- Annotation of the training and test corpora for the languages covered;
- definition of use cases;
- implementation of new functionalities: keyword extractor (at least 1000 keywords), glossary candidate detector;
- testing evaluation and feedback (2 cycles);
- possible optimisation of the functionalities after testing;
- documentation of the new functionalities.

4.3 WP 3: Enhancing eLearning with semantic knowledge

Ontologies will be adopted to structure, query and navigate through the learning objects which are part of the LMS.

Two groups of users will be taken into account: Educators and authors of teaching material who want to compile a course for a specific target group and who want to draw on existing texts, media etc., and learners who are looking for contents which suit their current needs, e.g for self-guided learning.

The ontology will play two roles. (i) In the classification of learning objects, each learning object will be connected to a set of concepts in the ontology. This classification will allow ontological search, i.e. search based on concepts and their interrelations within the ontology. (ii) In multilingual search for learning objects, the ontology acts as an interlingua between the different languages. Thus the user might specify the query in one language and retrieve learning objects in other language(s).

The main tasks in this WP are

- Definition of use cases;
- domain ontology development based on existing resources;
- creation of English vocabulary for the ontology;
- ontological annotation of learning objects for the various languages;
- creation of vocabularies for the various languages and their mapping to the ontology;
- multilingual retrieval: evaluation of problems and tuning;

4.4 WP 4: Integration of the new functionalities within ILIAS

ILIAS does not provide semantic web based functionalities but it does offer the possibility of reusing learning objects like media objects or glossary items in the process of creating learning material. Ontology-based retrieval of learning objects will considerably improve the task of reusing learning objects since ontologies will allow for intelligent searching and navigation in huge amounts of data.

Metadata annotation and ontology-driven search and navigation will allow for individual content assemblance for learners. Learners will be able to build individual learning paths by entering key terms of concepts they need to learn.

The work package includes

- Technical integration of the functionalities and ontology tools into the ILIAS LMS;
- testing of integrated functionalities
- technical and user documentation
- optimization of functionalities after testing and validation

4.5 WP 5: Validation of new functionalities in ILIAS

The aim of this work package is to validate the exctent to which LMSs are improved by adding new functionalities based on language technology tools and ontology-based retrieval. The objective is to assess the extent to which the integration of these new functionalities affects the effectiveness of the eLearning process. The work includes the following tasks:

- Development of a suitable validation methodology for eLearning;
- Preparation of experiments and questionnaires;
- Pilot experiments and questionnaires;
- Execution experiments and questionnaires;
- Report results.

5 Time Frame

In the table below, numbers have been rounded for the sake of readability.

WP	WP Name	Man Months	Malta MM	Start	Finish
1	Inventory	14	1.2	1	3
2	Metadata Generation	77	7.5	4	30
3	Semantic Retrieval	62	6.0	4	30
4	Integration	24	1.6	16	30
5	Validation	28	1.6	16	30
6	Dissemination	23	1.0	1	30
7	Management	18	0.2	1	30
	TOTAL	**250**	**18**	**1**	**30**

6 Partners

1. University of Utrecht (Coordinator)
2. University of Hamburg
3. University of Iasi
4. University of Lisbon
5. Charles University, Prague
6. Bulgarian Academy of Sciences
7. University of Tübingen
8. Polish Academy of Sciences
9. Zürich Academy of Applied Sciences, Winterthur
10. University of Malta
11. University of Cologne
12. Open University, UK

7 Conclusion

In this paper we have given a brief survey of the assumptions, goals, methodology and implementation underlying the LT4el project. At the time of writing, the project lies in the future, so there is a sense in which any conclusion has to be speculative. However, it will start soon. The projected kickoff date is October 2005, and therefore the proof of the pudding will soon be in the eating.

Apart from the intrinsic goals as described above, it is worth mentioning that the project provides a great opportunity to begin the hard task of integrating LT with other areas of CS that are of interest to the Department.

From this author's perspective, one of the success criteria will be measured in terms of the degree of synergy with other eLearning projects that are either under way or about to start in the Department.

References

1. S. Bird and G. Simons. OLAC overview. http://www.language-archives.org/documents/overview.html, 2001.
2. Foltz P. Laudauer T. and Laham D. Introduction to latent semantic analysis. Discourse Processes, pages 259{284, 1998}.
3. P. Wittenburg, D. Broeder, and B. Sloman. Eagles/isle: A proposal for a meta description standard for language resources, white paper. LREC Workshop, Athens, 2000.