

Language Technology for eLearning

Paola Monachesi¹, Lothar Lemnitzer², and Kiril Simov³

¹ Utrecht University, Uil-OTS
Trans 10, 3512 JK Utrecht, The Netherlands
{Paola.Monachesi}@let.uu.nl

² University of Tübingen
Wilhelmstr. 19, 72074 Tübingen, Germany
{lothar}@sfs.uni-tuebingen.de

³ LML, IPP, Bulgarian Academy of Sciences
Acad. G. Bonchev 25A, 1113 Sofia, Bulgaria
{kivs}@bultreebank.org

Abstract. Given the huge amount of static and dynamic content created for eLearning tasks, the major challenge for extending their use is to improve the effectiveness of retrieval and accessibility by making use of Learning Management Systems. The aim of the European project *Language Technology for eLearning* is to tackle this problem by providing Language Technology based functionalities and by integrating semantic knowledge to facilitate the management, distribution and retrieval of the learning material.

1 Introduction

In the *Language Technology for eLearning project* (LT4eL), we address one of the major problems users of ever expanding LMSs will be confronted with: how to retrieve learning content from an LMS. We tackle this problem from two different but related angles: from the content end and from the retrieval end.

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 content, 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 Language Technology based functionalities, such as a key word extractor and a glossary candidate detector. They allow for semi-automatic metadata annotation on the basis of a linguistic analysis of the learning material. We provide these functionalities for all the nine languages represented in our project, that is Bulgarian, Czech, Dutch, English, German, Maltese, Polish, Portuguese and Romanian.

On the retrieval side, the standard retrieval systems, based on keyword matching, only consider the queries. They do not really take into account the systematic relationships between the concepts denoted by the queries and other concepts that might be relevant for the user. In the *LT4eL* project, we 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 domain of *ICT and eLearning* is being developed as well as an English vocabulary and English annotated learning objects. Language specific vocabularies for the other languages of the project are created which will be linked to the ontology as well as to the metadata of the learning objects. The ontology should facilitate the multilingual retrieval of learning objects.

The functionalities developed within the *LT4eL* project could be integrated in any open source LMS, however, for validation purposes the ILIAS Learning Management System (www.ilias.de) has been adopted.

The contribution of the project consists thus in the introduction of new functionalities which will enhance the adaptability and the personalization of the learning process through the software which mediates it. In particular, the system enables the construction of user-specific courses, by semantic querying for topics of interest through the ontology. Furthermore, the metadata and the ontology are the link between user needs and characteristics of the learning material: content can thus be personalized. In addition, the functionalities allow for retrieval of both static (introduced by the educator) and dynamic (learner contribution) content within the LMS and across different LMSs allowing for decentralization and for an effective co-operative content management.

The project is in its initial phase since it started in December 2005 and it will last for 30 months.

2 Semi-automatic Metadata Generation Based on Language Technology

In eLearning, we deal with learning objects of varying granularity. A learning object should be accompanied by metadata which describes it: the "Learning Object Metadata" (LOM) [1] has become the most widespread and well-known standard for the encoding of the metadata. The aim of the *LT4eL* project is to improve the retrieval and accessibility of content through the identification of the learning material by means of descriptive metadata. To this end, we employ available Language Technology resources to develop functionalities which facilitate the semi-automatic generation of metadata from content. The result will be two modules: (1) semi-automatic selection of keywords; (2) selection of definitory contexts, to be used for glossary entries, which can be used stand-alone, as web services, or integrated into Learning Management Systems.

2.1 Key Word Detection and Extraction

Keywords which describe the topics and contents of a document are essential for the retrieval and accessibility of this document and are therefore a key feature of the metadata. A software assistant will support authors and content managers in selecting appropriate keywords.

The keyword selector can draw on quantitative and qualitative characteristics of keyword candidates. Keywords are supposed to characterize the topic(s) of a learning object. They therefore tend to appear more often in a document than can be expected if all words would be distributed randomly. Keywords tend to cluster in certain documents and will not appear at all in other documents. A statistics which is frequently used to model the distribution of words in texts is Poisson or, alternatively, a mixture of Poissons (cf. [2]). While the distribution of function words is close to the expected distribution under these models, good keyword candidates deviate significantly from it. The score of this deviation can be used as a statistics by which the lexical units are ranked (cf. [3]).

A second distributional characteristics of keywords is their *burstiness*. Good keywords tend to appear, within documents, in clusters. Once a keyword appeared in a text, it tends to appear in shorter intervals. After a while, the word might disappear and appear – and burst – again, if the topic is resumed. Reference to a certain topic is a local phenomenon. The concept of term burstiness can be formalized by measuring the gaps between the individual occurrences of each term in the text. Sarkar et al. use a combination of two exponential functions the distribution within burts and outside bursts. Through an iterative process the variables values in these functions are modified to optimize the fit between observed and expected distribution values. Optimal values for these three variables are derived once the fitting reaches a stable maximum. These values turn out to be a good indicator for the “keywordiness of words in texts ([4]).

In the project, we employ linguistically annotated texts since it has been shown that the results of the keyword selection, measured against the performance of human readers, improves significantly if the text is annotated linguistically (cf. [5]).

We will experiment with different statistics. These statistics are:

- tf.idf, a standard term weighting measure in Implementation Retrieval
- Residual Inverse Document Frequency (cf. [3]).
- Term burstiness (cf. [4])

Our experiments are based on the assumption that the distributional characteristics of good keywords are similar to that of good terms, in the sense used in Information Retrieval.

The output of the various statistics will be compared to manual keyword annotation, performed by experienced annotators. A comparison of high ranking keyword candidates with keywords selected by humans will approximate recall and precision values. The (mix of) statistics which we will use in the final system will be selected on the basis of these evaluation results.

2.2 Detection of Candidates for Glossary Entries

Research on the detection and identification of definitory contexts has been pursued mainly in the context of question answering systems, where finding answers to definitory questions is a particularly difficult problem (cf. [6], [7]). In

the field of eLearning, this work is relevant for the construction and maintenance of glossaries.

Glossaries are an important kind of secondary index to a text. They can be seen as small lexical resources which support the reader in decoding the text and understanding the central concepts which are conveyed. A glossary can be built on the definitory contexts which are presented in the learning objects themselves.

A glossary supports the non-linear reading of learning materials. The reader will not necessarily start at the place where the author defined the central terms and therefore be in need of a definition when he first encounters the term. An exploratory and self-guided learning style is supported (cf.[8]).

Glossaries should be derived from the learning objects in order to capture the exact definition which the author of these documents uses. This definition in many cases overrides a more general definition of the term.

The linguistic structure of definitory contexts is language specific. But within each language, there is not so much variation in the patterns of these contexts. In the project, definitory contexts are identified and learned in a bottom-up manner. First, a substantial amount of definitions are identified and annotated manually in the learning objects which are the asset of this project. From these examples, local grammars with the complexity of regular languages are abstracted. These language-specific local grammars are applied to a test set from the same language in order to estimate their coverage (cf. also [10]). A major issue will be the precision of these methods: it has been shown that too simple local grammars also capture text snippets which are not definitions (cf. [9]). This has to be taken into account when drafting and refining the local grammars for the involved languages.

3 Enhancing eLearning with Semantic Knowledge

An additional aim of the project is to enhance LMSs with semantic knowledge in order to improve the retrieval of the learning objects. We employ ontologies, which are a key element in the architecture of the Semantic Web, to structure the learning material. The ontology layer presents the appropriate level of abstraction over the meaning in general (upper ontologies) and in concrete domains (domain ontologies). We integrate the use of ontologies to structure, query and navigate through the learning objects which are part of the LMS. We take two groups of users into account: (1) Educators who want to compile a course for a specific target group and who want to draw on existing texts, media etc.; (2) Learners who are looking for contents which suit their current needs, e.g for self-guided learning.

The primary key for the access of contents are their metadata. In many cases, learners additionally want to consult the full text of a learning object. However, the search with a standard search engine too often leads to results which are undesirable. The set of documents might be too large and too unspecific (low recall, low precision of the retrieved documents). We improve the retrieval of the learning objects with the use of ontologies which will be integrated within the

LMS to structure the learning material. Each learning object will be indexed by an ontology ‘chunk’ that will allow for more detailed search. An ontology chunk is a part of the classes and relations encoded in the ontology which are considered relevant to the topic of the learning object — see [12].

In our work, the ontology is closer to a taxonomy, that is a set of concepts arranged in Is-a hierarchy. The ontology allows for the classification of learning objects since they 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. Furthermore, multilingual search for learning objects will be possible. In this case the ontology plays the role of Interlingua between the different languages. Thus the user might specify the query in one language and get learning objects in other language(s).

Within the *LT4eL* project, we are developing a domain ontology in the area of our sample learning materials, that is *ICT and eLearning*. Furthermore, vocabularies related to the languages in the project will be aligned to this ontology. The development of the domain ontology will be done as a specialization of an upper-level ontology — DOLCE (see [11]). As criteria for the selection of this upper-level ontology, we can point to the following: (1) the ontology should be constructed on rigorous basis; (2) it should be easy to represent it in an ontological language such as RDF or OWL; (3) there are domain ontologies constructed with respect to it; (4) it can be related to lexicons - either by definition, or by already existing mapping to lexical resource. All of the above points apply to DOLCE.

The alignment of vocabularies with the ontology will ensure an appropriate way of searching over the same learning materials in different languages. Inference mechanisms can be assumed for searching the appropriate learning objects. We expect that the integration of ontologies within an LMS will facilitate the construction of user specific courses, by semantic querying for topics of interests; will allow direct access of knowledge; will improve the creation of personalized content and will allow for decentralization and co-operation of content management.

4 Conclusions

With *LT4eL*, we want to use Natural Language Processing techniques and resources to enhance the effectiveness of eLearning systems and processes.

On the one hand, we want to support authors of learning materials in the tedious task of metadata generation. This part of the task should neither be avoided by the author, nor should it detract too much from the original task, the generation of high-quality learning material. On the other hand, we want to employ language resources, in particular ontologies, to help users find the best learning material they can get to satisfy their current information needs.

We consider the diversity of languages in the European context not as an obstacle to an international eLearning infrastructure, but as a challenge. By tackling multilingual issues we want to broaden the horizon of the learner beyond what is available in his native language and in English.

For the sake of validation, we will integrate the tools and resources into the ILIAS Learning Management System, but we will make them available for other LMS as well. In the validation process, we will take four dimensions into account in order to evaluate the success of the project: (1) the usability of the platform itself, and in what way it is affected by the integration of the new functionalities; (2) the pedagogical impact of integrating the functionalities; (3) the consequences of incorporating multilinguality; (4) the social impact on virtual learner communities - and crucially, how this is affected by multilinguality.

References

1. *Final Draft, Standard for Learning Object Metadata*, IEEE, 2002 – P1484.12.1
2. K. Church and W. Gale. 1995. *Poisson mixtures*. In: Natural Language Engineering. Vol. 1, No 2, pp 163–190.
3. K. Church, W. Kenneth and W. Gale. 1995. *Inverse Document Frequency (IDF): A Measure of Deviations from Poisson*. In: Proc. of Third Workshop on Very Large Corpora.
4. Avik Sarkar, Paul H. Garthwaite, and Anne De Roeck. 2005. *A Bayesian Mixture Model for Term Re-occurrence and Burstiness*. In: Proc. of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005). Ann Arbor, Michigan. ACL. pp 48–55.
5. Anette Hulth. 2003. *Improved Automatic Keyword Extraction Given More Linguistic Knowledge*. In: Proc. of the 2003 Conference on Empirical Methods in Natural Language Processing. pp 216–223.
6. S. Miliaraki and I. Androutsopoulos. 2004 *Learning to identify single-snippet answers to definition questions*. In: 20th International Conference on Computational Linguistics (COLING 2004). pp. 1360–1366
7. S. Blair-Goldensohn, K. McKeown and A.H. Schlaikjer. 2004. *Answering definitional questions: a hybrid approach*. In: New directions in question answering. pp. 47–58
8. B. Liu, C.W.Chin and H.T. Ng. 2003. *Mining topic-specific concepts and definitions on the web*. In: WWW 03. Proc. 12th Internat. Conf. on World Wide Web. pp 251–260
9. Ismail Fahmi and GosseBouma. 2006. *Learning to Identify Definitions using Syntactic Features*. In: Proceedings of the EACL 2006 workshop on Learning Structured Information in Natural Language Applications.
10. Judith Klavans and Smaranda Muresan. 2001. *Evaluation of the DEFINDER System for Fully Automatic Glossary Construction*. In: Proc. of AMIA Symposium 2001.
11. Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, Alessandro Oltramari, Luc Schneider. 2002. *WonderWeb Deliverable D17. The WonderWeb Library of Foundational Ontologies and the DOLCE ontology*. Preliminary Report (ver. 2.0, 15-08-2002).
12. Atanas Kiryakov and Kiril Simov. 1999. *Ontologically Supported Semantic Matching*. in the Proceedings of NoDaLiDa'99 (Nordic Conference on Computational Linguistics). Trondheim, Norway.