

Integrating Language Technology and Semantic Web techniques in eLearning

Paola Monachesi¹, Dan Cristea², Diane Evans³, Alex Killing⁴, Lothar Lemnitzer⁵, Kiril Simov⁶, Cristina Vertan⁷

Utrecht University_1, University “Al.I.Cuza” of Iasi and Romanian Academy - the Iasi branch_2, Open University_3, Eidgenössische Hochschule Zürich_4, University of Tuebingen_5, Bulgarian Academy of Sciences_6, University of Hamburg_7

Key words: *eLearning, Semantic Web, Language Technology, ontologies.*

Abstract:

In the LT4eL project, we will improve the retrieval of learning objects within Learning Management Systems by employing Language Technology resources and tools for the semi-automatic generation of descriptive metadata. We will thus develop new functionalities (key word extractor and glossary candidate detector), tuned for the various languages addressed in the project. Semantic knowledge, in the form of ontologies, will be integrated to enhance the management, distribution and searchability of the learning material.

1 Introduction

Given the huge amount of static and dynamic content created for eLearning tasks, the major challenge in facilitating their use is to improve the effectiveness of retrieval and accessibility by employing Learning Management Systems (LMS).

The available commercial and open-source LMSs offer different levels of metadata support for the retrieval of the learning objects. A preliminary investigation we have carried out has shown that ten out of thirteen popular open source learning management systems provide basic metadata support (e.g. for keywords) while five of the systems considered support metadata based on standards such as the Learning Object Metadata (LOM) [1] or DublinCore [2]. However, the usage of Language Technology based functionalities to enable more efficient metadata annotation or the adoption of Semantic Web techniques to improve the data retrieval is rarely found in popular learning management systems.

The aim of the European project *Language Technology for eLearning (LT4eL)* is to show that the integration of Language Technology based functionalities and Semantic Web techniques will enhance the management, distribution and retrieval of the learning material within Learning Management Systems. The functionalities will be developed for all the nine languages represented in our consortium that is Bulgarian, Czech, Dutch, English, German, Maltese, Polish, Portuguese and Romanian.

We will improve on the retrieval of the learning material by employing Language Technology resources and tools for the semi-automatic generation of descriptive metadata on the basis of a linguistic analysis of the content. To this end, we have collected a corpus of learning objects in the domain of *ICT and eLearning*. This learning material has been annotated with linguistic information and it plays a crucial role in the development of the functionalities as well as in the validation of the project.

Providing metadata should be common practice since it facilitates the retrieval of the learning object(s) and therefore enhance their visibility. However, this is not always the case, since it 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 in order to make the process semi-automatic.

The metadata provided should facilitate the retrieval of the objects. Existing Learning Management Systems employ keyword matching in order to retrieve the relevant material, the focus is thus only on the queries. The systematic relationships between the concepts denoted by the queries and other concepts that might be relevant for the user are not taken into account. In the *LT4eL* project, we use *ontologies* which are a key element in the architecture of the Semantic Web, as an instrument to express and exploit such relationships. This should provide better search results and more sophisticated ways to navigate through the learning objects. We integrate the use of ontologies to structure, query and navigate through the learning objects which are part of the LMS.

The new functionalities could be integrated in any (open-source) LMS, however, for validation purposes the ILIAS Learning Management Systems (LMS) (<http://www.ilias.de>) is adopted. The validation process will employ User Scenarios to assess the success of the project and will determine the pedagogical impact of the integrated LMS as well as the role of multilinguality and its social impact.

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

2 Automatic detection and extraction of metadata

As already discussed, metadata plays a crucial role in facilitating the retrieval of the learning objects. However, content providers are often not aware of the impact of metadata and they are reluctant to supply this additional information as part of the authoring process. One measure against this attitude and its consequences is to raise the awareness of the impact of metadata.

Another possibility is the approach we have taken in the *LT4eL* project, which is to assist authors in the generation of metadata, specifically content-related metadata. The functionalities which we develop will provide keywords candidates and definitory contexts

for the creation of a glossary which the authors will be able to choose and eventually manipulate.

Keywords play a relevant role also for learners who employ them to search for relevant material, however, the search for learning objects could be made more precise by the use of ontologies, as discussed in more detail in section 3. Keywords provide the terms which will lead to the definition of the concepts that populate the domain ontology employed in our system. Given the multilingual approach of our project, the keyword driven search as well the retrieval through the ontology will be crosslingual: users can extend their search to those languages in which they are able to read and understand the learning material.

Furthermore, learners who encounter a term which they do not fully understand within a learning object and which might have been defined elsewhere in the text will be able to refer to a glossary which is constructed from the definitory contexts found in all topic related learning objects.

2.1 Creation of an archive of learning objects

In order to develop these functionalities, i.e. a keyword extractor and a glossary candidate detector, as well as our domain ontology, we need a collection of learning objects which are annotated with linguistic information. Since the collection plays a crucial role also in the validation of the project, it is important to select material in a domain which could constitute an appropriate corpus both for the development of the functionalities and which would support plausible user scenarios.

Within the criteria to identify the domain, the following prevailed: processing, retrievability, concept intersectability, and Intellectual Property Rights (IPR) issues. From the point of view of processing, we are interested to develop similar, if not identical, processing chains for the annotation of the objects which have been collected for all the languages involved in the project. This configures a language independent technology that could also foster its transfer to other partners (and other languages) at the end of the project. From the point of view of the retrieval, it is interesting to work on materials which are closely connected, but not identical. The connectivity of domains over the documents encourages potential users to search cross-lingually for specific topics. However, the closeness of the domains over languages should not be taken to its extreme, namely to the point where the texts would be parallel, because this would make the cross-lingual retrieval an uninteresting task. The fact that the documents are similar in content, but not identical, assures also intersectability of their concepts, which is important for the creation of a domain ontology. Optimally, each concept present in our domain ontology should be referred at least once in some document of each language. Finally, it was agreed that, as far as possible, only documents under no constraints of property rights should be kept, in order to assure unrestricted access to them by any user.

2.2 Levels of processing of the learning objects

In order to facilitate the upload and replace activities during the acquisition of collections, versioning, as well as the chaining of transformations of the documents' format, a dedicated portal has been created. The design of the portal was based on the following requirements: the uploading/downloading procedures should be identical and handy for all partners/languages; the visual interface should make explicit the processing steps and the level of the linguistic annotation; automatic re-computation of the significant parameters (including sizes) of the collections and global reports available on request.

As the initial documents of the collection originate from different sources, their formats are diverse: doc (or rtf), pdf, txt (plain text), html or latex and other. For standardization reasons and compatibility with most processing technologies currently used in language technology and eLearning, we adopted an XML format for the representation and processing at the conceptual level. The UTF-8 encoding standard was agreed upon in order to secure representation of all languages character sets.

The learning objects have undergone three levels of processing which are reflected in the structure of the portal. The first layer accommodates initial documents in all possible original formats and the processing steps of this layer include conversion procedures from any of the original formats onto a simple XML document. This schema, called Base-XML, preserves only the plain text and very few formatting information considered important for the identification of the keywords as well as figures, tables and other objects which are of a different nature than pure text. Layer two is dedicated to linguistic annotation. It includes annotations of a document at the sub-syntactic levels: morphological annotation, tokens, part-of-speech, lemmas and noun-phrases. The linguistic annotation schema at the end of all linguistic processing steps is called WP2-XML. Finally, the last layer is dedicated to keywords and definitions and the corresponding schema is called K&D-XML. For the first time in the whole processing chain, all conversion of documents from WP2-XML to K&D-XML will be realized with the same processor, but equipped with specific linguistic resources to reflect language peculiarities. This uniformity of processing tools is in accordance with the methodology described in [3] for representing hierarchies of XML schemas and computing processing sequences triggered by them.

2.3 Extraction of keywords

The collection of learning objects constitute a corpus of linguistically annotated material which can be employed for the development of our functionalities: a keyword extractor and a glossary candidate detector. However, in order to extract appropriate keywords from text it is necessary to define what qualifies as a good keyword. It can be observed that there is a strong relationship between the meaning of a keyword and the topic or subtopic of a text which it characterizes. The functionalities which we develop will not be able to draw on a deep syntactic and semantic analysis of the texts but only on shallow linguistic information. More

specifically, they rely on:

- distributional patterns of keywords which reflect their importance for the text in which they occur;
- layout characteristics of the text which help defining the relevance of the keyword in the context.

In general, words which are supposed to represent the topic of a text tend to appear more often in that text than could be expected if all words were distributed randomly over a corpus. Appropriate keywords for a given text tend to appear frequently in certain documents and will not appear at all or rather infrequently in most other documents. A statistic which is used to model the expected distribution of words in texts is a Poisson distribution or a mixture of Poisson distributions [4].

One of the statistics which has proven to be particularly successful for our task is residual inverse document frequency (RIDF). RIDF measures the difference between inverse document frequency (IDF) which is to be expected for a word, given its collection frequency, under the assumption of a Poisson distribution of words, and the observed (i.e. real) distribution of the word in the data. While the observed distribution of function words is indeed close to the expected distribution, good keyword candidates deviate significantly from the expectation. The score of this deviation can be used as a statistics by which the lexical units are ranked [5]. Pure RIDF does not take into account the frequency of a term in a single document. This is, however, what is required by our use cases. The system should present appropriate keywords for one single document on the background of other documents of the same domain. Therefore, we need to adjust the RIDF figures by the term frequency in the document which is to be keyworded.

A second distributional characteristic of keywords is what has been called their *burstiness* (cf. [6]). While RIDF measures the distributional behaviour of terms across document, term burstiness measures the distributional behaviour within a document. Good keywords tend to appear in clusters within one document. It can be observed that, once a keyword appeared in a text, it tends to appear in the following section of the text in much shorter intervals as would be the case if all the occurrences of this word were distributed evenly throughout the text. When the burst ends, the frequency comes back to normal, until the keyword appears - and probably bursts - again. This intuition is captured by term burstiness, which is a fit of the observed gaps between occurrences of a term to a mixture of two exponential functions. The free parameters of these functions determine the keywordiness of a term. Sarkar et. al shows that the results they achieve resemble the results which have been achieved in earlier studies, e.g. by using RIDF (cf. [7]). It remains to be shown which of the two measures, or a mixture of both methods, yields the best results. At the moment, we are carrying out studies and experiments to verify the best methodology.

Additionally, appropriate keyword candidates can be characterized qualitatively. They appear typically in certain salient regions of the text. These are the headings and the first paragraphs after the headings as well as the abstract or summary. Words occurring in these areas of the text will be weighed higher than the rest. Relevant terms might also be highlighted or emphasized by the author, e.g. by using bold font or underlining the word.

2.4 Identification of definitory contexts

An additional functionality which will be provided within the *LT4eL* project is the possibility to create glossaries on the basis of definitory contexts identified in the learning objects. These contexts will be presented to the author of the learning object to support him in writing an accompanying glossary. Definitory contexts are also very useful to extract ontological relations from the texts, which is another task in the project as discussed in 3.

Previous research has shown that local grammars which match the syntactic structures of the definitory contexts are the most successful approaches if deep syntactic and semantic analysis of texts is not available (cf. [8], [9]).

For each language in the project, we will therefore proceed in three steps. We will:

1. mark the definitions in our training learning objects;
2. write local grammars on the basis of the syntactic patters of the definitions which have been marked;
3. apply these grammars to a test set of learning objects, in which candidates for definitory contexts are marked, extracted, and presented to the author of the learning object.

Since all our learning objects are linguistically annotated in XML, the rules of the local grammars will be applied to these structures. In short, these rules should express constraints over sequences of linguistic items, i.e. sentences, nominal chunks, and tokens. Our candidates are therefore general-purpose XML processing tools, like XQuery or XSLT, or linguistic processing tools which are built on these XML processing tools. Our first investigations showed that the former tools are not adequate to express constraints over sequences of elements and to extract or mark the sequences which match these constraints. Linguistic tools proved to be more adequate for these specific tasks and we have decided to employ them in our project; they allow easy and elegant expression of constraints of sequences of elements.

3 Semantic Web techniques and eLearning

The Semantic Web vision involves a set of new web technologies which have to ensure a semantic layer over the content of the World Wide Web. This layer has to provide meaningful access to the huge amount of information on the Web. Thus, the Semantic Web could support

eLearning in many respects, as outlined in [10]. In the *LT4eL* project, we investigate the techniques that the Semantic Web offers in order to improve the retrieval of the learning objects.

As discussed in the previous sections, metadata plays a crucial role in facilitating the access to the content of the learning objects. Within the Semantic Web initiative, metadata can be represented by employing ontologies which constitute a 'formal, explicit specification of a shared conceptualization' [11]. As such, the ontology (usually called domain ontology) represents common accepted concepts (classes) and relations (properties) in a domain. Thus, ideally the ontology is a source for relevant metadata in a specific domain. The metadata in this case could include a concept or several concepts with the relations among them. The choice depends on the concrete task.

We envisage the latter option as more relevant to the task of indexing learning objects in a LMS. The set of concepts and relations between them that we use for indexing is called 'ontology chunk' [12]. The ontology chunk plays several functions:

- it allows more detailed search without consulting the ontology;
- it represents the relevant information in the context of the learning object --- the author might use different ontology chunks for indexing depending on the concrete learning object;
- it does not hamper the general ontology search via navigation over the ontology.

In the current *LT4eL* architecture, the ontology chunks are assigned to the whole learning object, which is acceptable considering that the size of an average learning object is relatively small. We plan to provide a mechanism for the annotation of the content of the learning objects with ontology chunks. To this end, the user will be able to employ a mechanism to select a chunk on the basis of the concepts and relations to be included in the chunk and a navigation strategy over the ontology which fills the chunk with the appropriate data from the ontology. In addition, the content provider can use a predefined set of chunk patterns.

A domain ontology is currently being developed in the *LT4eL*'s sample learning materials area, which is *ICT and eLearning*. The development of the domain ontology is conceived as a specialization of an upper-level ontology --- DOLCE [13]. The following criteria have led us to the selection of this upper-level ontology: (1) the ontology should be constructed on rigorous basis; (2) it should be easy to be represented 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 some lexical resource.

The actual strategy for the creation of the domain ontology follows the definition in [14]:

- lexicon (vocabulary with natural language definitions);
- simple taxonomy;

- thesaurus (taxonomy plus related-terms);
- relational model (unconstrained use of arbitrary relations).
- fully axiomatized theory.

We have started with the construction of a terminological dictionary of the chosen domain. The entries of the terminological dictionary contain the term in English, a short definition in English and the corresponding translations of the term in the languages represented in the project. The next step is to formalize the definitions in such a way that they reflect the basic ontological relations like *is-a*, *part-of*, *used-for* and others that will be inferred from the upper ontology. Then the definitions will be translated into ontological definitions in OWL-DL. In our project, we will not achieve a fully axiomatized theory, but we will have a relational model of the domain. By connecting the domain ontology to the upper ontology, we will ensure the inheritance of the axiomatization of the upper ontology to the concepts in the domain ontology. We envisage the mapping on the upper ontology especially for senses which do not have a correspondent in our domain ontology. This situation can appear when including already existing semantic lexicons like WordNet where many senses of the words are not relevant for our domain [15]. As by-products of this approach are the vocabularies for several languages aligned to the ontology.

It should be noticed that the following problems might be encountered when mapping the lexicons of the various languages to the ontology:

1. one word in a language subsumes two or more concepts in the ontology (P1);
2. one word in a language subsumes two or more concepts in the ontology but only in relations with some other concepts (P2);
3. one word has a more restrictive meaning not present in the ontology (P3).

In case of (P1), we will define the lexical items in OWL-DL expressions: disjunction, or conjunctions of classes. For (P2), we will express the lexical items in OWL-DL using together with operations on classes also relations between the involved concepts. In case of (P3), we will insert new concepts in the ontology. If one word cannot be mapped directly on the ontology, we will look if a similar meaning can be retrieved in some other languages. If this seems not to be an isolated case then the new concept will be inserted in the ontology. In order to ensure consistency in the retrieval phase we will assign to each concept a label indicating the languages in which this concept is lexicalized.

4 Integration and validation of functionalities in ILIAS

The functionalities developed in the project could be integrated in any open source LMS, however for validation purposes, the ILIAS system has been chosen. The goal is to develop an

interface between the functionalities and the ILIAS learning management system.

ILIAS offers the typical learning management system features like creating, editing and publishing of learning materials, collaboration and communication tools, course management, test and assessment tools and user administration. It also includes basic LOM support, but lacks, as is the case for other LMSs, advanced techniques for more efficient metadata handling and learning object retrieval.

The basis of the interface will be a common communication standard like XML-RPC or SOAP in order to make the integration of the functionalities possible for a wide range of learning management systems. Most actual integration efforts of educational systems prefer web service based architectures, therefore SOAP and related standards will be probably adopted in the *LT4eL* project.

The basis for the integration of the functionalities within the LMS is constituted by the use cases. They show how the behaviour of the LMS changes through the use of the developed functionalities, especially how existing features of the system are improved and how new features have been made possible through their use. Examples of relevant use cases are:

- author annotates semi-automatically learning objects with keywords;
- author generates semi-automatically glossaries for learning objects;
- learner searches for learning objects.

The use cases constitute the basis for the definition of the interface between the LMS and the developed functionalities. They will also provide a starting point for the definition of validation scenarios in the validation phase of the project.

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. A suitable validation methodology is being developed which will be applied to the validation of the new functionalities as well as to their integrated set into ILIAS.

Our validation process will be centred on the development of a number of User Scenarios, this approach was used extensively in the EU funded Mobilearn project.[16] User Scenarios, which focussed on the role of teachers and learners, were found to be helpful in developing ideas about possible uses, enabling progression towards field studies, and also in influencing the development process by focusing on the role of users throughout different stages. User Scenarios were defined as 'a story focused on a user or group of users, which provides information on the nature of the users, the goals they wish to achieve and the context in which the activities will take place'. [17]

They are written in ordinary language, and are therefore understandable to various stakeholders, including users. They may also contain different degrees of detail.

In the context of the *LT4eL* project, scenarios are being developed which will focus on our

potential users who will be found among Course Creators, Content Authors or Providers, Teachers and Students. Our scenarios will be constructed to take the following four dimensions into account in order to evaluate the success of the project:

- the usability of the platform itself, and in what way it is affected by the integration of the new functionalities (D1);
- the pedagogical impact of integrating the functionalities (D2);
- the consequences of incorporating multilinguality (D3);
- the social impact on virtual learner communities - and crucially, how this is affected by multilinguality (D4).

The scenarios are still very much in their infancy and it is expected that they will be considerably enriched as the development of the functionalities progresses. The resulting dialogue between evaluators and developers will help to establish the possibilities for future use and subsequent scenario development and may also influence the development process.

5 Conclusions

In the *LT4eL* project, we will improve the retrieval of learning objects within LMSs by employing Language Technology resources and tools for the semi-automatic generation of descriptive metadata. We will thus develop new functionalities (key word extractor and glossary candidate detector), tuned for the various languages addressed in the project. Semantic knowledge, in the form of ontologies, will be integrated to enhance the management, distribution and searchability of the learning material.

Multilinguality is a central issue in the project and it will play a crucial role in the validation process: we have collected learning objects in nine languages belonging to different language families, we have normalized this material by means of a convertor which has taken into account multilingual issues, such as the handling of different diacritics types and of Cyrillic characters. However, we believe that the most innovative result of the project is the crosslingual retrieval of the learning objects (which will be integrated in the ILIAS LMS) with the help of the language independent ontology and the language specific lexicons.

References:

- [1] IEEE Standard for Learning Object Metadata, IEEE, 2002 - 1484.12.1 <http://ltsc.ieee.org/wg12/>
- [2] ISO Standard 15836-2003, ISO, 2003 <http://dublincore.org/documents/dces/>
- [3] D. Cristea, C. Forascu, and I. Pistol. 2006. Requirements-Driven Automatic Configuration of Natural Language Applications. In Bernadette Sharp (Ed.): Natural Language

Understanding and Cognitive Science , Proceedings of the 3rd International Workshop on Natural Language Understanding and Cognitive Science - NLUCS 2006, in conjunction with ICEIS 2006, Cyprus, Paphos, May 2006. INSTICC Press, Portugal.

[4] 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}.

[5] K. Church, W. Kenneth and W. Gale. 1995. Poisson Mixtures In: Natural Language Engineering 1(1995)2. pp.~{163--190}.

[6] S.M. Katz. 1996. Distribution of content words and phrases in text and language modelling. In: Natural Language Engineering 2(1996)1. pp.15--59.

[7] Avik Sarkar, Paul H. Garthwaite, and Anne De Roeck. 2005. A {B}ayesian 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. pp48--55.

[8] Judith Klavans and Smaranda Muresan. 2001. Evaluation of the DEFINDER System for Fully Automatic Glossary Construction. In: Proc. of AMIA Symposium 2001.

[9] 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. pp251--260

[10] Demetrios G. Sampson, Miltiadis D. Lytras, Gerd Wagner and Paloma Diaz. (Guest Editors) 2004. Special Issue on ``Ontologies and the Semantic Web for E-learning. *Journal of Educational Technology & Society*. Vol. 7, Issue 4.

[11] Thomas R. Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5, pp. 199-220

[12] Atanas Kiryakov and Kiril Simov. 1999. Ontologically Supported Semantic Matching. in the Proceedings of NoDaLiDa'99 (Nordic Conference on Computational Linguistics). Trondheim, Norway.

[13] 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).

[14] Nicola Guarino. 2000. Ontological Analysis and Ontology Design. A short course at Ontolex 2000. Sozopol. Bulgaria.

[15] Paul Buitelaar 2003, Ontology-based Semantic Lexicons: Mapping between Terms and Object Descriptions In: Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Oltramari, Alessandro Lenci, Laurent Prevot (eds.) *Ontologies and Lexical Resources for Natural Language Processing*. In Preparation.

[16] Project Website: <http://www.mobilelearn.org/>

[17] Evans, D., & Taylor, J., (2005) "The role of user scenarios as the central piece of the development jigsaw puzzle" in J. Attewell & C. Savill-Smith (eds.) *Mobile learning anytime everywhere*, published by Learning and Skills Development Agency, London, UK , ISBN 1-84572-344-9

Contact Author(s):

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