Language Technologies for an eLearning Scenario

Claudia Borg and Michael Rosner

Department of Aritificial Intelligence, University of Malta
{claudia.borg|mike.rosner}@um.edu.mt

Abstract. One of the problems with eLearning platforms when collating together documents from different resources is the retrieval of documents and their accessibility. By providing documents with additional metadata using Language Technologies one enables users to access information more effectively. In this paper we present an overview of the objectives and results achieved for the LT4eL Project, which aims at providing Language Technologies to eLearning platforms and to integrate semantic knowledge to facilitate the management, distribution and retrieval of the learning material.

1 Introduction

eLearning is the process of acquiring knowledge, information or skill through electronic means. One of the most popular gateways to eLearning is online via the Internet, often through Learning Management Systems (LMS). LMS allow tutors to manage collections of learning materials and monitor students' progress, whilst providing students with a structured way to access data. However, given the huge amount of static and dynamic learning content created for eLearning tasks, it becomes necessary to improve the effectiveness of retrieval and the accessibility of such documents through the LMS.

Language Technology can support the evolution of eLearning, especially when used to enhance LMS. From a content perspective, it would be ideal if Learning Objects (LOs) would contain additional information to facilitate the retrieval of such documents. Content creators would want to emphasis their efforts on the learning task, rather than manually selecting and entering metadata. In the project Language Technologies for eLearning (LT4eL) we assist content creators by providing tools, such as a keyword extractor and a glossary candidate detector, to produce useful metadata which can be included within the LO.

Standard retrieval systems tend to offer keyword-based searching, matching words present only in the query term. Some more advanced search techniques might take synonyms into account, yet most techniques do not take into account systematic relationships between concepts denoted in the query term and other related concepts which might be relevant for the user. Ontologies are instrumental in expressing such relations and could result in better and more sophisticated ways to navigate through the learning objects. LT4eL has developed an ICT domain ontology that also allows multilingual retrieval of LOs.

The functionalities developed by LT4eL can be integrated in any open source LMS. For the purpose of validation within the LT4eL Project, ILIAS[1] Learning Management System 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.
LT4eL is a 6th Framework Programme Project, with the aim to facilitate the retrieval of learning objects through the use of Language Technologies and semantic knowledge. The consortium is made up of 12 European Universities representing 9 languages including English and Maltese for which the University of Malta is responsible. This paper presents an overview of the work that has been carried out so far, problems encountered and the results achieved for both English and Maltese.

2 Setting the Scene

The initial proposal submitted for LT4eL was to create new functionalities that will improve the eLearning process. Thus, our goals included putting together a corpus of learning material which could be used for the project. We also needed linguistic tools that would provide us with the required information to enable us to use Language Technology tools within the project. The main tools identified as being most important were a part-of-speech tagger and a morpho-syntactic analyser. Each language had to provide its own tools, and optionally could also make use of a noun phrase chunker and other linguistic tools.
The document collection for English proved to be a relatively easy task, with many IPR-free (Intellectual Property Rights) documents available for download from the Internet. The target for the corpus was of 200,000 words for each language in the areas of ICT and eLearning, with the final English corpus consisting of over 1.2 million words. The situation for Maltese was quite the opposite, where no ICT documents and a negligible amount of eLearning documents are available in Maltese. One can assume that the reason behind this is that the education and examination of many subjects taught in Malta is mainly held in English. Thus it stands to reason that the content is created in English rather than in Maltese. With no corpus available for Maltese, the following sections focused on the work carried out by the University of Malta on the English content.
The functionalities proposed by LT4el are described in Sections 3 and 4. The integration of these functionalities for the purpose of validation is described in

[1] www.ilias.de

Section 5. Section 6 presents the possible ways for the Maltese language to be included in such a project, and what is being proposed to reach this end. Finally we present our conclusions in Section 7 with a view to possible future work outside of the project.

3 Language Technologies for Metadata Generation

The learning objects within the corpus tend to be written in a proprietary format, which does not allow easy manipulation and addition of metadata. In order to standardise the formats across the corpus all LOs were initially converted into HTML[2] to retain layout information. Additionally we included linguistic information to enable the application of Language Technologies. A part-of-speech tagger, a lemmatizer and a morpho-syntactic analyser were used to provide documents with the necessary additional linguistic metadata. The linguistic metadata together with the HTML files were combined into an XML[3]-based format conforming to the XCES[4] DTD, a specification for linguistically annotated corpora [IS02].

Once the corpus has all the linguistic metadata available, we manually identified and annotated a set of 1000 keywords and 450 definitions within our corpora to assist in the creation and evaluation of the tools described in Sections 3.1 and 3.2. Through this schema, all information within our corpus becomes easily extractable and machine readable. Below is a sample of the final annotation, including the marking of keywords and definitions.

```
<s id="s1501">
<definingText id="dt46" def="m281">
<markedTerm id="m281" dt="y">
<tok id="t20908" rend="b" ctag="NNP" base="datum" msd="N,SG,proper,vrbl">
     Metadata</tok>
</markedTerm>
<tok id="t20909" ctag="VBZ" base="be" msd="AUX,PRES,S,finite">is</tok>
<tok id="t20910" ctag="VBN" base="define" msd="V,PAST,ED,finite">
     defined</tok>
<tok id="t20911" ctag="IN" base="as" msd="CJ">as</tok>
...
</definingText>
</s>
```

3.1 Keyword Extraction

A keyword extract (described in [LD06]) was created as part of the deliverables of the project. The first task was to identify the characteristics of keywords, by analysing the manually annotated keywords for their linguistic features.

[2] Hypertext Markup Language

[3] eXtensible Markup Language

[4] Corpus Encoding Standard using XML

Generally keyword extraction techniques consider only word count, with more sophisticated techniques taking into consideration the frequency of a word not only in the document, but also in the whole corpus. The keyword extractor was implemented with three different statistical techniques (Inverse Document Frequency and Residual Inverse Document Frequency described by [CG95], and Term Burstiness [Kat96,SGD05]). The keyword extractor was then further improved by taking into consideration the Part-Of-Speech (POS) tags. A language model was created to reflect what type of linguistic classes would fall under single-word keywords (such as nouns) or multi-word keywords at any positions (such as verbs). Additional weighting is also given to those words which have particular layout information, such as bold or italic, which was retained from the original file.

3.2 Definition Extraction

In the case of definition extraction, the approach taken by the project was for each language partner to identify possible linguistic patterns that could extract definitions. An XML transducer, *lxtransduce* [Tob05], was used to match the defined patterns and a rewrite rule is then applied to the matched cases. In our case, definitions are left intact, surrounded with `definingText` tags. The following is an example of a grammar rule which looks for a determiner at the beginning of a sentence followed by a noun:

```
<rule name="det_S_noun_phrase">
 <seq>
  <query match="s/*[1][name()=’tok’][@ctag=’DT’]"/>
  <ref name="noun_group" mult="+"/>
 </seq>
</rule>
```

The set of 450 manually annotated definitions was split into three sets: (i) a training set; (ii) a testing set; and finally (iii) an evaluation set, each consisting of 150 definitions. The training set was used to extract possible patterns that can be commonly found in definitions. The rules were created through the manual observation of these sentence definitions, representing mainly the POS sequences noticed.

We observed that this task was a tedious one, and it was easy to overlook certain cases. A divide-and-conquer approach was adapted, and the definitions were split into categories. This reduced the complexity of the search space, whereby at each grammar identification attempt it is possible to focus only on one type of definition. The types of definitions observed in our texts have been classified as follows:

1. Definitions containing the verb “to be” as a connector.
   E.g.: ‘A joystick is a small lever (as in a car transmission gearshift) used mostly in computer games.’

2. Definitions containing other verbs as connectors such as “means”, “is defined”, “is called”.
   E.g.: ‘the ability to copy any text fragment and to move it as a solid object anywhere within a text, or to another text, usually referred to as cut-and-paste.’ In this case the term being defined is at the end of the sentence, and it is classified so by the use of ‘refer to’
3. Definitions containing punctuation features, usually separating the term being defined and the definition itself.
   E.g.: ‘hardware (the term applied to computers and all the connecting devices like scanners, modems, telephones, and satellites that are tools for information processing and communicating across the globe).’ where the definition is contained within brackets
4. Definitions containing particular layout style, similar to the punctuation feature, but separated through the use of a table (similar to the punctuation definition, however the term and definition would be placed in separate cells) or the defining term is a heading and the definition is the sentence below it.
5. Definitions containing a pronoun, usually referring to the defining term which would be placed outside the definitory context. This is common in cases where the definition is over more than one sentence, and the second sentence would refer to the defining term using a pronoun.
   E.g.: ‘This (Technology emulation) involves developing techniques for imitating obsolete systems on future generations of computers.’
6. Other definitions to capture those which do not fall in the above categories.
   E.g.: ‘information skills, i.e. their ability to collect and process the appropriate information properly in order to reach a preset goal.’ where the defining term and the definition are separated by ‘i.e.’

The above classification allows for the generalisation of rules to identify definitions in categories 1–5. However, the sixth categorisation does not facilitate the task of identifying a grammar for this category since it contains exceptional cases, and thus cannot be generalised.

Machine Learning for Definition Extraction Through the categorisation of definitions, we were able to improve results for certain categories, such as the is-a category. However, having achieved a high recall, precision was considerably low. This meant that whilst good definitions were being captured, a high number of incorrect definitions were also being included in the result set.

An other problem was that there was no ranking of the results as the extraction method used was a simple yes-no classification. This meant that definitions were presented to the user in no particular order. Since the system was intended to suggest definitions to a content creator for approval, having a few incorrect definitions was not deemed as a problem. However, it was desirable that the definitions are presented in a ranked order, so that those definitions with a higher confidence value are presented at the top of the results. We also observed that incorrectly classified definitions could be filtered out using post-processing filtering after the initial grammar was applied.

To tackle these tasks, a Machine Learning (ML) group within the project was created to research on possible ML techniques and to improve these results. The University of Malta is part of this group and its approach to this task is further reported in [Bor07,BRP07].

4 Enhancing Learning Objects with Semantic Knowledge and Multilinguality

Semantic knowledge provides additional useful information that can be utilised for enhancing of document retrieval. There are two types of users that we consider: (i) educators and content creators compiling a course from existing resources, and (ii) learners searching for content to suit their current needs. LT4eL aims to improve the retrieval of LOs with the use of ontologies, which will be integrated within the LMS to structure, query and navigate the LOs.

An ontology constitutes a formal representation of concepts (classes), and the relations (properties) between them. There are different approaches to ontology design. In our case we looked at a layered design, where generic concepts are represented in an upper-level ontology and more specific concepts are represented in a domain ontology. Through this approach we are able to re-use existing upper-level ontologies, and concentrate more on creating a domain ontology (described in [MLS07]). An analysis of upper-level domain ontologies was carried out to identify which would suit our requirements best. We concluded that DOLCE [MBG$^+$02] (Descriptive Ontology for Linguistic and Cognitive Engineering) suited best our requirements. DOLCE was built using formal ontological analysis and formal semantics, approached using ontological engineering practices. It is also modular and has an open license. These factors influenced our bais towards selecting DOLCE as the upper domain ontology.

In designing a domain ontology, we followed the strategy as specified in [Gua00]:

1. lexicon (vocabulary with natural language definitions);
2. taxonomy;
3. thesaurus (taxonomy plus related terms);
4. relational model (unconstrained use of arbitrary relations);
5. fully axiomatized theory.

We started by constructing a terminological dictionary, which contains the term in English, a short definition describing the term, and the translations of the term in the represented languages within the project. Then we formalised these terms by including basic ontological relations (is–a, part–of, used–for) which are inferred from the upper ontology. These will be translated in OWL–DL. Within the scope of the project, we aim to achieve a full relational model of the domain ontology. By connecting the ontology domain to the upper-level ontology, we will ensure inheritance of the axiomatization of the upper ontology to the concepts in the domain ontology.

The annotation of LOs with semantic knowledge can be based either on a concept or on several concepts and their relations. The latter is referred to as an ontology chunk. We envisage this to be more relevant to our task as it will allow (i) a more detailed search without consulting the ontology, (ii) represents the relevant information in the context of the LO, and (iii) it facilitates generic ontology searches by allowing navigation over the ontology.

The availability of a multilingual lexicon allows for retrieval of documents in several languages. For instance, a user can search using a Polish term and request to view both Polish and English documents where that concept is present. This is particularly useful in a LMS were learners would want to see all relevant information possible. In future, we intend to provide the facility for a content creator to annotate the content of an LO with ontology chunks. Thus the user will be able to employ a mechanism to select a chunk on the basis of the concepts and relations between them, and to include that chunk as part of a LO's metadata to facilitate the navigation over the ontology with respect to the content of the LO.

5 Integrating and Validating eLearning Scenarios

The functionalities described above had to be integrated to an open source LMS. For purposes of validation, ILIAS was chosen to develop an interface between the functionalities and the LMS. ILIAS offers typical LMS features such as creating, editing and publishing of materials, collaboration and communication tools, course management, test and assessment tools and user administration. It also includes basic LOM[5] support, but lacks, as is the case for other LMSs, advanced techniques for more efficient metadata handling and learning object retrieval. In order to make the functionalities available for any LMS, a web service based architecture was used to integrate these functionalities.

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 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 the 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.

[5] Learning Object Metadata

Our validation process will be centred on the development of a number of User Scenarios, which focus on the role of teachers and learners. User Scenarios are 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’ [ET05]. 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 being developed currently focus on course and content creators, teachers and students. The 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;
- the pedagogical impact of integrating the functionalities;
- the consequences of incorporating multilinguality;
- the social impact on virtual learner communities — and crucially, how this is affected by multilinguality.

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 dialog between evaluators and developers will help to establish the possibilities for future use and subsequent scenario development and may also influence the development process.

6 Maltese in Context of LT4eL

The inclusion of Maltese within LT4eL was challenging from its inception. We were aware of the lack of resources in the domain of ICT. Yet running in parallel with this project, the Maltese Lexicon Resource Server [RFAG06] aimed to create a corpus for Maltese, and to set out the framework for language tools to be created. Still the domain of ICT remained short of documents, and thus we were forced to limit the participation of Maltese in LT4eL.

We proposed alternative solutions which had to be discarded as they were deemed either unpractical or unrealistic for the scope of the project. Amongst these:

- to translate existing documents in ICT from English to Maltese — this was discarded because it would have been expensive, and the end result would be a translation — something that we wanted to avoid within the project.
- to use an ICT related corpus in Maltese that is not aimed for eLearning — here we suggested to use documents from the European Parliament website[6] which discussed ICT in general. This idea was also discarded as the corpus would cover to varied a domain from the one LT4eL had achieved (e.g. LOs on Word Processors).

[6] http://www.europarl.europa.eu/registre/recherche/ListeDocuments.cfm Last accessed 15th October 2007

Finally, we opted for the translation of the lexical dictionary described in Section 4. This translation task still remains a challenging one, since many ICT terms are not yet defined in Maltese. Once completed these terms will be integrated into the ontology. This will enable search for LOs in Maltese, and then retrieving English LOs. Of course, this is far from the ideal scenario. However, this discussion is beyond the scope of this paper.

7 Conclusion

The project LT4eL brings together existing research knowledge and tools in different areas and puts their use and application in an innovative way. Keyword-based retrieval of documents is used extensively in many search applications. However, apart from taking the normal statistic approach, we are also including linguistic knowledge to improve the results. We also propose both quantitative and qualitative approaches to the validation of the keyword extractor. In definition extraction, we not only look at grammar patterns that could form definitions, but we will also apply machine learning to this area to see the improvement of extraction. Semantic knowledge will give us increased meta-data that will enable improved document retrieval and navigation through the ontology. All these functionalities will be applied to a LMS to improve the eLearning experience. We believe that the most innovative result of LT4eL will be the crosslingual retrieval of learning objects with the help of language independent ontology and language specification lexicons.

References

[Bor07] Claudia Borg. Discovering grammar rules for Automatic Extraction of Definitions. In *Doctoral Consortium at the Eurolan Summer School 2007, Iasi, Romania.*, pages 61–68, 2007.

[BRP07] Claudia Borg, Michael Rosner, and Gordon Pace. Towards Automatic Extraction of Definitions. In *Proceedings of the Computer Science Annual Workshop, CSAW07, Malta*, 2007.

[CG95] Kenneth W. Church and William A. Gale. Inverse document frequency (idf): A measure of deviations from poisson. In *Proceedings of the 3rd Workshop on Very Large Corpora*, 1995.

[ET05] Diane Evans and Josie Taylor. The role of user scenarios as the central piece of the development jigsaw puzzle. In *Mobile learning anytime everywhere*. Learning and Skills Development Agency, 2005.

[Gua00] Nicola Guarino. Ontological Analysis and Ontological Design. In *A short course at Ontolex 2000, Sozopol, Bulgaria*, 2000.

[IS02] Nancy Ide and Keith Suderman. Xml corpus encoding standard document xces 0.2. Technical report, Department of Computer Science, Vassar College, and Equipe Langue et Dialogue, Vandoeuvre-lés-Nancy, 2002.

[Kat96] Slava M. Katz. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1):15–59, 1996.

[LD06] Lothar Lemnitzer and Łukasz Degórski. Language Technology for eLearning — Implementing a Keyword Extractor. In *the fourth EDEN Research Workshop, Research into online distance education and eLearning*, 2006.

[MBG⁺02] Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari. Wonderweb deliverable d18. ontology library. Technical report, Laboratory for Applied Ontology, ISTC–CNR, 2002.

[MLS07] Paola Monachesi, Lothar Lemnitzer, and Kiril Simov. Language Technology for eLearning. In *First European Conference on Technology Enhanced Learning*, 2007.

[RFAG06] Michael Rosner, Ray Fabri, Duncan Attard, and Albert Gatt. MLRS, a Resource Server for the Maltese Language. In *Computer Science Annual Workshop (CSAW)*, pages 90–98, 2006.

[SGD05] Avik Sarkar, Paul H. Garthwaite, and Anne DeRoeck. A bayesian mixture model for term re-occurrence and burstiness. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL)*, pages 48–55, 2005.

[Tob05] Richard Tobin. Lxtransduce a replacement for fsgmatch. Technical report, University of Edinburgh, 2005.