

Crosslingual retrieval in an eLearning environment

Cristina Vertan¹, Kiril Simov², Petya Osenova², Lothar Lemnitzer³, Alex Killing⁴, Diane Evans⁵ and Paola Monachesi⁶

¹ Natural Language Systems Division, Institute of Informatics, University Hamburg, Germany,

² LML, IPP, Bulgarian Academy of Sciences, Sofia, Bulgaria

³ Seminar für Sprachwissenschaft, Universität Tübingen, Germany

⁴ Center for Security Studies, ETH Zürich, Switzerland

⁵ The Open University, Milton Keynes, UK

⁶ UiL-OTS, Utrecht University, the Netherlands

Abstract. In this paper we are reporting about an ongoing project LT4eL (Language Technology for eLearning) aiming at improving the effectiveness of retrieval and accessibility of learning objects within a learning management system. We elaborate the process of building the domain ontology and present the multilingual support offered to the application.

1 Introduction

The aim of the Language Technology for eLearning project (LT4eL, www.lt4el.eu) is to enhance eLearning with Language Technology tools and resources as well as with semantic information [8], [9] in order to provide new functionalities which will enhance the adaptability and the personalization of the learning process through the software which mediates it.

An important objective of the project is to enhance LMSs with semantic knowledge in order to improve the retrieval of the learning objects. We take two groups of users 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, as well as learners who are looking for content which suit their current needs.

Ontologies, which are a key element in the architecture of the Semantic Web, have been adopted to structure, query and navigate through the learning objects which are part of the LMS. The ontology plays two roles. On the one hand, it is employed in the classification of the learning objects. Each learning object is connected to a set of concepts in the ontology. This classification allows for ontological search, i.e. search based on concepts and their interrelations within the ontology. On the other hand, it makes multilingual search for learning objects 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).

The ontology will be integrated in the ILIAS Learning Management System and we expect that the integration of ontologies within an LMS together with the appropriate tools for navigation, will facilitate the construction of user specific courses, by semantic querying for topics of interests; will allow direct access to ontological knowledge; will improve the creation of personalized content and will allow for decentralization and co-operation of content management.

The paper is structured as follows. In section 2, we present the developed methodology for building an ontology relevant for the domain under consideration, that is computer science for non experts. In section 3, we explain how multilingual material (lexicons and content) was linked to the ontology. In Section 4, we present the multilingual search scenarios which constitute the basis for the specification of the cross-lingual search engine. Section 5 concludes the paper and gives the directions for the future work.

2 Ontology development

We intend to use the ontology to support cross-lingual semantic searches in a repository of learning objects. Two aspects are critical for the the ontology in that context and will therefore be evaluated carefully, i.e. consistency of the taxonomical structure and domain coverage. Consistency of the taxonomy is established by using the OntoClean methodology [4] and by linking our domain ontology to upper ontologies. Domain coverage is evaluated in the process of annotating our learning objects with concepts from the ontology, see below.

In this section we describe the methodology for the constructing a domain ontology, linking it to upper ontologies and formalizing the concepts.

2.1 LT4eL Methodology

Within the LT4eL project the domain chosen is Computer Science for non-specialists. Since no ontology covering the domain was available, we decided to create our own ontology. In the following we describe the major steps of the ontology creation process.

Processing of the Keywords Once the learning objects for all languages involved were in place (cf. [7]), we started to build the ontology by annotating and extracting keywords from them using our keyword extractor(cf. [6], [8]).

The processing itself was performed in the following way:

- *Selection of appropriate keywords* We only considered those keywords which relevant for our domain. Admittedly, this domain is large or vaguely defined, and might contain concepts which are not, strictly speaking, related to computer science at large, but are nevertheless associated to it, such as the results of processes which involve computer, e.g. documents, and typical domains of computer use, e.g. distance learning, desktop publishing, and the web.

– *Collecting definitions*

The Internet was searched for definitions of the selected keywords. The reason behind that was to provide non-formal accounts of the meaning(s) of the keywords. If possible, we selected several definitions for a concept in order to reflect various aspect of the meaning and to have a textual basis from which later on relations between concepts will be derived. From these sets of collected definition we selected or compiled one canonical definition for each concept.

– *Polysemy*

For polysemous keywords, two or more senses and, consequently, concepts have to be established. For example the keyword *word* refers to the linguistic unit as well as to the Microsoft product. MPEG can refer to the organization as well as the standard. Words senses which are irrelevant for our domain have been not taken into account though.

Formalization of the senses The next step was to formalize definitions for the extracted concepts and relations in OWL-DL ([11]). For each meaning an appropriate class in the domain ontology was created. This resulted in an initial formal version of the ontology. This ontology was linked to upper ontologies.

Linking to upper ontologies Linking our ontology to an upper ontology has two advantages. First, we can inherit relations from the upper ontologies, second, as a side product of this manual process the consistency and validity of our ontology is checked and improved.

We considered the following criteria for the selection of the upper ontology: (1) The ontology has to be constructed on a rigorous basis and to suit our domain; (2) it should be represented in an adequate formal language, preferably OWL; (3) domain ontologies exist which have been constructed using this upper ontology, and (4) support is provided by the authors of the upper ontology. After an initial evaluation of the alternatives and consultation of other evaluations of upper ontologies (cf. [12] and [10]) we selected DOLCE. This decision does not aim at ruling out mappings to other ontologies, though.

In order to get appropriate taxonomic relations between the concepts in the ontology and to facilitate the mapping to an upper ontology, we mapped each concept to synsets of OntoWordNet [2], which is a version of WordNet 1.6 mapped to DOLCE ontology. The mapping was performed via the two plug-in relations of *equality* and *hypernymy*. Thus, we created the taxonomical backbone of our ontology. Later on we will add more types of relations. The connection of OntoWordNet to DOLCE allows an evaluation of the defined concepts with respect to meta-ontological properties as they are defined in the OntoClean approach – cf. [3], [14] and [5].

Our ontology was also mapped onto WordNet 2.0. This mapping provides additional benefits. WordNet 2.0 is larger than OntoWordNet and has a richer set of relations. Therefore, the mapping will enable us to derive more relations between the concepts in our ontology. Furthermore, WordNet 2.0 is aligned to

SUMO and thus we get an indirect mapping to another upper level ontology. Since those two mappings have been performed independently of one another, the outcomes can be used to cross-check the validity of each concept and relation.

Results Our ontology currently consists of 707 concepts / classes formalized in OWL and linked to WordNet and DOLCE ontology. In the following we present the concept ‘WebPage’ as an example.

```
<owl:Class rdf:about="http://www.lt4el.eu/CSnCS#WebPage">
  <rdfs:comment>A document (file) connected to the
  World Wide Web and viewable by anyone connected to
  the internet who has a web browser.</rdfs:comment>
  <rdfs:comment>Hyper CSnCS:
  http://www.lt4el.eu/CSnCS#TextFile</rdfs:comment>
  <rdfs:comment>Equal WN20: ENG20-05964213-n</rdfs:comment>
  <rdfs:comment>ID: id1757</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Class rdf:about="http://www.lt4el.eu/CSnCS#TextFile">
      </owl:Class></rdfs:subClassOf>
  </owl:Class>
```

3 Mapping multilingual lexicons and content on the ontology

The main aim of the ontology in the learning management system is to enable a conceptual search through the learning objects. Moreover we intend to offer the user the possibility to perform this search multilingually. The connection between content, user’s query and the ontology is realised through:

- mapping of lexicons in each involved language onto the ontology
- semantic annotation of the learning objects with ontology concepts.

In the following paragraphs we elaborate these two aspects.

3.1 Mapping of the lexicons onto the ontology

Terminological lexicons represent the main interface between the user’s query and the ontological search engine. The terminological lexicons were constructed on the basis of the formal definitions of the concepts within the ontology. In this approach of construction of the terminological lexicon we escaped from the hard task of mapping different lexicons in several languages as it was done in EuroWordNet Project [13]. The main problems with this approach of construction of terminological lexicons are that (1) for some concepts there is no lexicalized term in a given language, and (2) some important term in a given language has no appropriate concept in the ontology which to represent its meaning. In order to

solve these problems we, first, allow the lexicons to contain also non-lexicalized phrases which have the meaning of the concepts without lexicalization in a given language. Even more, we encourage the lexicon builders to add more terms and phrases to the lexicons for a given concept in order to represent as many ways of expressing the concept in the language as possible. These different phrases or terms for a given concept are used as a basis for construction of the regular grammar rules for annotation of the concept in the text. Having them, we could capture in the text different wordings of the same meaning. In order to solve the second problem we modify the ontology in such a way that it contains all the concepts that are important for the domain.

We could summarize the connection between the ontology and the lexicons in the following way: the ontology represents the semantic knowledge in form of concepts and relations with appropriate axioms; and the lexicons represent the ways in which these concepts can be realized in texts in the corresponding languages. Of course the ways in which a concept could be represented in text are potentially infinite in number, thus, we could hope to represent in our lexicons only the most frequent and important terms and phrases.

Here is an example of an entry from the Dutch lexicon:

```
<entry id="id60">
  <owl:Class rdf:about="http://www.lt4el.eu/CSnCS#BarWithButtons">
    <rdfs:subClassOf>
      <owl:Class rdf:about="http://www.lt4el.eu/CSnCS#Window"/>
    </rdfs:subClassOf>
  </owl:Class>
  <def>A horizontal or vertical bar as a part of a window,
    that contains buttons, icons.</def>
  <termg lang="nl">
    <term sheaf="1">werkbalk</term>
    <term>balk</term>
    <term type="nonlex">balk met knoppen</term>
    <term>menubalk</term>
  </termg>
</entry>
```

Each entry of the lexicons contains three type of information: (1) information about the concept from the ontology which represent the meaning for the terms in the entry; (2) explanation of the concept meaning in English; and (3) a set of terms in a given language that have the meaning expressed by the concept. The concept part of the entry provides minimum information for formal definition of the concept. The English explanation of the concept meaning facilitates the human understanding. The set of terms stands for different wordings of the concept in the corresponding language. One of the terms is the representative for the term set. This representative term will be used where just one of terms from the set is necessary to be used, for example as an item of a menu. In the example above we present the set of Dutch terms for the concept <http://www.lt4el.eu/CSnCS#BarWithButtons>. One of the term is non-

lexicalized - attribute `type` with value `nonlex`. The first term is representative for the term set and it is marked-up with attribute `shead` with value 1.

3.2 Semantic annotation of learning objects.

After having mapped the lexicons to the ontology, we proceeded with the annotation of the Learning objects with the relevant parts of the ontology. The annotation was made within the textual content of the learning objects. This annotation allows us better search for learning content not only as whole learning objects, but also as parts of learning objects, for example paragraphs. The annotation was performed as regular grammars within CLaRK System⁷, called annotation grammars. Ambiguous cases are resolved by using rules implemented as constraints within the system. For the purposes of the project the disambiguation was done manually and in this way we construct a gold standard corpus of LOs. As it was mentioned above in the lexicon we represent not only terms for the concepts in the ontology, but also non-lexicalized phrases. This allows us to construct better annotation grammars for corresponding concepts.

4 Crosslingual search and integration in the learning management system

As we mentioned in sections 2 and 3 the ontology and the lexicons constitute the backbone for the crosslingual search engine which we are integrating in the learning management system. The engine is currently under development. In this section we will present the user scenarios which constitute the basis of the specification for the multilingual search, as well as the architecture for the integration in the eLearning system. The user will specify within the learning management system the languages in which he intends to retrieve documents. The annotation described in section 3 has to be performed for all documents to be searched.

Under these assumptions a search scenario from the user's point of view implies following steps:

- Submit query. User submits a free text query.
- See document list. A list of documents is displayed with meta information like: title, length, original language, keywords and concepts that are common to both the query and the document, other keywords and concepts that are related to the document but not to the query.
- See concept browsing units. Each concept that is assumed to be related to the search query, is presented to the user together with its neighbourhood from the ontology (related concepts and relations between the concepts) called "browsing unit". If no concept related to the search query is found, the root of the ontology with its neighbourhood is chosen as the browsing unit.

⁷ CLaRK System is an XML-based system for corpora development: <http://www.bultreebank.org/clark/index.html>.

- View documents. User views the documents from the list.
- Browse ontology. User browses the ontology: starting from the presented concepts, he navigates to related concepts, and concepts that are related to those.
- Select concepts. User selects ontology fragments (sets of related concepts, possibly only indirectly related) from the presented browsing units.
- Select search option. this step will be detailed in the following paragraphs
- See new document list. A new list of documents is displayed, based on only ontological search
- See updated concept browsing units. As in step 3, but now, those concepts are presented, that are common to at least N of the found documents; this includes the concepts that were used as the search key but might include further concepts. The number of documents, specified by the parameter N, can be relative (a percentage of the number of found documents) or absolute.
- Repeat steps from step 6 (Select concepts). User selects another set of related concepts and submits it as the search key, etc.

With respect to the search option, we are implementing following four strategies for combining ontology fragments

- fully disjunctive search: find documents, in which any of the concepts from any of the selected fragments occur
- disjunctive within fragments, conjunctive between different fragments: find documents, in which from each ontology fragment at least one concept occurs
- conjunctive within fragments, disjunctive between different fragments: find documents, in which at least one ontology fragment fully matches: every concept of the ontology fragment must occur in the document
- fully conjunctive search: find documents, in which all of the selected concepts from all of the ontology fragments occur

The steps 1 to 10 are currently formalized in functional modules composing the search engine.

The crosslingual search will be integrated together with the other component developed in the project (keyword extractor, definitory context finder) in the eLearning Management system ILIAS. The bases for the integration process are the use cases defined for the keyword extractor, the definitory context finder and the ontology enhanced searching and browsing capabilities. The use cases have been the major input for the specification of a web service interface between the language technology tools and the learning management system. It is a major goal of the project to make the language technology based functionalities re-usable for other learning management systems. To make the integration of the tools as easy as possible, the interface of the tools will be well-documented, standards-based and technology independent. The implementation of the interface as web services should ensure that these goals are met.

The major components of the integration setup are the language technology server and the learning management system. The language technology server provides the keyword extractor, definitory context finder and ontology management

system functionalities. The tools are developed using the Java programming language and are hosted on a Java web server. The functionalities can be accessed directly on the webserver for test purposes or they can be used by the learning management system through the web service interface.

The fact that multiple developers are working on different parts of the overall structure has led to the decision to setup a Subversion server as a central code repository. The project partners have also decided to make the results immediately available to the general public and to give everyone the opportunity to join and collaborate with the project. The source code is available under an open source license and it is hosted on the SourceForge.net portal for open source projects at <https://sourceforge.net/projects/lt4e1/>.

5 Conclusions and Further work

In this paper we described a possible enhancement of search facilities in an learning management system through ontological support. Particulary we focused on the multilingual character of the problem.

We are currently working on the implementation of the multilingual engine and validation of the multilingual scenarios, as described in Section 5. The extension from monolingual search to multilingual search triggers additional problems like:

- Ranking throughout languages: The user is maybe less interested in receiving a list of documents classified per language. Another possibility is to display the complete list according to the same ranking criteria, and for each document indicate its language. In this way, the user can compare the relevance of two documents even if they are in a different language. A further refinement could be to include the language as a ranking criterion by giving a bonus which differs per language; the bonus could still be overruled by the annotation frequency criterion.
- Parameters: A number of decisions do not have to be taken when implementing the search functionality, but should be known at runtime and therefore treated as parameters:
 - Possible languages of search query (in which lexicons should we look?)
 - Retrieval languages
 - Show concepts that are shared in at least N of the found documents
 - If less than N documents are found for a certain concept: try with superconcept and subconcepts
- Documents annotation with relations: In the current approach, only concepts and no relations are annotated in the documents. Relations between concepts are only known to the ontology and serve as a connection between concepts. They are used to find related concepts, automatically as well as for manual ontology navigation. An extension could be to annotate in the documents those relations between the annotated concepts that are valid for the document. Then, for example, a user can search for documents that contain the

concept ‘computer memory’ only if it is described as ‘part-of’ another concept. We are currently investigating the possibility of introducing relations corresponding to some pedagogical criteria, like the ACM-Ontology.

Within the learning management system we already integrated a keyword search engine (Lucene). The evaluation of the cross-lingual retrieval engine will be based on precision and recall measures comparing these 2 approaches. We are defining now also quality measures to prove that the ontological search improves also the learning performance of users. We are intending also to compare the ontological search with statistical based methods like inverted index or LSA

References

1. Christiane Fellbaum. 1998. Editor. WORDNET: an electronic lexical database. MIT Press.
2. Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003. The OntoWordNet Project: extension and axiomatisation of conceptual relations in WordNet. International Conference on Ontologies, Databases and Applications of Semantics (ODBASE 2003), Catania, Italy.
3. Aldo Gangemi, Nicola Guarino, Claudio Masolo and Alessandro Oltramari. 2001. Understanding top-level ontological distinctions. Proc. of IJCAI 2001 workshop on Ontologies and Information Sharing.
4. Nicola Guarino. 2000. Ontological Analysis and Ontology Design, An invited tutorial at the First Workshop on Ontologies and Lexical Knowledge Bases - OntoLex 2000, Sozopol, Bulgaria.
5. Nicola Guarino and Christopher Welty. 2002. "Evaluating Ontological Decisions with OntoClean." Communications of the ACM, 45(2): 61-65.
6. Lothar Lemnitzer and Lukasz Degórski. 2006. Language Technology for eLearning – Implementing a Keyword Extractor. Paper presented at the EDEN 2006 conference, Casteldelfels, Spain.
7. Eelco Mossel and Lothar Lemnitzer and Cristina Vertan. 2007. Language Technology for eLearning – A Multilingual Approach from the German Perspective. Proc. GLDV-2007 Spring Conference. Tübingen, April 2007, pp. 125-134
8. Lothar Lemnitzer and Cristina Vertan and Alex Killing and Kiril Simov and Diane Evans and Dan Cristea and Paola Monachesi. 2007. Improving the search for learning objects with keywords and ontologies. Paper to be presented at ECTEL 2007, Crete, September 2007.
9. Paola Monachesi and Dan Cristea and Diane Evans and Alex Killing and Lothar Lemnitzer and Kiril Simov and Cristina Vertan. Integrating Language Technology and Semantic Web techniques in eLearning. Presented at ICL 2006, September 27 - 29, 2006, Villach, Austria.
10. Daniel Oberle et al. 2006. DOLCE ergo SUMO: On Foundational and Domain Models in SWIntO (SmartWeb Integrated Ontology), Submission to Journal of Web Semantics.
11. OWL. Web Ontology Language (Overview). <http://www.w3.org/TR/owl-features/>
12. Salim Semy, Mary Pulvermacher, Leo Obrst. 2004. Toward the Use of an Upper Ontology for U.S. government and Military Domains: An Evaluation. MITRE Technical Report 04B0000063, September.

13. Vossen Piek (ed.), EuroWordNet General Document. Version 3, Final, July 19, 1999, <http://www.hum.uva.nl/ewn>
14. Christopher Welty and Nicola Guarino. 2001. Supporting Ontological Analysis of Taxonomic Relationships. Data and Knowledge Engineering.