Automatic Annotation of Content Units in TANGRAM

Jelena Jovanović
FON-School of Business Administration,
University of Belgrade
POB 52, Jove Ilića 154, Belgrade,
Serbia and Montenegro
jeljov@gmail.com

Dragan Gašević
School of Interactive Arts and Technology,
Simon Fraser University Surrey
102 Ave., Surrey,
BC V3T 5X3, Canada
dgasevic@sfu.ca

Vladan Devedžić
FON-School of Business Administration,
University of Belgrade
POB 52, Jove Ilića 154, Belgrade,
Serbia and Montenegro
devedzic@fon.bg.ac.yu

ABSTRACT

The paper presents an approach to automatic annotation of learning objects' (LOs) content units that we tested in TANGRAM, an integrated learning environment for the domain of Intelligent Information Systems. The paper mainly reports on the content-mining algorithms and heuristics applied for determining values of certain metadata elements used to annotate content units. Specifically, the focus is on the following elements: title, description, subject (based on a domain ontology), and pedagogical role (based on an ontology of instructional context).

Categories and Subject Descriptors

I.7.5 [**Document and Text Processing**]: Document Capture—Document Analysis; K.3 [**Computing Milieux**]: Computers and Education

General Terms

Algorithms, Design

Keywords

Semantic annotation, content mining, metadata, ontologies

1. INTRODUCTION

Learning content represented in the form of reusable learning objects (LOs) promised to significantly reduce the time and cost of authoring high-quality learning materials, making them more affordable and readily available. Annotations of LOs with the standard-compliant metadata (e.g. IEEE Learning Object Metadata) are seen as the primary mean for fostering LOs reusability. However, very often a content author needs to reuse just some specific parts of a LO, rather than the entire LO - for example, just a couple of slides out of a slide presentation, or an image or a table out of a text document. Automating reuse of LOs' individual components can reduce the efforts that content authors put in preparation of learning materials. However, an approach to such a kind of automation is still an open question. We believe that each content unit should be semantically annotated in order to be more easily searchable and thus reusable.

In this paper we present our approach to automatic annotation of LOs' components in TANGRAM – an integrated learning environment for the domain of Intelligent Information Systems (IIS). TANGRAM leverages the automatically generated annotations of LOs' components to build new content out of needs of individual learners. Although the annotation principles we discuss are implementation-independent, their implementation in TANGRAM helped us reveal important practical details we were not aware of initially.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

2. ONTOLOGICAL FOUNDATION

In this section we briefly present each of the ontologies our automatic annotating approach is based upon. Note also that we have defined a profile of the IEEE LOM RDF Binding which we use to describe each content unit (CU). Specific metadata fields of the profile refer to these ontologies (e.g. *dc:subject* field refers to a concept from the domain ontology).

In our previous collaborative research efforts with the ARIADNE research group from K.U. Leuven, Belgium, we developed ALOCoM ontology [2], as a content structure ontology based on the Abstract Learning Object Content Model (ALOCoM) [4]. The ontology defines concepts and relationships that enable formal definition of the structure of a LO. However, our latest research led to a major revision of the ALOCoM ontology and its division into: ALOCoM Content Structure (ALOCoMCS) ontology and ALOCoM Content Type (ALOCoMCT) ontology. Being based on the common model, these two ontologies share the same root concepts: Content Fragment (CF), Content Object (CO) and Learning Object (LO). However, these basic types of CUs are considered from completely different perspectives – ALOCoMCS is about content structuring, whereas ALOCoMCT focuses on potential instructional/pedagogical roles of CUs.

The domain ontology is defined using the SKOS Core ontology (http://www.w3.org/2004/02/skos/core/). Each concept of the domain is represented as an instance of the *skos:Concept* class, whereas the conceptual scheme of the domain is represented as an instance of the *skos:ConceptScheme* class. Each identified domain concept is assigned one or more aliases (i.e., alternative terms typically used in literature when referring to a concept) using the *skos:prefLabel*, *skos:altLabel*, and *skos:hiddenLabel* properties. The generalization hierarchy is represented via the *skos:broader* and its inverse *skos:narrower* properties, whereas the *skos:related* property is used for representing semantic relations between concepts belonging to different branches of the hierarchy. One should note that the domain ontology does not contain any information regarding topics sequencing, in terms of the order in which the topics should be presented to students. That kind of information is stored separately in the Learning Paths ontology (not presented here due to the limited size of the paper).

3. ANNOTATION OF CONTENT UNITS

Whereas the majority of metadata required for annotation of a LO is directly (manually) supplied by the content author (i.e. LOs are semi-automatically annotated), annotation of the LO's components is fully automated. Peculiarities of the automatic annotation approach we apply can be summarized as follows:

- The values of some metadata elements (*dc:creator*, *dcterms:created*, and *dc:language*) are literally copied from LOs to their components;
- Some metadata elements of the TANGRAM LOM RDF profile are meaningful only when attached to a LO as a whole. Therefore, they are not assigned to the components smaller than LOs (e.g. *lom-cls:accessibilityRestrictions* referring to the learning styles that a LO is particularly suitable for);
- The values of the other metadata elements are mined from a component itself, its content and presentational context. Due to the limited size of the paper, in what follows, we only briefly explain automatic generation of values for some of the metadata elements from this category. For more details see [1].

dc:title metadata element is not assigned to all kinds of CUs defined in ALOCoMCS ontology, since for a large number of them this property is not applicable (e.g. *alocomcs:Paragraph*, *alocomcs:Link*, etc.). On the other hand, a LO of the type *alocomcs:SlidePresentation* is an example of a CU that naturally has a title, hence we generate the value for its *dc:title* metadata out of the content of the slide presentations' first slide (known as the title slide). CUs of the type *alocomcs:SlideBody*, *alocomcs:Slide*, *alocomcs:Image* are additional examples of CUs that are automatically assigned *dc:title* metadata. By attaching *dc:title* to an image we try to compensate for its frequently missing caption (we have noticed that authors of slide presentations very rarely use captions to describe the semantics of the included images). Accordingly, we generate a textual value that reflects the semantics of an image and can serve as its caption, using an appropriate template (e.g. "Figure <ordinal_num>. illustrating <title_of_the_slide>").

dc:subject metadata element. To semantically annotate a CO with concept(s) from the domain ontology we apply the following approach: the domain ontology is queried for concepts that are semantically related to the domain concepts that were manually assigned to the CO's parent LO. We assumed domain concepts as semantically related if they are interconnected via *skos:narrower*, *skos:broader* or *skos:related* properties. The retrieved concepts and their aliases, i.e. labels assigned to them as values of *skos:prefLabel*, *skos:altLabel* i *skos:hiddenLabel* properties, are stored in a hashmap and serve as the basis for the subsequent steps of the annotation process. Subsequently each component of the CO containing text is searched for the aliases stored in the hashmap, and if some of them are found, the component (i.e. CO or CF) is annotated with the domain concepts that the aliases refer to. Afterwards, we apply a *bottom-up* approach to annotate the CO with a union of concepts assigned to its components. However, if no concept can be mined from the CO's content, the CO is annotated with concepts manually assigned to the parent LO.

For CFs that do not contain text at all, like CFs of the *alocomcs:Image* type, this approach is not applicable. Currently, in the absence of a better solution, such CFs directly inherit the value of the *dc:subject* metadata from the COs they are aggregated in.

alocom-meta:type metadata element is aimed at capturing the pedagogical role of a CU, making a reference to a concept from the ALOCoMCT ontology. It is used for annotating LOs and COs, but not for CFs, as according to the ALOCoM model [4] an instructional role can not be assigned to a single CF.

Due to the lack of well defined formats for representing learning content of a certain instructional role (e.g. an explicit format for representing definitions), we opted for a heuristics-based approach to infer instructional role of CUs. The heuristics that we use are partially founded on our previous joint research efforts done with the ARIADNE group from K.U. Leuven, Belgium. Using the experience discussed in [3], we did some initial research aimed at defining patterns for recognizing CUs having instructional role of *alocomct:Definition*, *alocomct:Example* and *alocomct:Reference*. Besides this pattern-based approach, we apply some simple heuristics to determine the instructional role of slides (i.e. COs of type *alocomcs:Slide*). For example, if the content of the slide's title is one of the following terms/phrases: "Bibliography", "References", "Reference list", while the content of the slide's body is structured as a list, the instructional role of the slide is presumed to be of type *alocomct:Bibliography*. Additionally, each list item appearing in the slide's body is assumed to be of *alocomct:Reference* instructional type.

dc:description metadata element is generated out of the (known) values of other metadata elements and using predefined templates (one for LOs and the other for COs). For example, the following template is used to generate a description of a LO: "A <alocom-meta:type> with title: '<dc:title>' authored by <dc:creator>; creation date <dcterms:created>; evaluated by the author as being of <lom-edu:difficulty> difficulty level and treating issues of {<dc:subject>}".The metadata elements appearing in the angled brackets are replaced by their actual values. Curly brackets indicate that the enclosed element can have multiple values.

4. CONCLUSIONS

The general principles of ontology-based annotation of LOs' components identified in the paper are implemented in TANGRAM, our learning environment for the domain of Intelligent Information Systems. TANGRAM applies a highly structured approach to annotation and implements a number of semantic annotation heuristics, some of which are discussed in the paper. The initial evaluation of the TANGRAM's annotation subsystem, although limited in scope, helped us identify strengths and weaknesses of the current solution. A brief description and demonstration of TANGRAM as well as the ontologies referred to in the paper can be found at http://iis.fon.bg.ac.yu/TANGRAM/home.html.

Our future work will be directed towards improving existing functionalities of the TANGRAM's annotation subsystem and augmenting it with additional ones required for recognition of pedagogical roles not included in the current solution. Specifically, we intend to empower TANGRAM with advanced features of the latest frameworks for natural language processing and information extraction, such as GATE (http://gate.ac.uk) and KIM (http://ontotext.com/kim).

5. REFERENCES

[1] Jovanović, J., Gašević, D., Devedžić, V., "Ontology-based Automatic Annotation of Learning Content," *Int'l J. on Sem. Web and Inf. Sys.*, Vol. 2, No. 2, 2006 (forthcoming).

[2] Jovanović, J. et al, "Ontology of learning object content structure," *In Proc. of the 12th Int. AI-ED Conf.*, Amsterdam, The Netherlands, 2005, pp. 322-329.

[3] Liu, B., Chin, C.W. and Ng, H.T., "Mining Topic-Specific Concepts and Definitions on the Web," *In Proc. of the 12th Int. WWW Conference*, Budapest, Hungary, 2003, pp. 251-260.

[4] Verbert, K. et al, "Towards a Global Component Architecture for Learning Objects: an Ontology Based Approach," *In Proc, of the OTM 2004 Workshop on Ontologies, Semantics and E-learning,* Agia Napa, Cyprus, 2004, pp. 713-722.