ISSN:2320-0790

An International Journal of Advanced Computer Technology

Automatic Caption Generation for Electronics Textbooks

Veena Thakur[1], Trupti Gedam[2]

[1,2] Department of Computer Engineering, RMD Sinhgad School of Engineering, Pune, Maharashtra, India.

Abstract:
Automatic or semi-automatic approaches for developing Technology Supported Learning Systems (TSLS) are required to lighten their development cost. The main objective of this paper is to automate the generation of a caption module; it aims at reproducing the way teachers prepare their lessons and the learning material they will use throughout the course. Teachers tend to choose one or more textbooks that cover the contents of their subjects, determine the topics to be addressed, and identify the parts of the textbooks which may be helpful for the students it describes the entities, attributes, role and their relationship plus the constraints that govern the problem domain. The caption model is created in order to represent the vocabulary and key concepts of the problem domain. The caption model also identifies the relationships among all the entities within the scope of the problem domain, and commonly identifies their attributes. It defines a vocabulary and is helpful as a communication tool. DOM-Sortze, a framework that enables the semi-automatic generation of the Caption Module for technology supported learning system (TSLS) from electronic textbooks. The semiautomatic generation of the Caption Module entails the identification and elicitation of knowledge from the documents to which end Natural Language Processing (NLP) techniques are combined with ontologies and heuristic reasoning.

Keywords: Caption Module, DOM-Sortze, LDO, LO

1. INTRODUCTION

Advances in the last few years have greatly increased influence of new technologies in Information and Communication Technologies (ICT) in particular. On-line applications have become essential, they are continuously used for communication (e.g., instant messaging, mailing, phoning), consulting bank accounts, and so on. This revolution has also affected education, providing means than enhance both teaching and learning. Years of research have facilitated the development of different kinds of Technology Supported Learning Systems (TSLSs) such as Learning Management Systems (LMSs), Intelligent Tutoring Systems (ITSs), Collaborative Learning Systems or Adaptive and Intelligent Web-based Educational Systems [1].

The caption module is created in order to represent the vocabulary and key concepts of the problem domain. The caption module also identifies the relationships among all the entities within the scope of the problem domain, and commonly identifies their attributes. A caption module that encapsulates methods within the entities is more properly associated with object oriented models. The caption module provides a structural view of the domain that can be complemented by other dynamic views, such as use case models. The caption module uses the DOM-Sortze framework to test with an electronic textbook and the gathered knowledge has been compared with the Caption Module that instructional designers developed manually.

DOM-Sortze depends on ontologies for semi-automatic construction and heuristic reasoning which means thinking, understanding and decision-making take place in the real world, where there are usually time pressures and rarely a full range of information available to support a complete appraisal of the problem at hand. For instance, suppose you are buying a new washing machine. A good basis for the decision might include comparative data on reliability, ease of servicing, servicing and repair costs, ease of use, even noise levels during operation. The list could go on and on. Although sometimes data of this sort might be available, and sometimes it might be published in magazines, it is more likely that you will have to cut corners. In other words, you might not be able to obtain a machine that fulfils all of your desirable features, but you will instead settle for the closest that is available. Kahneman, Slovic and Tversky popularized the term heuristic reasoning for thinking and decision making that involves these types of short cuts [3]. They also suggested that these short cuts are so common that they should be considered part of the machinery of thought itself.

Ontology is an effective formal representation of knowledge used commonly in artificial intelligence, semantic web, software engineering, and information retrieval. In open and distance learning, ontologies are used as knowledge bases for e-learning supplements, educational recommenders, and question answering systems that support students with much needed resources. In such systems, ontology construction is one of the most important phases. Since there are abundant documents on the Internet, useful

learning materials can be acquired openly with the use of ontology.

1.1 Learning Object Metadata (LOM)

The capability of searching, evaluating and acquiring LOs is essential for the use (and the reuse) of LO's. Metadata, i.e., descriptive data about educational data and resources, allows cataloging the LOs so that they can be searched and retrieved [4]. The LOM defines the syntax and semantics of metadata for LOs. These include properties to describe general information about the resource, terms of distribution, technical requirements, teaching or interaction style, etc.

1.2 Learning Domain Ontology

Ontology learning is the automatic or semi-automatic creation of ontologies, including extracting the corresponding domain's terms and the relationships between those concepts from a corpus of natural language text, and encoding them with an ontology language for easy retrieval. As building ontologies manually is extremely labor-intensive and time consuming, there is great motivation to automate the process.

1.3 Learning Objects

Learning Objects (LOs) are considered the core notion for learning content. A learning object is a collection of content items, practice items, and assessment items that are combined based on a single learning objective the notes teachers use for their classes may be considered LOs since they can be referenced during the learning process, even though its reusability in an application is quite limited. Wiley (2000) instead recommends considering LOs as any digital resource that can be reused to support learning [5].

2. LITERATURE SURVEY

2.1 Semi Automatic Caption Module Generation Process

Gathering the caption knowledge of a TSLS from already existing documents in a semiautomatic way may considerably reduce the development cost. Artificial Intelligence methods and techniques such as Natural Language Processing (NLP) and heuristic reasoning can be applied to achieve the semi automatic generation of the Caption Module [6]. Following steps are carried out to generation of generation module.

Document preprocessing *:* First, the document must be prepared for the subsequent knowledge acquisition processes. The document preprocessing and the outcomes are then used to gather the two levels of knowledge encoded in the Caption Module. The outline of the document is suitable for the construction of the Learning Domain Ontology (LDO), while the content of the document is useful for both building the LDO and generating Learning Objects (LO's).

The Caption Module encodes knowledge at two different levels: (1) the knowledge to be learnt, including the topics and the pedagogical relationships that enable planning and determining the learning sessions, which is described by the Learning Domain Ontology (LDO) and (2) the set of LOs that will be used for each domain topic. Using ontology to describe the learning topics and the pedagogical relationships among the topics will facilitate reusing the described. The fig. 1 shows steps to be are carried out to develop the Caption Module.

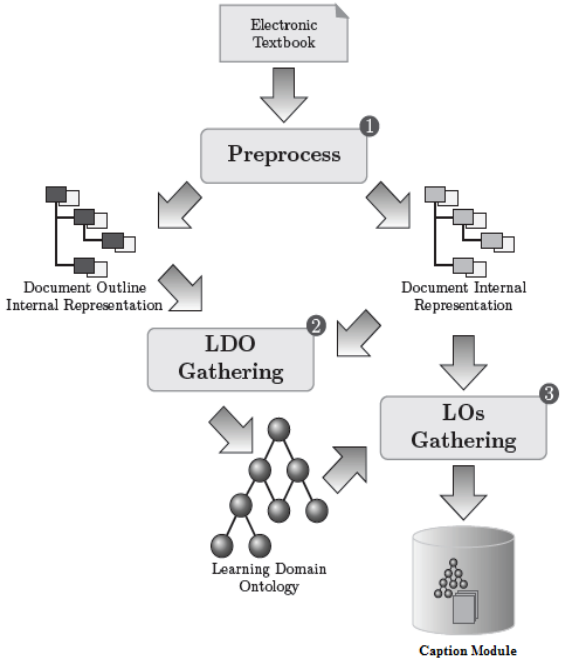

Fig. 1. Caption Module Building Process

Gathering the LDO*:* At this phase, the domain topics to be mastered, along with the pedagogical relationships among them, are identified and described in the LDO. The LDO ontology can be used in different ways for learning. On the one hand, instructive TSLSs would use this information to plan the learning sessions. On the other hand, the students can rely on the LDO to guide them during the learning process.

Gathering the LO's: At this stage the LOs - definitions, examples, exercises, etc. - to be used during the learning process are identified and generated.

2.2 Gathering the learning domain ontology

The Caption Module is described by means of the Learning Domain Ontology (LDO), which contains the main domain topics and the pedagogical relationships among them. Pedagogical relationships can be structural - isA and partOf - or sequential - prerequisite and next.

Outline Analysis: The outline analysis process consists of two phases:

- **Basic Analysis***:* In this task the main topics of the domain and the relationships among these topics are mined from the homogenized outline internal representation. In this approach, each index item is

considered as a domain topic. Besides, the structure of the document outline is used as a means to gather pedagogical relationships. A sub item of a general topic is used to explain part of it or a particular case of it. Therefore, structural relationships are defined between every outline item and all its sub items. In addition, the order of the outline items reflects the recommended sequence for learning the domain topics. Thus, an initial set of sequential relationships is identified from the order of the outline items.

- **Heuristic Analysis *:*** The results of the basic analysis are refined based on a set of heuristics that categories the relationships identified in the previous step and mine new ones, mainly prerequisite relationships [3]. The identified relationships are labeled with the inferred kind, the heuristic used, and the confidence in the inferred information. The heuristics entail the condition to be matched, and the post-condition, i.e., the relationships that are recognized.

Whole Document Analysis*:*At this stage, the initial LDO is enhanced with new topics and relationships gathered from the whole document. In order to achieve this goal, two processes are carried out: first, new topics are identified as and later new pedagogical relationships among the topics are identified.

- **Identifying New Top**ics*: This process aims at enhancing the LDO gathered in the previous phase with new domain topics. The whole document is analyzed to get such new topics. In the last few years, the use of hybrid methods that combine NLP techniques and statistic methods has prevailed in term extraction.*

  Many approaches use a set of patterns such as ((A|N) + |((A|N) (NP)?(A|N)*)N to get the set of candidate terms, where A is an adjective, N is a noun, and P is a preposition, and then apply some term hood measures to rank the set of candidate terms and filter the most appropriate ones [7].*

- **Identifying New Relationships *:*** This process allows the identification of new pedagogical relationships from the electronic document using a pattern-based approach. These patterns recognize pedagogical relationships between domain topics based on the syntactic structures found in the sentences in which the topics appear. Therefore, the internal representation of the document is first annotated to label any domain topic appearance. Then, nested domain topics, i.e., domain topic constructed on other domain topics are identified to propose isA relationships among them. For example, "Sirius star" contains the topic "star". Thus, it can be inferred that "Sirius star" is a "star". Finally, the document is given a grammar-driven analysis to identify a set of sentences which relate two or more domain topics. The grammar contains a set of rules describing syntactic structures corresponding to pedagogical relationships. The Constraint Grammar formalism, one of the most successful syntax analyzing and disambiguation systems, has been used to develop and apply the grammar on the documents [8].

Gathering learning objects from electronic textbooks

The generation of LOs from the electronic textbooks entails identifying and extracting the relevant DRs i.e., fragments of the document related to one or more topics with a particular educational purpose, their annotation with LOM and storage in the LOR. The gathered LOs are mainly text-based, although they may also contain some of the images used to illustrate the domain topics in the textbook. Los are gathered from the electronic textbook by carrying out the process described in below figure.

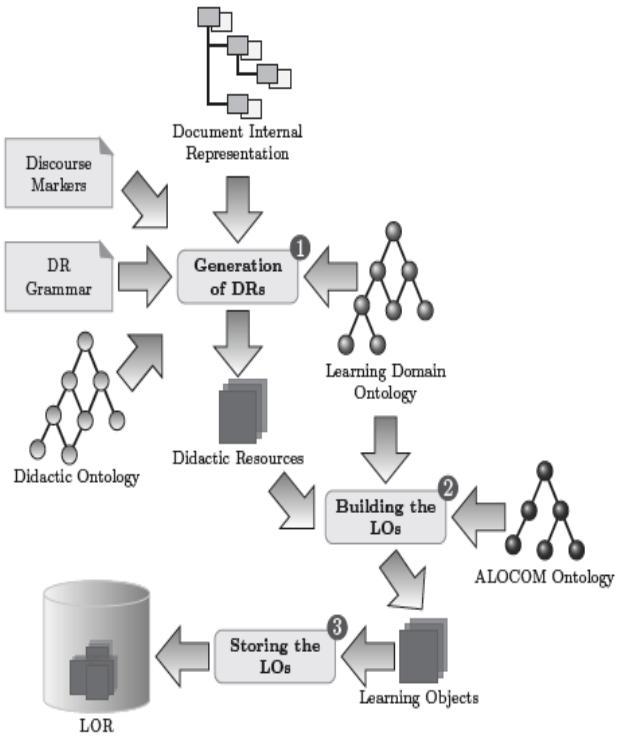

Fig. 2 Generation of Learning Objects from Textbooks

- **Generation of Didactic Resources (DR):**The DR grammar allows locating the sentences that use any of the identified syntactic structures referring to LDO topics. Text-based DRs are built from the sentences selected by the grammar. For each selected sentence, an atomic DR is built. The atomic DRs also contain the sentences that follow the selected one as long as they are not identified as other DR by the grammar, and they are content related, i.e., they are considered similar from the content perspective [9]. Content similarity is measured considering the domain topics referred in the text. Textbook authors may also include some sentences that do not necessarily include the domain topics but that connect different sentences that do refer to domain topics. An empirically

gathered number of consecutive sentences of this kind are also allowed while building the atomic DRs, with the aim of being as complete and coherent as possible. Besides, every image found in the textbook is also considered a DR that requires no deeper processing.

- **From didactic resources to learning objects:** The gathered DRs might be not only useful for the Domain Module being developed from the processed electronic textbook, but also for other Domain Modules. Thus, to facilitate their reuse, LOs, i.e. reusable digital educational resources, are built from the gathered DRs. Building reusable DRs entails two aspects: using an appropriate format to store and represent the content, and also describing it (annotating it) with LOM to allow searching in and retrieving those LOs from the LOR.

  The presentation format of the LO may also affect its reusability. Presentation formats such as html, pdf, doc, and odf are suitable for final presentation, but are not appropriate for flexible content reuse, as the components cannot be easily accessed. The generated DRs are stored in a ZIP file that contains the XML file for the LO, the referenced images or other Resources.

  After an analysis of the LOM elements, and considering the kind of documents being processed, these elements were classified and it was concluded that only some of them differed from one of the gathered LOs to another, while most of them had similar values. The initial metadata elements are automatically generated from the electronic textbook [10], using an automatic metadata generator. Then, the metadata is enhanced with more information that has been extracted during the DR generation to improve some elements (keywords or Learning Resource Type). Most keyword annotation applications use statistical methods and rely on the frequency of the terms in the analyzed text, but do not consider semantic relationships among the topics.

- **Learning Object Storage:** To make the LOs available, they are stored in LORs. Two repositories are employed, one for the LOs (both resources and metadata) and another one for keeping the preview files for the LOs. Once the LOs and their preview files have been generated, they are pushed to the LOR to allow their retrieval and use in new TSLSs [11]. The LO publishing service is based on the SPI specification Once the LOs and their preview files have been generated, they are pushed to the LOR to allow their retrieval and use in new TSLSs. The LO publishing service is based on the SPI specification When the LO is composed, its components are also annotated with LOM and stored in the LOR, as they might be useful in certain contexts [12].

3. DOM-SORTZE ARCHITECTURE

DOM-Sortze is a suite of applications and web-services that cope with different tasks of the Caption Module generating process. Below Figure shows the architecture of DOM-Sortze. The rounded boxes represent web services, while the applications or modules are represented by rectangular boxes. This web-service oriented approach makes DOM-Sortze flexible and platform-independent. Although it uses some platform-specific applications (mainly NLP tools), these are used only by the web-services. Therefore, the client side applications are platform independent.

DOM-Sortze entails four main applications – the Preprocessor and LDO Builder – that carry out the tasks for building the Domain Module. The first carry out the textbook processing task and the latter facilitates the intervention of the Caption Module authors, either instructional designers or teachers, to supervise the results. These applications take advantage of some web-services to perform their job [13].

The preview files might be helpful for the users to determine whether or not the LO fits their requests.

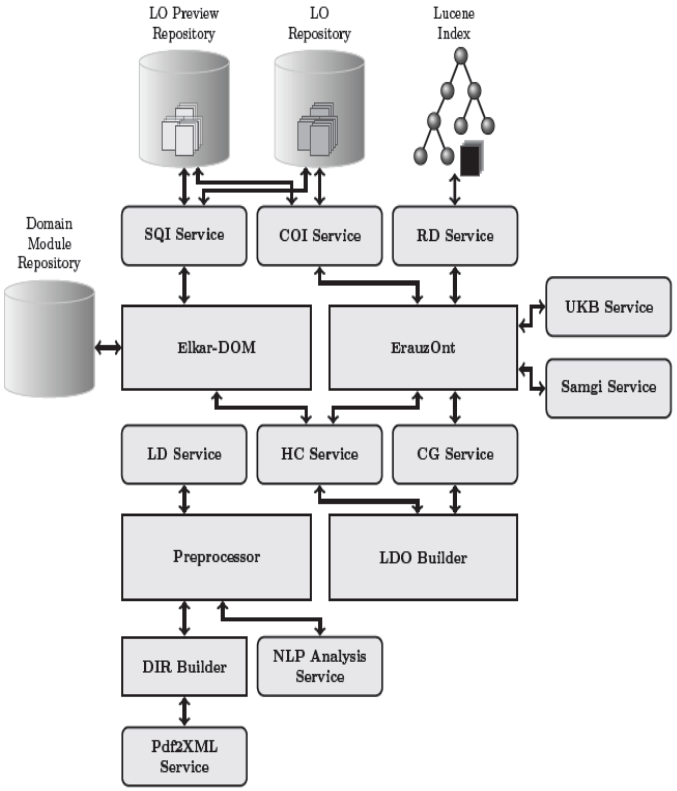

Fig. 3 Architecture of DOM Sortze

3.1 Preprocessor

Documents are available in different electronic formats (e.g., pdf, doc, html, etc.), although usually all the documents are structured in a hierarchical structure (chapters, section.). Besides, some documents have their outline at the beginning

of the document while others place it at the end. The Preprocessor is responsible for carrying out the preprocess of the document [13]. It relies on the DIR Builder module to gather the internal representation of both the textbook to be analyzed and its outline. The DIR Builder module allows the Preprocessor to perform its work independently of the format of the document. It currently supports pdf documents to which end it delegates on the Pdf2XML Service to build the internal representations of the document and its outline.

3.2 LDO Builder

The LDO is gathered by the LDO Builder from the internal representations of the electronic textbook. The topics of the LDO are gathered from the outline of the textbook and from the whole document. The identification of the topics from the whole document is conducted [14]. The identification of the pedagogical relationships is also achieved following a pattern-recognition approach. The LDO uses an XML-base formalism to describe the gathered LDO. Listing shows a fragment of an XML file that describes a LDO fragment of a LDO in which some topics and a relationship are described. As can be observed, the information about the heuristic used and the confidence on that heuristic are also included to facilitate the supervision process depicted later. The formalism for describing the LDO also supports the description of the kind of topic, the relevance of the topic, and the difficulty level, although these features are not currently elicited from the textbooks.

4. PROPOSED WORK

Future work on DOM-Sortze comprises improving the generation of the LDO. It is planned to enhance the grammar for identifying pedagogical relationships to increase the recall of the relationships. Alternative ways to gather prerequisite relationships, which have a very poor recall, will be also tested. Besides, attributes of the domain topics, such as the domain relevance or the difficulty, which might be estimated using term hood measures are aimed to be automatically gathered.

5. CONCLUSION

The generation of caption Module using DOM-Sortze entails several tasks:

Preparing the textbook for the process, gathering the LDO which is carried out by analyzing both the outline and the whole textbook, and the generation of LOs from the textbook. DOM-Sortze a system for the semiautomatic generation of the Domain Module from electronic textbooks. DOM-Sortze was developed incrementally, coping with the acquisition of the LDO from the outline of the textbooks first, later the extraction of LOs from the documents and, finally, the whole Domain Module construction process. At each stage of the development of DOM-Sortze, an evaluation was carried out to verify the correct performance and to measure how much it can help the authors when developing new Domain Modules [14]. DOM-Sortze has been tested using an electronic textbook and comparing the automatically generated elements with the Domain Module manually developed by instructional designers Every conducted evaluation has been performed using the Gold Standard approach using the LDOs and sets of DRs defined by instructional designers as reference.

In DOM-Sortze, the Domain Module entails an LDO, which contains the main domain topics and the pedagogical relationships among the topics, and the learning objects (LOs) that are used to enable mastering each domain topic.

REFERENCES

[1] B. Parsad and L. Lewis, "Distance Education at Degree-Granting Postsecondary Institutions: 2006-07," technical report, Nat'lCenter for Education Statistics, Inst. of Education Sciences, USDepartment of Education, 2008.

[2] Maritxalar, M. (2010). "Automatic Distractor Generation for Domain Specific Texts". In H. Loftsson, E. Rögnvaldsson, and S. Helgadóttir (Eds.),ICETAL-2010, 6233, pp. 27–38. Springer, Reykjavik, Iceland.

[3] Kahneman, Slovic and Tversky "Amos Tversky (eds.), Judgment Under Uncertainty:Heuristics and Biases" in Journal of Forecasting, Volume 3, Issue 2, April 1984, pages 235-239.

[4] C., McCalla, G.I., and Winter, M. (2005). "Flexible Learning Object Metadata". In L. Aroyo and D. Dicheva (Eds.), SW-EL'05: InternationalWorkshop on Applications of the Semantic Web for E-LearningTechnologies at the 12th International Conference on Artificial Intelligencein Education, AIED 2005, pp. 1–8. SW-EL, Amsterdam, TheNetherlands.

[5] Cardinaels, K. (2007). "A Dynamic Learning Object Life Cycle and its Implications forAutomatic Metadata Generation".Ph.D. thesis, FaculteitIngenieurswetenschappen,KatholiekeUniversiteit Leuven.

[6] Cardinaels, K., Meire, M., and Duval, E. (2005). "Automating Metadata Generation:the Simple Indexing Interface".In A. Ellis and T. Hagino (Eds.), Proceedings of the 14th International Conference on World Wide Web, WWW 2005.
ACM, Chiba, Japan.

[7] J.S. Justeson and S.M. Katz, "Technical Terminology: SomeLinguistic Properties and an Algorithm for Identification of Termsin Text," Natural Language Eng., vol. 1, no. 1, pp. 9-27, 1995.

[8] F.Karlsson, A.Voutilainen, and J. Heikkila "Constraint Grammar: Language-Independent System for ParsingUnrestricted Text," Natural Language Processing, , eds., no. 4, Mouton de Gruyter, 1995.

[9] Downes, S. (2001). "Learning Objects: Resources for Distance Education Worldwide" . International Review of Research in Open and Distance Learning,vol. 2(1).

[10] Meire, M., Ochoa, X., and Duval, E. (2007). "SAmgI: Automatic Metadata Generation v2.0". In C. Montgomerie and J. Seale (Eds.), Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications2007, ED-MEDIA 2007, pp. 1195–1204. AACE, Vancouver, Canada.

[11] S. Ternier, D. Massart, F.V. Assche, N. Smith, B. Simon, and E. Duval, "A Simple Publishing Interface for Learning ObjectRepositories," Proc. World Conf. Educational Multimedia, Hypermedia,and Telecomm. (ED-MEDIA '08), pp. 1840-1845, 2008.

[12] B. Simon, D. Massart, F.V. Assche, S. Ternier, E. Duval, S.Brantner, D. Olmedilla, and Z. Miklo´ s, "A Simple Query Interfacefor Interoperable Learning Repositories," Proc. 14th Int'l Conf.World Wide Web (WWW '05), pp. 11-18, 2005.

[13] S. Ternier, D. Massart, F.V. Assche, N. Smith, B. Simon, and E. Duval, "A Simple Publishing Interface for Learning ObjectRepositories," Proc. World Conf. Educational Multimedia, Hypermedia,and Telecomm. (ED-MEDIA '08), pp. 1840-1845, 2008.

[14] K.S. Jones, "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," J. Documentation, vol. 60, no. 5, pp. 11-21, 1972.