TOWARDS SEMI-AUTOMATIC GENERATION OF INTELLIGENT TUTORING SYSTEMS

M. LARRAÑAGA, J. A. ELORRIAGA, U. RUEDA, A. ARRUARTE,
I. FERNÁNDEZ DE CASTRO AND A. DÍAZ DE ILARRAZA

Department of Languages and Information Systems
University of the Basque Country 649 P.K., E-20080 Donostia, Basque Country
E-mail: jiblaolm@si.ehu.es

In this paper, a proposal for automating the acquisition of the domain for the Teaching-Learning Systems is presented. Nowadays, in the Information Age, a lot of information of any domain can be easily found in different electronic formats: handbooks, technical reports, web pages, etcetera. All this information can be used in order to build the desired Teaching-Learning System. Natural Language Processing (NLP) techniques will be used to gather from books and other electronic material the Domain Module. Domain topics and didactic resources can be gathered using both ontologies and NLP techniques. Didactic resources must be annotated and stored in order to facilitate knowledge reuse and thus, Teaching-Learning Systems building process.

1 Introduction

Although Teaching Learning Systems (TLS) have been successfully used in several areas many domains, such as Machine Tool, still remain unexplored. Machine Tool domain seems to be a promising area for the application of TLSs. On the one hand, resources (machinery) are very expensive and rare. On the other hand, accidents are quite usual and serious. Efficient instruction can be performed with low cost by using TLSs.

In addition, this domain presents several problems: learners have not enough motivation, mainly because most of them come from other studies on which they did not succeed. Existing teaching material, e.g. a handbook, is neither pedagogical nor motivating. Besides, the domain is in a continuous evolution, so new documents appear frequently. TLSs could overcome both problems by using rich interfaces, e.g. virtual reality or multimedia, and by updating their Domain Module according to new material.

IRIS [2] is an Authoring Tool for building Teaching-Learning Systems that has been developed with the aim of reducing human instructor's workload. The TLSs generated by IRIS have the classic architecture, which includes: Domain Module, Student model, Didactic Module and Interface. IRIS is based on a cognitive theory of instruction, the CLAI model [1]. This model combines human learning cognitive processes, and learning strategies together with aspects from teaching processes. TLSs built with IRIS use the following four elements to describe the domain:

- Teaching-learning contents are called Basic Learning Units (BLUs). Based on Merrill's Component Display Theory [11], IRIS uses four kinds of BLUs: *concepts, procedures, principles* and *facts*.
- Relationships among contents (Rs). Structural relationships – *is-a* and *part-of* – and pedagogic relationships – *prerequisite*, *corequisite* and *next* – are used to establish the sequence of BLUs in the learning process.
- Instructional Objectives (IOs) specify the skills to be reached on a particular BLU. The defaults IOs in IRIS are those of the widely accepted Bloom's taxonomy [5]: knowledge, comprehension, application, analysis, synthesis and evaluation.
- Didactic Resources (DRs): Any kind of IO needs a set of *presentation resources* or techniques which can be used to transmit the domain BLUs to the learner – *definitions, examples, analogies, etc.*– and *evaluation resources* or techniques for assessing domain contents – *tests, fill gaps, item sorting, etc.*–.

IRIS uses IKAT [8], an incremental knowledge acquisition tool, to build the domain module. IKAT allows the user to built ontologies incrementally by generating versions he/she can modify. IKAT can work either as an autonomous acquisition tool or integrated in another application such as IRIS.

Despite the benefits provided by authoring tools, human instructors still encounter great troubles building TLSs due to the amount of data needed and the difficulty of specifying the system requirements.

Nowadays, in the Information Age, a lot of information of any domain can be easily found in different electronic formats: handbooks, technical reports, web pages, etcetera. All this information can be used in order to build the desired TLS. In this context, the Domain Module can be obtained in a semi-automatic way from existing electronic material.

2 Procedure for acquiring the Domain Knowledge

The process starts with the analysis, by using NLP techniques, of a document, e.g. a handbook, getting the relevant topics of the domain. Once the relevant topics have been identified, these ones can be used to analyse the document again in order to generate the corresponding didactic resources. Finally, new documents can be analysed in order to enhance the Domain Module of the TLS.

Recently, Mizoguchi and several other authors [13] [7] have proposed the use of ontologies in different modules of the TLSs to profit from the advantages they offer (pedagogic domain reuse, didactic module standardization and so on). High Level ontologies, e.g.: WordNet [12], GUM [4], have successfully been used for Natural Language Processing (NLP) mainly because they offer a basic vocabulary for the analysis of texts. Thus, the use of ontologies and a set of NLP tools can facilitate the acquisition of the teaching Domain Ontology in a semi-automatic way. In this way, the TLS generating process of IRIS will be facilitated. This process is divided into three phases: Domain Module Structure acquisition, Generation of didactic material and Domain Module Maintenance.

2.1 First Phase: Domain Module Structure Acquisition

In this phase the structure of the Domain Module will be gathered. First, the documents will be automatically annotated in order to facilitate the analysis by using NLP techniques. As result of this phase, are obtained both: the Domain Module Structure and the Domain Ontology. The Domain Module Structure contains the domain topics and relationships in terms of IRIS elements and will be used in the TLSs that IRIS produces. The Domain Ontology contains a more general representation of the Domain which contains synonyms, acronyms and so on. The Domain Ontology will be used for didactic resource generation in the second phase and for the domain maintenance. Both, the Domain Module Structure and the Domain Ontology are incrementally built in the following steps: index analysis, whole document analysis, topic classification, statistical analysis and supervision:

- In the first step, the main domain topics, as well as the sequential and structural relationships among them, are identified by analysing document indexes using NLP techniques. The process starts analysing document indexes because they are usually well structured and contain the main topics of the domain. The identified topics will be traduced into BLUs of the Domain Module. Besides, the structure of the index can aid to infer the existence of structural relationships (e.g.: part-of, is-a) among the topics. Sequence among index items can be considered the starting point to set the pedagogic order among BLUs of the domain (next relationship). Figure 1 illustrates this step.
- On the second step, the analysis will be extended to the whole document. The main goal of this step is to obtain second level topics that have not been previously identified in the index. Some experiments with high-level ontologies and existing domain ontologies will be performed. The goal of these experiments is to know if ontologies can provide a better outcome or refine the results of the previous step. Ontologies will be used to look for unidentified domain topics. On the one hand, the method to determine the domain specific relevance of WordNet synsets (*Synonym Sets*) can be used [6]. In this work, occurrences for each term of the synset are quantified in order to get the domain relevance of the synset. The system must find the synsets corresponding to the topics identified in the first step. Once the relevant synsets have been identified, other synsets related to these ones will be checked in order to obtain unidentified topics. On the other hand, other ontologies such as Wordnet Domain [10], an extension of WordNet in which synsets have been clustered by means of domain labels, can be used to

complete the Domain Module Structure. New topics can be identified by looking for the synsets labelled with the current ontology in Wordnet Domain.

- The goal of the Topic Classification is to decide the BLU category of the identified topics. Part-of-speech and morphological information will be used for this purpose. For example, nouns or noun phrases will contain *Concept* BLUs while verbs will identify *Procedure* BLUs.
- On the statistical analysis, the occurrences of the identified topics in the text will be analysed for discovering new relationships between them. These relationships can be inferred from the co-occurrences of the terms. If some terms are frequently found close in the document it may be a possible relationship between those topics. This way other relationships such as *prerequisite* can be identified. The structural relationships that have been inferred in the first step can also be refined into the corresponding IRIS structural relationships (*is-a* or *part-of*). Besides, term occurrences may help to numerically assess the relevance of the topics and the strength of the relationships.
- In the last step human instructors must check the domain, which has been automatically built. In order to provide an intuitive interface, the Domain Module (Domain Module Structure and the Domain Ontology) will be presented to the human instructors by means of concept maps. CM-ED [3], a generic concept map editor, will be used to visualise these concept map. Human instructors can review and adapt the Domain Module and the changes will be automatically propagated to both the Domain Module Structure and the Domain Ontology.

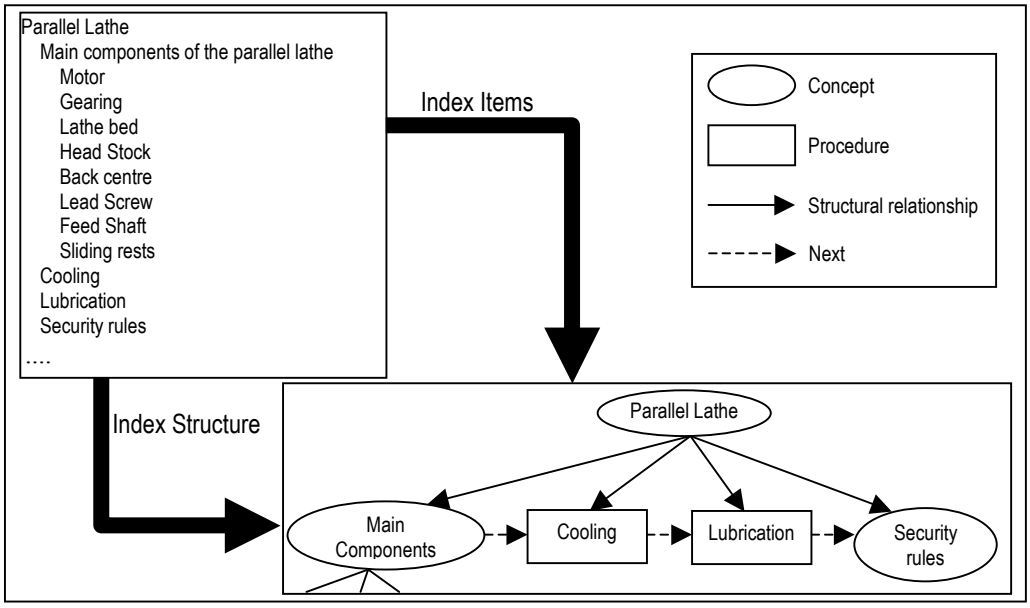

Figure 1: Acquisition of the main topics of the domain from the index of the document

2.2 Second Phase: Generation of didactic material

Once the Domain Module Structure is finished, the whole document will be analysed in order to identify fragments of texts with different learning goals such as definitions analogies and so on. These fragments (didactic resources) will be identified, attached to the BLUs related and stored in a Didactic Resource Repository for further use in new Teaching-Learning systems. Four steps can be identified in the process of generation of didactic material: fragment identification, didactic resource identification, supervision and didactic resource storing.

- Fragment identification: In this step a representation of the structure of the document will be obtained, i.e. sentences, paragraphs, chapters and so on. Besides, the domain ontology will be used to associate the fragments corresponding to every BLU. Occurrences of the topics identified in the first phase will be analysed and stored for the next step.

- Didactic Resource Identification: Once the fragments and the structure of the document have been identified, the didactic resources must be gathered. A pedagogic ontology (an ontology that contains pedagogic issues) and a library of patterns will be used to identify the kind of didactic resource, learning goal and so on. These patterns represent the different ways in which the didactic resources may appear in a document. In order to improve this pattern-recognition process, a high-level ontology such as Wordnet will be used to look for the synonyms of the keywords of the pattern. Obviously not all the didactic resources have the same size. A definition can be done in a small paragraph but a set of paragraphs may be used to present an analogy or a general explanation. Thus, some heuristics will be used to decide the scope of every didactic resource. Once the didactic resource identification has finished, the different presentation and evaluation forms will be added to the above-mentioned Domain Module.
- Supervision: The results will be presented to the user in order to be verified and validated. Again, concept maps will be used to present to the human instructor the state of the Domain Module (Domain Module Structure and Didactic Resources) so that he/she can supervise and adapt it.
- Didactic resource storing: Another important issue for enhancing the process of building TLSs is reusing existing didactic resources. It has been pointed out the need of annotating the pedagogical material using metadata in order to facilitate TLS reusability [14]. Therefore, didactic resources in IRIS will be annotated with the information obtained in the previous steps following a Learning Objects standard (e.g.: LTSC/LOM [9]) and added to a Didactic Resource Repository (see Figure 2). This Didactic Resource Repository (DRR) will aid the human instructors to add learning material to the TLSs they are building.

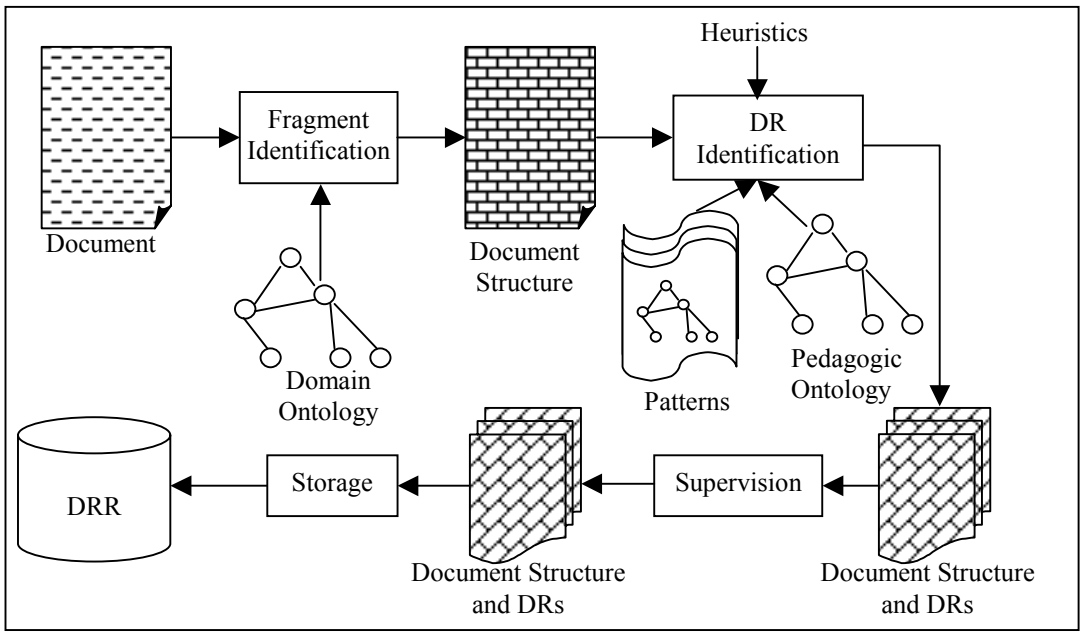

Figure 2: Didactic resource generation process

2.3 Third Phase: Domain Module maintenance

TLSs can not be static because domains change, so TLSs must adapt to domain evolution. Domains such as Machine Tool are strongly influenced by the evolution of the technology. New machinery appears and the staff must learn to use it. Machines are becoming more and more sophisticated, therefore new presentations and evaluation forms from recent documents must be added to the tutor. However, adding new documents

to the TLS is not trivial, their contents must be integrated. On the one hand, new didactic resources must be generated and added to the DRR. Some existing didactic resources must be labelled as obsolete. On the other hand, the recent texts may define new topics that must be added to the domain, so ontology integration problems and methodologies must be taken into account.

IRIS may profit from IKAT incremental development of ontologies through versions to adapt the Domain Modules to changes and additions in the domain. This update can be performed in two ways: using the interface of IRIS (human instructors) or using directly IKAT (knowledge engineers).

3 Conclusions

In this paper a procedure for semi-automating the generation of the Domain Module for the TLSs is presented. In the first phase of this process, the Domain Module Structure and the Domain ontology are gathered using NLP techniques. Next, the document is analysed by using the Domain Ontology, a Pedagogic Ontology and some patterns in order to generate the set of Didactic Resources for the TLS. These Didactic Resources are annotated and stored in a didactic resource repository for future use. Domain Module must be maintained update by adding new didactic resources and concepts. The proposed procedure ensures the knowledge reuse and facilitates the TLS building process.

4 Acknowledgements

The works presented in this paper are supported by the University of the Basque Country (UPV/EHU) (1/UPV 00141.226-T-13946/2001), the Economy Department of the Gipuzkoa Council (IRIS-D(II); RF 949/2001), the Basque Government (UE1999/36) and the CICYT (TIC99-0252).

References

1. Arruarte, A., Fernández, I., Greer, J. (1996). The CLAI model. A cognitive theory to guide ITS development. In *Journal of Artificial Intelligence in Education* 7(3/4), pp. 277-313.
2. Arruarte, A., Fernández, I., Ferrero, B., Greer, J. (1997). The IRIS Shell: How to build ITSs from pedagogical and design requisites. In *International Journal of Artificial Intelligence in Education* 8(3/4), pp. 341-381.
3. Arruarte, A., Elorriaga, J. A., Rueda, U. (2001). A Template Based Concept Mapping tool for Computer-Aided Learning. Okamoto, T. Hartley, R., Kinshuk, Klus, J. P. (Eds), *IEEE International Conference on Advanced Learning Technologies 2001*,IEEE Computer Society, pp. 309-312.
4. Bateman, J. A., Magnini, B., Fabris, G. (1995). The generalized upper model knowledge base: Organization and Use. In N. J. I. Mars (Editor), *Towards very large knowledge bases: knowledge building and knowledge sharing*, pp. 60-72. IOS press, Amsterdam, NL.
5. Bloom, B.S., Engelhart, M.D., Murst, E.J., Hill W.H. & Drathwohl, D.R. (1956) Taxonomy of Educational Objectives. *The Cognitive Domain*, Longmans.
6. Buitelaar, P., Sacaleanu, B. (2001). Ranking and Selecting Synsets by Domain Relevance. *NAACL 2001 Workshop Wordnet and other lexical resources: Applications and Customizations*.
7. Kay, J. (1999). Ontologies for Reusable and Scrutable Student Models. *AI-ED'99 Workshop on Ontologies for Intelligent Educational Systems. Le Mans, France.*
8. Larrañaga, M., Elorriaga, J. A. (2002). IKAT: A tool for incremental development of ontologies through versions. *Intelligent Information Processing (IIP) of the 17th edition of the IFIP World Computer Congress (WCC2002)*, pp. 65-76. Montreal, Canada.
9. LTSC. (2001). IEEE P1484.12 Learning Object Metadata Working Group homepage [On-line]. http://ltsc.ieee.org/wg12/

10. Magnini, B., Gavaglià, G. (2000). Integrating subject field codes into Wordnet. *In Proceedings of LREC-2000, second International Conference on Language Resources and Evaluation*. Athens, Greece.
11. Merrill, M.D. (1983). Component Display Theory, C.M. Reigeluth (eds.), *Instructional-Design Theories and Models: an overview of their current status*, Lawrence Erlbaum Associated, pp. 279-333.
12. Miller, G. A. (1990). WORDNET: an online lexical database. *International Journal of lexicography*, 3(4), pp. 235-312.
13. Mizoguchi, R., Bordeau, J. (2000). Using Ontological Engineering to Overcome Common AI-ED Problems. *International Journal of Artificial Intelligence in Education*, Vol. 11, pp. 107-121.
14. Ranwez, S., Crampes, M., Leidig, T. (1999). Description and Construction of Pedagogical Material using an Ontology based DTD. *In AI-ED 99 Workshop on Ontologies for Intelligent Educational Systems*. Le Mans, France.