Enhancing ITS building process with semi-automatic domain acquisition using ontologies and NLP techniques

Larrañaga, M.

University of the Basque Country (UPV/EHU)
649 P.K., E-20080 Donostia, Basque Country
jiblaolm@si.ehu.es

In this paper, a proposal for automating the acquisition of the domain for the ITSs is presented. Natural Language Processing (NLP) techniques will be used to gather from books and electronic material the Domain Module Structure in a semi-automatic way and to build the ontology that describes it. Main domain topics and relationships among topics will be identified using NLP techniques. On the one hand, this ontology will be used to analyse the document in depth in order to identify the fragments corresponding to each particular learning unit and to generate the didactic resources for those units. On the other hand, the ontology can also facilitate the management and the adaptation of the ITS according to the evolution of the domain.

Keywords: Intelligent Tutoring Systems, Automatic Domain Acquisition, Ontologies, versioning

Context

ITSs have been successfully used in several areas but they still remain many domains, such as Machine Tool, to be explored. Machine Tool domain seems to be a promising area for the application of ITSs: 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 ITSs.

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. ITSs 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 ITSs 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. ITSs built with IRIS use the following four elements to describe the domain:

- Teaching-learning contents are called Basic Learning Units (BLUs). Four kinds of BLUs are used in IRIS: *concepts, procedures, principles* and *facts*.
- Relationships among contents (Rs). Structural relationships – *is-a* and *part-of* – and pedagogic relationships – *prerequisite, corequisite, postrequisite* 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 *knowledge, comprehension, application, analysis, synthesis* and *evaluation*.
- Didactic Resources (DRs) are needed to achieve any IO both to present BLUs to the learner and to asses domain contents.

IRIS uses IKAT [7], 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 ITSs 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 ITS. In this context, the Domain Module can be obtained in a semi-automatic way from existing electronic material. 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 ITS.

Research Proposal

Recently, Mizoguchi and several other authors [11] [6] have proposed the use of ontologies in different modules of the ITSs to profit from the advantages they offer (pedagogic domain reuse, didactic module standardization and so on). High Level ontologies, e.g.: WordNet [10], GUM [4], have successfully been used for Natural Language Processing (NLP) 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 from the books and the electronic material. In this way, the ITS generating process of IRIS will be facilitated. This process will be divided into three phases: Domain Module Structure acquisition, Generation of didactic material and Domain Module Maintenance.

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, both are obtained: 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 ITSs that IRIS produces. The Domain Ontology contains a more general representation of the Domain. 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.

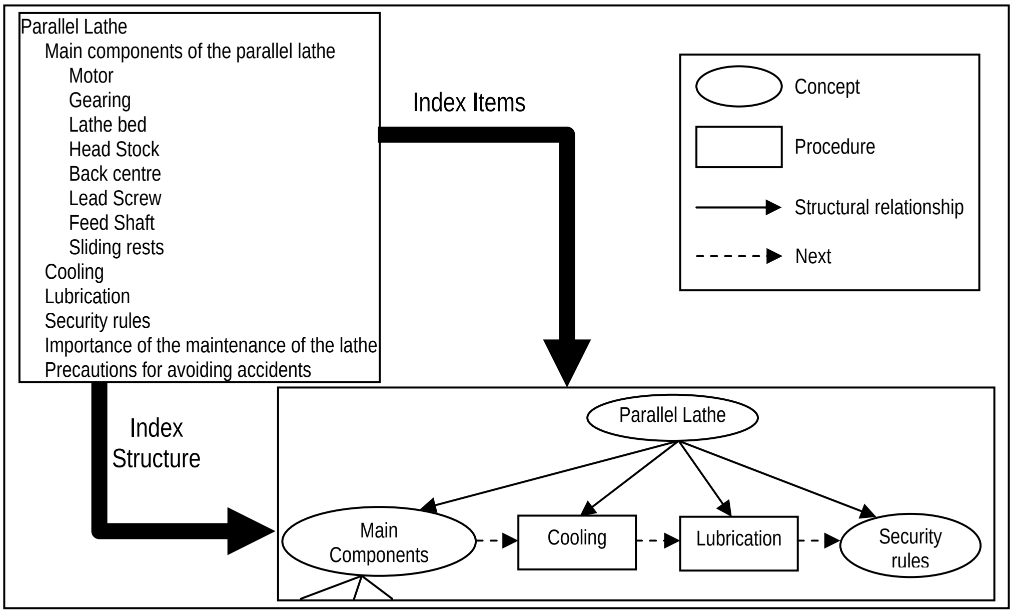

Figure 1: Acquisition of the main topics of the domain from the index of the document

- 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 [5]. 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 [9], 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 asses 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 Ontology will be presented to them in a concept map. CM-ED [3], a generic concept map editor, will be used to visualise the concept map. They can review and adapt the Domain Ontology and the changes will be propagated to both the Domain Module Structure and the domain ontology.

Second Phase: Generation of didactic material

Once the Domain Module Structure is finished, didactic resources, such as definitions, examples, analogies, must be provided and associated with each BLU and their instructional objectives. The whole document will be analysed in order to identify the didactic resources. Text fragments corresponding to different BLUs will be identified with NLP techniques and the domain ontology built in the previous phase. A Pedagogic ontology (an ontology that contains pedagogic issues) will be used in order to infer the didactic goal in every fragment of the documents. References to other BLUs will be also annotated in every didactic resource.

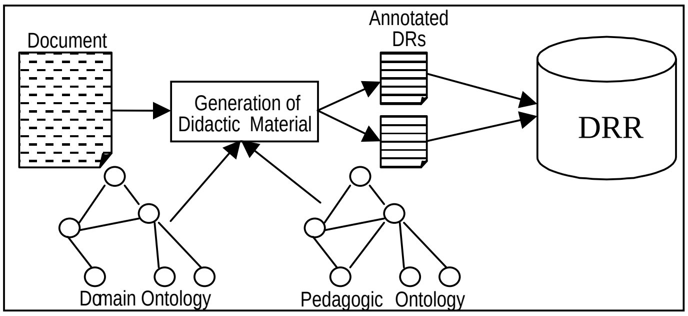

Figure 2: Didactic resource generation process

Once this second phase analysis is finished, the different presentation and evaluation forms will be added to the above-mentioned Domain Module and stored in IRIS format. 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.

Another important issue for enhancing the process of building ITSs (or any other kind of Teaching-Learning Environment) is reusing existing didactic resources. It has been pointed out the need of annotating the pedagogical material using metadata in order to facilitate its reusability [12]. Therefore, didactic resources in IRIS will be annotated following a Learning Objects standard (e.g.: LTSC/LOM [8]) 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 ITSs they are building.

Third Phase: Domain Module maintenance

ITSs cannot be static because domains change, so ITSs must adapt to domain evolution. Domains such as Machine Tool are strongly influenced by the evolution of the technology. New machinery appears and thus the staff must learn to use these new machines. These 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 ITS 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 existing ITSs to new domain or user requirements. This update can be performed in two ways: using the concept map based interface of IRIS (human instructors) or using directly IKAT (knowledge engineers).

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. Buitelaar, P., Sacaleanu, B. (2001). Ranking and Selecting Synsets by Domain Relevance. *NAACL 2001 Workshop Wordnet and other lexical resources: Applications and Customizations*.
6. Kay, J. (1999). Ontologies for Reusable and Scrutable Student Models. *AI-ED'99 Workshop on Ontologies for Intelligent Educational Systems*. Le Mans, France.
7. Larrañaga, M., Elorriaga, J. A. (2002). IKAT: A tool for incremental development of ontologies through versions. To appear in *Proccedings of the 17th edition of the IFIP World Computer Congress (WCC2002).* Montreal, Canada.
8. LTSC. (2001). IEEE P1484.12 Learning Object Metadata Working Group homepage [On-line]. http://ltsc.ieee.org/wg12/
9. 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.
10. Miller, G. A. (1990). WORDNET: an online lexical database. *International Journal of lexicography*, 3(4), pp. 235-312.
11. 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.
12. 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.