Index Analysis: A Means to Acquire the domain module Structure

Mikel Larrañaga, Urko Rueda, Jon A. Elorriaga, Ana Arruarte
University of the Basque Country (UPV/EHU)
{olagaray, jibrumou, elorriaga, jiparlaa}@si.ehu.es

Abstract: The work here presented aims to acquire semi-automatically the domain module of Teaching Learning Systems. This paper focus on acquiring the domain module structure by analysing the index of existing documents. The system uses Natural Language Processing techniques and Heuristic Reasoning in order to identify the domain topics and the pedagogical relationships between them.

1 Introduction

Teaching Learning Systems have proved to be very useful in many learning situations: distance learning, training, etc.. Different types of approaches (intelligent tutoring systems, e-learning systems, collaborative learning systems,…) profit from new technologies in order to better educate different kinds of students. All those systems require the representation of the domain to learn, i.e., the domain module. However, building these systems is not easy because of the amount of the data that must be represented. The domain module is the one that needs more data exchange with the instructional designer. Murray [6] pointed out the need of tools that facilitate the construction of the domain module in a semi automatic way.

The IRIS [4] approach for representing the domain is based on the following elements. *Basic Learning Units* (BLUs) represent the teaching learning topics; any domain can be represented using four kinds of BLUs (*concepts, procedures, principles* and *facts*). *Pedagogic Relationships* among topics are used to establish the sequence of BLUs in the learning process. There are two groups of pedagogical relations: structural relationships - is-a and part-of - and sequential relationships – *prerequisite, corequisite* and *next* -.

The aim of this research work is to extract the domain knowledge of a Teaching Learning Systems from existing documents in order to lighten its developing cost [5]. The system here presented uses Artificial Intelligence methods and techniques like Natural Language Processing (NLP) and Heuristic Reasoning. In addition, ontologies will provide a solid basis to the process of automating the acquisition of the domain. However, the acquisition of this knowledge requires the collaboration of instructional designers in order to get an appropriate representation of the domain module. This paper focuses on the analysis of the document indexes in order to extract the domain structure together with additional information.

2 Index Analysis

Indexes are useful sources of information for acquiring the domain module in a semi-automatic way. Furthermore, textbook indexes are usually well-structured and contain the main topics of the domain. Besides, they are quite resumed so a lot of useful information can be extracted with a low cost process. The authors of texts have previously analysed the domain and decided how to organise the content according to pedagogical principles. They use the indexes as the basis for structuring the domain module. So, the implicit pedagogical relations can be inferred by using NLP techniques and a collection of heuristics. The index analysis process is composed of the next five phases:

Index Pre-Process: The indexes are usually human made text files and therefore, they may contain different numbering formats and some inconsistencies such as typographic errors, format errors, etc. In order to run an automatic analysis process the indexes must be error-free, so they have to be corrected and homogenized before the analysis. In the pre-process step, the numbering of the index items is filtered and replaced by tabulations with the aim of sharing the same index structure. This is performed automatically. However, the correction of inconsistencies can hardly be performed automatically. So, this task is performed manually by the users. The result of this step is a text file in which each title of section is in one line (index item) and the level of nesting of the title is defined by the number of tabulations.

Linguistic Process: This work has been performed with documents written in Basque language. Basque is an agglutinative language. As prepositional functions are realised by case suffixes inside word-forms, Basque presents a relative high power to generate inflected word-forms. This characteristic is particularly important because the words in Basque contain much more part-of-speech information than words in other languages. These characteristics make morphosyntactic analysis very important for Basque. Thus, for the index analysis, the lemmas of the words must be extracted so as to the gather correct information. This process is carried out using EUSLEM [2], a lemmatizer/tagger for the Basque. Noun phrases, verb phrases and multiword terms are detected by ZATIAK [1]. The result of this step is the list of lemmas and the chunks of the index items. This lemmas and chunks constitute the basis of the domain ontology that will be completed in the analysis of the whole document. The morphosyntactic analysis is performed by EUSLEM, which annotates each word with the lemma and morphosyntactic information. Later, entities, postpositions are extracted. ZATIAK extracts the noun and verb phrases.

Basic Analysis: In this task the main topics of the domain and the relationships among these topics are mined from the homogenized index. In this approach, each index item is considered as a domain topic (BLU). Two kinds of pedagogical relationships are detected: structural and sequential. Structural relations are inferred among an item and its subitems (nested items). A subtitem of a general topic is used to explain a part of that issue or a particular case of it. Sequential relations are inferred among concepts of the same nesting level. The order of the items establishes the content sequence in the learning process.

Heuristic Analysis: After carrying out a previous analysis of a significant sample of document indexes, a collection of heuristics (see table 1) to mine general information about BLUs and relations has been identified. The heuristics are classified in two different groups: *heuristics for BLUs* and *heuristics for relationships*. The first ones are useful for extracting characteristics of the BLU (relevance, difficulty and type) using basic information like number of pages or subitems. Regarding the relationships, the basic analysis gets the general category of each relation and it should be refined into *part-of*, *is-a*, *prerequisite*, *corequisite* and *next*. As the empirical analysis proved that the most common structural relation is *part-of* relation, by default, the structural relations will be traduced into part-of relation. Some heuristics have been identified to detect *is-a* relation or to reinforce the hypothesis that the structural relations is certainly *part-of*. For example, entity names are used to identify particular examples, so when the subitems are entities *is-a* relation is the most probable. However, index items do not always share the same linguistic structure. Therefore, different heuristics apply in the same of set of index subitems. The system combines the information provided by the heuristics that can be applied in order to refine the structural relationships. Regarding sequential relations, *next* is the most common one, so, by default, any sequential relation is traduced into *next*. However, the *prerequisite* relation can be also found in the indexes. For example, if the text of the index has references to other item probably there is a prerequisite relation between them.

Table. 1. Identified heuristics.

Heuristics for BLUs	***BLU characteristic***
Number of pages	BLU relevance
Number of subitems and nesting levels	BLU difficulty
Part-of-speech information and patterns	BLU type
Heuristics for Structural relationships	***Structural Relationship***
MultiWords	is-a
Entity Names	is-a
Acronyms	is-a
Possessive Genitives	part-of
Heuristics for Sequential Relationships	***Sequential Relationship***
References	prerequisite
Possessive Genitives	prerequisite

Supervision of the results: The results of the analysis are presented to the user in a graphical way by means of concept maps using the CM-ED tool [3]. The user can supervise and modify the structure of the domain to some particular needs. These modifications can be performed on the concept maps by adding, removing or modifying both concepts and relations.

4. Conclusions

The aim of this work is to facilitate the building process of Teaching Learning Systems by

acquiring the domain module from textbooks and other existing documents. The semi-automatic acquisition of the domain knowledge will significantly reduce the instructional designers' workload when building the Teaching Learning Systems. This paper has presented a system for generating the domain module structure from the analysis of indexes of textbooks. The domain module structure includes the topics of the domain and the pedagogical relationships among them. The system performs the analysis using Natural Language Processing (NLP) tools and Heuristic Reasoning. The system applies EUSLEM and ZATIAK, which are two robust NLP tools that have been tested with large documents obtaining very good results. In addition, some heuristics have been implemented in order to identify pedagogical relations between topics. These heuristics provide additional information about the topics and the relations such as the relevance of the topics or the type of the pedagogical relations.

References

1. Aduriz, I., Aranzabe, M. J., Arriola, J. M., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A., Urizar, R. (2003). Methodology and steps towards the construction of a Corpus of written Basque tagged in morphological, syntactic, and semantic levels for the automatic processing (IXA Corpus of Basque, ICB). In proceedings of Corpus linguistics 2003. Lancaster. United Kingdom, 10-11.
2. Aduriz, I., Aldezabal, I., Alegria, I., Artola, X., Ezeiza, N., Urizar, R. (1996). EUSLEM: A Lemmatiser / Tagger for Basque. In Proceedings of the *EURALEX'96*, Part 1. Gothemburg (Sweden) , 17-26.
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 *ICALT 2001*, IEEE Computer society, 309-312.
4. Arruarte, A., Fernandez, I., Ferrero, B., Greer, J. (1997). The IRIS Shell: How to build ITS from pedagogical and design requisites. *International Journal of Artificial Intelligence in Education* 8(3/4), 341-381.
5. Larrañaga, M. (2002). Enhancing ITS building process with semi-automatic domain 5acquisition using ontologies and NLP techniques. In Proceedings of the Young Researches Track of the Intelligent Tutoring Systems (ITS 2002). Biarritz (France).
6. Murray, T. (1999). Authoring Intelligent Tutoring Systems: an Analysis of the State of the Art. *International Journal of Artificial Intelligence in Education*, 10, 98-129.

Acknowledgements. This work is partly supported by the University of the Basque Country (UPV00141.226-T-14816/2002) and the CICYT (TIC2002-03141).