ErauzOnt: a Framework for Gathering Learning Objects from Electronic Documents

Mikel Larrañaga, Iñaki Calvo, Jon A. Elorriaga, Ana Arruarte
Department of Computer Languages and Systems
University of the Basque Country
Donostia, Basque Country
e-mail: mikel.larranaga@ehu.es

Katrien Verbert, Erik Duval
Department of Computer Science
Katholieke Universiteit Leuven
Leuven, Belgium
e-mail: katrien.verbert@cs.kuleuven.ac.be

Abstract—Retrieving and reusing Learning Objects (LOs) can lighten the workload of constructing new on-line courses or Technology Supported Learning Systems (TSLs). This paper presents ErauzOnt, a framework for the automatic generation of new LOs from electronic documents using domain ontologies and NLP techniques.

Keywords: Learning Objects, Knowledge Acquisition, Content authoring

I. INTRODUCTION

Advances in last years have highly increased the influence of new technologies in general, and Information and Communication Technologies (ITCs) in particular, in nowadays society. On-line applications have become essential; they are continuously used for communication (e.g., instant messaging, e-mailing and phoning), etc.

This evolution has also affected education, providing means that enhance both teaching and learning. Learning Management Systems (LMSs) such as Moodle or WebCT are currently being broadly used in universities, institutes and schools [1, 2]. The use of such systems allows human instructors to concentrate on defining and choosing their favourite learning resources without having to deal with technical issues [3, 4].

Content authoring is known to be time and effort consuming. However, relying on the retrieval and reuse of existing Learning Objects (LOs) can significantly reduce the workload of the development of Technology Supported Learning Systems (TSLs). The work presented in this paper aims to go one step further, as its main goal is to automate the generation of new LOs from existing electronic documents. This paper presents ErauzOnt, a framework for identifying and gathering LOs from electronic documents using Natural Language Processing (NLP) techniques and ontologies. This framework aims to be applicable on any document no matter the domain it relates to. None of its components relies on implicit domain-specific knowledge. All the used domain-specific knowledge are the domain topics and the relationships among them described on the domain ontology, which is the input for the LO extraction process together with the document to be analysed. These domain ontologies are built semi-automatically from the document. The approach has been presented in our earlier work [5].

This paper is organized as follows: first, the LO gathering process of ErauzOnt is described. Section III presents the underlying architecture of the ErauzOnt framework. Scalability and extendability characteristics of the framework are discussed in Section IV. Then, the experience using ErauzOnt is presented to finish with some conclusions and future work.

II. BUILDING LOS USING NLP AND ONTOLOGIES

In this work, building LOs entails identifying pieces of text or images corresponding to Didactic Resources (DRs) and their enrichment with metadata to facilitate their reuse. From now on, we will refer as DR to a fragment of educational material extracted from the document, while a LO will be a DR annotated with metadata to support content reuse that is stored in a LO Repository (LOR).

To identify such pieces of text, the process illustrated in Figure 1 is followed. First, an internal tree-like representation is built from the electronic document. The linguistic analysis is performed on the tree-like document representation to get the part-of-speech information. While this information might be not so necessary for some languages such as English, where regular expressions can be used to identify examples and definitions [6, 7], it is essential for other languages such as Basque. Basque is an agglutinative language, i.e., for the formation of words the dictionary entry takes each of the elements needed for the different functions (syntactic case included). More specifically, the affixes corresponding to the determiner, number and declension case are taken in this order independently of each other. As prepositional functions are realised by case suffixes inside word-forms, Basque presents a relative high power to generate inflected word-forms, which makes morphosyntactic analysis very important for being able to extract information from text fragments.

The next step consists on improving the part-of-speech information by labelling the occurrences of the domain topics listed in the domain ontology. Such ontology specifies the topics and the relationships among them.
Once the references to domain topics have been marked, DRs, such as definitions, examples, problem-statements, etc., are identified. The identification of DRs is performed by matching patterns commonly used to define contents, present examples, etc. The approach is exemplified in [8].

The output of this step is a set of atomic DRs that are enhanced next in two ways: it has been observed that human instructors combine different DRs to get more accurate material (e.g., an example can complement and reinforce the knowledge provided by a definition). Therefore, consecutive DRs are combined if they are similar or close enough, according to the content and the kind of DR. For sake of brevity, no more details on the composition of DRs is provided. Details on the composition of the DRs can be found in [9].

In order to assure the cohesion of the gathered DRs, they are post-processed. The cohesion maintenance relies on the use of discourse markers (e.g., on the one hand, ... on the other hand,...) and references to previous sentences. The system deals differently with each kind of discourse markers. If the DR starts with the second part of a complex discourse marker (e.g., on the other hand,...), it will add all the necessary sentences until the initial part (the one hand,...) is included. References usually regard to up to three previous sentences; thus, at most three sentences are added in this case.

Finally, the generated DRs are enriched with metadata. The original electronic document is analysed by the Samgi Service (cf. Section III) to automatically get the initial metadata elements set. These elements are used as the basis to annotate every new LO, and it is enhanced with information that can be extracted from each particular resource. The approach is detailed in the next Section.

III. ARCHITECTURE OF THE ERAUZONT FRAMEWORK

ErauzOnt has been developed in order to carry out the process of extracting LOs from electronic documents and aiming to be domain-independent. Besides, the framework is not tightly coupled to a particular language or to an electronic document format.

The architecture of the ErauzOnt framework, which is illustrated in Figure 2, entails a client application (Didactic Resource Extractor and Generator), some web services (represented with rounded boxes) and two local modules (Pdf2Tree Builder and LO File Builder).

This web-service oriented approach makes ErauzOnt flexible and platform-independent. Although it uses some platform-specific applications (mainly NLP tools), they are used only by the web-services, so these limitations only apply on the server side and have no effect on the client side. ErauzOnt employs two repositories, one that stores the LOs (both, resources and metadata) and another one for keeping the preview files for the LOs. Both repositories build on the ARIADNE Knowledge Pool System [10] for storage and retrieval of resources and their metadata.

The ErauzOnt framework consists of several web services that provide core background functionalities of the
in the analysed text, but do not consider other relationships among the topics. For example, a keyword extractor may identify Earth, Mars, Mercury and Venus in a fragment of text if they appear in it, but it would not consider that all of them are planets, and therefore it would not infer planet as a keyword, as it is not aware of the semantic relationships among these topics.

ErauzOnt profits from a domain ontology to perform LO extraction and this ontology can also be used to better infer keywords for a particular LO and overcome this problem. Thus, ErauzOnt uses the UKB Service to get a probabilistic weight distribution of the topics of the domain for the analysed LO in which the semantic relationships are taken into account and produces more accurate keyword annotation.

The current version of the framework supports the transformation of PDF documents. The Pdf2XML Service is used to get a XML representation for a pdf document. It returns a ZIP file that contains the XML version of the original document and all the extracted images.

The NLP Analysis Service provides the part-of-speech information for a text fragment. As mentioned above, NLP information is not only useful, but necessary in many languages (agglutinative languages such as Basque). Thus, the NLP Analysis Service has been included with the aim of supporting either agglutinative or non-agglutinative languages.

The Constraint Grammar Service (CG Service) is the responsible of identifying DRs, such as definitions, examples, problem-statements, etc. The identification of DRs is performed by matching patterns commonly used to define contents, present examples, etc. The Constraint Grammar formalism [14-16] is used to cope with this task.

The UKB Service provides a means to measure similarity/relatedness using ontologies. The main contribution is that it does not only consider topics, but also the semantic relationships among them, which gives more accurate results. It uses the UKB algorithm and suite of programs [17-19] available at http://ixa2.si.ehu.es/ukb.
The **LO File Builder** generates the final archive for the generated LOs, as well as the preview files. Presentation formats such as *pdf*, *doc* and *odf* are suitable for final presentation, but are not appropriate for flexible content reuse, as the components cannot be easily accessed (manual *copy&paste* has been used so far). The ALOCOM framework [8] was developed to decompose composite LOs and make those components available for *on-the-fly* content reuse. This framework relies on the ALOCOM ontology [20], which represents a content model for LOs and their components. The generated LOs are stored in a ZIP file that contains the XML-based representation of the LO based on the ALOCOM formalism, as well as the referenced images or resources. The XML-based ALOCOM document format allows content reusing and even versioning [21].

The **Didactic Resource Extractor and Generator** is the core of the ErauzOnt framework, and is responsible of generating new LOs from the electronic document using the server-side web-services and the client-side modules. It converts the document to be processed into the internal tree-like document representation using the *Pdf2Tree Builder*, which in turn invokes the *Pdf2XML Service* to extract the information from the input document. As mentioned, above, previously processed documents or fragments are not processed again. Once the tree-like structure is built, it is processed by the **Didactic Resource Extractor and Generator** to identify DRs using the information provided by the **NLP Analysis**, and **CG services**. Then, these DRs are later enhanced to get more accurate and cohesioned DRs (cf. Section II) relying on the information provided by the **UKB Service**. Next, the final archive files and preview files for the LOs are generated invoking the **LO File Builder**. Finally, the identified LOs are annotated and stored in the LOR.

The next Section describes with more detail scalability and extendability issues of the framework.

IV. EXTENDABILITY OF THE ERAUZONT FRAMEWORK

So far, the processed electronic documents are *pdf* files written in Basque language. However, the ErauzOnt framework is flexible enough to provide support for new languages and document formats without major effort. The only document-format dependent step is the construction of the tree-like document structure. Document formats such as *doc*, *docx* and *odf* use XML internally to store the information. XML files have an inherent tree-structure, so taking this into account, and the availability of libraries to work with this kinds of documents, tree builders for other document formats can quite easily be plugged to ErauzOnt.

To support a new language, the discourse markers for that language and a new version of the Didactic Resources Grammar must be developed. The discourse markers are stored in a simple XML file, and the **Didactic Resource Extractor and Generator** determines which discourse markers has to apply depending on the language the document is written on. The **CG Service** also determines the grammar to apply according to the language, and it has been observed/tested that very similar patterns apply in other languages, particularly in English.

Currently, the **NLP Analysis Service** only supports Basque Language, which is carried out using the EUSLEM lemmatiser/tagger for Basque [22]. However, this web-service can be improved to use new NLP parsers for the new languages with minor effort. Once again, the language of the document will determine which tool would be used. Using Freeling [23] will result on supporting English, Spanish and Catalan. The Stanford Parser [24, 25] has been used for English, German, Arabic and even Chinese [26], so it can also significantly contribute to support new languages.

VI. EXPERIENCE USING ERAUZONT

ErauzOnt is currently being exploited on basic textbooks for primary school, most of them related to Nature Sciences. The textbooks are provided by the Gipuzkoako Ikastolen Elkartea (GIE - http://www.ikastola.net/), one of the main publishers for Basque-medium education at all educational levels and in all subjects in compulsory education at the Basque Country. As the approach here presented aims to be domain-independent, very formalised domains such as mathematics have been initially discarded as they may result on biased approaches for such domains.

The domain ontologies used to gather the LOs from these electronic documents have been developed following two approaches. Some of them have been manually developed while others were gathered from the electronic documents following the approach described on [5] and supervised using CM-DOM [27].

The analysed textbooks tackle subjects such as cosmology, geology and anatomy. They are being used to build the domain module, i.e., a course, for Technology Supported Learning Systems (TSLs) such as web-based intelligent tutoring systems or e-learning systems. The domain module entails the domain ontology and the LOs for the domain topics. This domain ontology differs from the one used to gather the LOs from the electronic documents on that this one also related the domain topics with their corresponding LOs.

V. CONCLUSIONS AND FUTURE WORK

This paper has presented ErauzOnt, a framework for identifying and gathering LOs from electronic documents using ontologies and Natural Language Processing (NLP) techniques. This framework is intended for human instructor or instructional designer no matter their computing mastering level.

The framework relies on the identification of commonly used patterns to identify fragment of the document containing educational material. LOs are generated from the identified fragments.

Although ErauzOnt currently supports only *pdf* documents written in Basque language, it can be easily enhanced to support either new document formats or languages (cf. Section IV).

Future work includes adding support for new languages such as Spanish and English, as well as new document formats.

Another planned task is to evaluate the performance and effect of the generated TSLs on the learners learning process.

Multicultural and multilingual issues are also being considered, and thus the generation of LOs in many languages is also planned. Both Belgium and Basque Country have many official languages, so the automatic
generation of LOs in many languages might be really helpful. Machine Translation might be used to generate LOs in other languages or at least to facilitate the retrieval of similar LOs in other languages. The metadata should be enhanced to contain references to “equivalent” LOs in other languages.

ACKNOWLEDGMENT

This work is supported by the Univ. of the Basque Country (UE09/09), the MEC (TIN2009-14380) and the Gipuzkoa Council in an EU Program. Also, we want to thank the collaboration of the Gipuzkoako Ikastolen Elkartea. Katrien Verbert is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO).

REFERENCES

The references are incomplete - proceedings title/conference name is for instance missing