Slicepedia: Automating the Production of Educational Resources from Open Corpus Content

Killian Levacher[1], Seamus Lawless[2], and Vincent Wade[3]

Trinity College, Dublin, Ireland

Abstract. The World Wide Web (WWW) provides access to a vast array of digital content, a great deal of which could be ideal for incorporation into eLearning environments. However, reusing such content directly in its native form has proven to be inadequate, and manually customizing it for eLearning purposes is labor-intensive.
This paper introduces Slicepedia, a service which enables the discovery, reuse and customization of open corpus resources, as educational content, in order to facilitate its incorporation into eLearning systems. An architecture and implementation of the system is presented along with a preliminary user-trial evaluation suggesting the process of slicing open corpus content correctly decontextualises it from its original context of usage and can provide a valid automated alternative to manually produced educational resources.

1 Introduction

Educational Adaptive Hypermedia systems (EAH) have traditionally attempted to respond to the demand for personalized interactive learning experiences through the support of adaptivity, which sequences re-composable pieces of information into personalized presentations for individual users. While their effectiveness and educational benefits have been proven in numerous studies [1], the ability of EAH systems to reach the mainstream audience has been limited [2]. This is in part due to their reliance upon large volumes of educational resources available at high production costs, incurred by labor-intensive work [3].

This dependency has been extensively studied by the research community, which has addressed this issue mainly by improving either the discovery [4] or the reuse [5] of existing educational resources. Solutions proposed so far however, do not address the fundamental problem which is the labor intensive manual production of such resources.

In parallel with these developments, the field of Open Adaptive Hypermedia (OAH) has attempted to leverage the wealth of information, which has now become accessible on the WWW as open corpus information. However, open corpus reuse and incorporation has been achieved so far, using either manual [6] or at best traditional information retrieval (IR) approaches [7]. Even when retrieving relevant open web information, these IR techniques suffer because they

only provide one-size-fits-all, untailored, document level, delivery of results, with limited control over topics, granularity, content format or associated meta-data.

This results in limited and restricted reuse of such resources in OAH. Open corpus material, in its native form, is very heterogeneous. It comes in various formats, languages, is generally very coarse-grained and contains unnecessary noise such as navigation bars, advertisements etc. Hence, there remains a significant barrier to automatically convert native open corpus content into reusable educational resources meeting specific content requirements (topic covered, style, granularity, delivery format, annotations) of individual EAH.

We believe that the cost intensive, manual production and/or adaptation of educational resources must be augmented or replaced by the automated repurposing of open corpus content into such resources. This transformation will make it possible to provide on-demand automated production and right-fitting of educational resources in large volumes to support re-composition and personalization within EAH systems.

Contribution: This paper presents Slicepedia, a service that leverages content[1] from open corpus sources to produce large volumes of right-fitted educational resources at low cost. This novel approach leverages complementary techniques from IR, Content Fragmentation, Information Extraction (IE) and Semantic Web to improve the reuse of open corpus resources by converting them into information objects called slices.

- An implementation of the system architecture is presented, which has been applied in an authentic educational user-trial scenario.
- Initial results, of an evaluation currently underway, investigating the quality of automated open corpus reuse and its suitability within an educational context, are presented in this paper.

2 The Web Converted as Slices

The Slicepedia service enables the automated reuse of open corpus content through slicing. Slicing [8] is the process of automatically harvesting, fragmenting, semantically annotating and customizing original web resources into recomposable information objects called slices, tailored for consumption by individual EAH systems. The system is available as a fully autonomous service, composed of successive and easily pluggable components, and provides slices according to the formats (LOM, SCORM etc...) and meta-data requested.

Harvesting: The first component of a slicer pipeline acquires open corpus resources, from the web, in their native form. Standard IR systems[2] or focused crawling techniques [9] are used to gather relevant documents, which are then cached locally for further analysis.

[1] In order to deal with the significant technical challenges of right sizing and reuse, some specific aspects are deemed beyond the scope of this paper; namely copyright and digital rights management issues.

[2] http://developer.yahoo.com/search/web/V1/webSearch.html

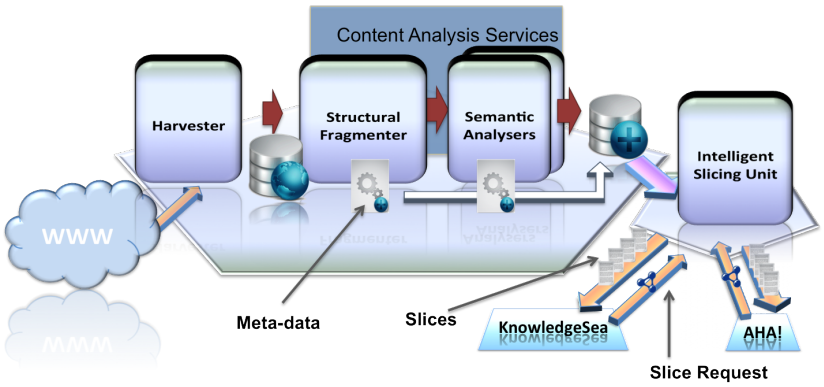

Fig. 1. Slicepedia Architecture

Structural Decontextualisation: Resources harvested are then fragmented into individual structurally coherent sections (such as menus, advertisements, main article). Structural meta-data, such as the location of each fragment within the original resource, is extracted and stored in the meta-data repository. This phase is critical since, maximising the reuse potential of a resource involves the ability to identify specific reusable parts of pages from any clutter content present within the original document. Densitometric fragmentation [10] was selected for this slicer implementation.

Semantic Analyser: Once decontextualised, each resulting fragment, available as linked-data in a Sesame store[3], is annotated with rdf semantic labels, using a list of pre-selected algorithms. A boilerplate detection algorithm[4] annotates to what degree a fragment of a page can be reused or not. Concepts mentioned within each fragment are identified using the AlchemyApi concept tagging service[5] and tagged as Dbpedia concepts[6]. Reading level difficulty[7] of fragments, expressed as Flesh Reading scores and finally, the part of speech, noun phrase and verb phrases are also identified and annotated[8] with their relevant linguistic attributes.

Slice Creation: The fourth step finally analyses slice requests received from EAH slice consumers, which are then matched with all possible fragments/meta-data combinations available in the system. Fragment combinations obtaining the closest match are delivered to EAH systems in their format of choice.

A slice request could consist of the following: *"Slices should be written in Portuguese, have a granularity ranging from 3 sentences up to 3 paragraphs, should*

[3] http://www.openrdf.org/
[4] http://code.google.com/p/boilerpipe/
[5] http://www.alchemyapi.com/api/
[6] http://dbpedia.org/About
[7] http://flesh.sourceforge.net/
[8] http://gate.ac.uk/

cover the topics of whale migration, atlantic ocean and hunting. Their Flesh reading score range from 45 to 80 and should not contain any tables or bullet points lists. They should be scoped on the specified topics (i.e. exclude content not on these topics). Slices should contain between 7 and 15 annotations consisting of verbs conjugated in the past perfect continuous and should be delivered as LOM objects".

3 Evaluation

Although the reuse of open corpus material is ultimately aimed at re-composition and personalization, the preliminary results presented in this paper focus, as a first step, in evaluating the automated reuse of individual open corpus slices. The purpose of the evaluation presented below hence was to investigate whether the approach to automate the production of educational resources via open corpus content reuse using slicing techniques could:

- H1: Correctly decontextualise preliminary open corpus resources from original associated clutter
- H2: Offer a suitable alternative to manually generated educational resources from a user perspective.

The experiment compared automated and manual reuse of arbitrary selected open corpus content using a real life user-trial. A simple online e-assessment application (available in English, Spanish and French), built for this experiment, presented users with traditional gap filler grammar exercises, built using different sets of open corpus resources converted (manually or automatically) into grammatical e-assessment exercises. Verb chunks conjugated at specified tenses were removed and replaced by gaps, which users had to fill according to particular infinitives and tenses specified for each gap. The answers provided were compared to the original verb chunks and users were assigned a score for each grammar point.

In order to guarantee open corpus resources harvested, represented a truly random set of resources, a group of five independent English teachers were asked to arbitrarily select a total of 45 pages (content batch CBN) of their choice from the web, from any source/topic of their choice. These teachers were then asked to arbitrarily select fragments (of any size) of pages harvested, which they felt were adequate for grammar exercises, and manually annotate tenses encountered within these extracts, to produce content batch CBM. The entire collection of pages was then harvested from the web in their original form by the slicer and sliced in order to produce a set of automatically generated resources CBO, with similar characteristics as their manual counterparts, This resulted in 3 content batches consisting of CBN (open corpus pages in their native form), CBM (annotated fragments manually produced) and CBO (annotated fragments automatically produced). All of the extracts produced were subsequently converted into grammar e-assessment pages.

The slice consumer application represents an excellent evaluation platform for the purpose of this experiment since it was necessary to select a reuse vehicle where the user needs were very sensitive to: (i) the accuracy of annotations (i.e.: verbal annotations) and (ii) the visual layout (i.e.: content formatted correctly). After performing exercises on each content batch (presented using a Latin Squared Distribution), users were asked to rate each set of pages presented to them using a 10 point Likert scale.

4 Results

A total of 41 users, divided into two groups consisting of Experts (63%) and Trainees (37%), performed the experiment.

H1: As pointed out in Section 2, a part of maximizing the reuse potential of a previously published resource requires the ability to decontextualize this content from its original setting. Hence, users were asked directly, for each content, whether *in addition to the main content, a lot of material displayed on the page was irrelevant to the task (such as advertisement, menu bar, user comments..).* Results obtained for both the manual and automatically produced content were very similar (Mean CBM=2.13, Mean CBO=2.53) with paired t-tests considering mean differences as insignificant (p=0.423). These results indicate that although users did notice a difference in average between the decontextualisation of preliminary open corpus resources carried out by the slicer, the difference was statistically insignificant.

H2: Following a correct decontextualisation of open corpus resources, the overall re-purposing performance of both content batches, with respect to their ability to provide adequate e-assessments, was measured. The number of grammar mistakes (E=29.74%, T=42.61%) measured upon content created automatically, appears to be higher than for the content produced manually (E=23.92%, T=35.13%) for both groups of users. This appears to suggest that automatically generated content occasioned users to perform more errors during the e-assessments tasks. Although the difference in errors between content batches was slightly higher for the trainees in comparison to the experts group, an independent t-test considers this difference as insignificant (Mean Percentage Difference: E=5.80%, T=7.48%, p=0.891, Equal Variances Assumed), which would indicate that although the automatically generated content did induce users to answer erroneously some assessment units, users from the expert group didn't appear to use their language skills to compensate differences between content batches used. When trainees were asked whether, for content batch CBO, *"the number of erroneous assessment units presented was tolerable"*, a mean score of 7 on the liker scale was measured. When asked whether *"Overall, I felt this content was adequate to perform a grammar exercise"* both content achieved very similar scores (CBM=8.33, CBO=8.57, p=0.536) with t-tests suggesting any difference observed as insignificant. Hence, these result appear to indicate that although users appear to achieve lower performances on assessments automatically generated in comparison to those manually produced, this tendency didn't appear to

affect Trainees more than the Experts group of users, nor did it appear to decrease the perceived usefulness of the content for the assessment task performed.

5 Conclusion

The preliminary results presented in this paper appear to indicate that the process of automatically slicing content can correctly decontextualise individual portions of open corpus documents into structurally coherent and independent information objects.

Initial results obtained with respect to the suitability of content automatically generated by this approach, as an alternative to manually produced educational resources, suggest a slight difference in appropriateness. However, user perception appears to consider both content types as interchangeable. Taking into account the low production cost and high volume of educational objects such a slicing service could provide, a slight decrease in content quality could more than likely be tolerable in many educational use cases. An experiment investigating the reading quality decrease, re-composition and personalization of slices in an independent third party AHS is currently in progress.

References

1. Lin, Y.l., Brusilovsky, P.: Towards Open Corpus Adaptive Hypermedia : A Study of Novelty Detection Approaches. In: UMAP. (2011) 353–358
2. Armani, J.: VIDET : a Visual Authoring Tool for Adaptive Tailored Websites. Educational Technology & Society **8** (2005) 36–52
3. Meyer, M., Hildebrandt, T., Rensing, C., Steinmetz, R., Ag, S.A.P., Darmstadt, C.E.C.: Requirements and an Architecture for a Multimedia Content Re-purposing Framework. In: EC-TEL. (2006) 500–505
4. Drachsler, Koper, R.: ReMashed - An Usability Study of a Recommender System for Mash-Ups for Learning. Int. Journal of Emerging Technologies in Learning **5** (2010)
5. Meyer, M., Rensing, C., Steinmetz, R.: Multigranularity reuse of learning resources. Multimedia Computing, Communications, and Applications (2011)
6. Henze, N., Nejdl, W.: Adaptation in Open Corpus Hypermedia. International Journal (2001)
7. Zhou, D., Truran, M., Goulding, J.: LLAMA: Automatic Hypertext Generation Utilizing Language Models. In: Int. Conf. on Hypertext and hypermedia. (2007)
8. Levacher, K., Lawless, S., Wade, V.: Slicepedia: Providing Customized Reuse of Open-Web Resources for Adaptive Hypermedia. In: HT'12: Proc. of the 23rd Conf. on Hypertext and Social Media. (2012)
9. Lawless, S., Hederman, L., Wade, V.: OCCS : Enabling the Dynamic Discovery , Harvesting and Delivery of Educational Content from Open Corpus Sources. In: Int. Conf. on Advanced Learning Technologies. (2008)
10. Kohlschütter, C., Nejdl, W.: A Densitometric Approach to Web Page Segmentation. In: CIKM, Proceeding of the 17th international conference on Information and knowledge management. (2008) 1173–1182