Building a Semantic Virtual Museum: From Wiki to Semantic Wiki using Named Entity Recognition

Alain Plantec
LISyC, Complex System Lab
U.B.O, Brest, France
plantec@univ-brest.fr

Vincent Ribaud
LISyC, Complex System Lab
U.B.O. Brest, France
ribaud@univ-brest.fr

Vasudeva Varma
Search and Information Extraction Lab
IIIT Hyderabad, India
vv@iiit.ac.in

Abstract

In this paper, we describe an approach for creating semantic wiki pages from regular wiki pages, in the domain of scientific museums, using information extraction methods in general and named entity recognition in particular. We make use of a domain specific ontology called CIDOC-CRM as a base structure for representing and processing knowledge. We have described major components of the proposed approach and a three-step process involving name entity recognition, identifying domain classes using the ontology and establishing the properties for the entities in order to generate semantic wiki pages. Our initial evaluation of the prototype shows promising results in terms of enhanced efficiency and time and cost benefits.

Categories and Subject Descriptors
I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods – *Semantic networks.*

***General Terms**:* Documentation, Languages, Standardization.

***Keywords**:* Information extraction, ontology, semantic wiki.

1. Introduction

Scientific patrimony of a university or a scientific museum is constituted of all scientific equipments used for research and teaching together with the knowledge associated with their use – from user manuals to experience protocols. The aim of this research work is to provide a starting path for universities (or scientific museums) looking for a semantic Web site intended to gather and present scientific instruments knowledge.

The semantic Web site was designed with two main directing ideas: to extract knowledge of existing Web site (typically Wikipedia) – providing a bootstrap of the semantic Web site – and to favor a collective building of new knowledge by the scientific actors themselves – a condition required to maintain accurate and "living" knowledge. Thus, we choose a semantic wiki technology – Semantic MediaWiki (SMW) - that provides a good compromise between formal internal structure and easy collective use [1].

RDF [2] is the most widely language used to represent semantic annotations in the form <subject, property, value>. Semantic MediaWiki pages can be seen as "normal" wiki pages (including Wikipedia pages) annotated with values of semantic properties. From an RDF point of view, the subject of the page is the subject of many triples - pairs of property/value. The conventional process of converting normal wiki pages to SMW pages manually is very tedious and time consuming and it requires a good understanding of the underlying ontology. Our attempt is make this process simple and computer aided to the extent possible.

Hence, the problem can be stated in two directions: identifying the value (and related information type) and selecting the property in a suitable ontology. The former problem may be addressed through Named Entity Recognition (NER) and should be automated as much as possible. The latter problem relies on class recognition and property disambiguation; it uses automated annotation tools but may require more human participation.

In this paper, we present a semi-automated solution for transforming normal wiki pages to Semantic MediaWiki pages using a three step process. The steps include recognizing the named entities within the normal wiki page and then recognizing the types of named entities as classes of the CIDOC-CRM (CIDOC Conceptual Reference Model - our domain ontology [3]) and finally, after a human assisted process to disambiguate the possible properties, generating the SMW pages. In our initial evaluation of the proposed model, we have experienced reduced time and increased efficiency in creating SMW pages.

2. Related Work

Though there exists literature for using Wikipedia as a usable set of data for information extraction [4], Semantic Wikis propose a more elaborate way of structuring the knowledge (as a fully typed hypertext network) that may prove to be more powerful [5]. However, we have not come across any work that makes use of information extraction techniques such as NER to transform wiki pages to semantic MediaWiki pages.

3. Proposed Solution: An Example

Let us see a typical Wikipedia page fragment, such as

```
An '''ammeter''' is a [[measuring instrument]] used to measure the [[electric current]] in [[ampere]]s (A), hence the name. The earliest design is the [[Jacques-Arsène d'Arsonval|D'Arsonval]] [[galvanometer]] or '''moving coil''' ammeter.
```

This fragment uses some format directives (between quotes) and links to other pages (between double brackets). If we wish to semantically annotate this fragment with the CIDOC CRM, this fragment should be transformed such in

```
An '''ammeter''' is a [[measuring instrument]] used to measure the [[electric current]] in [[P39B.P40 observed dimension::ampere]]s (A), hence the name. The earliest design is the [[P14 carried out::Jacques-Arsène d'Arsonval|D'Arsonval]] [[P130 shows feature
```

OOPSLA 2009, October 25-29, 2009, Orlando, FL, USA.
ACM 978-1-60558-768-4/09/10.

of::galvanometer]] or '''moving coil''' ammeter. [[Category:E24 Physical Man-Made Thing]]

The last statement indicates that this semantic fragment is related to an instance of the *E24 Physical Man-Made Thing* (a class of the ontology). This instance is the subject of several triples represented as property/value pairs using the syntax [[Px property name::value]], such as [[P130 shows feature of::galvanometer]] indicating that an ammeter is linked to a galvanometer through the *P130 shows feature of* property. Px is an identifier of the property in the CRM and provides cross-language reference. A special case occurs when an intermediate node has to be created to correctly translate a link between two pages (e.g. ammeter and ampere): if a direct property between the two pages' types does not exist, it requires the use of two or more properties (e.g. P39B was measured by, P40 observed dimension) and the creation of intermediary nodes.

This translation process is performed in three steps: named entity recognition (NER), CIDOC-CRM class recognition, and CIDOC-CRM property disambiguation. The two former steps are automatic but require human validation to ensure that named entities or CRM classes were correctly recognized. The last step is a computer-aided process.

4. Prototype and Evaluation

We have built a prototype implementation of the proposed approach and the simplified flow diagram is given in Figure 1.

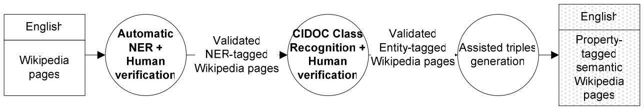
English
Wikipedia pages
Automatic NER + Human verification
Validated NER-tagged Wikipedia pages
CIDOC Class Recognition + Human verification
Validated Entity-tagged Wikipedia pages
Assisted triples generation
English
Property-tagged semantic Wikipedia pages

Figure 1. Simplified diagram of the three-step process.

The major components of the prototype system are Named Entity Recognition, CIDOC CRM class recognition, and assisted triples generation modules:

NER: We are using Conditional Random Fields (CRFs) based NER system. In this system, NER task is modeled as a sequence-labeling task [6] where for a given sequence of words, the NER system constructs a label sequence in which each label represents a predefined set of classes for named entities. For example, the predefined classes include names of people, organizations, places and the domain specific named entities such as instrument, part of instrument, instrument material, and measurement. The final label sequence is the one that has highest probability among all the possible label sequences occurring for a given word sequence.

CRM Class Recognition: We have used ICOM/CIDOC Conceptual Reference Model (CIDOC CRM), an ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information and an ISO 21117 standard since 2006 as the base structure for organizing information [3]. CIDOC classes are mapped to the tags used by the NER. One-to-one mapping is easy to process, but complex mapping (e.g. Measurement) requires a language to define transformation rules, ideally performed as automated processes.

Assisted triples generation: RDF is a property-centric (rather than record-centric) approach to representation. The domain of a property is used to indicate that a particular property applies to a designated class [2]. The range of a property is used to indicate that the values of a particular property are instances of a designated class [2]. Thanks to NER and class recognition, domain and range of each potential triple is known. A suitable representation of the CIDOC CRM ontology is required to provide the user with the possible choices.

We use Platypus (http://cassoulet.univ-brest.fr/mme), a model-driven engineering tool, to specify source (NER tagging) and target (CIDOC CRM) meta-models and transformation rules.

We have studied the utility of the solution prototype with the help of human experts involved in creation of the semantic wiki pages. We have observed that the efficiency of creating SMW pages is increased by more than 60% on an average in terms of the time taken. Without the help of our prototype the human experts needed approximately 650 minutes for converting each page (a typical page consists about one hundred links or entities to deal with), which is now reduced to 275 minutes. Considering the tedious nature of this work of manually converting pages to semantic wiki pages, the value provided by the prototype tool is very well appreciated.

5. Discussion and Future Work

In this paper we have proposed a semi-automatic information extraction based approach to transform wiki pages into semantic media wiki pages in the domain of scientific museums. This work is a beginning of an exciting interdisciplinary research that combines ontology engineering, information extraction and museum sciences. We have reported the approach and promising initial subjective evaluation results. We are in the process of improving the prototype and conduct more objective evaluation of the system.

We plan to focus on multilingual aspect in the immediate future. We plan to exploit the structural aspect of the wiki and SMW pages to optimize the conversion process so that the new languages added to the system take much less time and effort.

6. References

[1] Krötzsch, M. and al. 2007. Semantic Wikipedia. Web Semantics: SSA on the WWW. 5, 4 (Dec. 2007), 251-261

[2] W3C. 2004. RDF/XML Syntax Specification (Revised). http://www.w3.org/TR/rdf-syntax-grammar/ (June 07, 2th)

[3] ISO/IEC 21127:2006, "A reference ontology for the interchange of cultural heritage information", ISO, Geneva.

[4] Richman, A. E. and Schone, P. 2008. Mining Wiki Resources for Multilingual Named Entity Recognition. In Proceedings of ACL'09: HLT, 1-9.

[5] Kousetti, C., Millard, D. E. and Howard, Y. 2008. A Study of Ontology Convergence in a Semantic Wiki. In Proceedings of the WikiSym 2008. ACM, New York, NY.

[6] Lafferty, J., McCallum, A. and Pereira, F. 2001. Conditional random: Probabilistic models for segmenting and labeling sequence data. In 18th Proc. of the Int. Conf. on Machine Learning, 282-289. Morgan Kaufmann, San Francisco.