Preserving meanings in multilingual text mining for Cultural Heritage

Michel Généreux and David Arnold
Computing, Mathematical and Information Sciences (CMIS)
University of Brighton, United Kingdom
{M.Genereux,D.Arnold}@brighton.ac.uk

Extended abstract

In an area of unprecedented wealth of information on almost anything available on-line at no cost, it is important to develop ways of harnessing these sources to our needs. This is even more important as the Cultural Heritage (CH) domain is likely to adopt CIDOC-CRM as its ontological model to classify that information. Despite digitalisation becoming pervasive and people getting more and more used to exchange information in a compact manner, Natural Language (NL) remains the favourite way of communication among humans. For the domain expert, NLs provide the redundancy to express subtleties ; for the novice, it removes the constraints imposed by formal languages. We first report on experiments showing how to extract CIDOC-CRM triples from NL documents in English and, as this state of affairs is somewhat complicated by the additional mapping required to cater for different languages, we set out to extend our approach to include more than one language. Results from this work are important to evaluate the effectiveness of the extraction method and to assess how portable it is across different languages.

CIDOC-CRM is a new standard for encoding a wide range of information for CH. The CIDOC-CRM ontology aims at accommodating a wide variety of data from the CH domain, but it sheer complexity may make it difficult for non-expert to learn it quickly, let alone use it efficiently. For others, it may even be simpler to find a way to translate automatically their data from the storage mechanism already in place into CIDOC-CRM. For practitioners unfamiliar with tight formalisms, it may be more natural to describe collections in NL (e.g. English). At present, existing CH collections are stored using all sorts of formats, sometimes proprietary, often defined roughly, which makes it difficult to share or access heterogeneous information among the CH community. To add to complexity, these diverse collections can be described in various languages, which makes comparative evaluations of mapping tools into CIDOC-CRM a difficult problem. This paper is an attempt to tackle this evaluation problem, by building up on our previous work in using Language Technology (LT) to extract CIDOC-CRM triples from texts.

Part I : Triple extraction from texts Wouldn't it be practical to be able to describe a collection of artifacts in plain English, with little or no knowledge of the CIDOC-CRM formalism, and use LT take over and produce a CIDOC-CRM database ? There is a need for a tool to build fragments of a CIDOC-CRM database directly from informal descriptions in NL, as the CH community may be reluctant to switch to new formats of data entry. Therefore, the first part of this presentation focus primarily on the mapping of CH data described in NL into CIDOC-

CRM triples, the building blocks of the full CIDOC-CRM ontology. The method exploits the propositional nature of CIDOC-CRM triples. Using WORDNET as a lexical database and the WEB as corpus, we first extract triples from examples provided in the CIDOC-CRM literature, and then from text describing the medieval city of Wolfenbüttel. We show the strong points of the system and point out where and how it could be improved. Although the triples extracted automatically from texts do not provide a full picture of the CIDOC-CRM structure buried in the textual description, our results indicate that it provides a sound initial working basis for the mapping/translation process, saving time on what would otherwise have to be done by hand. The method is based on the idea that triples have a predicative nature, which is structurally consistent with the way NLs are build. According to the documentation on the CIDOC-CRM website[1] :

> The domain class is analogous to the grammatical subject of the phrase for which the property is analogous to the verb. Property names in the CRM are designed to be semantically meaningful and grammatically correct when read from domain to range. In addition, the inverse property name, normally given in parentheses, is also designed to be semantically meaningful and grammatically correct when read from range to domain.

A triple is defined as :

DOMAIN PROPERTY RANGE

The domain is the class (or entity) for which a property is formally defined. Subclasses of the domain class inherit that property. The range is the class that comprises all potential values of a property. Through inheritance, subclasses of the range class can also be values for that property. Example 1 illustrates how triples can be extracted from NL.

(1) *Rome* *identifies* *the capital of Italy.*
DOMAIN E41 PROPERTY P1 RANGE E1
E48 :Place Name P1 :identifies E53 :Place
'Rome identifies the capital of Italy.'

The task of the NL processing tool is to map relevant parts of texts to entities and properties in such a way that triples can be constructed. In a nutshell, the Noun Clauses (NC) *Rome* and *the capital of Italy* are mapped to *Entity 48* and *Entity 53* respectively, themselves subclasses of the domain E41 and range E1 respectively, while the Verb Clause (VC) *identifies* is mapped to *Property P1*.

Part II : Comparative evaluation for other languages Extending the approach to other languages, we must provide multilingual thesauri for each CIDOC-CRM entities and properties, as well as a set of rule mappings for constituents between languages. For the comparison to be valid, we limit ourselves to semantically equivalent texts in languages for which the quantity and quality of resources are similar. This yields to a set of three languages : English, French and German. We wish to compare how well meaning is preserve for languages other than English, given that CIDOC-CRM triples are, at the origin, built around the English language.

[1] See http ://cidoc.ics.forth.gr/