Structured and Unstructured: Extracting Information from Classics Scholarly Texts

Romanello, Matteo
matteo.romanello@kcl.ac.uk
King's College London

The poster presents an ongoing PhD research project that applies Digital Humanities to Classics. The project is focussed on the extraction of information from modern scholarly texts (i.e. “secondary sources”), namely all the modern publications about ancient works written in Greek or Latin (being our so-called “primary sources”). The project addresses both the problem of extracting information fom scholarly texts in an automatic and scalable way, and that of providing users (i.e. scholars) with advanced and meaningful entry points to information rather than just search engine-like functionalities over an electronic corpus of texts.

1. Background

A currently ongoing project such as the Million Book Library drew considerable attention to the characteristic features of the next generation of digital libraries and to the consequences of a change of scale on the practice and the results of the research itself (Crane 2006). This is a big chance for Humanities and Classics particularly given the extended “shelf-life” of humanities relative to scientific publication. However, once those secondary sources are digitized this does not mean that information contained within them will be immediately accessible. Issues that we need to address concern the accuracy of our electronic resources, such as encoding of Greek text, inaccurate OCR transcription due to low image quality, problem of missing pages (Boschetti 2009), as well as the scalability of ways to provide access to information, given that we cannot afford to correct manually every scanned page.

2. Motivations

Secondary sources are without any doubt intrinsically valuable for Classicists, as they shape the scholarly discourse of the discipline. Printed citations and references – and even mentions of names or geographical places – can be considered as being already a form of hypertext as they virtually create links between texts. However in the currently available digital libraries – except for the Perseus Digital Library[1] – primary and secondary sources are scarcely interconnected, despite the fact that one of the main advantages of digital libraries is the way in which they represent the hypertextual nature of text collections.

In particular, Classics scholars are interested in named entity mentions inside texts, as is reflected by the widespread use of different kinds of indexes and concordances in this field that basically organize in a systematic manner references to text passages when a given entity is mentioned. Translating the problem into computational terms, we are faced with the task of automatic entity extraction from a corpus of unstructured texts to develop a discipline-specific system of semantic information retrieval.

3. Related Work

In addition to all the unstructured information being made available on the web, several projects in the Humanities have produced over the last decade an increasing amount of structured information that was then stored in a wide range of data formats (i.e. databases, XML fles, etc.). The approach we undertake is to reuse information contained within those structured data sources as training data for a supervised system that extracts semantic information fom an unstructured corpus of texts. A likely approach is suggested for instance by a research recently conducted by IBM to investigate the automatic creation of links between structured data sources (i.e. database containing product information) and unstructured texts (i.e. emails of complaint about products) (Bhide et al. 2008). More generally, the problem implied by our approach of determining when two bits of information refer to the same entity has been thoroughly explored in the AI (Artifcial Intelligence) field (Li et al. 2005).

4. Method

The very first phase of the project is devoted to the task of building our corpus of unstructured texts. So far we considered two corpora of texts: the papers contained in the open archive Princeton/Stanford Working Papers in Classics[2] (Josiah Ober et al. 2007) and the articles published by the journal Lexis[3] available online under an open access policy. Although the texts are already available in electronic format, some pre-processing is needed – particularly for the sequences of Greek text contained – before we can start extracting information.

In the second phase we integrate different structured data sources into a single knowledge base. As far as this task is concerned, it is possible to observe at least two main categories of lack of interoperability. The first category consists of cases where entities that are similar from an ontological perspective (e.g. a geographical place name) are encoded using different data structures (i.e. the same place name could be encoded using elements belonging to diferent XML dialects). The second category covers cases where chunks of information common to more than one collection are described with different degrees of depth and precision. For instance, the name “Alexandria” inside an inscription is just marked up as a place name where instead a collection of geographical data offers many details for the city of Alexandria such as coordinates, orthographical variants of the name, denomination in diferent languages etc. For this purpose we mean to apply high level ontologies to aggregate information related to the same entity but spread over diferent data sources. Among the most suitable ontology vocabularies are worth to be mentioned FOAF,[4] CIDOC CRM,[5] FRBRoo[6] and YAGO.[7]

At a further stage we are taking into account how to automatically extract information from the corpus. We are mainly interested in extracting: 1) named entities; 2) bibliographic references; 3) canonical references, namely references to ancient texts expressed in a concise form and characterized by a logical reference scheme (e.g. based on references to books or lines of a work instead of page numbers).

The very first step in named entities processing is the recognition and identification of named entities within texts. Once identified the named entities should be classified and then disambiguated on the basis of the context. This task will be accomplished through the comparison of semantic spaces using methods and algorithms developed in the field of Latent Semantic Analysis (LSA) (Rubenstein & Goodenough 1965; Sahlgren 2006). In particular the semantic spaces of all the contexts where a given named entity appear will be compared with each other in order to determine which resources are really referring to the same entity.

For the information extraction task we use mainly tools based on machine learning methods, such as Conditional Random Fields (CRF) or Support Vector Machine (SVM). Some of them are already available as open source sofware and just need to be trained to work with our data, while others are being specifcally developed such as a Canonical Reference Extractor (Romanello et al. 2009). Information contained in the knowledge base is meant to be used as training material for those sofware components.

5. Further Work

This ongoing project is expected to prove a scalable approach to extract entity references from unstructured corpora of texts, such as OCRed materials. In addition to this, it will prove how to reuse several data sources of structured information to train text mining components and it will show to what extent those data sources can be made interoperable with each other. Finally we plan to evaluate with some experts how effective such an information retrieval system is in helping Classicists with their research.

Acknowledgement
This research project is supported by an AHRC (Arts and Humanities Research Council) award.

References

Bhide, M. et al. (2008). 'Enhanced Business Intelligence using EROCS'. *Data Engineering, 2008. ICDE 2008. IEEE 24th International*

Conference on Data Engineering, 2008. Pp. 1616-1619. http://ieeexplore.ieee.org/iel5/4492792/4497384/04497635.pdf?tp=&arnumber=4497635&isnumber=4497384.

Boschetti, F. (2009) (2009). 'Improving OCR Accuracy for Classical Critical Editions'. *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*. Springer.

Crane, G. (2006). 'What Do You Do with a Million Books?'. *D-Lib Magazine*. **12(3)**. http://www.dlib.org/dlib/march06/crane/03crane.html (accessed March 19, 2009).

Josiah Ober et al. (2007). 'Toward Open Access in Ancient Studies: The Princeton-Stanford Working Papers in Classics'. http://www.atypon- link.com/ASCS/doi/abs/10.2972/hesp.76.1.229 (accessed July 15, 2009).

Li, X., Morie, P., Roth, D. (2005). 'Semantic integration in text: from ambiguous names to identifiable entities'. *AI Mag*. **26(1)**: 45-58. http://portal.acm.org/citation.cfm? id=1090494 (accessed March 12, 2010).

Romanello, M., Boschetti, F., Crane, G. (2009). 'Citations in the digital library of classics: extracting canonical references by using conditional random fields'. *NLPIR4DL '09: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*. Morristown, NJ, USAAssociation for Computational Linguistics, pp. 80–87.

Notes

1. http://www.perseus.tufts.edu/hopper/
2. http://www.princeton.edu/~pswpc/
3. http://lexisonline.eu/
4. http://www.foaf-project.org/
5. http://cidoc.ics.forth.gr/
6. http://cidoc.ics.forth.gr/frbr_inro.html
7. http://www.mpi-inf.mpg.de/yago-naga/yago/