THE HUMANITIES RESEARCH PORTAL: Human Language Technology Meets Humanities Publication Repositories

Massimo Poesio Uni Essex, UK Uni Trento, Italy Eduard Barbu Francesca Bonin Fabio Cavulli Egon W. Stemle Uni Trento, Italy

Asif Ekbal Sriparna Saha IIT Patna, India Christian Girardi Filippo Nardelli FBK, Italy Expert Systems / Cogito Italy

Abstract

The Portale della Ricerca Umanistica Trentina (Humanities Research Portal) will be a one-stop search facility for repositories of articles on about Humanities subjects about the History, History of Art, and Archaeology of Trentino that relies on automatically extracted entity-, spatial- and temporal- metadata to provide entity-based, spatially-based and temporallybased access to the articles. In this article we discuss the aims of this project and the current state of work.

1 Introduction

Content Extraction methods are playing an increasing role in digital libraries, particularly in disciplines like Computer Science and Computational Linguistics where research publications are routinely included in public repositories such as the ACL Anthology¹ or at least they are accessible from the researchers' web pages so that they can be analyzed e.g., to extract the citation statistics collected by services such as Google Scholar. More advanced techniques such as semantic tagging of text with NE tags (e.g., identifying temporal expressions) or tagging text with Wikipedia 'concepts' are not yet widely used but are the subject of great interest.

Such techniques ought to play an even greater role in digital libraries for the humanities, as the types of browsing that content extraction is meant to support play an even more important role there. For instance, the type of spatial browsing that geotagging could to support is not very relevant to index publications in Computer Science or Computational Linguistics, but it is critical for domains like Archeology where **sites** are one of the entities of scholarly relevance, illustrated by the APSAT / ALPINET portal.² At present however this metadata is typically added by hand. Another example is temporal browsing on the basis of the temporal metadata extracted from the content of the publication (as opposed to temporal browsing based on the date of publication). Again this type of browsing can be of limited utility when searching scientific or engineering publications but it is fundamental when searching for historical publications covering a specific period.

There are however several problems preventing the application of content extraction techniques to humanities publication repositories. For one thing, repositories offering the possibility to carry out even basic searches of research publications are not as widespread in the Humanities as they are in Computer Science in general and Computational Linguistics in particular. The current situation in Trentino is fairly typical: institutions such as the Museo Tridentino di Scienze Naturali host small or medium size collections of publications (typically published by the institution itself) which are often only partly digitized, but by and large the collections are still in paper form. In addition, developing content extraction techniques for the humanities requires developing domain adaptation techniques as the entities in these domains, apart from temporal and spatial, tend to be non-standard.

The Portale della Ricerca Umanistica in Trentino³ (Humanities Research Portal, PRU) is a pilot project under development in Trentino to set up a one-stop search facility for repositories of research articles and other types of publications in the Humanities. The portal uses content extraction techniques to automatically extract citations and metadata including temporal, spatial, and entity references from the publications being uploaded. It currently provides access to the Archeological articles in the APSAT / ALPINET repository; work is underway to provide access to a second repository of Middle Ages History publications. This paper will discuss the functionalities offered by the portal as well as the content extraction techniques employed.

¹http://acl.ldc.upenn.edu/

²http://alpinet.mpasol.it/webgis/

³http://www.portalericercaumanistica.org

The structure of the paper is as follows. In Section 2 we discuss the repositories in the process of being indexed. In Section 3 we discuss the functionality of the portal. Finally, in Section 4 we discuss how semantic metadata are automatically extracted from the text of the articles in the repositories.

2 Repositories of Humanities Research in the Trentino

Trentino is in the fortunate position of having a great number of institutions–Museums, publishers, and academic centres–operating in various areas of the Humanities. These institutions hold, in addition to extensive archives of primary sources, collections of journals and / or other scholarly publications, often published by the institutions themselves, each with a focus reflecting the interests of the holding institution. These collections are also very diverse in terms of their digitalization and / or accessibility through computer-based search, ranging from the purely paper-based to collections taking full advantage of modern IT technology, such as the APSAT / ALPINET portal.

This wealth and diversity of material is extremely attractive, but would clearly benefit from a one-stop point of access, and in particular for one establishing link between the content of such collections. The Portale della Ricerca Umanistica in Trentino, an initiative of the University of Trento's Dipartimento di Filosofia, Storia e Beni Culturali, is aimed at providing this point of access and linking facilities through content extraction.

Clearly realizing this vision will require addressing a number of technical, administrative, and political issues; we decided therefore to start with the easiest cases– the repositories which are already completely digitized– while resolving the issues required to make other repositories accessible as well.

2.1 The APSAT / ALPINET Portal and Collection

The first repository whose documents have been made accessible through the PRU is the collection of articles in the Archaeological domain in the APSAT / ALPINET digital library.

The APSAT / ALPINET portal is a pilot Spatial Humanities project developed by the Anonymous Archaeology Lab and allowing scholars to visualize Archaeological sites in the Alps through a Web GIS interface, through which Scholars can examine an area in general to find which sites are present, or look in detail at the features of a particular site. Through the portal, scholars also can access a collection of Archaeological articles about these sites, which can be searched either through keywords or by clicking on a site through the Web GIS interface.

In particular, the library contains a complete collection

of the journal *Preistoria Alpina* published by the Museo Tridentino di Scienze Naturali. We will focus on this collection in the present work. The collection is multilingual, containing articles written in English, French, German and Italian; in fact, as typical of the Humanities, many articles are themselves multilingual, in that they contain, in addition to text in the main language, an abstract, keywords, and occasionally captions in a second language, often but not always English.

3 Functionality

The portal will use content extraction techniques to automatically extract citations and semantic metadata including temporal, spatial, and entity references from the publications in those repositories. This information will then be used to offer visitors to the portal two main functionalities: **content-based search and browsing** and **semantic uploading**.

Besides standard keyword-based search, the PRU will also offer **entity-based search**. Two types of browsing will be possible: **spatial** and **temporal** browsing. Entity search allows users to retrieve all documents that discuss a particular entity irrespective of the way it's called–e.g., all Archaeological documents that discuss sites in which a particular shellfish was found irrespective of whether it's called in the document *Spondylus sp.* or *Spondilo*. Spatial browsing allows users to retrieve the publications that mention a particular locality in Trentino by visualizing a map of Trentino and clicking on the appropriate location, as illustrated in Figure 1.

Temporal browsing (currently under development) will allow users to retrieve all scholarly articles discussing a particular period (i.e., the temporal organization is not according to the time of publication of the articles, as in Google timeline for instance).

These novel types of searching and browsing will be supported by a **semantic upload function**: registered scholars and / or curators of the collections will be able to upload publications that will then be processed by the PRU pipeline discussed below to automatically extract both metadata and information about the publication to be inserted in the catalogue of the repository after being checked by the curator.

One of our goals was to facilitate the process of digitalization by automatizing the process of upload so that whenever a librarian uploads an article in this digital library, title, authors, abstract, keywords, semantic metadata, and bibliographic references from the article are automatically identified. The implemented portal already incorporates information extraction techniques that are used to identify in the 'content' part of the output of the pipeline temporal expressions, locations, and entities such as archeological sites, cultures, and artifacts. This



Figure 1: Spatial Browsing

information is used to allow spatial, temporal, and entitybased access to articles.

We are in the process of enriching the portal so that title and author information are also used to automatically produce a bibliographical card for the article that will be entered in the PRU Library Catalog, and bibliographical references are processed in order to link the article to related articles and to the catalog as well. The next step will be to modify the pipeline (in particular, to modify the Named Entity Recognition component) to include in the library articles from other areas of research in the Humanities, starting with History. There are also plans to make it possible for authors themselves to insert their research articles and books in the Portal, as done e.g., in the Semantics Archive.⁴.

4 Content Extraction

4.1 A Structure-Sensitive, Multi-Lingual Pipeline

The articles to be made accessible through the PRU are processed by a pipeline that tokenizes, POS-tags, and NE tags the text in order to extract semantic indices (Poesio et al., 2011). The pipeline, accessible as a Web service, is based on the TEXTPRO pipeline⁵ (Pianta et al., 2008) and is in most respects similar to other HLT pipelines except for two distinguishing features.

First, it is **structure sensitive**, in the sense that it includes a module that identifies the structure of a document to find citations and the like, in the manner of the FlyBase pipeline (Briscoe, 2011). Second, it is **constituent-level multilingual**, in that each constituent of the document structure is first run through a language identifier in order to find which version of the TEXTPRO system should be run on that constituent. (English and Italian are supported at the moment.) The first version of the pipeline included the default TEXTPRO NE tagger, ENTITYPRO, trained to recognize the standard ACE entity types. The objective of this work was to create a corpus that could be used to train a new NE tagger able to recognize the relevant entities in the APSAT / ALPINET collection.

The architecture of the system is presented in figure 2. Each module except the first one takes as input the output of the previous module in the sequence.

4.2 Domain Adaptation

Many of the entities mentioned in collections of scholarly articles in subjects such as Archaeology, History, or History of Art do not belong to the types found in the news corpora on which Computational Linguistics work has focused, such as the MUC and ACE corpora. For instance, the most important entity types found in archaeological texts are Culture, Site, and Artefact. In some such domains, even if more familiar types such as Person play an important role, it is essential to distinguish between their subtypes. E.g., in History of Art articles, it is not enough to classify an entity as a Person; it is also crucial to recognize if a particular individual was a Painter, a Sculptor, an Architect, etc. Hence, dedicated resources need to be created to train Named Entity (NE) recognizers for these domains; training on news corpora is of limited use to extract semantic content from such articles.

However, creating resources is always expensive, and Humanities projects tend not to have lots of funding for these purposes. In addition, collections of articles in the Humanities tend to be fairly small. It is therefore essential to use the limited funding available wisely, and to maximize the benefit to be obtained from the data. In other words, this is a domain for which **active learning** techniques (Settles, 2009), already used for NE tagging by, e.g., Vlachos (2006), seem ideally suited. The pipeline used for the archaeological domain in the current version of the PRU uses a CRF-based NE tagger trained by using active annotation. The current version of the NE tagger achieves an accuracy of around .7 (F value) on the Archaeology tagset.

5 Related Projects

There is widespread interest in new ways of using ICT to improving fruition of scholarly articles in Humanities subjects and / or linking these with primary sources.

The APSAT / ALPINET portal is but one example of a number of portals taking advantage of Geographic Information Systems (GIS) to offer scholars in fields such as Archaeology, but also in other fields, such as History,⁶ the possibility to recreate historical events in new and illuminating ways, leading some to talk of an emerging field of **Spatial Humanities**.⁷ In some projects, such as Europeana 4D,⁸ spatial and temporal perspectives are simultaneously displayed.

One project using content extraction techniques to provide access to resources in the Humanities is the *Panora-maFIRB* project at the Computer Science Department of the University of Pisa,⁹ that is developing a web service and search engine to support semantic and multilingual access to Italian Culture on the Web. This service automatically searches high quality websites selected by domain experts for relevant domain-related terms, and relevant named entities are extracted in order to offer entity based search.

⁴http://semanticsarchive.net/

⁵http://textpro.fbk.eu/

⁶www.virginiaexperiment.com/podcast/

speakerSeries041708.mp3

⁷http://spatial.scholarslab.org

⁸http://tinyurl.com/e4d-projct

⁹http://panoramafirb.it/



Figure 2: The pipeline of the system for PDF article processing in the Archeology Domain

6 Future Developments

We hope that the work already carried out on providing access to the APSAT / ALPINET material will facilitate extending the portal to provide access to additional collections, in particular to those that are already digitized. It is our hope that in many cases, the main work required will be to retrain our NE taggers as done for the Archaeology domain.

But another good reason for applying content extraction techniques to Humanities research is that in many domains of this type it is possible to use semantic metadata link publications and (textual) primary sources. This is going to be one of the foci of the research in the near future. Pilot studies have been carried out to provide access to a digital collection on Trentino's History in the Middle Ages which also includes digitized versions of a number of primary sources.

Acknowledgments

This work was in part supported by the Provincia Autonoma di Trento, Grande Progetto *LiveMemories*

References

- E. et al. Briscoe. 2011. Intelligent information access from scientific papers. In J. Tait et al, editor, *Current Challenges in Patent Information Retrieval*. Springer.
- E. Pianta, C. Girardi, and R. Zanoli. 2008. The textpro tool suite. In *Proc. of 6th LREC*, Marrakech.
- M. Poesio, E. Barbu, E. Stemle, and C. Girardi. 2011. Structure-preserving pipelines for digital libraries. In *Proc.* of *LaTeCH*, Portland, OR.
- B. Settles. 2009. Active learning literature survey. In In Computer Sciences Technical Report 1648 University of Wisconsin-Madison.
- A. Vlachos. 2006. Active annotation. In *Proc. EACL 2006 Workshop on Adaptive Text Extraction and Mining*, Trento.