**POIBEAU — Visualizing**

<0 images>

**Visualizing Data for Digital Humanities: Producing Semantic Maps with Information Extracted from Corpora and Other Media**

Poibeau, T., Tellier, I. and Terras, M.

**Brief Description of the Content and Its Relevance to the Digital Humanities Community**

The goal of this proposal is to explore efficient methods to extract information from texts and produce meaningful and readable representations. The three-hour session (half a day) will include formal presentations, demonstrations, and discussion with the audience. It will thus be halfway between a workshop and a tutorial (see the structure of the session in the next section).

It is well known that we are now facing an information deluge, and experts in different domains, especially social sciences and literary studies, have long acknowledged that available texts—and more generally the mass of data available through different media—now constitute one of the important sources of knowledge. However, computers are unable to directly access information encoded through texts: this information must first be extracted, normalized, and structured in order to be usable. Moreover, meaningful representations are needed in order for people to understand it. This process is not trivial, and more and more groups have to face this dilemma: information is here, available on the Web or in more remote databases, but its manipulation is hard since it requires a complex process that most of the time is out of the hands of social scientists and literary studies experts. See, for example, this quotation (from the Médialab in Paris) that is typical of the current situation:

Qualitative researchers [. . .] arrive at the médialab bringing rich data and longing to explore them. Their problem is that qualitative data cannot be easily fed into network analysis tools. Quantitative data can have many different forms (from a video recording to the very memory of the researcher), but they are often stored in a textual format (i.e. interview transcriptions, field notes or archive documents . . .). The question therefore becomes: how can texts be explored quali-quantitatively? Or, more pragmatically, how can texts be turned into networks? (Venturini et Guido, 2012)

The goal of this workshop is to practically address the question. It will include three presentations detailing some challenges and solutions. A large part of the workshop will be devoted to the presentation of practical tools and to discussion with the audience.

**Structure of the Workshop**

The workshop will be chaired by Thierry Poibeau, Melissa Terras, and Isabelle Tellier. It will consist of talks by Pablo Ruiz, Steven Gray, and Glenn Roe.

Pablo Ruiz will address the concept of entity linking. In natural language processing, entity linking is the task of determining the identity of entities mentioned in text. It includes named entity recognition (NER) and the identification of their reference (via links to DBPedia or other structured databases). The bases of entity linking will be presented along with the demonstration of a platform combining different entity linking systems so as to obtain an efficient and robust coverage across text genres and application domains. Lastly, the production of maps from the result of the entity linking analysis will be presented.

Steven Gray will address the collection and visualisation of real-time online datasets. By utilising the Big Data Toolkit (a toolkit that specialises in mining social media data), the session will focus on the collection of data from various APIs (Application Programming Interfaces) and the mapping of the data by using Google Maps’ online mapping APIs. The session will conclude by focussing on visualising this real-time textual data for the Web, building an interactive discovery tool for geospatial social media data.

Glenn Roe will address the identification and visualisation of text reuse in unstructured corpora. Identifying text reuse is a specific case of the more general problem of sequence alignment—that is, the task of identifying regions of similarity shared by two strings or sequences, often thought of as the longest common substring problem. This technique is widely applied in the field of bioinformatics, where it is used to identify repeated genetic sequences. This talk will outline several different approaches to sequence alignment techniques for humanities research, as well as two recent projects aimed at visualising the output of alignment comparisons between texts and the alignment process itself using visual analytics.

The structure of the workshop will be as follows: each talk will consist of 20 minutes for presentation, 15 minutes for demonstration, and 15 minutes for discussion. A final slot (30 minutes) will be devoted to a general discussion on the topic with the audience.

**Workshop Leaders**

**Thierry Poibeau**

Laboratoire LATTICE

Ecole Normale Supérieure & CNRS

1, rue Maurice Arnoux 92120 Montrouge FRANCE

Mail: thierry.poibeau@ens.fr

Web: http://www.lattice.cnrs.fr/Thierry-Poibeau

*Thierry Poibeau*is a director of research at CNRS and the head of the CNRS Lattice laboratory at Ecole Normale Supérieure. He is also an Affiliated Lecturer at the Department of Theoretical and Applied Linguistics (DTAL) of the University of Cambridge. His main interests are natural language processing (NLP), and its application to digital humanities. He is a recognized expert in information extraction, question answering, semantic zoning, knowledge acquisition from text, and named entity tagging. He is the author of three international patents and has published two books and more than 80 papers in journal and conferences.

**Melissa Terras**

Director, UCL Centre for Digital Humanities

Vice Dean of Research (Projects), UCL Faculty of Arts and Humanities

Professor of Digital Humanities

Department of Information Studies

Foster Court

University College London

Gower Street

WC1E 6BT

*Melissa Terras*is director of the UCL Centre for Digital Humanities, professor of digital humanities in UCL’s Department of Information Studies, and vice dean of research (Projects) in UCL’s Faculty of Arts and Humanities. With a background in classical art history, English literature, and computing science, her doctorate (engineering, University of Oxford) examined how to use advanced information engineering technologies to interpret and read Roman texts. Publications include *Image to Interpretation: Intelligent Systems to Aid Historians in the Reading of the Vindolanda Texts* (2006, Oxford University Press) and *Digital Images for the Information Professional* (2008, Ashgate), and she has co-edited various volumes, such as *Digital Humanities in Practice* (Facet, 2012) and *Defining Digital Humanities: A Reader* (Ashgate, 2013). She is currently serving on the Board of Curators of the University of Oxford Libraries, and the Board of the National Library of Scotland. Her research focuses on the use of computational techniques to enable research in the arts and humanities that would otherwise be impossible. You can generally find her on twitter @melissaterras.

**Isabelle Tellier**

Laboratoire LATTICE

Ecole Normale Supérieure & CNRS

1, rue Maurice Arnoux 92120 Montrouge FRANCE

Mail: isabelle.tellier@univ-paris3.fr

Web: http://www.lattice.cnrs.fr/sites/itellier/

*Isabelle Tellier*is a professor of natural language processing and digital humanities at Université Paris 3 Sorbonne Nouvelle. She has published more than 50 papers in international venues in computational linguistics. Her main research topic is developing advanced machine learning techniques adapted to Natural Language Processing issues. She is also an expert in the application of these techniques to domain-specific corpora, especially in the digital humanities field.

**Workshop Main Speakers**

*Pablo Ruiz*is a research associate in digital humanities at LATTICE, a research laboratory of Ecole Normale Supérieure in Paris. He has designed and developed several applications in the domain of natural language processing, including, for example, a prototype for the lexical normalization of Spanish microblog (tweets), and a sequence alignment engine used in automatic subtitling. He currently works on text analytics technologies for digital humanities, especially entity recognition and linking, so as to extract structured and meaningful pieces of information from unstructured texts.

*Steven Gray*is a teaching fellow in Big Data Analytics and the UCL Centre for Advanced Spatial Analysis. With over 10 years’ professional software engineering under his belt, and a background in computing science and human computer interaction, he has built multiple award-winning systems, and his work has been featured in various worldwide media outlets (CNN, BBC, etc.). In recent years he has specialised in building mobile applications and systems that open up the world of data visualisation, mining, and analysis to the masses. His current research focuses on distributed high-performance computing, machine learning, and analysing large datasets in real-time while visualising the results. He currently serves as a Google Developer Expert for Google Maps where he evangelises best practises using cloud computing and online mapping platforms. You can regularly find him either on Twitter, @frogo, or on Google Plus as +StevenGray.

*Glenn Roe*is lecturer in digital humanities at the Australian National University. In 2011–2013 he held a Mellon Post-Doctoral Fellowship in digital humanities at the University of Oxford. Prior to that, he spent eight years as a senior project manager for the University of Chicago’s ARTFL Project (American and French Research on the Treasury of the French Language), one of the older and better-known North American research and development centres for computational text analysis. Glenn’s research agenda is primarily located at the intersection of new computational approaches with traditional literary and historical research questions. Glenn has presented and published widely on a variety of scholarly subjects, from French literary and intellectual history, to the design and use of new digital methodologies for literary research, and the implications of large-scale digital collections on humanities scholarship.

**Target Audience**

The target audience is large. Practitioners in social sciences are of course the key target, but experts in literary studies, information technology, and library management are concerned as well. In brief, the workshop will be of interest for anybody concerned with the extraction of information from large corpora and producing relevant maps to visualize key information.

**Special Requirements for Technical Support**

None, except a conference room projector.