AHRC ICT Methods Network Workshop

TEXT MINING FOR HISTORIANS

UNIVERSITY OF GLASGOW, 17 – 18 JULY 2007

Report by Ian Anderson and Zoe Bliss

Background

Texts are central to historical research, however the use of computer assisted methods and tools remains remarkably underutilised by historians. Despite a long-standing interest in computer-aided text analysis historians continue to benefit only indirectly, largely through work conducted in other disciplines. As evidenced by the successful Methods Network workshop in Historical Text Mining at Lancaster University in July 2006 (see http://www.methodsnetwork.ac.uk/activities/act6report.html) the tools and methods being developed and used by corpus linguists have become increasingly sophisticated. At the same time a larger than ever body of historical texts are becoming available in electronic format. Building upon the Historical Text Mining Workshop this workshop aimed to publicise these tools and techniques to historians to encourage their use and to also explore how they could, or needed to, be adapted to more effectively meet their needs.

Introduction

This two-day workshop was held on 17 and 18 July 2007 at the University of Glasgow. It was organised by Zoe Bliss from AHDS History at the University of Essex and Ian Anderson convenor of the Association for History and Computing (UK) based at the University of Glasgow.

The workshop was attended by twenty researchers including both linguists and historians with the latter predominating and whose research interests ranged from early medieval British History to twentieth century Britain.

Content of the workshop

The first day of the two day programme comprised of presentations from invited speakers followed by discussions. The second day consisted of demonstrations of particular tools and hands-on tutorials for these tools, followed by a plenary session considering the strengths and weaknesses of these techniques for historians.

Day One

The first presentation was from Paul Rayson (University of Lancaster) and Dawn Archer (University of Central Lancashire) who provided an 'Introduction to Corpus annotation and retrieval' that stressed the importance of the representativeness and size of corpora, the way that corpus linguists develop and use corpora to investigate synchronic and diachronic variation, syntax, semantics, pragmatics, lexicography, and dialects. They also outlined techniques used both in terms of annotation and retrieval and some of the problems these techniques might pose to historians such as spelling variations, archaic terms and forms of words and the need to use period (and subject) specific taxonomies or thesauri. They also provided introduction to the Variant Spelling Detector (VARD) being developed at the University of Lancaster which is a tool that which would automatically regularise spelling variants with a text to their modern forms so that historical corpora becomes more amenable to further annotation and analysis. In addition they provided information on historical corpora already in existence and provided an example of research on analysis of early modern courtroom discourse which employs corpus linguistic techniques to illuminate nature of seventeenth and eighteenth century trials.

Following this presentation which conceived Historical Text Mining as an extension of textual analysis techniques were two presentations that linked Historical Text Mining more firmly with data mining techniques and the extraction of 'facts' as well as patterns from texts. The first of these were from Clare Llewellyn and Rob Sanderson from The National Centre for Text Mining (NaCTeM). Their presentation 'Language Independent Textual Correlation Analysis' outlined the current work being carried out at NaCTEM. Currently the project has largely focused on the Biosciences and has developed a number of tools to help researchers in this community to apply text mining techniques to problems in their areas of interest. Some of these tools could be utilised by historians, particularly to identify patterns in text. They also outlined how historians could possibly use Association Rule Mining (ARM), a technique used in industry to analyse shopping behaviour, with historical texts to identify and interrogate word collocations. However, whilst historians could adapt theses tools and techniques to identify patterns, using data mining techniques to reveal new 'facts' is far more complicated and would require sophisticated natural language processing and development of tools able to deal with the 'language' of historical texts and access to sufficiently large collections of appropriate texts.

Mark Greengrass's (University of Sheffield) presentation 'Armadillo: Data Extraction Across Multiple Text Datasets for Arts and Humanities Research' gave an overview of the Armadillo: Historical Data Mining Project. He outlined how this project aimed to overcome the problem of searching large, heterogeneous and multi-format historical texts by using some of the technologies developed for the semantic web, particularly ontologies.   The project had made it possible to cross search five very different eighteenth century sources, the Sun and Royal Exchange fire insurance policies, the Prerogative Court of Canterbury Wills, the Settlement Examination for the parish of St Martin-in-the-Fields, the Westminster Historical Database and the Old Bailing Proceedings Online. The project had also utilised automatically tagging of one source from historical/semantic information contained in others. This process, however, whilst certainly speeding up the encoding of texts, still requires quite a lot of manual checking and input.

Lastly, Christian Kay and Jean Anderson provided an overview of Historical Thesaurus of English a project that contains vocabulary of English from the earliest written records to the present, with first, and where appropriate, last recorded dates of usage and demonstrated how historical corpora, variant spellings, changes in meaning and word origin can provide a valuable resource for historians as well as linguists.

Day Two

The focus of the second day was practical hands-on tutorials in computer labs where participants were given the opportunity to try out a number of different software packages.

Historical Thesaurus of English and Using Wordsmith

In the first of the tutorials delegates were shown how the Historical Thesaurus of English (http://www.arts.gla.ac.uk/SESLI/EngLang/thesaur/homepage.htm) could be used to look at words in particular areas of meaning or to collect words from particular periods to use in searching texts. Participants were also introduced to Wordsmith (http://www.lexically.net/wordsmith/) and how to use the programme to create wordlists, concordances and collocations.

Introducing Basic Corpus Linguistics techniques using the British National Corpus with View.

In this session attendees were introduced to VIEW (http://corpus.byu.edu/bnc/) which allows searching of a wide range of words and phrases of English in the 100 million word British National Corpus (which represents modern

English of the late 20th century). The session stressed that whilst the tools are useful for locating patterns human interpretation was essential to make sense of these patterns.

Introducing key words and key domains with Wmatrix (Case Studies: (1) Political Party Manifestos; (2) Opening speeches in the nineteenth century courtroom

The third and fourth sessions of the day focused on Wmatrix (http://ucrel.lancs.ac.uk/wmatrix/). Initially Paul Rayson introduced Wmatrix and its capabilities and a case study comparing Liberal Democrat and Labour Party Manifestos for the 2005 UK General Election were used to illustrate its capabilities. Subsequently Dawn Archer demonstrated its use on the opening speeches of the Trial of William Palmer in 1856.

Historical spelling variation detection with VARD

In the final hands on session of the day participants were provided with an overview of VARD (http://ucrel.lancs.ac.uk/VariantSpelling/) which is tool being developed to deal with historical spelling variations and which allows for the detection and normalisation of variants to their modern equivalent whilst retaining the original spelling in the text. Participants were particularly enthusiastic about the way that the tool could be trained to deal with particular types of text and the collaborative research that this both necessitated and promoted.

Outcomes

The workshop was well received and provided the researchers present with a range of different interpretations of what text mining is and an introduction to a number of different tools that they could potentially use in their own research. In the plenary sessions of both days it became clear that despite having different aims and outcomes for their research, significant common ground was established between linguists and historians in their concern for the provenance, reliability and authenticity of texts, the challenges of variant spelling and meaning and the opportunities that text mining tools provided both communities.

Outputs

One of the specific aims of the workshop was to encourage a network expertise. As a result of the workshop a Text mining group has been set up on Digital Arts and Humanities website (http://www.arts-humanities.net/text_mining). In addition the worksheets from the practical sessions will shortly be made available via AHDS History and AHC (UK) websites. Initially it was envisaged that the workshop would lead to production of an AHDS History Guide to Good Practice however the withdrawal of AHRC funding for the AHDS has now made this impossible. However initial discussion have been had about repurposing the presentations from both days of the workshop into a introductory ‘How to Guide’ which will outline current and potential research made possible by text mining techniques and to provide downloadable tutorials to some of the tools that are available.