AHRC ICT Methods Network Workshop

HISTORICAL TEXT MINING

LANCASTER UNIVERSITY, 20 – 21 JULY 2006

Introduction

This workshop was held on 21 and 22 July 2006 in the Infolab21 building at Lancaster University. It was organised and chaired by Paul Rayson (Lancaster University) and Dawn Archer (University of Central Lancashire).

The workshop was a cross-disciplinary forum designed to encourage discussion across a number of different academic communities. Participants' research interests were based in English language and literature, linguistics, and computer science.

Participants were mostly from UK Higher Education institutions and represented a range of experience, from postgraduate students to professors. Representatives from international universities, electronic publishing and a national library also took part.

Content of the workshop

The two-day programme comprised research presentations, discussion groups, tutorials and software demonstrations.

The initial presentations identified current developments in the field with reference to work at the National Centre for Text Mining (NaCTem). Speakers introduced three tools that are currently used in the computational and corpus linguistics communities: GATE (General Architecture for Text Engineering); Wordsmith; and WMatrix. During a hands-on session after the tutorials, participants had the opportunity to explore these tools themselves.

Subsequent presentations considered the difficulties of applying such tools to historical data; and possible solutions to these problems. Some of the projects discussed were:

- a search engine for historical documents developed at Universitat Duisburg-Essen that does not expect users to be language experts
- the potential of using the Historical Thesaurus of English to facilitate text mining using contemporary words
- extensions created for the Corpus of Early English Correspondence (CEEC) that can test the applicability of socio-linguistic methods on historical data
- the easy-to-use interface of nora text mining tools
- the VIEW tool which is an example of the advantages of using a relational database approach to retrieve modern and historical corpora

For the full text of presentations given at the workshop and details of other workshop resources, please see http://www.methodsnetwork.ac.uk/activities/act6.html

Outcomes of the workshop

New directions for research

Participants' comments in the evaluation of the event indicate that the workshop raised awareness among scholars working with historical data about existing text-mining techniques that are applicable to their research needs. Many participants commented that the event would have a significant impact on the future direction of their research; either through the use of new tools and techniques or refining and improving their methodology.

Exchanging knowledge with other communities

The presentations and demonstrations raised awareness about techniques and tools used and developed by researchers working within different fields. Participants already using particular techniques and tools were keen to share their expertise, while those who were involved in tools development gained a better understanding of researchers' needs

Building collaborations

Participants were able to make new contacts and identify possible future opportunities for information-sharing and collaborative work.

What next?

To build on these new community networks, the workshop organizers have begun to develop a network of scholars interested in 'Historical Text Mining'. The network draws on expertise from various fields, including: text mining and e-Science; corpus development and annotation; historical linguistics; dialectology; and computational linguistics.

The recent 'Text Mining for Historians' workshop, funded by the AHRC ICT Methods Network, and which came about as an immediate result of the Historical Text Mining workshop, introduced historians to tools employed by corpus linguists.

Future work will focus on determining what more needs to be done to improve historical text mining. Further exchange of expertise, tools and techniques between the computational and corpus linguistics fields, and historical linguistics and historians was thought to be of primary importance.

The event also confirmed the need to open and maintain dialogues between the various academic and non-academic communities of data users and providers, and software users and developers working in this field in order to improve understanding and ensure continuing collaboration**.**