ISHER – Integrated Social History Environment for Research

http://www.nactem.ac.uk/DID-ISHER/
Sophia Ananiadou

Social historians and other researchers rely to a large extent on text data for their research. These data are increasingly available in electronic form, but researchers can be seriously hampered in discovering information and answers to research questions, given the inadequacy of many existing exploratory tools that facilitate the location of relevant information in the vast numbers of available digital documents. This means that, as before the age of digitization, much manual effort is required to research questions and consequently, many questions remain un(der)answered. In response to this, ISHER has developed a digital humanities toolkit to facilitate *search* in *social unrest and strikes.* Text mining-based search facilitates the exploration and discovery of facts in primary historical sources originating from the digitised historical newspaper archives of the New York Times (NYT) and the National Library of the Netherlands (KB). ISHER facilitates sophisticated semantic searching over the historical newspaper archives to provide social historians and social scientists with the means to detect and associate events, trends, people, organisations and other entities of relevance to their research goals. This functionality is based on the extraction of rich semantic metadata through the application of a number of text mining methods.

Text Mining Annotations enriching data

A number of text analytics tools have been combined to index the data archives for the following types of information:

1. *Named Entities*: a wide variety of named entities are used to index NYT, e.g. Person, Location, Time, Organisation, etc. Automatic semantic metadata derivation from named entities facilitates search.
2. *Events:* events related with social unrest and strikes e.g. *boycott, conflict, threat, attack, riot, arrest,* etc. have been automatically extracted, by domain adapting event extractors such as EventMine[1] trained on the ACE corpus.
3. *Discourse annotations.* Once the events have been extracted, then they have been further categorised in terms of discourse attitudes or metaknowledge (e.g. speculation, plan, intension, hypothesis, desire, promise, subjectivity (positive, negative), source).
   a. Hypothetical event*: It could swell to $500 billion if* ***we go to war in Iraq****.*
   b. Event phase (before, during, after): The taxi driver strike **that ended last Wednesday** …
   c. Speculative analysis made by 3rd party*:* ***John Paul II might retire at the end of this year****, a Belgian cardinal says*

Interoperability for Text Analysis

Semantic metadata concerning the above types of annotations is added to articles in the historical collection through the application of a pipeline of a number of text mining tools. Our aim was to allow text mining pipelines to be reused/customised in various tasks. Interoperable platforms ensured that individual tools could be substituted with minimum effort, since they share input and output formats. Text mining pipelines have been constructed and evaluated using interoperable text mining platforms, such as U-Compare[2] and Argo[3]. U-Compare and Argo are built on top of the widely used UIMA framework, for which a large library of interoperable text mining tools exists.

[1] http://www.nactem.ac.uk/EventMine/
[2] http://nactem.ac.uk/ucompare/
[3] http://argo.nactem.ac.uk

Search System

The ISHER search system, http://nactem.ac.uk/ISHER-NYT/ provides an advanced search engine over 1.8 million NYT articles:

- Articles retrieved by an initial search are clustered according to semantic similarity, and each cluster is assigned a representative label.
- Clusters are visualised to illustrate the distance of each document from the central cluster concept (centroid), and to see which documents belong to multiple clusters.
- Automatic multi-document summaries can be generated for each cluster.
- Original metadata about the articles can be used to focus the search, e.g. author, news desk or keywords fund within the articles.
- Combinations of entities (e.g. people, locations, organisations) and/or information about events (e.g., attacks, arrests) can be used to filter search results to those containing very specific information of interest.
- Discourse-related information (meta-knowledge) can be used to further narrow down event-based search results. Information includes the tense of the event, whether it describes a real situation or an abstract (e.g. speculated) situation, and whether or not the event is negated.

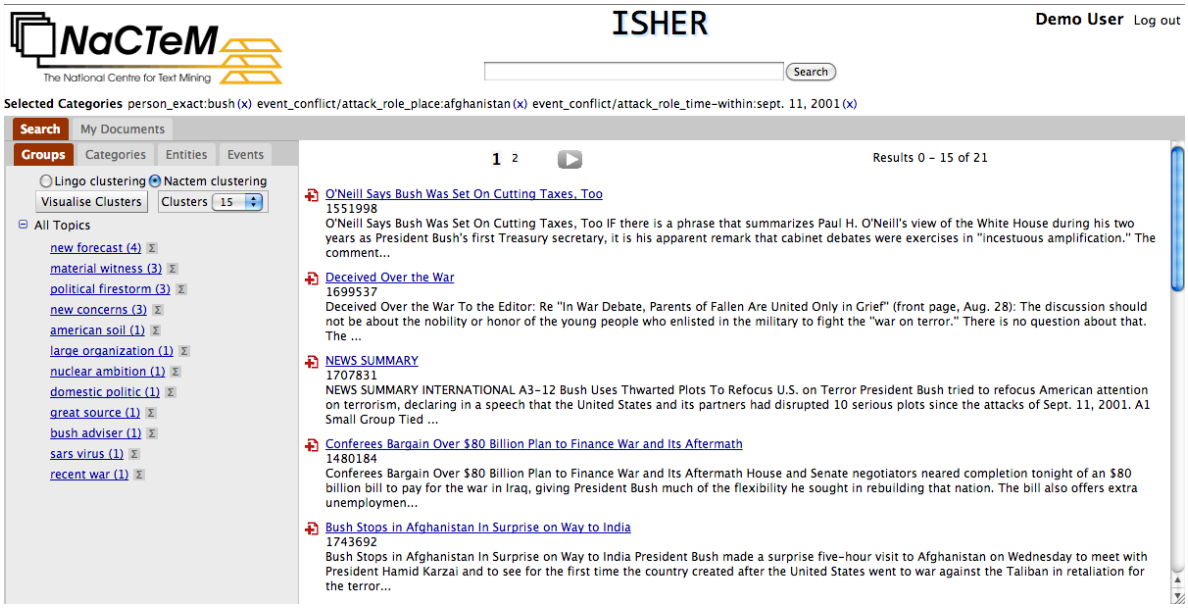

Figure 1: Search based on semantic clustering and events

References

- Ananiadou, S., Thompson, P. and Nawaz, R. (2013) Enhancing Search: Events and their Discourse Context, *Lecture Notes in CS,* Vol. 7817, 318-334
- Mihăilă, C. [...],Korkontzelos, I., and S. Ananiadou: Towards a Better Understanding of Discourse: Integrating Multiple Discourse Annotation Perspectives Using UIMA ACL, 2013
- Batista-Navarro, R, [...] Korkontzelos, I., and Ananiadou, S. (2013) Facilitating the Analysis of Discourse Phenomena in an Interoperable NLP Platform, *Lecture Notes in CS*, Vol. 7817
- Zervanou, K., Korkontzelos, I., van den Bosch, A. and Ananiadou, S. (2011) Enrichment and Structuring of Archival Description Metadata. *ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.

Partners

1. University of Manchester, National Centre for Text Mining, UK (Sophia Ananiadou)
2. Cognitive Computation Group (CCG), Department of Computer Science, University of Illinois at Urbana-Campaign, IL, USA (Dan Roth)
3. Radboud University Nijmegen, Centre for Language Studies, NL (Antal van den Bosch)