From Big Data Text Repositories to Argument Analysis

McAlister, S., Allen, C., Ravenscroft, A., Reed, C., Lawrence, J., Borner, K. & Bourget, D.

The Digging by Debating team, diggingbydebating.org

The Digging by Debating (DbyD) project *aims to* identify, extract, model, map and visualise philosophical arguments in very large text repositories e.g. Hathi Trust. ***Summarising*****,** our project has:

1. developed a method for visualizing points of contact between philosophy and the sciences;
2. used topic modelling to identify the volumes, and pages within those volumes, which are 'rich' in a chosen topic;
3. used a semi-formal discourse analysis technique to identify key arguments in the selected pages to incrementally expose and represent argument structures and used the OVA argument mapping tool to represent and map the key identified arguments and provide a framework for comparative analysis;
4. devised and used a novel analysis framework applied to the mapped arguments covering role, content and status of propositions; and, the importance, context and meaning of arguments.

These methods have allowed us to generate a list of the key arguments extracted from each text. It is significant that, in applying the methods above, a non-expert with limited or no domain knowledge of philosophy has identified the volumes of interest from a key 'Big Data Set' (Hathi Trust) AND identified key arguments within these texts. Thus providing several key insights about the nature and form of arguments in historical texts, and a proof-of-concept design for a tool that will be usable by scholars.

Specifically, we have prototyped a set of tools and methods that allow scholars to move easily between macro-scale, global views of the distributions of philosophical themes in such repositories, and micro-scale analyses of the arguments appearing on specific pages in texts belonging to the repository. Here we report on a 4-step process to link the macro-scale to the micro-scale. Our approach spans bibliographic analysis and science mapping, and LDA topic modelling conducted at Indiana University and machine-assisted argument markup into Argument Interchange Format (AIF) using the OVA (Online Visualization of Argument) tool from the University of Dundee. The latter has been used to analyse and model arguments by the team based at the University of East London. This work is currently being articulated as potential extensions to the repository – PhilPapers – designed by other members linked to the University of London.

In this project we use semi-formal discourse analysis linked to modelling techniques to better understand the distribution and nature of arguments in the texts of interest. The methods for moving between large scale analyses of entire repositories and close analyses of specific texts have not been fully automated, and are unlikely to be so in the near future without additional funding. Nevertheless, we can demonstrate that automation of certain key parts of the process affords considerable advantages to scholars in the humanities, that justifies continued development of these methods.

The final outcomes of the present project will allow us to decide whether relevant arguments can be suitably mapped using OVA, and therefore support and inform additional development of automated analysis algorithms to do the same. Or, alternatively, whether understanding and mapping arguments in our texts is particularly complicated, requiring complex human interpretation, and therefore tool development should focus on the development of visual interfaces allowing people to better understand, identify and represent arguments. Most likely, both options should be pursued simultaneously. The four interrelated and overlapping steps of the DbyD project has implemented the following component processes that are currently being articulated as proof-of-concept tools that can be incorporated into a repository and social media platform called PhilPapers.

1. Large Scale Visualisation of Science-Philosophy Interactions

A method for overlaying philosophical topics extracted from the Stanford Encyclopedia of Philosophy (SEP) onto the USCD Map of Science, making it possible to visualize the overlap between bibliographic citations in SEP article categories and the 13 disciplines and 554 subdisciplines represented in the UCSD map which is based upon article level data from Thompson Reuters' Web of Science and Elsevier's Scopus. Of the 554 subdisciplines, 275 contained a journal cited by at least one SEP article. By overlaying the citations from the bibliographies of SEP articles onto the regions of the UCSD map, it is possible to visualize which areas of science have received relatively more or relatively less attention from philosophers.

2. Selection of HathiTrust Volumes and Pages through Topic Modeling

Semantic modeling of the entire HathiTrust collection of more than 2.5 million books is currently not computationally feasible. We chose instead to focus on philosophical disputes about animal psychology in the period 1880-1908, drawing on expertise within the project group and a period of intense activity during the separation of psychology as a scientific discipline from its previous home in philosophy. An initial subset of 1315 volumes was obtained by a very broad keyword search for ‘darwin OR romanes OR comparative psychology’. LDA topic modeling of these 1315 made it possible to rapidly exclude irrelevant volumes such as scholarly monographs on religion and college course catalogues. Topics of high relevance were identified using an algorithmic comparison of the expert-provided terms 'anthropomorphism', 'animal', and 'psychology' to the topics in the model. The most appropriate topic among these was used to select a smaller set of 86 volumes, which were then modeled in a second round of (fine-grained) LDA modeling where individual pages rather than entire volumes were treated as the documents of interest. A simple algorithmic ranking of volumes by the relevance and quantity of rated pages led to the selection of the top three suitable volumes for the next stage of analysis.

3. Selection of Argument, Markup and Modelling of Content

Guided by the rated pages in the three volumes we manually identified argumentative sections for the next stage of the analysis, limiting the selection to a maximum of 40 pages from each book. The content was marked up in OVA+, an argument mapping application which links blocks of text with argument nodes. Each block of argumentative content, as selected in the previous section, was marked up as a set of propositions with links to the propositions that they support or undercut. In this way each argument was mapped as a conclusion with premises that support it. In some cases the text of the mapped arguments was reduced or modified to avoid verbosity, and this is noted.

4. Analysis and Evaluation of the Modelled Arguments

Firstly, analysing at a propositional level, content (e.g. evidence, criticism, explanation), status (e.g. reference, quotation, assertion) and role (e.g. support, counter, conclusion) were identified and classified according to an analysis scheme. Secondly, the representativeness, type importance, meaning and context of the whole argument were identified with respect to the topic of Animal Consciousness and Anthropomorphism. The propositional analysis scheme was found to discriminate well between the three authors’ styles, with even this small sample offering statistically significant differences in content and status of propositions. We have attempted to quantify the degree of variability among argument signifiers in these late 19th / early 20th C. texts and hypothesize that there is much less variability in modern scientific English, pending further analysis.

This project is showing for the first time how big data text processing techniques can be combined with deep structural analysis to provide researchers and students with navigation and interaction tools for engaging with the large and rich resources from datasets provided by the Hathi Trust Research Center, including parts of Google Books, and PhilPapers.