Cross-Lingual Searching and Visualization for Greek and Latin and Old Norse Texts

Jeffrey A. Rydberg-Cox
Lara Vetter
Department of English
University of Missouri Kansas City
Kansas City, MO 64110
[rydbergcoxj, vetterl]@umkc.edu

Stefan Rüger
Daniel Heesch
Department of Computing
Imperial College London
London SW7 2AZ, England
[s.rueger, daniel.heesch] @imperial.ac.uk

ABSTRACT
We explore approaches to multi-lingual information retrieval for Greek, Latin, and Old Norse texts and an innovative visualization facility for the results.

Categories and Subject Descriptors
H.3.7.: Information Storage and Retrieval – *Digital Libraries*

General Terms
Design, Standardization, Languages, Experimentation

Keywords
Cross-Lingual Information Retrieval, Digital Library

1. Cross-Lingual Query Generation

Cross-lingual information retrieval is a particularly intriguing technology for students and scholars of Ancient and Early-Modern Greek and Latin or Old Norse. This technology can be extremely useful for non-specialist scholars and students who are somewhat familiar with these languages, but who do not know enough to form a mono-lingual query for a search engine.

The problem of multi-lingual information retrieval is essentially one of machine translation on a very small scale. There have been two dominant approaches to this problem: 1) dictionary translation using machine-readable multi-lingual dictionaries and 2) automatic extraction of possible translation equivalents by statistical analysis of parallel or comparable corpora. The needs and nature of our user community of students and scholars in a humanities digital library suggest that we can profitably adopt both of these approaches.

The search facility begins with a simple interface that allows users to enter their search terms in English, to select the sources that will be used for query translation, and to restrict their results to words that appear in works written by a particular author. After entering query terms, the user is presented with an interface with detailed information to allow them to construct the best translation of the word for their needs. This process can range from the simple elimination of obvious ambiguities and mistakes to a careful consideration of every term. The interface provides a list of translation equivalents for the word or words that the user entered along with an automatically abridged English definition of the word, a link to the full definition for each word, a list of authors who use the words, and data about the frequency of each word in works by the selected authors. The tool also provides users with other possible query terms related to the exact matches returned by their initial query using a simple similarity coefficient to find related dictionary definitions.

JCDL'04, June 7–11, 2004, Tucson, Arizona, USA.
ACM 1-58113-832-6/04/0006.

2. Visualizing Results

After users translate their queries with these tools, the search is passed to a monolingual search engine with several visualization front ends. These front ends are alternatives to the traditional ranked list view of search results and are based on the on-the-fly calculation of keywords for the documents returned by the query.

These interfaces group related documents visually and label each group with the most appropriate keyword. The first visualization interface is a tree view that represents documents as the nodes of a binary tree flattened into a circular pattern. The second visualization generates a Sammon map that provides users with a visual landscape for navigation. In this interface, each cluster is represented as a circle and is labeled with its highest frequency keyword. The radius of the circle indicates the relative size of each of the clusters, while the distance between the circles represents the relative similarity of the different clusters. The third display offers a radial visualization in which the twelve highest ranked keywords in the returned search results are displayed in a circle. Each document in the returned set is represented as a point in the middle of the circle with its placement determined by the relative pull of each of the keywords distributed around the circle.

In each visualization, mousing over visual nodes in the display provides the user with a menu showing the number of documents and all of the keywords associated with that cluster and drill down into the associated subclusters. Further, the user is able to eliminate keywords from the search results, view fragments of every document in the collection, and follow a link to the complete document within the digital library.

3. ACKNOWLEDGMENTS

Funding for this work was provided by the National Science Foundation International Digital Libraries Program (Grant number IIS-0122491) and the European Commission Information Society Technologies program (Project number IST-2001-32745).