Computational Linguistics for Metadata Building: Enhancing image access through linguistic techniques

Judith Klavans, Jimmy Lin, Rebecca Passonneau, Eileen Abels, Carolyn Sheffield, Dagobert Soergel, Tandeep Sidhu

1. Identify relevant text for an image and suggest terms

   a) Many catalog records contain minimal subject-oriented access points
   b) CLiMB extracts information from scholarly texts for subject fields

2. Conceptual analysis and term extraction

   Conceptual Analysis: One Paragraph, Three Semantic Categories

   [Historical Context] Of the great projects built by Akhenaten hardly anything remains . . . . Through his choice of masters, he fostered a new style. [Implementation] Known as the Amarna style, it can be seen at its best in [Image Content] a sunk relief portrait of Akhenaten and his family.

3. Mapping terms to thesauri

   Problems in image indexing and access
   - too many images available with only limited access points
   - subject-oriented cataloging is expensive, time-consuming, and requires domain expertise

   Solution
   - apply computational linguistic techniques over text about images to identify potential subject terms
   - populate existing catalog records with these subject terms for enhanced access

   Methodology
   - create a cataloger’s toolkit for mining scholarly art historical texts for subject-oriented metadata
   - identify relevant text segments and conceptual function
   - disambiguate terms and phrases by mapping to the appropriate sense in the integrated thesaurus
   - use mined terms to guide cataloger’s process of creating subject terms
   - test the efficacy and effectiveness from the perspectives of both catalogers and end users

4. Export terms to catalog record

   Catalogers: use thesaurus for controlled vocabulary
   End users: expand or narrow a search through hierarchies

www.umiacs.umd.edu/~climb