Computational Linguistics for Metadata Building

Judith L. Klavans
jklavans@umd.edu

Jimmy Lin
jimmylin@umd.edu

University of Maryland College Park:
the iSchool at Maryland and the
University of MD Institute for
Advanced Computer Studies
(UMIACS)

Carolyn Sheffield
csheffie@umd.edu

Tandeep Sidhu
tsidhu@umd.edu

ABSTRACT
In this paper, we describe a downloadable text-mining tool for enhancing subject access to image collections in digital libraries.

Categories and Subject Descriptors
D.3.3 [**Programming Languages**]: Language Contructs and Features – *abstract data types, polymorphism, control structures.* This is just an example, please use the correct category and subject descriptors for your submission. The ACM Computing Classification Scheme: http://www.acm.org/class/1998/

General Terms
Algorithms, Design, Experimentation, Human Factors.

Keywords
Computational linguistics, image access, text mining, metadata mining, ontologies, disambiguation, text categorization.

1. CLiMB: A CATALOGERS' TOOLKIT

The Computational Linguistics for Metadata Building (CLiMB) research project uses text mining to address the need for high quality, filtered metadata for improving access to images in digital libraries. CLiMB addresses the specific need for enhanced subject access to digital image collections in the domains of art history and architecture. The CLiMB catalogers' workbench processes text associated with an image through natural language processing, categorization using machine learning, and disambiguation techniques to identify, filter, and normalize high-quality subject descriptors.

The CLiMB catalogers' workbench combines new and pre-existing technologies in a flexible, client-side architecture. To extract subject terms and associate them with an image, the system requires an image, minimal metadata (e.g. image, name, creator), and text. To identify disambiguation issues which arise with specialized vocabularies, we selected six image and text collections with which to test the underlying algorithms.

Conference'04, Month 1–2, 2004, City, State, Country.

The first stage of CLiMB's processing pipeline segments text into topical portions and associates these with relevant images. The next phase, Linguistic Analysis, consists of several subprocesses. A part-of-speech (POS) tagger labels the function of each word within a sentence. Complete noun phrases are identified by the NP chunker based on patterns. CLiMB uses the Stanford tagger[1] to provide sentential analysis of syntactic constructions. The output of Linguistic Analysis consists of XML-tagged terms which contain part of speech and syntactic parsed labels. Noun phrases are then input into the disambiguation algorithm, which enables sense mapping to an ontology. Currently, we map to the Getty Art and Architecture Thesaurus (AAT), Union List of Artist Names (ULAN), and Thesaurus of Geographic Names (TGN).[2] The Getty Vocabularies are well-established multi-faceted thesauri for cataloging art and architecture materials.

The CLiMB interface enables catalogers to select subject terms while viewing an image, its metadata, and associated texts. Within the text, common and proper nouns are highlighted for quick identification of potential terms. By clicking on any given term, the cataloger can view potential matches in the Getty resources. The CLiMB disambiguation algorithms highlight the most likely sense first and followed by other possible matches. Through the interface, catalogers can also view definitions and hierarchy positions for the Getty terms. As catalogers select terms, they populate a window within the interface for review before exporting. Export functionalities under development include mapping to catalog records in several standard metadata schemas, including the Visual Resources Association Core 4.0, MARC, and standard XML.

2. ACKNOWLEDGMENTS

We acknowledge: the entire CLiMB staff including Dagobert Soergel (University of Maryland), Rebecca Passonneau (Columbia University), and Eileen Abels (Drexel University); the Mellon Foundation for their continued support; Murtha Baca, director of the Getty Vocabulary Program, for providing us with research access to resources; collections partners, including Jeff Cohen (Bryn Mawr College);,Jack Sullivan (University of Maryland); the Senate Museum and Library, and ARTstor.

[1] Both the tagger and parser are available at:
http://nlp.stanford.edu/software

[2] Getty resources can be accessed at:
www.getty.edu/research/conducting_research/vocabularies/aat