Computational Linguistics for Metadata Building (CLiMB) Text Mining for the Automatic Extraction of Subject Terms for Image Metadata

Judith L. Klavans[1,2], Tandeep Sidhu[1], Carolyn Sheffield[1], Dagobert Soergel[1], Jimmy Lin[1,2], Eileen Abels [3], Rebecca Passonneau [4]
[1]College of Information Studies (CLIS)
[2]University of Maryland Institute for Advanced Computer Science (UMIACS)
University of Maryland, College Park, Maryland
[3]College of Information Science and Technology, Drexel University, Phila PA,
[4]Center for Computational Learning Systems, Columbia University, NY NY

Abstract. In this paper, we present a fully-implemented system using computational linguistic techniques to apply automatic text mining for the extraction of metadata for image access. We describe the implementation of a workbench created for, and evaluated by, image catalogers. We discuss the current functionality and future goals for this image catalogers' toolkit, developed in the Computational Linguistics for Metadata Building (CLiMB) research project.[1] Our primary user group for initial phases of the project is the cataloger expert; in future work we address applications for end users.

1 The Problem: Insufficient Subject Access to Images

The CLiMB project addresses the existing gap in subject metadata for images, particularly for the domains of art history, architecture, and landscape architecture. Within each of these domains, image collections are increasingly available online yet subject access points for these images remain minimal, at best. In an initial observational study conducted with six image catalogers, we found that typically 1 – 8 subject terms are assigned, and many legacy records lack subject entries altogether.

The literature on end users' image searching indicates that this level of subject description may be insufficient for some user groups. In a study of the image-searching behaviors of faculty and graduate students in American history, Choi and Rasmussen [3], found that 92% of the 38 participants considered the textual information associated with the images in the Library of Congress' American Memory Collection inadequate. The number of subject descriptors assigned to an image in this collection is comparable to or exceeds those found in the exploratory CLiMB studies. Furthermore, these searchers submitted more subject-oriented queries than known artist and title searches. Similar results demonstrating the importance of subject retrieval have been reported in other studies, including Keister [6], Collins [4], and Chen [2].

[1] This project was first funded by the Mellon Foundation to the Center for Research on Information Access at Columbia University.

2 Solutions

The CLiMB project was initiated to address the subject metadata gap under the hypothesis that automatic and semi-automatic techniques may enable the identification, extraction and thesaural linking of subject terms. The CLiMB Toolkit processes text associated with an image through Natural Language Processing (NLP), categorization using Machine Learning (ML), and disambiguation technologies to identify, filter, and normalize high-quality subject descriptors. Like Pastra et al. [10] we use NLP techniques and domain-specific ontologies, although our focus is on associated texts such as art historical surveys or curatorial essays rather than captions.

For this project, we use the standard Cataloging Cultural Objects (CCO) definition of subject metadata[2]. According to this definition, the subject element of an image catalog record should include terms which provide "an identification, description, or interpretation of what is depicted in and by a work or image." The CCO guidelines also incorporate instructions on analyzing images based on the work of Shatford-Layne (formerly Shatford). Shatford [12], building on Panofsky [8], proposed a method for identifying image attributes, which includes analysis of the generic and specific events, objects, and names that a picture is "of" and the more abstract symbols and moods that a picture is "about". Panofsky describes the pre-iconographic, iconographic, and iconologic levels of meaning found in Renaissance art images. Shatford's generic and specific levels correspond to Panofsky's pre-iconographic and iconographic levels, respectively, and encompass the more objective and straightforward subject matter depicted in an image. The iconologic level (Shatford's about) addresses the more symbolic, interpretive, subjective meanings of an image. To aid user access, catalogers are encouraged to consider both general and specific terms for describing the objective content of an image as well as to include the more subjective iconologic, symbolic, or interpretive meanings. Iconologic terms may be the most difficult for catalogers to assign but occur often in texts describing images.

3 Preparatory Studies of Cataloging

In order to get a better sense of the cataloging process and to inform our system design, we conducted studies on the process of subject term selection by image catalogers. Our goal was to collect data on the process as a whole in order to improve both our system function (either through rules or statistical methods) and our system functionality (i.e. how to incorporate our results into an existing workflow and how to perhaps replace a portion of the workflow with automatic techniques). In this section, we discuss two of these formative studies.

The first study was designed to identify the types of subject terms a cataloger may assign to a given image. Identifying these expert term assignments will help guide

[2] http://vraweb.org/ccoweb/cco/parttwo_chapter6.html.

the development of heuristic rules for automatically identifying high-quality descriptor candidates and filtering out term types which are rarely assigned manually. Participants were given four stimuli:

1) a hypothetical query for an image;
2) an image;
3) another image—this time with associated text; and
4) an image paired with a list of CLiMB-extracted terms.

For the first two stimuli, catalogers were asked to generate subject terms. For the third and fourth stimuli, catalogers were asked to select terms from the associated text or list of terms. We selected four image/text pairs from the National Gallery of Art Collection. To control for varying textual content which may occur with different image types, we chose one landscape, one portrait, one still life, and one iconographic image; we employed a Latin Square design. Twenty image catalogers recruited through the Visual Resource Association participated in the study. Through a combined quantitative and qualitative approach, we analyzed

- the number of terms assigned per task,
- the types of terms assigned, and
- the level of agreement between catalogers in terms used for the same concept (to be discussed in a future publication).

Table 1: Distribution of term assignments by category.

Terms assigned for landscape image, Task 2	Category
Gauguin	artist name
pea green	Color
Orange	Color
black and white	Color
Cow(s) / dairy cows / cattle	Figures/Objects
stacks of hay / mounds of hay/ bales / hay	Figures/Objects
Crops	Figures/Objects
Herding	Figures/Objects
woman herding	Figures/Objects
capped woman	Figures/Objects
Poppies	Figures/Objects
Rocks	Figures/Objects
white dress	Figures/Objects
19th Century	Period
fields/vegetable field	Place
France	Place
Brittany	Place
Dutch landscape/Dutch countryside; paintings and landscapes / paintings and countrysides	Type
sea/seascape or canal	Type

In analyzing the types of terms catalogers assigned, we identified seven categories (in order of frequency): figure/object, place, artist names, period/date, type, style and color. Table 1 above shows a subset of results from catalogers completing just one of the tasks for the landscape image. Results for landscape art across all four tasks yielded 13 terms for figure/object, 9 for place, 8 for artist names, 7 for period/date, 6 for type, and 4 for style and color. This distribution will help guide the priorities placed on term selection.

For the second study, we took a broader look at the overall image-indexing workflow, including standards, local policies, and actual practices, to determine how the CLiMB Toolkit fits into the cataloging process as a whole. This study not only enabled us to define interface parameters and necessary functionality, it also confirmed the lack of subject access currently provided by human indexers. We examined the similarities and differences in image cataloging practices both within a single institution and across three separate institutions. By observing catalogers as they indexed images from their respective collections, we also investigated the number and types of subject terms added per catalog record. Within and across these academic visual resource centers, we found that general practices and workflow patterns varied little, and that the number of subject terms entered per catalog record varied but typically fell somewhere between one and eight. One of the primary differences across institutions was the use of different software and metadata schemas, some of which were locally developed. These results indicate that, with flexible export functionality built in to a generic workbench, the CLiMB Toolkit should integrate smoothly with existing practices and different work environments, with little or no tailoring required.

4 CLiMB Architecture and Interface

This section describes the techniques we have developed to semi-automatically identify terms which qualify as potential subject descriptors. Our techniques exceed simple keyword indexing by:

- applying advanced semantic categorization to text segments,
- identifying coherent phrases,
- associating terms with a thesaurus, and
- applying disambiguation algorithms to these terms.

CLiMB combines new and pre-existing technologies in a flexible, client-side architecture which has been implemented into a downloadable toolkit, and which can be tailored to the user's needs. Figure 1 shows the overall architecture of the CLiMB Toolkit. The upper left shows the input to the system, an image, minimal metadata (e.g. image, name, creator), and text. To date, we have input six testbed collections, described more fully in Section 5.

The first stage of CLiMB's processing pipeline associates portions of the input text with images. Note that this requires segmentation, and association of segmented text with the image being described. In clear cases, such as online image captions or in

exhibition catalogs, association of image with text is a given. However, in cases where there is a more diffuse relationship between text and image (as in art history texts, for example), it is a computational challenge to ensure that text is accurately associated with the correct image, and not with an image in close proximity (which may or may not be described by the text). This logic creates associations between text and image based on explicit references in the text, rather than taking any text in proximity of an image.

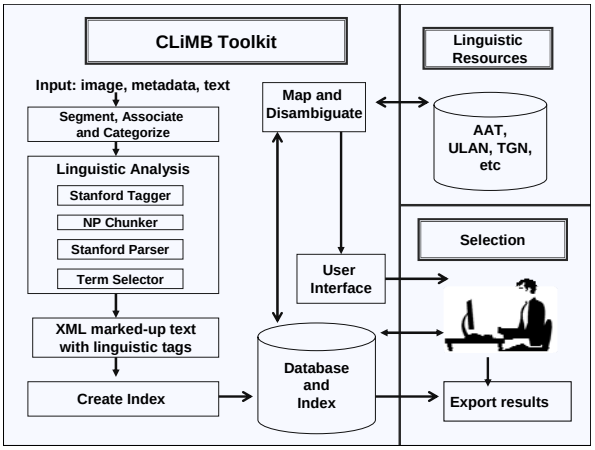

Figure 1: CLiMB Architecture.

In addition to segmentation, we are developing methods to categorize spans of text (e.g., sentences or paragraphs) as to their semantic function in the text. For example, a sentence might describe an artist's life events (e.g. "during his childhood", "while on her trip to Italy", "at the death of his father") or the style of the work ("impressionism".) A set of seven categories--Image Content, Interpretation, Implementation, Historical Context, Biographical Information, Significance, and Comparison--has been initially proposed through textual analysis of art survey texts. These categories have been tested through a series of labeling experiments. Full details are available in Passonneau et al. [9].

The next phase, Linguistic Analysis, consists of several subprocesses. After sentence segmentation, a part-of-speech (POS) tagger labels (i.e. tags) the function of each word in a text, e.g., noun, verb, preposition, etc. Complete noun phrases can then be identified by the NP chunker based on tag patterns. For example, a determiner, followed by any number of adjectives, followed by any number of nouns, is one such pattern that identifies a noun phrase, as in "the impressive still life drawing". The tagger used for CLiMB, the Stanford tagger[3] provides sentential analysis of syntactic constructions, e.g., verb phrases, relative clauses. The output of Linguistic Analysis consists of XML-tagged words which contain substantial part of speech and syntactic parsed labels. Lucene is used to create an index for these tagged words.[4]

[3] Both the tagger and parser are available at: http://nlp.stanford.edu/software.
[4] Lucene is a search engine library: http://lucene.apache.org.

At this point, the noun phrases stored in the index are input to the disambiguation algorithm, which then enables sense mapping, so that the proper descriptor can be selected from a controlled vocabulary. Words and phrases often have multiple meanings which correspond to different descriptors. The ability to select one sense from many is referred to as lexical disambiguation. Currently, we map to descriptors from the Getty Art and Architecture Thesaurus (AAT), the Getty Union List of Artist Names (ULAN), and the Getty Thesaurus of Geographic Names (TGN).[5] The AAT is a well-established and widely-used multi-faceted thesaurus of terms for the cataloging and indexing of art, architecture, artifactual, and archival materials. AAT has 31,000 such records and among those, there are 1,400 homonyms, i.e., records with same preferred name. For example, the term "wings" has five senses in the AAT. Each sense falls under distinct but separate subdomains of art and architecture, ranging from building divisions and theater spaces to costume accessories, furniture components, and visual works components.

Table 2 shows the breakdown of the AAT vocabulary by number of senses with a sample lexical item for each frequency. As with most dictionaries and thesauri, most items have two to three senses, and only a few have more.

Table 2: Scope of the disambiguation problem in the AAT Thesaurus.

# of Senses	# of Terms	Example
1	29,576	Scaglioni
2	1097	Bells
3	215	Painting
4	50	Alabaster
5	39	Wings
6 -7	14	Boards
8+	9	Emerald Carmine

First, we use all modifiers that are in the noun phrase to find the correct AAT record (Lookup Modifier). We search for the modifiers in the record description, variant names, and the parent hierarchy names of all the matching AAT senses. If this technique narrowed down the record set to one, then we found our correct record. For example, consider the term "ceiling coffers." For this term we found two records: "coffers" (coffered ceiling components) and "coffers" (chests). The first record has the modifier "ceiling" in its record description, so we were able to determine that this was the correct record. Next, we use SenseRelate to help select the correct WordNet

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

sense of the noun phrase (or its head noun). Using that sense definition from WordNet, we next examined which of the AAT senses best matches with the WordNet sense definition. For this, we used a word overlapping technique which takes senses of WordNet for each polysemous term in AAT and selects the highest value of word overlaps. If none of the AAT records received any positive score (above a threshold), then this technique could not find the best match. Other techniques, Best Record Match and Most Common Sense, are presented in Sidhu et al. [13].

For evaluation of the disambiguation model, we followed standard procedure in word sense disambiguation tasks (Palmer et al. [7]). Two labelers manually mapped 601 subject terms to a controlled vocabulary. Inter-annotator agreement for this task was 91%, providing a notional upper bound for automatic system performance (Gale et al. [5]) and a dataset for evaluation. We used SenseRelate (Banerjee and Pederson [1], Patwardhan et al. [11]) for disambiguating AAT senses. SenseRelate uses word sense definitions from WordNet 2.1, a large lexical database of English words.[6] The impact of using a general vocabulary such as WordNet compared to specialist vocabularies is an empirical issue which we are examining in current research.

Table 3 shows results of running different techniques on this data. Row 1 shows how few terms were mapped by the lookup modifier technique; only one was mapped for the Training Set. Rows 2 and 3 show that the SenseRelate technique was most successful in labeling terms, followed by the Best Record Match technique. The Most Common Sense technique (Row 4) was also poor. An analysis of results and errors shows that our overall accuracy is between 50-55% compared to 70% common in general disambiguation. In future work, we will explore re-implementing concepts behind SenseRelate to directly map terms to the AAT and additional approaches using hybrid techniques (including machine learning) for disambiguation. Currently, we are awaiting results from manual disambiguation tests with human catalogers before refining and integrating the module. Our plan is to use results to rank and select a sense for mapping that the user will confirm; once we collect enough feedback, we can apply learning to eliminate senses with greater confidence than at present.

Table 3: Breakdown of AAT mappings by Disambiguation Technique.

	Technique Name	Training (n=128)	Test (n=96)
1	Lookup Modifier	1	3
2	SenseRelate	108	63
3	Best Record Match	14	12
4	Most Common Sense	5	18

Figure 2 shows a screen shot of the CLiMB user interface, after having performed a search over the National Gallery of Art collection, and having run the text through the Toolkit.[7] Note that the center top panel contains the image, so the user can look at

[6] http://wordnet.princeton.edu/

[7] In the interest of space, we have included a full screen shot, accompanied by text explanations. If reviewers prefer, this can be enlarged or split into two Figures.

the item to be described. The center panel contains the input text, with proper and common nouns highlighted. Under this is the term the user has selected to enter. The right-hand panel gives the thesaural information. At the top of the right are the two senses for the word "landscape" with an indication of where they occur in the AAT hierarchy. Next is the text description of the sense selected. Finally, the entire hierarchy is displayed, bottom right, for the user to view and identify any related terms.

As part of the evaluation of the CLiMB approach, we have established a series of test collections in the fields of art history, architecture, and landscape architecture. These three domains were selected in part because of the existing overlap in domain specific vocabulary. Testing with distinct but related domains enables us to test for disambiguation issues which arise in the context of specialized vocabularies. For example, the AAT provides many senses of the term "panel" which apply to either the fine arts, architecture, or both, depending on context. In the context of fine arts, "panel" may refer to a small painting on wood whereas in the context of architecture, the same term may refer to a distinct section of a wall, within a border or frame.

We are currently working with five image-text sets and one image collection for which we are conducting experiments with dispersed digital texts. These six collections will be used for different phases of evaluation, discussed under Future Work. The texts and images for two of the collections, the National Gallery of Art (NGA) Online Collection and the U.S. Senate Catalogue of Fine Arts, can be found online and are in the public domain. For three of the other image collections, The Vernacular Architecture Forum (VAF)[8], The Society of Architectural Historians (SAH)[9], and The Landscape Architecture Image Resource (LAIR)[10], we have secured digital copies of relevant texts along with permissions for use in our testing. The final collection is the Art History Survey Collection, made available to us through ARTstor[11].

5 Conclusions and Future Work

We are working in a challenging domain with a highly specialized vocabulary. Currently we depend on the external program SenseRelate to perform much of the disambiguation. Furthermore, SenseRelate maps terms to WordNet and we then map the WordNet sense to an AAT sense. This extra step is overhead, and it causes errors in our algorithm. We are looking to incorporate additional domain-specific vocabularies for future testing, rather than more general resources which add noise. Sources under consideration include ICONCLASS and the Library of Congress' Thesaurus for Graphic Materials I & II.

[8] http://www.vernaculararchitectureforum.org/
[9] www.sah.org/
[10] www.lair.umd.edu/
[11] www.artstor.org

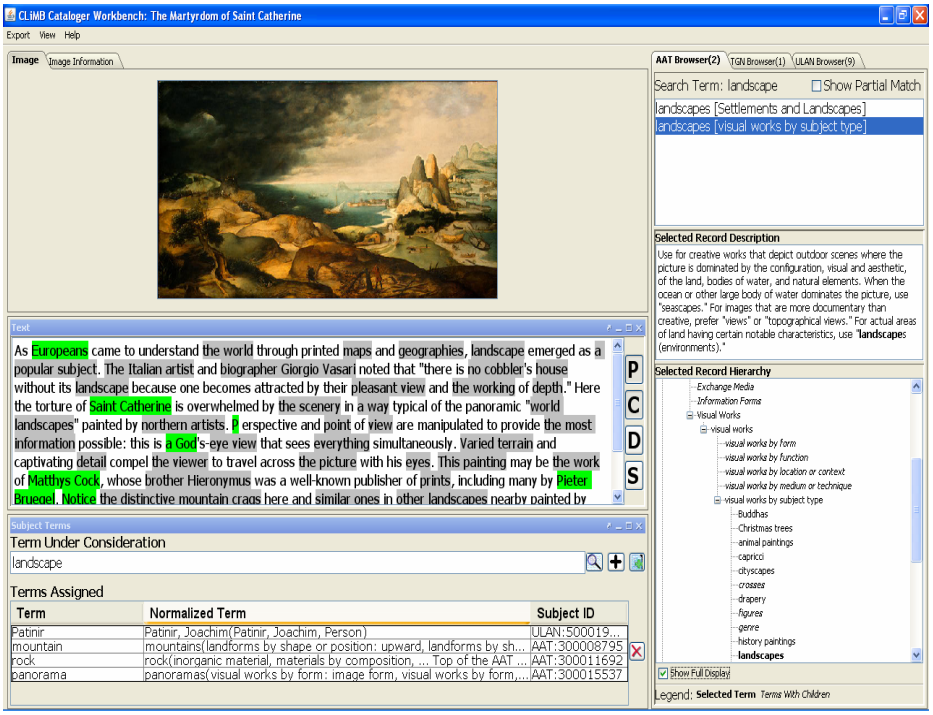

Figure 2: **CLiMB user interface for term "landscape".**

For future work, we have also designed a series of studies to test the toolkit in situ. We have partners from several museums and libraries, mentioned in Evaluation, that will test CLiMB with their cataloging staff, and then work with us to design evaluations of Toolkit success in three areas:

1) staff perception on Toolkit ease of use for cataloging within their collections;
2) end user satisfaction with these enhanced records; and
3) several component evaluations, including the named entity recognizer, the noun phrase selector, and the disambiguation component.

The proverbial tradeoff between precision and recall may vary for different sectors of the image community; we believe our research in different venues will provide insights on this critical issue. Finally, we intend to explore new directions for integrating CLiMB with current social networking technologies, including social tagging, trust-based ranking of tags, and recommender systems. These technologies address end user needs and offer CLiMB the potential to achieve more personalized results.

6 Acknowledgements

We acknowledge: the Mellon Foundation; Dr. Murtha Baca, director of the Getty Vocabulary Program and Digital Resource Management, Getty Research Institute for providing us with research access to resources; collections partners, including Jeff

Cohen, Bryn Mawr College and University of Pennsylvania; Jack Sullivan, University of Maryland; the Senate Museum and Library and ARTStor.

References

1. Banerjee, S., Pedersen, T.: Extended Gloss Overlaps as a Measure of Semantic Relatedness. In: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, (2003) 805–810 [7]
2. Chen, H.: An Analysis of Image Retrieval Tasks in the Field of Art History. Information Processing & Management, Vol. 37, No. 5 (2001) 701-720
3. Choi, Y., Rasmussen, E. Searching for Images: The Analysis of Users' Queries for Image Retrieval in American History. Journal of the American Society for Information Science and Technology, Vol. 54 (2003) 498-511
4. Collins, K.: Providing Subject Access to Images: A Study of User Queries. The American Archivist, Vol. 61 (1998) 36-55
5. Gale, W., Church, K., Yarowsky, D.: A Method for Disambiguation Word Senses in a Large Corpus. Computers and Humanities, Vol. 26 (1993) 415-439.
6. Keister, L.H.: User Types and Queries: Impact on Image Access Systems. In: Fidel, R., Hahn, T.B., Rasmussen, E., Smith, P.J. (eds.): Challenges in Indexing Electronic Text and Images. Learned Information for the American Society of Information Science, Medford (1994) 7-22
7. Palmer, M., Ng, H.T., Dang, H.T.: Evaluation. In: Edmonds, P., Agirre, E. (eds.): Word Sense Disambiguation: Algorithms, Applications, and Trends. Text, Speech, and Language Technology Series, Kluwer Academic Publishers (2006)
8. Panofsky, E. Studies in Iconology: Humanistic Themes in the Art of the Renaissance. Harper & Rowe, New York (1962)
9. Passonneau, R., Yano, T., Lippincott, T., Klavans, J. Functional Semantic Categories for Art History Text: Human Labeling and Preliminary Machine Learning. 3rd International Conference on Computer Vision Theory and Applications, Workshop on Metadata Mining for Image Understanding (2008)
10. Pastra, K., Saggion, H., Wilks, Y.: Intelligent Indexing of Crime-Scene Photographs. In: IEEE Intelligent Systems: Special Issue on Advances in Natural Language and Processing, Vol. 18, Iss. 1. (2003) 55-61.
11. Patwardhan, S., Banerjee, S., Pedersen, T.: Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City (2003)
12. Shatford, S.: Analyzing the Subject of a Picture: A Theoretical Approach. Cataloging & Classification Quarterly, Vol. 6, Iss. 3 (1986) 39-62
13. Sidhu, T., Klavans, J.L., Lin, J.: Concept Disambiguation for Improved Subject Access Using Multiple Knowledge Sources. In: Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTech 2007), 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic (2007)