Functional Semantic Categories for Art History Text: Human Labeling and Preliminary Machine Learning ⋆

Rebecca J. Passonneau[1], Tae Yano[2], Tom Lippincott[3], and Judith Klavans[4]

[1] Center for Computational Learning Systems, Columbia University
becky@ccls.columbia.edu
[2] Department of Computer Science, Carnegie Mellon University
taey@cs.cmu.edu
[3] Department of Computer Science, Columbia University
tom@cs.columbia.edu
[4] College of Information Studies, University of Maryland
jklavans@umd.edu

Abstract. The CLiMB project investigates semi-automatic methods to extract descriptive metadata from texts for indexing digital image collections. We developed a set of functional semantic categories to classify text extracts that describe images. Each semantic category names a functional relation between an image depicting a work of art historical significance, and expository text associated with the image. This includes description of the image, discussion of the historical context in which the work was created, and so on. We present interannotator agreement results on human classification of text extracts, and accuracy results from initial machine learning experiments. In our pilot studies, human agreement varied widely, depending on the labeler's expertise, the image-text pair, the number of labels that could be assigned to one text, and the type of training, if any, we gave labelers. Initial machine learning results indicate the three most relevant categories are machine learnable. Based on our pilot work, we implemented a labeling interface that we are currently using to collect a large dataset of text that will be used in training and testing machine classifiers.

1 Introduction

The work presented here was developed in the context of the Computational Linguistics for Metadata Building (CLiMB) research project, which has been investigating methods for automated support to image catalogers and other image professionals [5]. The CLiMB project is developing a Toolkit for image catalogers that would allow them to access electronic versions of texts in order to harvest descriptive metadata. Here we address how to classify, and ultimately to tag, text extracted from electronic documents into functional semantic categories that are relevant for indexing images of artworks. The tagging would be included in the preprocessing phase prior to importing texts into the Toolkit.

⋆ We thank the current and past members of the project members; numerous advisors who reviewed our labeling categories and tested the interface; and especially, volunteer labelers from University of Maryland, Columbia University, Drexel University, and Indiana University.

Figure 1 shows an image taken from the ARTstor Art Images for College Teaching collection (AICT): `http://www.arthist.umn.edu/aict/html/ancient/EN/EN006.html`. It depicts a relief portrait of Akhenaten and his family. Also shown is an extract from an art history survey text describing an image of the same work. If the terms **Akhenaten** and **shrine** were used to index the image, it would not be clear whether the image depicts Akhenaten or a shrine or both. Consider that the word Akhenaten occurs in a sentence about Akhenaten's role in fostering the Amarna style, and in another sentence indicating that he is depicted in the work. The word shrine occurs in a sentence indicating how the depicted work was used. Our goal is to automatically tag sentences like these prior to semi-automatic or automatic extraction of the bold face terms, and for the extracted terms to be associated with tags corresponding to our semantic categories; see the right side of Figure 1. This could permit terms to be filtered or prioritized during the term selection process, depending on the semantic tag. It could also facilitate image search; for a user who wants an image of a shrine, it would be possible to exclude cases where *shrine* does not come from text tagged as *Image Content*.

Fig. 1. Illustration of an image and associated descriptive text

Historical Context: Akhenaten
Image Content: Akhenaten
Historical Context: shrine

Of the great projects built by **Akhenaten** hardly anything remains . . . . Through his choice of masters, he fostered a new style. Known as the Amarna style, it can be seen at its best in a sunk relief portrait of **Akhenaten** and his family. The intimate domestic scene suggests that the relief was meant to serve as a **shrine** in a private household.

In consultation with experts, we developed a set of seven categories to apply to paragraphs or sentences extracted from art history survey texts, where the text extracts are about a specific image. A larger number of categories would lead to much sparser data; a smaller number would lead to categories that are less distinct. Two of the categories, for example, are **Image Content**, defined as text that describes the content of the image, and **Implementation**, text that describes how the work was created.

In our four pilot studies, interannotator agreement among humans varied widely, depending on the labeler's expertise, the image-text pair under consideration, the number of labels that could be assigned to one text, and the type of training, if any, we gave labelers. Human agreement improved if annotators could select multiple labels, which is consistent with our previous results on a lexical semantic annotation task [10]. We

also found that agreement was higher among domain experts, and that the consistency of the labeling depended heavily on the image/text pair under consideration.

Our seven semantic categories vary in relevance, frequency and distinguishability. Thus we do not anticipate attempting to apply machine learning to every category. Using texts labeled during our pilot studies, we have initial results on three of the classes. For example, using a separate naive Bayes classifier for each category, we have been able to achieve 80% accuracy. This indicates that large scale machine learning is feasible.

In section 2, we summarize related work on interannotator agreement among humans. Section 3 describes our datasets. Section 4 describes our pilot studies on human labeling, and our current large scale effort. We present preliminary learning results in section 5 and our conclusions in section 6.

2 Related Work

There are relatively few discussions of interannotator agreement or inter-indexer consistency for image indexing and classification tasks. Two works that address the topic deeply and broadly are [8] and [4]. In the twenty plus years since Markey's analysis of forty years of inter-indexer consistency tests, no comparable review has appeared, and her observations still hold. Although her goal was to use the conclusions from previous work to sort through the issues involved in indexing visual material, all the tests referenced in her paper were on indexing of printed material. She notes significant variability, with accuracy (percent agreement) results ranging from 82% to a low of 4%.

Giral and Taylor [4] looked at indexing overlap and consistency on catalog records for the same items in architectural collections, including an analysis of subject descriptors. On large ($\geq$ 1400) samples of records from the Avery Index to Architectural Periodicals and the Architectural Periodicals Index, they compare proportions of items in their samples that match according to a variety of criteria, and compute 90% confidence intervals. Only 7% of items match entirely, and they find some element of overlap in descriptors in only about 40% of the remaining cases ($\pm$ 3%).

Markey noted two features of documents that affect inter-indexer consistency: document length, and the complexity of the document, which is difficult to quantify. Our image/text pairs, which correspond to Markey's documents, are quite short. We did not attempt to measure complexity, but we did find wide variation in labeling consistency depending on the image/text pair being labeled. This indicates that the image/text pairs have inherent properties that make them more or less difficult for humans to agree on.

Markey found no significant difference between the indexers with or without experience in using such schemes. She found no higher levels of inter-indexer consistency among subject specialists, as compared with non-specialists. This is in contrast to our results. In our pilot studies, the two developers of the categories (the first two co-authors) were the most familiar with them, and had the highest interannotator agreement. In our current large-scale labeling effort, the highest agrement is found among the most expert pairs of labelers. We also found that across studies, agreement increased when we provided more training. Markey found that using a standardized scheme led to higher inter-indexer consistency, ranging from 80% to 34% (in contrast to 4%; see above). This is roughly the range we find, using a different metric but a similar scale.

3 Texts

The domain of digital images and texts we focus on parallels the ARTstor *Art History Survey Collection (AHSC)*. ARTstor is a Mellon funded non-profit organization developing digital image collections and resources. The AHSC is a collection of 4,000 images. One of our motivations for focusing on the AHSC is that it is based on thirteen standard art history survey texts, thus there is a strong correlation between the images and texts that describe them. The AHSC images all have metadata providing the name of the work, the artist, date, and so on, but very few have subject matter metadata.

We are currently using two of the texts from the AHSC concordance of thirteen art history survey volumes. Both books cover a broad time range, from Neolithic art to late 20th century, and have a similar lineup of chapter topics. Each text contains roughly thirty chapters (approximately five megabytes), with twenty to forty color images each.

For research purposes, we created electronic versions of the two texts, encoded in TEI compliant xml. TEI is a widely used interdisciplinary standard of text representation. (See http://www.tei-c.org/Lite/teiu5_split_en.html for more detail of this schema.) Chapters, subdivisions, and paragraphs (but not sentences) have distinctive xml tags.

To construct image/text pairs for our text labeling experiments, we started with a software module we developed for importing text into our image indexer's Toolkit. This software module relies primarily on the relative position of xml tags for image plates, major text divisions, and paragraph boundaries. It takes a chapter as input, and produces a list of all the plates in the chapter, with each plate number associated with a sequential list of associated paragraph numbers. We manually correct the output before importing the data into our labeling interface. Using Google image search, we locate non-copyrighted images of the works depicted in the book plates.

4 Text Labeling Experiments

4.1 Semantic Category Labels

Our current guidelines give four pieces of information per semantic category: the category name, one or two questions the labeled text should answer, one or two paragraphs describing the category, and four image/text pairs that exemplify each category. For the Image Content category (or label), the questions are *Does the text describe what the art work looks like? What conventional use of symbols does the artist rely on?*

Over a period of four months, we developed a set of functional semantic categories for classifying paragraphs and sentences in our art history survey texts. Three criteria motivated the classification. Most important, we did not attempt to develop an independent set of categories based on existing image indexing work. We took the texts as our starting point. Second, the set of classes were designed to apply to all chapters regardless of time period, and to allow most extracted paragraphs or sentences to fall into a specific category (i.e., to constitute a "covering" of the text), rather than to a default *Other* class. Finally, we worked with an image librarian at Columbia University and a metadata expert to arrive at a relevant set.

Table 1 summarizes our seven semantic categories. The column on the left indicates the name of the label, and the column on the right gives a highly abbreviated description

of the type of textual content that should be assigned a given label. The labels appear here in the same order that they appear in the interface. It puts the most central category first (Image Content), and lists categories that have a similar focus together. Thus the first three categories are all about the depicted art work (form, meaning, manner); Biographic and Historical Context are both about the historical context.

Table 1. Seven Functional Semantic Categories for Labeling Text Extracts

Category Label	Description
Image Content	Text that mentions the depicted object, discusses the subject matter, and describes what the artwork looks like, or contains.
Interpretation	Text in which the author provides his or her interpretation of the work.
Implementation	Text that explains artistic methods used to create the work, including the style, any technical problems, new techniques or approaches, etc.
Comparison	Text that discusses the art object in reference to one or more other works to compare or contrast the imagery, technique, subject matter, materials, etc.
Biographic	Text that provides information about the artist, the patron, or other people involved in creating the work, or who have a direct and meaningful link to the work after it was created.
Historical Context	Text describing the social or historical context in which the depicted work was created, including who commissioned it, or the impact of the image on the social or historical context of the time.
Significance	Text pointing to the specific art historical significance of the image.

During the first month, we arrived at a provisional set of six categories consisting of everything in Figure 1 apart from the italicized category, which now has the name *Implementation*, and developed our first set of guidelines. We added the seventh category after a month or so of pilot work. During the remaining three months we revised the category names and definitions.

4.2 Materials: Datasets, Annotation Constraints, Annotators, and other Task Parameters

We created three sets of image/text pairs for use in the experiments listed in Table 2. Set 1 consisted of thirteen images and 52 associated paragraphs. Set 2 consisted of nine images and 24 associated paragraphs. Set 3 consisted of ten images taken from two new chapters, and was used for sentence labeling (159 sentences) and paragraph labeling (24 paragraphs). Up to four paragraphs were associated with each image, but in most cases there were one or two paragraphs.

Labelers were recruited from the team of project researchers, their acquaintances, and colleagues at other institutions involved in image indexing.

The two parameters of most interest for comparing the experiments appear in columns five (Labels/Par) and six (Annotators). For the first two experiments, the first two coauthors were the annotators, and the number of labels that could be assigned to a single

Table 2. Annotation Task Parameters

Exp	Set	Images	Units	Label Set	Labels/Par	Annotators
1	1	13	52	6	any	2
2	2	9	24	7	any	2
3	2	9	24	7	two	5
4a	3	10	24	7	one	7
4b	3	10	159	7	one	7

paragraph was unrestricted (any). In experiment 1, the maximum number of labels for a single paragraph was three; each annotator used three labels twice; 99% of the labelings consisted of one or two labels. In experiment 2, 71% of all labels from both annotators were one or two labels; the maximum of four labels occurred once per annotator.

Due to the relative infrequency of more than two labels in experiments 1 and 2, we added a restriction in experiment three that only two labels could be used. In experiment four, we restricted annotators to a single label per text extract, but expanded the task to include sentences.

For experiments 1 through 3, the labeling was done with pen and paper. For experiment 4, we implemented a browser-based labeling interface that included the guidelines, training materials, and labeling task. Since then, we have developed a much more flexible web-based labeling interface using the Django python environment.

In all our studies and data collection, labelers worked independently at remote sites, and could suspend and resume work at will. After experiment 3, labelers were required to go through a training sequence (approx. one hour). Paragraphs were presented one at a time along with the corresponding image. When we began using sentences as well as paragraphs, labelers would first select a paragraph label; then the labeler would be presented with the same paragraph in a sentence-by-sentence format. Labelers had an opportunity to review and revise their choices.

4.3 Evaluation Metrics

We report interannotator agreement using Krippendorff's α [6], which factors out chance agreement. It ranges from 1 for perfect agreement to values close to -1 for maximally non-random disagreement, with 0 representing no difference from chance distribution. An advantageous feature of α is that instead of treating agreement as a binary distinction, it permits the use of a distance metric to weight the degree of agreement from 0 to 1. Because annotators could make multiple selections, we used a distance metric we refer to as MASI [9]. It is intended for set-based annotations, and gives partial agreement credit when the annotators' sets overlap. Our experiments typically allowed annotators to assign multiple labels to the same text. If one annotator assigns the single label {**Image Content**} to the same text that another annotator labels {**Image Content**, **Implementation**}, a non-weighted agreement measure would assign a score of 0 for non-agreement. In contrast, MASI would assign a weighting of $\frac{1}{3}$ (see [9] for details).

4.4 Human Labeling Pilot Studies

Table 3. Interannotator consistency of paragraph labeling under multiple conditions

Exper.	Dataset	Label Set	#Choices	#Labelers	α_{MASI}
1	Set 1	6	any	2	0.76
2	Set 2	7	any	2	0.93
3	Set 2	7	two	5	0.46
4a	Set 3	7	one	7	0.24
4a'	Set 3	7	merge 4b	7	0.36
4b	Set 3	7	one	7	0.30

Results for the four pilot experiments appear in Table 3. Experiment 2, with the final labeling set of seven labels, the first two co-authors as the sole annotators, and any number of label choices, had the best results. It improved on experiment 1, which used an earlier, less well-defined set of labels. It had a larger set of units (52 rather than 24 paragraphs) from two texts, rather than from a single text.

Experiment 3 was the first attempt to use a larger set of annotators. We hypothesized that with each new annotator, the number of distinct combinations of labels would increase, with the result that a large number of annotators would result in a large set of distinct classes, and correspondingly sparser data. In order to guard against this possiblity, we restricted the number of labels that annotators could apply to two. The resulting α_{MASI} score of 0.46 reflects the relative unfamiliarity of a majority of the five annotators with the labeling categories and domain. When we computed interannotator consistency for all combinations of annotators from the set of five, we found that the two experienced annotators had values on the three measures (0.88, 0.88, 0.90) that were consistent with the results of experiment 2.

We collected sentence labelings for the first time in experiment 4: 4a pertains to the paragraph labels, and 4b to the sentence labels. For experiment 4a', we computed agreement on paragraphs based on merging the sentence labels. We created a relatively short label consisting of each distinct type of label applied to any sentence in the paragraph. If three sentences of a five-sentence paragraph were labeled Image Content and two were labeled Historical Context, the paragraph level label we compute is the multilabel consisting of these two labels.

Experiments 4a and 4b yielded the poorest results, which we attribute to the constraint that annotators could only apply one label. The seven labelers consisted of the first two co-authors, plus five new annotators. As in experiment 3, we computed interannotator agreement metrics for all combinations of annotators in experiment 4a. For all 21 pairs of annotators, agreement ranged from a low of 0.15 to a high of 0.32.

In addition to much variation across annotators, we find wide variation depending on the individual units consisting of a single image and all text asscoiated with the image. For the ten units, agreement ranged from 0.12 to 0.40.

4.5 Initial Results of Large Scale Human Labeling

A key feature of our new labeling interface is that labelers can work concurrently on distinct labeling tasks. We plan to collect data on between six and ten datasets. We have currently collected labelings from six annotators on the first dataset consisting of 25 images (45 paragraphs, 313 sentences).

In the new interface, annotators can choose any number of labels. We recruited four new labelers, and used one previous labeler (not a co-author). Results for the first dataset, which consists of 25 images and 48 associated paragraphs (313 sentences), are better than experiments 3 and 4 where we also used multiple annotators. We believe the improvement is due to the training provided in the interface, and the lack of constraint on the number of labels annotators could pick.

As in experiment 4, sentence labeling had a higher agreement than for paragraphs. For sentences the overall α measure was 0.45, compared with 0.40 for paragraphs. For all combinations of 2 to 5 coders, paragraph labeling agreement ranged from 0.56 to 0.27, and ranged from 0.55 to 0.33 for sentences. Again, the two coders who are experts in the area of image indexing had the highest interannotator agreement. As in the pilot studies, there was a significant variation in agreement, depending on the unit, ranging from a high of 0.70 to a low of 0.16.

The most frequent label combination for both paragraphs and sentences was the single label Image Content. There were 47 distinct combinations of labels for sentences, of which 34 were label pairs and five were triples; the remaining 8 unigram labels were the seven labels plus the default "Other". There were 38 combinations for paragraphs: 7 singletons, 20 pairs, 10 triples, and 1 combination of four labels.

5 Preliminary Machine Learning Results

Using data from our pilot studies of human labeling, augmented by an additional set of images labeled by one of the co-authors, we investigated the learnability of three categories: Image Content, Historical Context and Implementation. There were insufficient examples from the other categories. All learning was done using WEKA [13], a Java-based toolkit that implements a wide range of machine-learning algorithms using a standard input format.

A typical text classification task involves topic or subject matter identification. In contrast, our task involves functional categories that are more abstract than topic. The novel nature of our classes (and source texts) warranted an exploratory phase prior to conducting large scale machine learning experiments. Our pilot machine learning dataset permitted us to explore features for text representation, and to experiment with a range of learning techniques.

To establish a baseline, we started with a classic text classification approach: Naive Bayes classifiers using unigram word counts to represent the input texts, with separate classifiers for each semantic category. We used chi-square independence tests to prune the word vectors in order to reduce the noise in the training data, which smaller corpora such as ours are sensitive to. The Image Content classifier performed best, presumably due to the relatively greater amount of training data.

We experimented with two other types of text representation; One drew on hand-selected features from approximately half a dozen semantic domains. For example, we created sets of key words and phrases characteristic of the art history domain, (e.g., *masterpiece*), and other words and phrases deemed discriminative by our experts and annotators, such as terms denoting parts of the human body. Input texts were represented using raw counts for these hand-picked features. Our third text representation combined hand-picked features with the unigram word count representation.

We tested several learning methods in addition to Naive Bayes (NB), including support vector machines and tree-based learners. NB performed best overall. On train-and-test validation, the highest classification accuracy was on Image Content using the third type of text representation, with 83% accuracy. Using the same representation and learner, accuracy was 63% for Historical Context and 53% for Implementation. The best Historical Context classifier was NB with the first text representation, where accuracy was 70%. The NB classifiers performed poorly for our third category, Implementation. For this category, performance improved significantly using Random Forest, one of the tree-based learning methods we tried. In more recent work, we have achieved similar performance on the same dataset using a wider range of automatically derived features. For example, we now use part of speech tagging to identify present tense, which helps discriminate Image Content from the other two classes. We are currently experimenting with WordNet's semantic distance metrics to weight terms with respect to various semantic domains, such as appearance, physical dimension, and so on.

6 Conclusions and Future Work

We have presented a detailed analysis of the development of a functional semantic labeling for art history texts, and have identified some of the problems that arise in achieving consistently high agreement scores across multiple annotators. One issue, the variance across texts, is more difficult to address. The other key issue is that annotators with expertise are much more consistent with each other than non experts. As we continue collecting data, and updating our training with the expert consensus on previously labeled examples, we hope to learn something about training and experts. However, we have found that we can still achieve high accuracy with machine learning. As pointed out in [12], the relationship between interannotator agreement and learnability is not a predictable one.

We believe the initial learning results are quite promising. One difficulty for learning functional semantic categories is that many of the content words are not relevant features, since they will be different for descriptions of different images. In contrast, for topical text classification, content words are often sufficient for automatic classification, which is the intuition behind approaches such as latent semantic indexing. By using features such as verb tense, which distinguishes the **Image Content** class from others, we have achieved high results on relatively small datasets. On the other hand, since our categories are functional, they may transfer more easily to texts that are substantially different from our training and test materials.

As illustrated in the introduction, we anticipate that classifying text into functional semantic categories can provide more control over selection of metadata. Our categories

have a rough correspondence with categories discussed in the image indexing literature [7, 3, 2]. As a result, it should be possible to map between our categories and the types of controlled vocabularies used in university visual resource centers. The external knowledge sources our project has examined include the three Getty resources (Art and Architecture Thesaurus, Thesaurus of Geographic Names, Union List of Artist Names), the Library of Congress Authorities and Library of Congress Thesauri for Graphic Materials, and ICONCLASS, a library classifi cation for art and iconography.

References

1. R. Artstein and M. Poesio. Kappa^3 = alpha (or beta). Technical Report NLE Technote 2005-01, University of Essex, Essex, 2005.
2. M. Baca. *Practical Issues in Applying Metadata Schemas and Controlled Vocabularies to Cultural Heritage Information*. The Haworth Press, Inc., 2003. Available through Library Literature, last accessed July 25, 2006.
3. H. Chen. An analysis of image queries in the field of art history. *Journal of the American Society for Information Science and Technology*, pages 260–273, 2001.
4. A. Giral and A. Taylor. Indexing overlap and consistency between the Avery Index to Architectural Periodicals and the Architectural Periodicals Index. Library Resources and Technical Services 37(1):19-44, 1993.
5. J. Klavans. Using computational linguistic techniques and thesauri for enhancing metadata records in image search: The CLiMB project. Article in preparation.
6. K. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA, 1980.
7. S. S. Layne. Some issues in the indexing of images. Journal of the American Society for Information Science, pages 583–8, 1994.
8. K. Markey. Interindexer consistency tests: a literature review and report of a test of consistency in indexing visual materials. Library and Information Science Research, pages 155–177, 1984.
9. R. Passonneau. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), 2006.
10. R. Passonneau, N. Habash and O. Rambow. Inter-annotatator agreement on a multilingual semantic annotation task. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), 2006.
11. R. J. Passonneau, D. Elson, R. Blitz, and J. Klavans. CLiMB Toolkit: A case study of iterative evaluation in a multidis ciplinary project. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), 2006.
12. D. Riedsma and J. Carletta. Reliability measurement: there's no safe limit To appear in Computational Linguistics.
13. I. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann: San Fransisco, 2000.