Relation between Agreement Measures on Human Labeling and Machine Learning Performance: Results from an Art History Domain

Becky Passonneau, Columbia University

Tom Lippincott, Columbia University

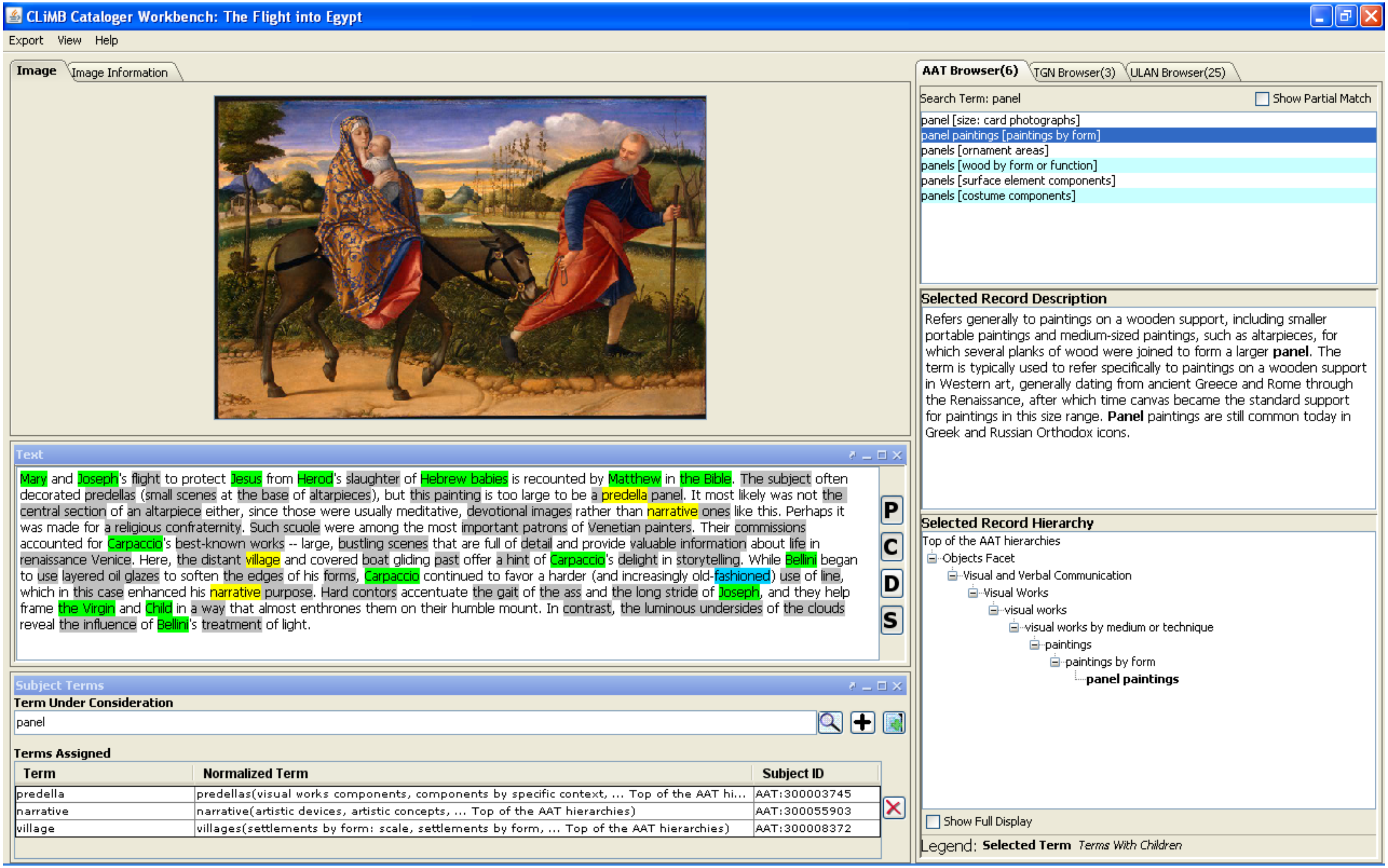
Tae Yano, Carnegie Mellon University

Judith Klavans, University of Maryland

FSC Image/Text Set: AHSC

- Images: *ARTstor Art History Survey Collection;* 4000 works of art and architecture
- Texts: two from a concordance of a dozen art history surveys used in creating the AHSC
- Meets our criteria: Curated, minimal cataloging, image/text association
- Characteristics of the texts:
  - Neolithic art to 20th century
  - About 30 chapters each; 20-40 platesper chapter (surrogate images freely available on the web)
  - Document encoding: TEI Lite
  - One to four paragraphs per image

Image Indexer’s Workbench



Example

Ram and Tree. Offering stand from Ur. c. 2600 B.C.

A far more realistic style is found in Sumerian sculpture . . . put together from varied substances such as wood, gold leaf, and lapis lazuli. Some assemblages . . . roughly contemporary with the Tell Asmar figures, have been found in the tombs at Ur . . . including the fascinating object shown in an offering stand in the shape of a ram rearing up against a flowering tree.

<p>
<semcat type="**implementation**">. . . substances such as wood, gold leaf, and lapis . </semcat>
<semcat type="**historical_contex**t"> . . . contemporary with the Tell Asmar figures . . . </semcat>
<semcat type="**image_conten**t"> . . .offering stand in the shape of a ram rearing up against a flowering tree.</semcat> . . .</p>

Motivation

Ram and Tree. Offering stand from Ur. c. 2600 B.C.

- Allow indexer's to choose what type of metadata to look for
  - Add descriptors about the work
  - Add descriptors about provenance
- Allow end user's to constrain the semantics of a search term
  - OF: Tell Asmar figures
  - Same Period: Tell Asmar figures

Functional Semantic Categories

Category Label	Rough Description
Image Content	**Describes the appearance or other objective features of the depicted object**
Interpretation	The author provides his or her interpretation of the work
Implementation	**Explains artistic methods/materials used in the work, including style, techniques**
Comparison	Comparison to another art work in order to make/develop an art historical claim
Biographic	Information about the artist, patron or other people involved in creating the work
Historical Context	**Description of historical, social, cultural context**
Significance	Explanation of art historical significance

Table of Results from Pilot Annotations

Exp	Dataset	#Labels	#Anns	Alpha (MASI)
1	I: 13 images, 52 paragraphs	any	2	0.76
2	II: 9 images, 24 paragraphs	any	2	0.93
3	II: (ditto)	two	5	0.46
4a	III: 10 images, 24 paragraphs	one	7	0.24
4b	III: 10 images, 159 sentences	one	7	0.30

- Comparable range to previous work

Summary of IA Results

- Semi-controlled study
  - IA decreases when restricted to one label per item
  - IA decreases with more annotators
- Pairwise IA for experiments varied widely
  - For 4a, 0.46 to -0.10 (7 annotators)
  - For 4b, same range
- IA varied greatly with the image/text unit
  - High of 0.40 for 7 annotators in 4a (units 1, 9)
  - Low of 0.02 for 7 annotators in 4a (unit 5)

Conclusions from Pilot Annotation Experiments

To optimize annotation quality for our large scale effort (50-75 images and 600-900 sentences):

- Allow multiple labels
- Develop annotation interface (with online training)
- Use many annotators, post-select the highest quality annotations
- Partition the data in many ways

Specific Questions

- Does ML performance correlate with IA among X annotators on class labels?
  - Compute IA for each class
  - Rank the X classes
- Does ML performance correlate with IA across Y annotators on a given class?
  - Compute Y-1 pairwise IA values for each annotator
  - Rank the Y annotators
  - *Swap in* each next annotator's labels

Data

- Three binary classifications, IA per class
  - Historical Context: 0.39
  - Image Content: 0.21
  - Implementation: 0.19
- Training data: 100 paragraphs labeled by D
- Test data: Single label per annotator
  - 24 paragraphs labeled by six remaining annotators in Exp 4
  - 6 paragraphs labeled by two annotators in Exp 2

Annotators’ Average Pairwise IA, for all FSC labels

Annotator	Avg. Pairwise IA (sd)	IA Year 1, Year 2
A	0.32 (0.12)	
A’	0.31 (0.10)	0.34
A’’	0.28 (0.13)	
B	0.21 (0.15)	0.88
C	0.17 (0.11)	
D	0.14 (0.14)	
E	0.10 (0.16)	

Machine Learning

- Naïve bayes, binary classifiers
  - Performs better than multinomial NB on small datasets
  - Performs well when independence assumption is violated
- Three feature sets
  - Bag-of-words (BOW)
  - Part-of-speech (POS): 4-level backoff tagger
  - Both

Annotator *Swap* Experiments

- For each classifier *and f*or each feature set
  - Disjoint training/testing data
    - Train on same 100 paragraphs, annotated by D
    - Test by swapping in annotations of 24 paragraphs by A, A', A", B, C, E (plus the 6 paragraph training set)
  - 10-fold cross validation on 130 paragraphs
    - For the 24 paragraph set, swap in each next annotator
- Correlate:
  - Average ML performance on 3 classes with per-class IA
  - Individual learning runs with individual annotators

Average ML per Condition Correlates with per-Class IA

- 6 runs X 3 feature sets X 2 evaluation paradigms
- Average learning performance correlates with IA among 6 annotators on bow and both, not on pos

	Train 100/Test 30			10-Fold Crossval 130		
	bow	pos	both	bow	pos	both
Historical Cont.	0.71	0.68	0.71	0.75	0.69	0.77
Image Content	0.57	0.72	0.57	0.63	0.69	0.63
Implementation	0.59	0.44	0.59	0.60	0.59	0.60
Correlation	**0.98**	0.46	**0.98**	**1.00**	0.58	**1.00**

Individual ML Runs do not Correlate with Annotator Rank

Train100/Test30					
Historical Context		Image Content		Implementation	
bow	0.05	bow	-0.25	bow	-0.43
pos	0.18	pos	-0.75	pos	-0.01
both	0.59	both	0.42	both	-0.43
Crossval 130					
bow	0.11	bow	-0.06	bow	-0.77
pos	-0.87	pos	0.07	pos	0.46
both	0.71	both	0.14	both	-0.87

Details: Individual Annotators/ML Runs

- Annotator A
  - Highest ranked annotator
  - Often the low(est) ML performance
- Annotator B
  - Mid-ranked
  - Often near top ML for Image Content and Implementation
- Annotator E
  - Lowest ranked annotator
  - Occasionally has highest ranked runs

Details: Feature Sets

- BOW: high dimensionality, low generality
- POS: low dimensionality, high generality
- Whether BOW/POS/Both does well depends on
  - Which classifier
  - Which annotator’s data
- POS > BOW for Image Content on average
- BOW > POS for Historical Context on average

Conclusions

- We need to repeat experiment on larger dataset
- Semantic annotation requirements
  - No *a priori* best IA threshold
  - More qualitative analysis of label distributions
- ML correlated with per-class IA
- ML did not correlate with individuals' IA

Discussion

- When using human labeled data for learning:
  - Data from a single annotator with high IA does not guarantee good learning data
  - Data from an annotator with poor IA does not guarantee the data is not good learning data
  - Different annotations may lead to different feature sets
- Learners should learn what a range of annotators do, not what one annotator does

Current and Future Work

- Large-scale annotation effort: 5 annotators
  - Done: 50 images/600 sentences from two texts, same time period (Ancient Egypt)
  - To do: 50 images/600 sentences from two new time periods (Early Medieval Europe; other)
- Redo annotator swap experiment on larger datasets
- Multilabel learning
- Learning from multiple annotators
- Feature selection