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

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

[1] Columbia University
New York, NY, USA
(becky|tom)@cs.columbia.edu

[2] Carnegie Mellon University
Pittsburgh, PA, USA
taey@cs.cmu.edu

[3] University of Maryland
College Park, MD, USA
jklavans@umd.edu

Abstract

We discuss factors that affect human agreement on a semantic labeling task in the art history domain, based on the results of four experiments where we varied the number of labels annotators could assign, the number of annotators, the type and amount of training they received, and the size of the text span being labeled. Using the labelings from one experiment involving seven annotators, we investigate the relation between interannotator agreement and machine learning performance. We construct binary classifiers and vary the training and test data by swapping the labelings from the seven annotators. First, we find performance is often quite good despite lower than recommended interannotator agreement. Second, we find that on average, learning performance for a given functional semantic category correlates with the overall agreement among the seven annotators for that category. Third, we find that learning performance on the data from a given annotator does not correlate with the quality of that annotator's labeling. We offer recommendations for the use of labeled data in machine learning, and argue that learners should attempt to accommodate human variation. We also note implications for large scale corpus annotation projects that deal with similarly subjective phenomena.

1. Introduction

We conducted a series of pilot annotation studies in the context of identifying specifications for marking up textual input for an image cataloger's toolkit (Klavans et al. 2008). Given an image, and a text extract that describes the work depicted in the image, we aimed to identify the semantic functions of the text. By semantic function, we mean the type of information provided; for example, a description of the work depicted in the image, versus biographical background on the artist (Passonneau et al. 2007). Interannotator agreement (IA) on semantic or pragmatic annotation tasks such as ours is typically difficult to achieve [see (Artstein & Poesio 2005) for a brief review]. Because, variation within and across individuals is an inherent feature of language use, we decided to investigate how this variation affects learning performance.

In consultation with domain experts, we developed a set of seven functional semantic categories to apply to paragraphs or sentences associated with specific images. Our categories were derived from what we observed in the texts, but have a loose correspondence with categories of information discussed in the image indexing literature (Layne 1994; Chen 2001; Baca 2003). By marking up electronic text with these categories, catalogers can select the type of information they want to see in searching for metadata. Figure 1 in the next section illustrates three of the seven categories that we focus on in this paper.

Our goals in conducting our pilot annotation studies were to understand why previous investigators have found such a wide range of agreement on similar tasks (Giral & Taylor 1993; Markey 1984), and to develop annotation specifications for our large scale study. We conducted four experiments under a variety of annotation constraints to guide the design of a large scale annotation effort. Our goal with respect to contributing to the image cataloger's toolkit is to find a set of one or more labels that would be useful to image catalogers, and that an automatic classifier can apply with high reliability to a comparable corpus of art history survey texts. However, it has been a continuing concern in our previous work to understand the impact of human variation (e.g., on discourse in Passonneau & Litman 1997; on summarization in Nenkova et al. 2007). Our pilot data presented us with the opportunity to examine the relation between learning performance and which annotator's data we select.

If an annotator fails to agree well with other annotators and makes non-systematic choices, machine learning performance can be expected to be relatively lower than for annotators with good agreement. However, if an annotator makes idiosyncratic choices that are nevertheless based on reasoned criteria, there is no reason a machine learning algorithm should fail to discover patterns in that annotator's choices. Our goal here is to explore on several levels whether IA correlates with learning performance.

Using very simple text representation features, we find relatively good learning performance despite low IA. We also find that on average, learning performance for a given functional semantic category correlates with the IA for that category, measured across seven annotators. For

example, we find the highest IA among seven annotators on Historical Context, and we find that the Historical Context classifiers do best. This is a reassuring result. However, we also find that learning performance on a particular annotator's data does not correlate with the annotator's ranking, based on averaging the annotator's pairwise agreement with other annotators. This is less reassuring, and suggests that factors in addition to an annotator's IA should be considered in selecting data for machine learning.

In the next section, we present a brief example of our functional semantic categories, followed by a section discussing related work. In section 4, we discuss the results of our annotation experiments. In section 5, we present results on the impact of selecting data from different annotators for machine learning. We discuss these results in section 6, and conclude in section 7 with recommendations regarding learning from inherently subjective data.

2. Brief Example

The domain of digital images and texts we focus on for our study of functional semantic categories parallels the ARTstor *Art HistorySurvey Collection (AHSC)*, a Mellon funded collection of 4,000 images. The AHSC is based on thirteen standard art history survey texts, thus there is a strong correlation between the images and the texts. The AHSC images all have tombstone metadata (e.g., the name of the work, the artist, date, the location of the work), but very few have subject matter metadata. As input for our studies, we are currently using two texts from the AHSC concordance, scanned and encoded in TEI-Lite (http://www.teic.org/Lite/teiu5 split en.html).

Using the TEI-Lite markup as input, we developed a simple algorithm to associate a sequence of one or more paragraphs with each image caption. The descriptive information a paragraph provides about an image can be categorized into types depending on the semantic function of the text. Figure 1 illustrates text from the first part of a few paragraphs associated with an image of a relief portrait of Akhenaten and his family. The image comes from the ARTstor Images for College Teaching: http://www.arthist.umn.edu/aict/html/ancient/EN/EN006.html. The text fragment is from one of the texts in the concordance to the ARTstor Art History Survey Collection. Here we have separated several sentences from the paragraph into labeled text spans exemplifying the three categories we performed machine learning on. As illustrated here, a single sentence can have subparts with distinct semantic functions (e.g., **Implementation** and **Image Content** for the sentence beginning *Known as the Amarna style,…*).

If we can automatically classify sentences into functional semantic categories, we can add this information to the electronic document markup, as shown in the provisional XML representation in Figure 1. The categories appear as values of a semcat attribute. Note that the XML shows a sentence-level assignment; we do not attempt to label subspans within sentences. However, in both the manual annotation and machine classification, we allow multiple labels per text span.

The full list of functional semantic labels includes the three shown in Figure 1, plus four more: **Interpretation, Biographic, Significance** and **Comparison**. Annotators could choose **Other** when none of the above applied. The labels, definitions, and examples appear on a set of web pages replicating the guidelines we provided in our labeling interface: http://www1.ccls.columbia.edu/~beck/.

The image/paragraph pair in Figure 1 is drawn from our first human labeling experiment (see Table 1 below). Two of the co-authors independently labeled all paragraphs, with the option of selecting multiple labels. The instructions were to pick a single label for each paragraph if possible, and to pick multiple labels only if the functions were equally balanced. The two labelings assigned to the full paragraph excerpted in Figure 1 were (**Image Content, Historical Context**) and (**Image Content, Significance**), thus both labelers viewed the paragraph has having two relatively equal semantic functions. They agreed on one of the functions, and disagreed on the other. In section four, we refer to the use of a weighted agreement coefficient to treat such cases as partial agreement.

3. Related Work

There have been no studies we know of that look at all three issues we address, namely the factors affecting IA on a semantic labeling task, machine learning performance on the same data, and the relation between the two. We briefly review each issue taken separately.

In the twenty some years since Markey's (1984) comprehensive summary and comparison of forty years of inter-indexer consistency tests, no comparable review has appeared, and her observations still appear to 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. The agreement scores, using accuracy (percent agreement), range from 82% to a low of 4%. Markey noted that greater inter-indexer consistency was attained when indexers employed a standardized classification scheme, comparable to a controlled vocabulary. However, even with controlled vocabularies, percent agreement ranges from 34% to 80%.

Percent agreement has the weakness that it is highly sensitive to the number and absolute frequency of categories assigned. If two categories are used, one of which is extremely frequent, percent agreement will necessarily be high (Artstein & Poesio 2005). While we use more robust methods for quantifying inter-annotator agreement, we find a similar range of values across four labeling experiments we conducted.

Giral and Taylor (1993) looked at indexing overlap and consistency on catalog records for the same items in architectural collections; they examined record data for title, geographic place names, and so on, including an analysis of subject descriptors. On large samples (>1400)

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 based on a formula for binomial proportions. For subject descriptors, only 7% match entirely, and they find overlap in descriptors in only about 40% of the remaining cases (+/- 3%).

The text classification task we address differs from many NLP classification tasks in that the type of text we look at has not been widely studied, and the categories are orthogonal to topic. Three recent studies that also pertain to atypical texts and non-topical classification tasks are Teufel and Moens (2002), Hachey and Grover (2004), and Argamon et al. (2007). On the task of categorizing sentences from scientific articles into argumentative classes, Teufel and Moens get IA values of between 0.70 and 0.80 from pairs of well-trained annotators, and between 0.35 and 0.72 on briefly-trained annotators. We find a similar range but lower absolute values on texts which are inherently more subjective. For labeling of legal arguments, Hachey and Grover get IA of 0.83 for one pair of well-trained annotators. Argamon et al. propose stylistic features derived from semantic functions of lexical items, and evaluate their features on a variety of tasks, such as authorship or sentiment evaluation.

The work most related to ours is a forthcoming article by Riedsma and Carletta who report on simulations of learnability from data with different levels of agreement. They present evidence that performance of machine learners does not correlate directly with agreement levels, and argue that systematic or patterned disagreements can lead to spurious learning, and are more harmful than disagreements that represent noise.

4. Human Labeling

Given the wide range of measures of human agreement on a related task where librarians classify documents with respect to an existing set of categories (Markey 1984) (Giral & Taylor 1993), we wanted to understand what factors might lead to variations on IA in our task. We conducted four pilot studies on the labeling where we varied the number of labels that could be assigned to a single item (one, two or unrestricted), the size of the text fragment being labeled (paragraph or sentence), the number of annotators (two to seven), and the type of training for annotators (none, static examples presented to trainees). Experiments one through three were done on paper and pencil or electronic editors; for four a and four b we implemented a labeling interface.

To measure IA, we use Krippendorff's *Alpha*[1] (Krippendorff 1980) along with MASI, a set-based distance measure (Passonneau 2006). MASI allows partial credit when the set of labels chosen by one annotator overlaps another's set. Used in the context of Alpha, MASI weights the comparison of every pair of annotators' choices for a given unit in a way that takes into account the relative sizes of the set of labels chosen by each annotator, and the type of overlap between the two sets. To quantify the relative size of the overlap, MASI incorporates Jaccard (1908), which is the ratio of the size of the set intersection to the set union. Jaccard does not take into account whether two sets are in a subsumption or difference relation; the former is monotonic, thus represents a less serious semantic conflict. MASI incorporates an equi-distant 4 point scale from 0 to 1 corresponding to the four possibilities of set identity, set subsumption, set difference and set disjunction. The example pair of labels from section 2, (**Image Content, Historical Context**) and (**Image Content, Significance**) would get a MASI distance of (1/2 x 1/3) for partial agreement rather than 0 for complete agreement or 1 for complete disagreement; see (Passonneau 2006) for details. Because IA coefficients do not directly capture the quantity of matches across annotations, we also report the average F measure taking each next annotation as the target of comparison. As we see in Table 1, the average F measure is about the same for rows 3 and 4, although IA is lower in row 4.

Annotation efforts typically aim for agreement measures above a threshold of 0.67, due to Krippendorff (1980). We have previously argued that because IA coefficients do not have a known probability distribution, and because they are applied to many kinds of data and for many types of annotator judgements, there is no one ideal threshold (Passonneau 2006; Passonneau et al 2006). Instead, we suggest that interpretation of IA values is an empirical question that depends in part on how the data will be used. It can be investigated in many ways, for example relating measures of tasks in which the annotations are used to the observed agreement levels.

For the four pilot studies, annotators were presented with images, the associated texts, and annotation guidelines. Table 1 shows the four experiments with the size of the data set, how many labels annotators could select, how many annotators were used, and the Alpha MASI values. IA varies widely across experiments. Comparison of rows 1-3 with 4 and 5 in Table 1 indicate that IA is higher when annotators can select multiple labels, which is consistent with our previous results on a lexical semantic annotation task (Passonneau et al., 2006). The biggest drop in IA, at experiments 4a and 4b, is due to the constraint that labelers select a single label.

We also found that IA varies widely between annotators, suggesting that the task is inherently easier for certain individuals. The two annotators from experiments 1 and 2 are included in all the experiments, and always have high IA with each other. In looking at IA for all pairs of annotators in experiments 3 through 4b, we find conflicting results regarding experience and training. In experiment three, the novices had lower average pairwise agreement while in experiments 4a and 4b the three annotators with the highest IA values were new to the task. This may be due in part to the fact that there was little training in experiment 3, which was done using the annotator's favorite text editor, compared with experiment

[1] Multi-annotator weighted Kappa (Arstein & Poesio forthcoming) gives us almost identical IA values.

4, where we provided a labeling GUI with more detailed guidelines that incorporated four training examples. We find a very wide range of overall agreement depending on the unit being labeled, where a unit is an image and all paragraphs associated with it (4 paragraphs per image in experiment 1, 2.7 in exps. 2-5). Across the ten units in experiment 4a, IA ranged from 0.40 to 0.02. Finally, sentences had higher agreement than paragraphs.[2]

Table 1 indicates that IA varies across experimental conditions, and we found that pairwise IA varies across individuals. We also looked at consistency within the same individual. The annotation for experiment 4a was originally performed in January, 2007, and included three of the co-authors. Two of the co-authors reannotated the same data in March, 2008; alpha was 0.88 for one (referred to as B below) and 0.34 for the other (referred to as A' below). For B, twenty two out of twenty four items had the same label in the two annotations; for A' only fourteen out of twenty four were the same.

We conclude from the IA results that the labeling task is very subjective, yet the reception to our categories from image catalogers, visual resource professionals, and other domain experts has been uniformly positive. In addition, the annotators who have participated in our studies find the judgements difficult to make, but find the categories meaningful and relevant for distinguishing among types of information provided about an image.

For the machine learning experiments where we look at the relation to agreement, we use the labelings from experiment 4a. One important advantage to using this data, despite the fact that overall interannotator agreement is relatively low here, is that annotators chose a single label. This simplifies the computation of IA, and the interpretation of results.

We assume that an annotator who is more consistent with other annotators makes judgements that are less idiosyncratic and more representative of the linguistic community. In line with this assumption, we ranked annotators by their average pairwise IA. First, we collapsed all labels other than the three of interest (Image Content, Historical Context, Implementation) into a single Other category. Then we averaged the six alpha values for each annotator paired in turn with the other annotators. Pairwise IA ranged from 0.46 to -0.10. The annotator averages ranged from 0.32 to 0.10. Annotator A', who had low self-consistency, had a relatively high average pairwise IA of 0.31 (sd=0.10); annotator B had high self-consistency and a moderate average pairwise IA of 0.21 (sd=0.15).

5. Machine Learning

[2] We have recently completed an annotation effort on a larger dataset consisting of 50 images and 600 sentences. We implemented a new interface, and provide true training examples with feedback. Five annotators participated, and independently labeled all sentences. We have better results, with very similar patterns of agreement. IA is highest among domain experts, and varies across image/text units.

We investigated the learnability of three of our functional semantic categories: **Image Content, Historical Context** and **Implementation.** There were insufficient examples from the other categories. All learning was done using WEKA (Witten and Frank, 2005). Due to the small size of our dataset, and the similarities of the categories, we used Naïve Bayes, which can perform well even when the independence assumption is violated. We trained a binary NB classifier for each semantic category; it performs better with smaller corpora than multinomial NB (Sebastiani, 2002). Also, we wanted to investigate the relation of IA to learning for each semantic category independent of the others.

5.1. Data sets

To look at the relation between IA and machine learning performance, we conducted three pairs of experiments, one pair for each semantic category. The first experiment in each pair has disjoint training and test data, and the second uses ten-fold cross-validation.

In the first of each pair (train100test30 in Tables 2-4), we train using 100 paragraphs taken from a single chapter about Egyptian art and architecture, labeled by a single annotator (the same one in all experiments). We test using thirty randomly selected paragraphs from chapters covering the art and architecture of ancient Egypt, Romanesque Europe, and twentieth century Western art and architecture. For twenty four of the thirty test paragraphs we have labelings from seven annotators. The remaining six paragraphs consist of the training examples from our labeling interface, which were drawn from our earliest pilot studies. The second set of experiments in each pair uses all 130 paragraphs (crossval 130).

5.2. Feature Sets and Learning Algorithm

To facilitate comparison of results when we swap data from different annotators, we used very simple feature sets: bag-of-words (BoW), part of speech tags (POS), and a combination of the two. Before we select terms as features for the training data, we apply preprocessing that tokenizes the text, lowercases words, removes punctuation, then passes the token through WordNet's "morphy" function to return a lemma. The part-of-speech tagger is a 4-level backoff tagger: bigram then unigram, both trained on the Brown corpus, then a regexp tagger followed by the assignment of a single default tag for unknown parts of speech.

5.3. Results

Tables 2 through 4 show the machine learning results when we vary which annotator's labelings we swap in. Following the recommendation of Huang et al., we report results using the area under the receiver operating characteristic (ROC) curve, which they have argued on theoretical (2003a) and empirical (2003b) grounds to be more reliable, consistent, and discriminative than accuracy.

Each table shows the performance for each of the three feature sets (bow, pos, both) under the two conditions

(train100/test30, crossval130) using labelings from the seven annotators (who). The top three annotators have nearly the same average pairwise IA, so we label them A, A' and A". The remainder we label in alphabetical order of their descending rank (B through E). The last two rows of each table show the mean roc scores for all seven runs and their standard deviations.

We computed IA on three datasets for **Historical Context, Image Content** and **Implementation** where anything other than the label of interest was mapped to **Other**. The overall IA among seven annotators for **Historical Context** versus **Other** was Alpha=0.39; this is higher than for the full dataset with eight labels, where IA was 0.24 (see Table 1). For **Image Content** versus **Other** IA was 0.21, and for **Implementation** IA was 0.19.

There was higher average performance for **Historical Context**, which had the highest IA. For example, in the 10-fold cross validation condition, the average performance for **Historical Context** classifiers using both features is 0.77 (sd=0.05) compared with 0.63 (sd=0.05) for Image Content and 0.60 (sd=0.03) for Implementation. The higher sd for the **Historical Context** classifiers indicates a wider range in performance values.

Because our classes represent function rather than content, we believe the optimal features we find in our future work will differ from other text classification tasks. Note that the part-of-speech features do well, particularly for the **Image Content** class. This is not the case for typical text classification tasks, which generally do best with bag-of-words (cf. Forman 2003). The best performing pos feature for **Image Content** is present tense, which corresponds to cases we observe where the image is described as a visual "tour" in present tense.

The key issue of interest here is how the order in each "who" column compares with the alphabetical order representing the ranking of annotators. In general, the ranking of machine learning runs differs from the annotator ranking. The 100-paragraph training set was labeled by D; unsurprisingly, D's runs score relatively high. If we exclude D, the two orderings are most similar for the Historical Context classifiers, and most dissimilar for Implementation. To quantify the observation that individual IA scores do not line up with learnability, we used Pearson's correlation coefficient (excluding D). For example, for each annotator in the second to last column of Table 2 (**Historical Context**, crossval130, both), we replaced the annotator symbol with that annotator's average pairwise IA score, and computed the correlation with the last column. As shown in Table 5 the correlation for this case was 0.59, which was the best overall. The correlation for the condition (**Historical Context**, train100test30, both) was only 0.05. In general, correlations were quite poor, and there were two cases with a high inverse correlation of -0.87: (**Implementation**, crossval130, both) and (**Historical Context**, crossval130, bow).

Four summary observations of interest regarding the disparity between annotator rank and learning performance are:

1. Annotator A is the highest ranked annotator but runs using this data are often at the lowest or next-lowest performance ranking.
2. Annotator B is only the fourth best annotator, but runs using this data are often in the top two for Image Content and Implementation classifiers.
3. Annotator E is the lowest ranked annotator but occasionally has the highest ranked runs.
4. A comparison of the ordering in the *who* column across feature sets for a given classifier and evaluation method (train100test30 vs. crossval130) shows that whether a feature performs well depends on whose data was used.

Points 1 through 3 above suggest that IA scores do not predict how well machine learners can perform on an annotator's data. Machine learning performance can be in an inverse relation to the annotator's average pairwise IA, or can be non-predictive in a less extreme way. Point 4 suggests that feature selection performance can be contingent on who annotates the data.

We also consider whether annotator consistency has an impact on learning performance. Annotator B, whose self-agreement was 0.88, was a mid-ranked annotator. For the Historical Context runs, the B runs were also mid-ranked. For the other two classifier runs, the B runs often ranked first or second, except for the Implementation using POS features. Annotator A' was a high-ranked annotator with low self-agreement of 0.34. The runs for A' were generally much lower than would be expected given the annotator rank, except on the Implementation classifiers on the 100train30test trials. It is tempting to infer that annotator B relied on strategies that were more stable both across time and within the corpus, and that annotator A' relied on strategies that were less so. However, we believe further work on such issues is required to support such a conclusion.

6. Discussion

We presented the results of four human labeling experiments. One of the most significant factors affecting IA was the number of labels that could be assigned, thus in our current large-scale effort, we have returned to allowing an unrestricted number of labels, as in our first two pilot experiments. Another key factor was the specific image/paragraph set being labeled; we have not yet determined whether we can raise IA on the more difficult texts. The two most expert labelers (two of the co-authors) continued to have high IA with each other throughout the four experiments. Novices seemed to perform much better once we provided a labeling interface and training. Although it is possible we could achieve higher IA, for example by revising the categories or increasing the training, the more general conclusion we draw is that the judgements we have elicited are inherently subjective.

IA for a given semantic function appears to correlate with overall learning performance for the corresponding classifier. As we have relatively little data for **Implementation**, and results on only three of the seven

classifiers, we plan to run a similar experiment on our new, larger dataset to see how results compare.

IA for a given annotator appears not to correlate with overall learning performance. For example, choice of annotator among three who have virtually the same average pairwise IA does not result in similar learning. The results differ both with respect to how well the learner performs, and which features yield the highest performance. In (Riedsma & Carletta forthcoming) we find an explanation for how this might occur. They use simulated data to argue that a quantitative measure of agreement does not provide insight into the qualitative type of agreement. In our real dataset, annotator A, whose runs yield low performance scores for the Historical Context classifier, may have patterns of disagreement that confuse the learning algorithm

7. Conclusion

Because we are conducting IA studies in tandem with machine learning, we can investigate the relationship between the two on a complex annotation task pertaining to functional semantic categories. Our results bear out the simulation study presented in Riedsma and Carletta, that good learning performance can occur when agreement is less than the 0.67 threshold proposed by Krippendorff (1980). This does not mean, of course, that good learning performance never requires higher levels of agreement. Instead, it shows that in richly annotated datasets such as this one, where we have attempted to develop a set of fully covering categories, IA and learnablity interact with the distributions of various categories in the labeled data, and with the strategies employed by individual annotators as reflected in their annotation decisions.

We draw the following tentative conclusions regarding the use of manually annotated data in machine learning:

1. We need to interpret machine learning results informed by an understanding of how IA varies across the classes to be learned;
2. We should perform machine learning on data labeled by more than one annotator in order to determine whether annotators are using different strategies, as reflected in differential performance across feature sets;

More generally, we believe more research is needed on the interaction between conventional IA metrics, or other ways of evaluating annotator reliability, and machine learning performance on manually annotated data.

Implicit in the range of results seen here is a much deeper methodological issue. The prevalent paradigm for machine learning in NLP that uses manually labeled data, and the prevalent paradigm for assembling manually annotated datasets, is to arrive at a single labeling, often referred to as a gold standard. Deviation from the gold standard is viewed as problematic, rather than as an inherent property of language use. If we train learners on data that fails to capture the natural variation we see across human language users, we risk reaching an impasse regarding the phenomena that are meaningful precisely because they are the least widely agreed upon.

8. Acknowledgements

This work was conducted as part of the project, Computational Linguistics for Metadata Building (CLiMB), supported by an award from the Mellon Foundation to the University of Maryland. The authors extend enormous thanks to Roberta Blitz, David Elson, Angela Giral, and Dustin Weese, who helped provide expert feedback on the functional categories, on our initial labeling interface and labeling guidelines, and on the issue of consistency among image catalogers. We also thank the annotators who helped label the pilot dataset: James Masciuch, Adam Goodkind, Justin Cranshaw and another we know only as Ginger.

9. References

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., and Levitan, S. (2007) Stylistic text classification using functional lexical features: Research Articles. *J. Am. Soc. Inf. Sci. Technol.* 58, 6 (Apr. 2007), 802-822.

Artstein, R. and M. Poesio. (2005) Kappa3 = Alpha (or Beta). Technical Report NLE Technote 2005-01, University of Essex, Essex, 2005.

Baca, M. (2003) *Practical Issues in Applying Metadata Schemas and Controlled Vocabularies to Cultural Heritage Information*. The Haworth Press, Inc.

Chen, H. (2001) An analysis of image queries in the field of art history. *Journal of the American Society for Information Science and Technology*, pages 260–273.

Hachey, B. and C. Grover. (2004) Sentence classification experiments for legal text summarisation. In *Proceedings of the 17th Annual Conference on Legal Knowledge and Information Systems (Jurix)*.

Forman, G. (2003) An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning. Research* 3 (Mar. 2003), 1289-1305.

Giral, A. and A. Taylor. (1993) 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.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise des Sciences Naturelles* 44:223-270.

Klavans, J.; Sidhu, T.; Sheffield, C.; Soergel, D.; Lin, J.; Abels, E.; Passonneau, R. (2008) Computational Linguistics for Metadata Building (CLiMB) Text Mining for the Automatic Extraction of Subject Terms for Image Metadata. *International Conference on Computer Vision Theory and Applications, Workshop 3: Metadata Mining for Image Understanding*, Funchal, Portugal.

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA.

Layne, S. S. (1994) Some issues in the indexing of images. *Journal of the American Society for Information Science*, pages 583–8.

Markey, K. (1984) 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.

Passonneau, R; Yano, T.; Lippincott, T.; Klavans, J. (2008). Functional semantic categories for art history text: human labeling and preliminary machine learning. *International Conference on Computer Vision Theory and Applications, Workshop 3: Metadata Mining for Image Understanding*, Funchal, Portugal.

Passonneau, R. (2006). 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)*. Portugal.

Passonneau, R.; Habash, N.; Rambow, O. (2006) Inter-annotatator agreement on a multilingual semantic annotation task. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC).* Portugal.

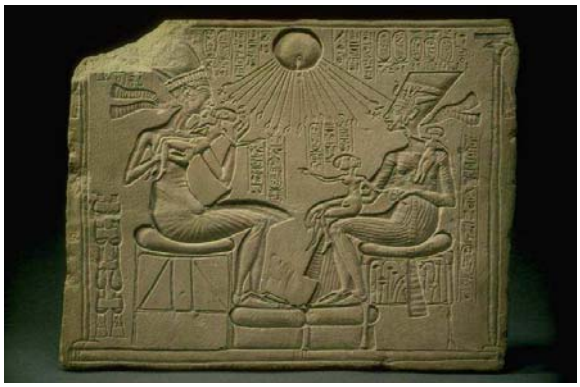
Riedsma, D. and Carletta, J. (Forthcoming) Reliability measurement: there's no safe limit To appear in *Computational Linguistics*.

Sebastiani, F. (2002) Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, March 2002.

Teufel, S. and M. Moens. (2002) Summarising scientific articles – experiments with relevance and rhetorical status. *Computational Linguistics*, pages 409–445.

Witten, I. H. and E. Frank (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition. Morgan Kaufmann, San Francisco, 2005.

Yang, Y. and J. O. Pedersen. (1997) A comparative study on feature selection in text categorization. In the *Proceedings of the 14th International Conference on Machine Learning (ICML)*, pp. 412-420.



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 a sunk relief portrait of Akhenaten and his family. The intimate domestic scene suggests that the relief was meant to serve a shrine in a private household.
Image Content	
Historical Context	

Enhanced XML representation:

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

Figure 1. Semantic classification of text extracts

Exp.	Dataset	#labels/text	#annotators	Alpha-MASI	Mean F
1	I: 13 images, 52 paragraphs	any	2	0.76	0.80
2	II: 9 images, 24 paragraphs	any	2	0.93	0.87
3	II: 9 images, 24 paragraphs	two	5	0.46	0.47
4a	III: 10 images, 24 paragraphs	one	7	0.24	0.41
4b	III: 10 images, 159 sentences	one	7	0.30	0.43

Table 1. Interannotator consistency under various conditions

Train100/Test30						10-Fold Crossval130					
who	bow	who	pos	who	both	who	bow	who	pos	who	both
D	0.976	D	0.889	D	0.976	A''	0.822	A'	0.709	A''	0.830
A''	0.781	A''	0.789	A''	0.781	A'	0.795	D	0.703	A'	0.802
B	0.744	A'	0.741	B	0.744	B	0.778	A''	0.696	C	0.779
A'	0.699	B	0.659	A'	0.699	D	0.771	E	0.694	D	0.777
C	0.663	A	0.625	C	0.663	C	0.765	B	0.684	B	0.775
A	0.625	E	0.563	A	0.625	A	0.647	C	0.668	A	0.755
E	0.472	C	0.505	E	0.479	E	0.680	A	0.654	E	0.660
Avg	*0.709*	Avg	*0.682*	Avg	*0.710*	Avg	*0.751*	Avg	*0.687*	Avg	*0.768*
sd	*0.154*	sd	*0.134*	sd	*0.153*	sd	*0.063*	sd	*0.020*	sd	*0.053*

Table 2. Historical Context Classifiers

Train100/Test30						10-Fold Crossval130					
who	bow	who	pos	who	both	who	bow	who	pos	who	both
D	0.625	E	1	D	0.625	B	0.710	B	0.73	B	0.712
B	0.571	B	0.757	B	0.571	D	0.662	D	0.727	D	0.657
A	0.563	D	0.734	A	0.563	A''	0.636	A	0.705	A''	0.643
C	0.563	A	0.711	C	0.563	A'	0.613	A''	0.687	A'	0.620
A''	0.559	A''	0.681	A''	0.559	C	0.604	C	0.668	C	0.596
A'	0.556	C	0.648	A'	0.556	A	0.600	E	0.661	A	0.594
E	0.543	A'	0.537	E	0.543	E	0.590	A'	0.633	E	0.585
Avg	*0.569*	Avg	*0.724*	Avg	*0.569*	Avg	*0.631*	Avg	*0.687*	Avg	*0.630*
sd	*0.026*	sd	*0.141*	sd	*0.026*	sd	*0.043*	sd	*0.036*	sd	*0.045*

Table 3. Image Content Classifiers

Train100/Test30						10-Fold Crossval130					
who	bow	who	pos	who	both	who	bow	who	pos	who	both
B	0.722	A'	0.674	B	0.722	E	0.643	A''	0.695	E	0.632
A'	0.700	D	0.613	A'	0.7	B	0.631	A	0.639	B	0.612
D	0.625	A''	0.603	D	0.625	C	0.616	D	0.631	C	0.603
E	0.625	B	0.444	E	0.625	D	0.598	C	0.559	A''	0.601
C	0.614	E	0.388	C	0.614	A''	0.596	E	0.531	A'	0.596
A''	0.516	C	0.205	A''	0.516	A'	0.589	B	0.527	D	0.593
A	0.326	A	0.130	A	0.326	A	0.546	A'	0.515	A	0.551
Avg	*0.590*	Avg	*0.437*	Avg	*0.590*	Avg	*0.603*	Avg	*0.585*	Avg	*0.598*
sd	*0.134*	sd	*0.210*	sd	*0.134*	sd	*0.032*	sd	*0.070*	sd	*0.025*

Table 4. Implementation Classifiers

Historical Context			Image Content			Implementation		
train100test30	bow	0.05	train100test30	bow	-0.25	train100test30	bow	-0.43
	pos	0.18		pos	-0.75		pos	-0.01
	both	0.59		both	0.42		both	-0.43
crossval130	bow	0.11	crossval130	bow	-0.06	crossval130	bow	-0.77
	pos	-0.87		pos	0.07		pos	0.46
	both	0.71		both	0.14		both	-0.87

Table 5. Correlations of roc scores on learning with average pairwise IA