Beyond Flickr: Not All Image Tagging Is Created Equal

Judith L. Klavans[1], Raul Guerra[1], Rebecca LaPlante[1], Robert Stein[2], and Edward Bachta[2]

[1]University of Maryland College Park, [2]Indianapolis Museum of Art
{jklavans, laplante}@umd.edu, raulguerrap@gmail.com, {rstein, ebachta}@imamuseum.org

Abstract

This paper reports on the linguistic analysis of a tag set of nearly 50,000 tags collected as part of the steve.museum project. The tags describe images of objects in museum collections. We present our results on morphological, part of speech and semantic analysis. We demonstrate that deeper tag processing provides valuable information for organizing and categorizing social tags. This promises to improve access to museum objects by leveraging the characteristics of tags and the relationships between them rather than treating them as individual items. The paper shows the value of using deep computational linguistic techniques in interdisciplinary projects on tagging over images of objects in museums and libraries. We compare our data and analysis to Flickr and other image tagging projects.

Challenges of Tags

Identifying linguistic traits of tags provides some unique challenges. Linguistic analysis of words or phrases within the context of full text draws upon context to provide clues to the characteristics of the individual components. With tags, especially those affiliated with images, the contextual environment is minimal or non-existent.

Image tagging has been studied from many perspectives from the use of tags for training algorithms to their use in constructing folksonomies for describing image sets and to aiding users in search and access. The focus of this paper is on the use of tagging in the museum context, a specific application of image-tagging geared towards museum visitors, both virtual and real.

We present results on the semantic and linguistic properties of the steve.museum tagset, demonstrating how this tagset differs from others. The fundamental research questions driving this research are:

(1) How can a set of social tags describing museum images be analyzed using (a) computational linguistic tools, such as morphological analyzers, part of speech taggers; (b) online lexical resources such as WordNet (Miller 1995) or the Art and Architecture Thesaurus (Getty 2010), and (c) clustering (Becker, Naaman, and Gravano 2010) to characterize an image?
(2) What are the optimal linguistic processes to normalize tags since these steps have impact on later processing?
(3) In what ways can social tags be associated with other information to improve users' access to museum objects?

Description of the Tag Data

The steve project (http://www.steve.museum) is a multi-institutional collaboration exploring applications of tagging in museums. The project seeks to improve collection documentation and public access to online object collections. Initial research showed that user tags in the steve.museum project enhance existing object documentation, providing information that is not currently recorded in the museums' formal documentation (Trant, Bearman, and Chun 2007; Trant 2009). The T3: Text, Tags, Trust project (http://t3.umiacs.umd.edu) builds on this research and is developing open source software that applies techniques from computational linguistics that enhance the usability of the collected set of social tags.

In this paper, we have used the steve.museum original dataset of nearly 50,000 tags applied to 1,785 works. Trant (2009) describes the analysis of the tags collected by token. However, applying computational linguistic processing to the data as in Klavans et al. (2011) reveals significantly different results.

Related Work

Social tags have been computationally analyzed from many perspectives. Unlike other image tagging projects, such as Flickr or Google Image Labeler or Tag Cow, the data in this project was collected within a highly controlled environment over a carefully selected set of images with participation from 18 museum partners interested in the use of social media for museums, a rapidly growing area of interest for the museum community (Vogel 2011).

Peekaboom (von Ahn, Liu, and Blum 2006) gathers user-generated input for locating objects in images to train computer vision algorithms. However, unlike our data, these images are harvested from Web pages and contain little associated metadata. The Visual Dictionary Project (VDP) (Torralba, Fergus, and Freeman 2008) has collected user input to create training data for vision recognition systems (http://groups.csail.mit.edu/vision/TinyImages/).

The combination of visual features and tags (Aurnhammer Hanappe, and Steels 2006) is related in that tags need to be analyzed in terms of their semantics. Begelman, Keller, and Smadja (2006) explore the use of clustering over tags for the same image to identify semantically related tags and thus help users in the tagging experience. This research is relevant to the tag collection process.

Few tag analysis projects have undertaken deep computational linguistic analysis of tags. For example, Lee and Schleyer (2010) use the basic Porter Stemmer for normalization (van Rijsbergen, Robertson, and Porter 1980) and Google tools for spell-checking and compound word separation. They examine mapping between tags and the controlled subject headings from MeSH terms, showing (as did Trant 2007) that there is little overlap. In our research, we are taking this analysis further to examine the differences between morphological analyzers, parts of speech asssignment, semantic categorization and disambiguation.

The categories of tags identified in this study were compared to information gathered through other studies on a variety of different user groups and image collections. Overall, there are few similarities found between the types of tags assigned to images of art objects and those assigned to other image collections, showing that tag assignment is domain-specific. It may also reflect Golder and Huberman's (2006) finding that a significant amount of tagging is done for personal use rather than public benefit, so the nature of the tagging task may impact tag type.

Computational Linguistic Analysis of Tags

Morphological Analysis

Klavans et al. (2011) explore various processes needed to normalize the steve.museum dataset in a pipeline architecture. These preprocessing techniques include handling the range of anomalous characters occurring in tags, among them white spaces, character returns, and punctuation. In addition, stop words were removed. The Morphy lemmatizer from Natural Language Toolkit (NLTK) (Bird, Klein, and Loper 2009), was used to conflate tags.

Different preprocessing and normalization methods yield different output. Simple preprocessing to conflate tags dramatically reduces the number of tags by type compared with the number of tags by token. The majority of tags (79% by token, 52% by type) consist of one word, followed by those consisting of two words (15% by token, 33% by type). Only a small percentage of tags (6% by token, 15% by type) are longer than 2 words. Since basic tag frequency is used by many subsequent analyses, the impact of conflation cannot be underestimated.

Part of Speech Analysis

One of the original contributions of this paper is to provide insight on the role of part of speech (POS) tagging in tag normalization and analysis. Operations like morphological analysis may in fact depend on POS tagging as discussed below. Similarly, domain-dependent factors may influence standard POS tagging approaches.

The purpose of undertaking a comparison of POS tagging is to determine which is the most accurate for this type of input data. POS tagging is an essential step in the process of disambiguation (Lin 1997) . For example, given a tag "wing", there is no way to tell out of context if this is a noun or verb. Furthermore, it is difficult to determine if this is "wing of a bird", "wing of a chair", "wing in a theater" or any of the other senses; there are 11 senses for "wing" in WordNet and 20 senses in the Art and Architecture Thesaurus (AAT). In the case of tagging, where most tags are one word, the tag cloud serves as the context for interpretation, rather than a full phrase.

For evaluation, we created a gold standard for Part-Of-Speech tagging (POS) consisting of 850 randomly chosen distinct tags: 497 one token tags, 201 two token tags, 102 three token tags, 50 four token tags. However, both the gold-standard and the entire dataset are similar in that they are both dominated by one word tags, then to a lesser degree by two word tags, and so on. Of these 850 tags, there are 78 (9%) that are foreign words, misspellings, or symbols and punctuation such as "??". These tags were removed manually.

When developing the gold standard, we found that a given tag or tag phrase may have more than one possible POS. We decided to keep track of this ambiguity in the gold standard. This has the effect of increasing the total number of tag / POS combinations possible. For example, if the one word tag "painting" can be considered to be both a singular noun (NN) and a gerund verb (VBG) it was reflected both in the number of NN POS tags and the number of VBG POS tags present in the collection.

Most POS taggers use lexical probability, trained on a particular corpus, along with contextual probability derived from lexical and syntactic context. We tested three systems from NLTK: a Maximum Entropy tagger an affix tagger, and an n-gram tagger. The Maximum Entropy tagger has two big advantages over the other taggers. First it is a probabilistic classifier so for a given token it can give a probability distribution over all the POS tags. This is useful

to maintain the ambiguity discussed before. The second advantage is that a Maximum Entropy model works by extracting features from the input and combining them linearly. Thus the classifier can be tuned to use whatever features we think useful and it is not as limited to token occurrence like the previous taggers. For the Stanford MaxEnt tagger we utilized the bidirectional-distsim-wsj-0-18.tagger model shipped with the Stanford POS tagger. According to the documentation this model was trained on WSJ sections 0-18 using a bidirectional architecture and including word shape and distributional similarity features.

The next tagger we used was the Affix tagger. We used the Affix Tagger that shipped with the NLTK. The Affix Tagger learns POS for prefixes and suffixes from the training data. Thus this tagger can learn from the very little context that tags have. We also decided to combine the Affix tagger with the sequential Bigram tagger by using the Affix tagger as the Unigram's back-off tagger. The Affix tagger then backs-off to the default tagger that tags everything with "NN". This gave better performance than the bigram or the affix taggers did individually.

The most straight forward tagger is the n-gram tagger. We used the n-gram tagger shipped with the NLTK. More specifically we used a bigram tagger that backs-off to a unigram tagger, which backs-off to a default tagger that tags everything as "NN". All these taggers were trained using the CONLL2000 training+ testing data. An n-gram tagger chooses a token's POS tag based on the given token and the preceding tokens. The bigram tagger we used and n-gram taggers in general perform well with n-grams seen before, but perform poorly on n-grams not seen before, in other words, those out of its vocabulary (OOV).

Number of words in tag	Stanford MaxEnt Tagger	Sequential Bigram Tagger	Affix Tagger	Bigram Tagger
1	79.86	71.53	71.06	62.73
2	77.60	53.13	45.83	50.00
3	63.27	46.94	40.82	41.84
4	68.00	52.00	40.00	48.00

Table 2. Results of comparing three tagging algorithms

The results of these different attempts have brought us closer to the answer of one of the fundamental research questions driving this project: to figure out how to best handle the normalization of tags since this could impact basic statistical issues, such as frequency values. Further down the analysis pipeline, processes such as clustering and similarity detection rely on frequency.

Note that one of the major challenges of POS tagging of the dataset is that most items are one word (e.g. "blue", "wind", "squares".) As a result, there is little information in a tag itself to help decipher the nature of the words within that short string. Other tags on the same object may provide some context. For example, "blue" in the context of "sad" or "lonely" indicate the meaning of "blue" as "saddened"; the example of part of speech for "wind" was given above.

However, since tags can reflect a wide variety of characteristics, such as subject matter (woman), biographical data (painted by Pablo Picasso), or opinion (scary), there may be a loose relationship between an individual tag and the set of tags on the same object. For example, "sad" and "lonely" might apply to one of Picasso's blue period paintings, which are predominantly blue in color. There is no unambiguous way of knowing which sense of "blue" is intended.

Once we determined that the MaxEnt tagger appeared to perform better than others, we ran the entire dataset through the tagger. These results are shown in Table 2:

Rank	POS Tag	Freq. by Token	POS Tag	Freq. by Type
1	NN	25205	NN	6706
2	JJ	6319	NN_NN	1713
3	NNS	4041	JJ_NN	1194
4	NN_NN	2257	JJ	921
5	JJ_NN	1792	NNS	757
6	VBG	1043	JJ_NNS	303

Table 2. The top 6 POS patterns ranked by frequency by token and frequency by type.

After all the tags have been assigned a POS, then an analysis of patterns can be performed. (n=6319) and the NNS, plural noun (n=4041). The next most frequent patterns are for two word phrases, NN-NN, noun-noun compound, and then JJ-NN, adjective-noun. Again, given the context of museum objects and images of these objects, this is to be expected. At the same time, a deeper analysis of results is needed to confirm that labeling is as expected, since typically noun compounds in English are ambiguous. The next category is VBG, which are gerunds such as "sitting" or "beating". Our initial examination of these VBG's shows that approximately 60% are used as nominals, but this is the focus of future research. Similarly, VBN's are usually used adjectivally, so that the nominal VBG's could be conflated with NN's and VBN's with JJ's. Proper nouns, ordinal numbers with nouns, and (unexpectedly) adjectives with plural nouns are the next three categories in frequency.

The graph shows that the frequency of the POS patterns for tags follows a power law (Zipf 1949); in other words, the frequencies of the POS patterns decrease exponentially

so that a POS pattern is inversely proportional to its position in the list.

Part of speech tagging is integral to most NLP pipelines, since this step is a precursor to parsing. However, for social tags, parsing is not a meaningful step. Therefore, by studying the POS properties of tagsets in and of themselves, there is an opportunity to understand the nature of this kind of descriptive tagging. Linking POS data with other lexical resource information, and with semantic information may contribute ultimately to a deeper understanding of the nature of social tagging as linguistic data, and to the utilization of these tags in the museum context. The leap between using tags for access and understanding tags as a set of linguistic entities is the purpose of this research, so we are addressing relevant parts of this question in this paper.

Theory-Driven Semantic Disambiguation by Domain

The second novel contribution of this paper is in the semantic disambiguation of tags by theory-driven distinctions. Identification of the subject matter expressed through social tags can provide an additional tool to understand, and thus control and manage, the noise created through the collection of this type of unstructured information. LaPlante, Klavans, and Golbeck (n.d.) are undertaking a study to examine a set of 100 images of two-dimensional, representational paintings with 2909 unique tags in this specific collection.

While there are many theoretical approaches to categorizing the way an image can be described, from identifying a broad range of attributes (Jörgensen 1998) to showing a hierarchical structure with levels of generality (Rorissa 2008), there is still no consensus on the best approach to use (Stewart 2010). To address this challenge, LaPlante, Klavans, and Golbeck (n.d.) are using a two-dimensional matrix based on the work of Shatford (1986) that reflects both the depth and breadth of information available about an image (Armitage and Enser 1997). One axis of the matrix describes specificity, or an individual's depth of knowledge about the content of an image. The elements of this axis are:

- Generic (G), or a very basic knowledge about an image,
- Specific (S), or a more detailed knowledge about an image, and
- Abstract (A), or a sophisticated understanding of an image.

The second facet describes the type of subject matter expressed, and includes:

- Who (1), or people or things,
- What (2), or events, actions, conditions, and emotions,
- Where (3), or locations, and
- When (4), or time periods.

This core matrix was modified to include a visual elements category (V) to capture information on shapes, colors, and forms, as well as an unknown category (U) to capture information not related to subject matter, such as the artist's name, the title of a piece, or an undecipherable tag.

Individuals from the museum community as well as project staff have categorized the tags assigned to these images using this two-dimensional matrix. Coders agreed on the categorization of 2284, or 79% of the tags. Of these 2284 tags, G1 (generic person or thing) is the most frequently assigned category at 48%, followed by A2 (emotion or abstraction) at 10%, and G2 (generic event, action, condition) at 10% (Table 2).

G1: Generic who	**G2: Generic what**	**G3: Generic where**	**G4: Generic when**
1095	227	161	32
48%	10%	7%	1%
S1: Specific who	**S2: Specific what**	**S3: Specific where**	**S4: Specific when**
33	5	37	62
1%	0.2%	2%	3%
A1: Abstract who	**A2: Abstract what**	**A3: Abstract where**	**A4: Abstract when**
27	236	3	2
1%	10%	0.1%	0.1%
V: Visual Elements	**U: Unknown**		
148	216		
6%	9%		

Table 2. Categorization of tags.

A comparison of our results with those on flickr indicate that the largest category is G1. However, Table 3 shows that the majority of tags (52% for our data and 54% for Flickr) are of different semantic types:

Art Objects		Flickr (Chung and Yoon 2009)	
G1	48%	G1	46%
A2	10%	S3	14%
G2	10%	G3	10%
U	9%	A2	7%
G3	7%	Flickr-specific	6%

Table 3. Top 5 Tags Assigned to Art Objects and Flickr Images

The importance of this analysis is that knowledge of this type of information can assist with managing the volume of unstructured tag information provided by users. It can help weigh the likelihood of different parts of speech in a tag set thus providing help in disambiguation. For example, the preponderance of tags expressing the who of an image would suggest that tags that are ambiguous such as gerunds are more likely to be nouns. It can also help visualize the type of information found in a tag set associated with art objects. For instance, this tag set can provide a substantial amount of generic information on things or events, but little valuable data on specific periods of time.

Original Contributions

Our overall research program addressed three questions, stated in Section 1. The novel contributions of this paper cross-cut these three questions. We have shown:

- Basic computational linguistic processing can impact tag analysis by token and type which will in turn affect down-stream tag analysis;
- Morphological and part of speech analysis impacts how tags are clustered and viewed;
- Computational linguistic tools can reduce some of the "noise" in tagsets;
- Theory-driven semantic analysis of tags reveals categories useful for disambiguation.

Future Work

Our future work addresses other aspects of the research questions set out in Section 1. As in Agichtein et al. (2008), we will be combining high quality content from museum sites with social tags. We will use the output of a toolkit to identify named entities and noun phrases in texts associated with these images, provided by museum partners. Mapping information from existing text resources along with social tags raises challenges in concept relationships, disambiguation and then in sifting and filtering to improve object access.

For disambiguation, we plan to analyze the temporal order of tagging based on a user session to see if any patterns arise when looking at an individual user or at an individual session. For example, if in a given tagging session, a user tags one image with the words "red", "purple", and "green", can we use that information to disambiguate a less clear tag such as "gold" which could refer to either a color or a metal? Similarly, if we know that users tend to tag with nouns first, can we use that information to disambiguate tags in other tagging sessions?

In addition to these more general questions, there are some domain-specific questions that would be valuable to examine to help cultural heritage organizations manage large collections of tags. For instance, are there distinctions between the linguistic characteristics of tags provided based on object type, such as paintings or photographs? Similarly, are there distinctions between two- and three-dimensional objects or abstract and representational works of art? Based on initial observations, it appears that there are many lexical properties of tags that could be inferred from using information about object type, but this hypothesis is yet to be confirmed.

Acknowledgments

We thank the members of the Museums Working Group of the T3: Text, Tags, and Trust project, Susan Chun, Independent Museum Consultant. This research is supported by the Institute of Museum and Library Services (IMLS).

References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. 2008. Finding High-Quality Content in Social Media. In *Proceedings of the International Conference on Web Search and Web Data Mining* (WSDM '08), 183-194. New York, NY:ACM.

Armitage, L.H. and Enser, P.G.B. 1997. Analysis of User Need in Image Archives. *Journal of Information Science* 23(4): 287-299.

Aurnhammer, M., Hanappe, P., and Steels, L. 2006. Integrating Collaborative Tagging and Emergent Semantics for Image Retrieval. In *WWW2006: Proceedings of the Collaborative Web Tagging Workshop*, Edinburgh, UK.

Becker, H., Naaman, M., and Gravano, L. 2010. Learning Similarity Metrics for Event Identification in Social Media. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (WSDM '10), 291-300. New York, NY:ACM.

Begelman, G., Keller, P., and Smadja, F. 2006. Automated Tag Clustering: Improving Search and Exploration in the Tag Space. In *WWW2006: Proceedings of the Collaborative Web Tagging Workshop*, Edinburgh. UK.

Bird, S., Klein, E., and Loper, E. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolset*. Sebastopol, CA: O'Reilly Media.

Corston-Oliver, S. and Gamon, M. 2004. Normalizing German and English Inflectional Morphology to Improve Statistical Word Alignment. In *Proceedings of the Conference of the Association for Machine Translation in the Americas*, 48-57. Springer-Verlag.

Golder, S.A. and Huberman, B.A. 2006 The Structure of Collaborative Tagging Systems. *Journal of Information Science* 32: 198-208.

Hsu, M. and Chen, H. 2008. Tag Normalization and Prediction for Effective Social Media Retrieval. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Volume 01 (WI-IAT '08)*, 770-774. Washington, DC: IEEE Computer Society.

J. Paul Getty Trust. 1988-2010. *Art & Architecture Thesaurus (AAT).* Retrieved from www.getty.edu/research/conducting_research/vocabularies/aat/.

Jörgensen, C. 1998. Attributes of Images in Describing Tasks. *Information Processing & Management* 34(2-3): 161-174.

Klavans, J., Stein, R., Chun, S., and Guerra, R. 2011. Computational Linguistics in Museums: Applications for Cultural Datasets. In *Museums and the Web 2011: Proceedings*. Philadelphia, PA: Museums and the Web.

LaPlante, R., Klavans, J., and Golbeck, J. n.d. *Subject Matter Categorization of Tags Applied to Images of Art Objects*. In progress.

Lee, D.H. and Schleyer, T. 2010. A Comparison of meSH Terms and CiteULike Social Tags as Metadata for the Same Items. In *Proceedings of the 1st ACM International Health Informatics Symposium (IHI '10)*, 445-448. New York, NY: ACM.

Lin, Dekang. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (EACL '97). Association for Computational Linguistics.

Miller, G.A. 1995. WordNet: a lexical database for English. *Communications of the ACM* 38(11): 39–41.

Rorissa, A. 2008. User-generated Descriptions of Individual Images Versus Labels of Groups of Images: A comparison Using Basic Level Theory. *Information Processing & Management* 44(5): 1741-1753.

Shatford, S. 1986. Analyzing the Subject of a Picture: A Theoretical Approach. *Cataloging & Classification Quarterly* 6(3): 39-62.

Stewart, B. 2010. Getting the Picture: An Exploratory Study of Current Indexing Practices in Providing Subject Access to Historic Photographs / Se Faire une Image: Une Exploration des Pratiques D'indexation Courantes Dans la Fourniture de L'accès Par Thème à des Photographies Historiques. *Canadian Journal of Information and Library Science* 34(3): 297-327.

Torralba, A., Fergus, R., and Freeman, W.T. 2008. 80 Million Tiny Images: A Large Dataset for Non-parametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(11): 1958-1970.

Trant, J. 2009. *Tagging, Folksonomy, and Art Museums: Results of steve.museum's research*. Available from http://conference.archimuse.com/blog/jtrant/stevemuseum_research_report_available_tagging_fo.

Trant, J., Bearman, D., and Chun, S. 2007. The Eye of the Beholder: steve.museum and Social Tagging of Museum Collections. In *Proceedings of the International Cultural Heritage Informatics Meeting - ICHIM07*, Toronto, Canada.

von Ahn, L., Liu, R., and Blum, M. 2006. Peekaboom: A Game for Locating Objects in Images. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (CHI '06), 55-64. New York, NY: ACM.

van Rijsbergen, C.J., Robertson, S.E., and Porter, M.F. 1980. *New Models in Probabilistic Information Retrieval*. London: British Library. (British Library Research and Development Report, no. 5587).

Vogel, C. 16 March, 2011. Four to Follow. *The New York Times*, F24. Retrieved on March 29, 2011 from http://www.nytimes.com/2011/03/17/arts/design/four-innovating-for-museums-online.html.

Zipf, G.K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, Mass.: Addison-Wesley Press.