Apples to Oranges: Evaluating Image Annotations from Natural Language Processing Systems

Rebecca Mason and **Eugene Charniak**
Brown Laboratory for Linguistic Information Processing (BLLIP)
Brown University, Providence, RI 02912
{rebecca,ec}@cs.brown.edu

Abstract

We examine evaluation methods for systems that automatically annotate images using co-occurring text. We compare previous datasets for this task using a series of baseline measures inspired by those used in information retrieval, computer vision, and extractive summarization. Some of our baselines match or exceed the best published scores for those datasets. These results illuminate incorrect assumptions and improper practices regarding preprocessing, evaluation metrics, and the collection of gold image annotations. We conclude with a list of recommended practices for future research combining language and vision processing techniques.

1 Introduction

Automatic image annotation is an important area with many applications such as tagging, generating captions, and indexing and retrieval on the web. Given an input image, the goal is to generate relevant descriptive keywords that describe the visual content of the image. The Computer Vision (CV) literature contains countless approaches to this task, using a wide range of learning techniques and visual features to identify aspects such as objects, people, scenes, and events.

Text processing is computationally less expensive than image processing and easily provides information that is difficult to learn visually. For this reason, most commerical image search websites identify the semantic content of images using co-occurring text exclusively. But co-occurring text is also a noisy source for candidate annotations, since not all of the text is visually relevant. Techniques from Natural Language Processing help align descriptive words and images. Some examples of previous research use named-entity recognition to identify people in images (Deschacht and Moens, 2007); term association to estimate the "visualness" of candidate annotations (Boiy et al., 2008; Leong et al., 2010); and topic models to annotate images given both visual and textual features (Feng and Lapata, 2010b).

Image annotation using NLP is still an emerging area with many different tasks, datasets, and evaluation methods, making it impossible to compare many recent systems to each other. Although there is some effort being made towards establishing shared tasks[1], it is not yet clear which kinds of tasks and datasets will provide interesting research questions and practical applications in the long term. Until then, establishing general "best practices" for NLP image annotation will help advance and legitimitize this work. In this paper, we propose some good practices and demonstrate why they are important.

2 Image Annotation Evaluation in CV and NLP

In this section, we first review related work in image annotation evaluation in computer vision, specific challenges, and proposed solutions. We then relate these challenges to the NLP image annotation task and some of the specific problems we propose to address.

[1] http://imageclef.org/

2.1 Related Work in Computer Vision

The work of Müller et al. (2002) is one of the first to address the issue of evaluation for image annotation systems. While using the exact same annotation system, dataset, and evaluation metric, they dramatically improve the apparent performance of their system by using dataset pruning heuristics.

Others have criticized commonly-used CV datasets for being too "easy" — images with the same keywords are extremely similar in low-level features such as orientation, lighting, and color; while differences between images with different keywords are very clear (Westerveld and de Vries, 2003; Ponce et al., 2006; Hervé and Boujemaa, 2007; Tang and Lewis, 2007). These features are unwittingly exploited by certain algorithms and obscure the benefits of using more complex techniques (Ponce et al., 2006). The problem is further exacerbated by evaluation metrics which essentially prefer precision over recall and are biased towards certain keywords. Annotations in test data might not include all of the "correct" keywords, and evaluation metrics need to account for the fact that frequent keywords in the corpus are safer guesses than keywords that appear less frequently (Monay and Gatica-Perez, 2003).

New baseline techniques, evaluation metrics, and datasets for image annotation have been developed in response to these problems. Makadia et al. (2008; 2010) define a basic set of low-level features, and propose new baselines for more complex systems to evaluate against. Barnard et al. (2003) present a normalized loss function to address the preference towards precision in evaluation metrics. New datasets are larger and provide more diverse images, and it is now easy to obtain multiple human-annotations per image thanks to distributed services such as Amazon's Mechanical Turk, and the ESP game (von Ahn and Dabbish, 2004). Hanbury (2008) provides an overview of popular CV annotation datasets and methods used for building them.

2.2 Image Annotation using Natural Language Processing

Many of the problems from CV image annotation are also applicable to NLP image annotation, and bringing NLP to the task brings new challenges as well. One of these challenges is whether to allow infrequent words to be pruned. In CV annotation it is typical to remove infrequent terms from both the keyword vocabulary and the evaluation data because CV algorithms typically need a large number of examples to train on. However, using NLP systems and baselines one can correctly annotate using keywords that did not appear in the training set. Removing "unlearnable" keywords from evaluation data, as done in (Boiy et al., 2008; Feng and Lapata, 2010b), artificially inflates performance against simple baselines such as term frequency.

Nearly all NLP annotation datasets use naturally-occurring sources of images and text. A particularly popular source is news images alongside captions or articles, which are collected online from sources such as Yahoo! News (Berg et al., 2004; Deschacht and Moens, 2007). There are also domain-specific databases with images and descriptions such as the art, antiques, and flowers corpora used in Boiy et al. (2008). Wikipedia has also been used as a source of images and associated text (Tsikrika et al., 2011). These sources typically offer well-written and cleanly formatted text but introduce the problem of converting text into annotations, and the annotations may not meet the requirements of the new task (as shown in Section 3.1). Obtaining data via image search engines is a common practice in CV (Fei-Fei et al., 2004; Berg and Forsyth, 2006) and can also be used to provide more challenging and diverse instances of images and co-occurring text. The additional challenge for NLP is that text content on many websites is written to improve their rank in search engines, using techniques such as listing dozens of popular keywords. Co-occurring text for retrieved images on popular queries may not be representative of the task to be performed.

3 Datasets

In this paper, we examine two established image annotation datasets: the BBC News Dataset of Feng and Lapata (2008) (henceforth referred to as *BBC*), and the general web dataset of Leong et al. (2010) (henceforth referred to as *UNT*). These datasets were both built to evaluate image annotation systems that use longer co-occurring text such as a news article or a webpage, but they use data from differ-

Dataset	**BBC**	**UNT**
data instances	article, image, and caption from a news story	image and text from a webpage
source of data	scraped from BBC News website	Google Image Search results
candidate keywords or collocations for annotation	descriptive unigram words from training data	$n \leq$ 7-grams extracted from co-occurring text; collocations must appear as article names on Wikipedia
gold annotations	descriptive words from held-out image captions	multiple human-authored annotations for each image
evaluation metric	precision and recall against gold annotations	metrics adapted from evaluation of lexical substitutions (SemEval)
number of train instances	3121 instances of related news article, image, and caption	none (train using cross-validation)
number of test instances	240 instances of news article and related image	300 instances of webpage with text and image
preprocessing procedure	lemmatize tokens, remove from dataset all words that are not descriptive or that appear fewer than five times in training articles	stem all tokens
average number of text tokens after preprocessing	169 word tokens per article, 4.5 per caption	278 word tokens per webpage
average document title length	4 word tokens	6 word tokens
total vocabulary after preprocessing	10479 word types	8409 word types

Table 1: Comparison of the BBC and UNT image annotation datasets.

ent domains, different sources of gold image annotations, different preprocessing procedures, and different evaluation measures.

Table 1 provides an overview of the datasets; while this section covers the source of the datsets and their gold annotations in more detail.

3.1 BBC

The BBC Dataset (Feng and Lapata, 2008)[2] contains news articles, image captions, and images taken from the BBC News website. Training instances consist of a news article, image, and image caption from the same news story. Test instances are just the image and the article, and hold-out the caption as a source of gold image annotations.

Using news image captions as annotations has the disadvantage that captions often describe background information or relate the photo to the story, rather than listing important entities in the image. It also fails to capture variation in how humans describe images, since it is limited to one caption per image.[3] However, captions are a cheap source of data; BBC has ten times as many images as UNT.

To address the problem of converting natural language into annotations, a large amount of preprocessing is performed. The established preprocessing procedure for this dataset is to lemmatize and POS-tag using TreeTagger (Schmid, 1994) then remove all but the "descriptive" words (defined as nouns, adjectives, and certain classes of verbs). This leaves a total text vocabulary of about 32K words, which

[2]Downloaded from `http://homepages.inuf.ed.ac.uk/s0677528/data.html`

[3]The Pascal Sentences dataset (`vision.cs.uiuc.edu/pascal-sentences`) provides multiple captions per image, but they are not naturally-occurring.

is further reduced by removing words that appear fewer than five times in the training set articles. Table 1 shows the number of word tokens and types after performing these steps.[4]

3.2 UNT

The UNT Dataset (Leong et al., 2010)[5] consists of images and co-occurring text from webpages. The webpages are found by querying Google Image Search with frequent English words, and randomly selecting from the results.

Each image in UNT is annotated by five people via Mechanical Turk. In order to make human and system results comparable, human annotators are required to only select words and collocations that are directly extracted from the text, and the gold annotations are the count of how many times each keyword or collocation is selected. The human annotators write keywords into a text box; while the collocations are presented as a list of candidates and annotators mark which ones are relevant. Human annotators tend to select subsets of collocations in addition to the entire collocation. For example, the gold annotation for one image has "university of texas", "university of texas at dallas", "the university of texas", and "the university of texas at dallas", each selected by at least four of the five annotators. Additionally, annotators can select multiple forms of the same word (such as "tank" and "tanks"). Gold annotations are stemmed after they are collected, and keywords with the same stem have their counts merged. For this reason, many keywords have a higher count than the number of annotators.

[4] We are unable to reproduce work from Feng & Lapata (2008; 2010a; 2010b) and Feng (2011). Specifically, our vocabulary counts after preprocessing (as in Table 1) are much higher than reported counts, although the number of tokens per article/caption they report is higher than ours. We have contacted the authors, who confirmed that they took additional steps to reduce the size of the vocabulary, but were unable to tell us exactly what those steps are. Therefore, all system and baseline scores presented on their dataset are of our own implementation, and do not match those reported in previous publications.

[5] Downloaded from `http://lit.csci.unt.edu/index.php?P=research/downloads`

4 Baselines

We run several baselines on the datasets. Term frequency, tf*idf, and corpus frequency are features that are often used in annotation systems, so it is important to test them on their own. Document Title and tf*idf are both baselines that were used in the original papers where these datasets came from.

Sentence extraction is a new baseline that we propose specifically for the BBC dataset, in order see if we can exploit certain properties of the gold annotations, which are also derived from sentences.

4.1 Term Frequency

Term frequency has been shown to be a powerful feature in summarization (Nenkova and Vanderwende, 2005). Words that appear frequently are considered more meaningful than infrequent words. Term frequency is the number of times a term (excluding function words) appears in a document, divided by the total number of terms in that document. On the UNT dataset we use the stopword list included with the MALLET[6] toolkit, while the BBC dataset doesn't matter because the function words have already been removed.

4.2 tf*idf

While term frequency baseline requires the use of an *ad hoc* function word list, tf*idf adjusts the weights of different words depending on how important they are in the corpus. It is a standard baseline used for information retrieval tasks, based on the intuition that a word that appears in a smaller number of documents is more likely to be meaningful than a word that appears in many documents.

tf*idf is the product of term frequency and inverse document frequency – $idf(t_i) = \log \frac{N}{n_i}$ where N is the number of documents in the corpus, and n_i is the number of documents that contain the term t_i. For the BBC Dataset, we base the idf weights on the document frequency of the training articles. For UNT, we use the reported tf*idf score which uses the British National Corpus to calculate the idf scores.[7]

[6] `mallet.cs.umass.edu`

[7] We also ran tf*idf where for each document we recalculate idf using the other 299, but it didn't make any meaningful difference.

4.3 Corpus Frequency

Image annotations in both NLP and CV tend to be distributed with a relatively small number of frequently occuring keywords, and a long tail of keywords that only appear a few times. For UNT, we use the total keyword frequency of all the gold annotations, except for the one document that we are currently scoring. For BBC, we only measure the frequency of keywords in the training set captions, since we are specifically interested in the frequency of terms in captions.

4.4 Document Title

For BBC, the news article headline, and for UNT, the title of the webpage.

4.5 Sentence Extraction

Our baseline extracts the most central sentence from the co-occurring text and uses descriptive words from that sentence as the image annotation. Unlike sentence extraction techniques from Feng and Lapata (2010a), we determine which sentence to extract using the term frequency distribution directly. We extract the sentence with the minimum KL-divergence to the entire document.[8]

5 BBC Dataset Experiments

5.1 System Comparison

In addition to the baselines, we compare against the Mix LDA system from Feng and Lapata (2010b). In Mix LDA, each instance is represented as a bag of textual features (unigrams) and visual features (SIFT features quantized to discrete "image words" using k-means). A Latent Dirichlet Allocation topic model is trained on articles, images, and captions from the training set. Keywords are generated for an unseen image and article pair by estimating the distribution of topics that generates the test instance, then multiplying them with the word distributions in each topic to find the probability of textual keywords for the image. Text LDA is is the same model but only using words and not image features.

[8]One could also think of this as a version of the KLSum summarization system (Haghighi and Vanderwende, 2009) that stops after one sentence.

5.2 Evaluation

The evaluation metric and the source of gold annotations is described in Table 1. For the baselines 4.1, 4.2, 4.3 and the Mix LDA system, the generated annotation for each test image is its ten most likely keywords. We also run all baselines and the Mix LDA system on an unpruned version of the dataset, where infrequent terms are not removed from training data, test data, or the gold annotations. The purpose of this evaluation is to see if candidate keywords deemd "unlearnable" by the Mix LDA system can be learned by the baselines.

5.3 Results

The evaluation results for the BBC Dataset are shown in Table 2. Clearly, term frequency is a stronger baseline than tf*idf by a large margin. The reason for this is simple: since nearly all of BBC's function words are removed during preprocessing, the only words downweighted by the idf score are common – but meaningful – words such as *police* or *government*. This is worth pointing out because, in many cases, the choice of using a term frequency or tf*idf baseline is made based on what was used in previous work. As we show here and in Section 6.3, the choice of frequency baseline should be based on the data and processing techniques being used.

We use the corpus frequency baseline to illustrate the difference between *standard* and *include-infrequent* evaluations. Since including infrequent words doesn't change which are most frequent in the dataset, precision for corpus frequency doesn't change. But since infrequent words are now included in the evaluation data, we see a 0.5% drop in recall (since corpus frequency won't capture infrequent words). Compared to the other baselines, this is not a large difference. Other baselines see a larger drop in recall because they have both more gold keywords to estimate and more candidate keywords to consider. tf*idf is the most affected by this, because idf overly favors very infrequent keywords, despite their low term frequency. In comparison, the term frequency baseline is not as negatively affected and even improves in precision because there are some cases where a word is very important to an article in the test set but just didn't appear very often in the training set (see Table 3 for examples). But the base-

	Standard			Include-infrequent		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Term Frequency	13.13	27.84	17.84	13.62	25.71	17.81
tf * idf	9.21	19.97	12.61	7.25	13.52	9.44
Doc Title	17.23	13.70	15.26	15.91	11.86	13.59
Corpus Frequency	3.17	6.52	4.26	3.17	6.02	4.15
Sentence Extraction	16.67	15.61	16.13	18.62	16.83	17.68
Mix LDA	7.30	16.16	10.06	7.50	13.98	9.76
Text LDA	8.38	17.46	11.32	7.79	14.52	10.14

Table 2: Image annotation results for previous systems and our proposed baselines on the BBC Dataset.

Cadbury increase **contamination testing level**

malaria parasite **spread** mosquito

Table 3: Examples of gold annotations from the test section of the BBC Dataset. The bolded words are the ones that appear five or more times in the training set; the unbolded words appear fewer than five times and would be removed from both the candidate and gold keywords in the standard BBC evaluation.

lines with the best precision are the Doc Title and Sentence Extraction baselines, which do not need to generate ten keywords for every image.

While sentence extraction has a lower recall than term frequency, it is the only baseline or system that has improved recall when including infrequent words. This is unexpected because our baseline selects a sentence based on the term frequency of the document, and the recall for term frequency fell. One possible explanation is that extraction implicitly uses correlations between keywords. Probabilities of objects appearing together in an image are not independent; and the accuracy of annotations can be improved by generating annotation keywords as a set (Moran and Lavrenko, 2011). Recent works in image captioning also use these correlations: explicitly, using graphical models (Kulkarni et al., 2011; Yang et al., 2011); and implicitly, using language models (Feng and Lapata, 2010a). In comparison, sentence extraction is very implicit.

Unsurprisingly, the Text LDA and Mix LDA systems do worse on the include-infrequent evaluation than they do on the standard, because words that do not appear in the training set will not have high probability in the trained topic models. We were unable to reproduce the reported scores for Mix LDA from Feng and Lapata (2010b) where Mix LDA's scores were double the scores of Text LDA (see Footnote 4). We were also unable to reproduce reported scores for tf*idf and Doc Title (Feng and Lapata, 2008). However, we have three reasons why we believe our results are correct. First, BBC has more keywords, and fewer images, than typically seen in CV datasets. The BBC dataset is simply not suited for learning from visual data. Second, a single SIFT descriptor describes which way edges are oriented at a certain point in an image (Lowe, 1999). While certain types of edges may correlate to visual objects also described in the text, we do not expect SIFT features to be as informative as textual features for this task. Third, we refer to the best system scores reported by Leong et al. (2010), who evaluate their text mining system (see section 6.1) on the standard BBC dataset.[9] While their f1 score is slightly worse than our term frequency baseline, they do 4.86% better than tf*idf. But, using the baselines reported in Feng and Lapata (2008), their improvement over tf*idf is 12.06%. Next, we compare their system against frequency baselines using the 10 keyword generation task on the UNT dataset (the *oot normal* scores in table 5). Their best system performs 4.45% better

[9]Combined model; precision: 13.38, recall: 25.17, f1: 17.47. Crucially, they do not reimplement previous systems or baselines, but use scores reported from Feng and Lapata (2008).

than term frequency, and 0.55% worse than tf*idf.[10] Although it is difficult to compare different datasets and evaluation metrics, our baselines for BBC seem more reasonable than the reported baselines, given their relative performance to Leong et al's system.

6 UNT Dataset Experiments

6.1 System Comparison

We evaluate against the text mining system from (Leong et al., 2010). Their system generates image keywords by extracting text from the co-occurring text of an image. It uses three features for selecting keywords. *Flickr Picturability* queries the Flickr API with words from the text in order to find related image tags. Retrieved tags that appear as surface forms in the text are rewarded proportional to their frequency in the text. *Wikipedia Salience* assigns scores to words based on a graph-based measure of importance that considers each term's document frequency in Wikipedia. *Pachinko Allocation Model* is a topic model that captures correlations between topics (Li and McCallum, 2006). PAM infers subtopics and supertopics for the text, then retrieves top words from the top topics as annotations. There is also a combined model of these features using an SVM with 10-fold cross-validation.

6.2 Evaluation

Evaluation on UNT uses a framework originally developed for the SemEval lexical substitution task (McCarthy and Navigli, 2007). This framework accounts for disagreement between annotators by weighting each generated keyword by the number of human annotators who also selected that keyword. The scoring framework consists of four evaluation measures: *best normal*, *best mode*, *oot* (out-of-ten) *normal*, and *oot mode*.[11]

The two *best* evaluations find the accuracy of a single "best" keyword generated by the system[12]. *Best normal* measures the accuracy for each system annotation a_j as the number of times a_j appears in the R_j, the multi-set union of human tags, and averages over all the test images.

$$Bestnormal = \frac{\sum_{i_j \in I} \frac{|a_j \in R_j|}{|R_j|}}{|I|}$$

In *oot normal*, up to ten unordered guesses can be made without penalty.

$$ootnormal = \frac{\sum_{i_j \in I} \frac{\sum_{a_j \in A_j} |a_j \in R_j|}{|R_j|}}{|I|}$$

where A_j is the set of ten system annotations for image i_j.

The *best mode* and *oot mode* metrics are the same as the *normal* metrics except they only evaluate system annotations for images where R_j contains a single most frequent tag. We use the scoring software provided by SemEval[13] with the gold annotation file provided in the UNT Dataset.

6.3 Results

The results of the lexical substitution evaluation on the UNT Dataset are shown in Table 5. The results from the *normal* show support for our earlier idea that the relative performance of term frequency vs tf*idf depends on the dataset. Although the term frequency baseline uses a stopword list, there are other words that appear frequently enough to suggest they are not meaningful to the document – such as copyright disclaimers.

Recall that the *mode* evaluation is only measured on data instances where the gold annotations have a single most frequent keyword. While running the evaluation script on the gold annotation file that came with the UNT dataset, we discover that SemEval only identifies 28 of the 300 instances as having a single mode annotation, and that for 21 of those 28 instances, the mode keyword is "cartoon". Those 21/28 images correspond to the 75% *best mode* score obtained by Corpus Frequency baseline. Given the small number of instances that actually

[10] And as we stated earlier, the relative performance of term frequency vs tf*idf is different from dataset to dataset.

[11] Both the original framework and its adaptation by Leong et al. (2010) give precision and recall for each of the evaluation measures. However, precision and recall are identical for all baselines and systems, and only slightly different on the upper bound (human) scores. To preserve space, we only present the metric and scores for precision.

[12] In contrast to the original SemEval task, where systems can make from zero to many "best" guesses, penalized by the total number of guesses made.

[13] http://nlp.cs.swarthmore.edu/semeval/tasks/task10/data.shtml

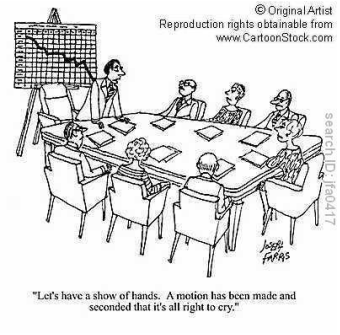		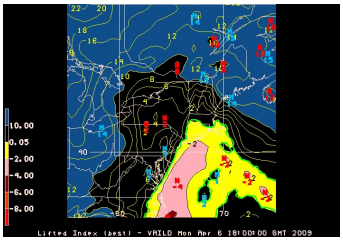
cartoon(6), market(5), market share(5), declin(3), imag(3), share(3), pictur(1), illustr(1), cartoonstock(1), origin(1), artist(1), meet(1), jfa0417(1), meeting-copyright(1)	cartoon(6), bill gate(5), gate(4), monopoli(4), pearli gate(4), bill(3), imag(3), caricatur(2), pictur(2), illustr(1), copyright(1), artist(1), own(1), pearli(1)	lift index(5), gener(3), index(3), condit(2), comput(2), comput gener(2), unstabl(2), zone(2), area(1), field(1), between(1), stabl(1), encyclopedia(1), thunderstorm(1), lift(1), free encyclopedia(1), wikipedia(1)

Table 4: Examples of gold annotations from the UNT Dataset.

	Best		Out-of-ten (oot)	
	Normal	**Mode**	**Normal**	**Mode**
Term Frequency	5.67	14.29	33.40	89.29
tf * idf	5.94	14.29	38.40	78.57
Doc Title	6.40	7.14	35.19	92.86
Corpus Frequency	2.54	75.00	8.22	82.14
Flickr Picturability	6.32	78.57	35.61	92.86
Wikipedia Salience	6.40	7.14	35.19	92.86
Topic Model (PAM)	5.99	42.86	37.13	85.71
Combined (SVM)	6.87	67.49	37.85	100.00

Table 5: Image annotation results for our proposed baselines, the text mining systems from (Leong et al., 2010)

count towards these metrics, we conclude that *mode* evaluation is not a meaningful way to compare image annotation systems on the UNT dataset.

That said, the number of cartoons in the dataset does seem to be strikingly high. Looking at the source of the images, we find that 45 of the 300 images were collected from a single online cartoon library. Predictably, we find that the co-occurring text to these images contains a long list of keywords, and little other text that is relevant to the image. We looked at a small sample of the rest of the dataset and found that many of the other text documents in UNT also have keyword lists.

Including this types of text in a general web corpus is not necessarily a problem, but it's difficult to measure the benefits of using complex techniques like topic modeling and graph similarily to find and extract annotations when in so many cases the annotations have already been found and extracted. This is shown in the *normal* evaluation results, where the combined system is only slightly better at selecting the single best keyword, and no better than tf*idf for the *out-of-ten* measure.

7 Conclusion

The intent of this paper is not to teach researchers how to inflate their own results, but to encourage better practices. With that purpose in mind, we make the following suggestions regarding future work in this area:

Get to know your data. The ability to quickly and cheaply collect very large – but very noisy – collections of data from the internet is a great advance for both NLP and CV research. However, there still needs to be an appopriate match betwen the task being performed, the system being proposed, and the dataset being used; and large noisy datasets can hide unintended features or incorrect assumptions about the data.

Use relevant gold annotations. Do not convert other sources of data into annotations. When collecting human annotations, avoid postprocessing steps such as merging or deleting keywords that change the annotators' original intent. Keep an open dialogue with annotators about issues that they find confusing, since that is a sign of an ill-formed task.

Preprocessing should be simple and reproducable. The use of different preprocessing procedures affects the apparent performance of systems and sometimes has unintended consequences.

Use strong baselines and compare to other work only when appropriate. Systems developed for different tasks or datasets can make for misleading comparisons if they don't use all features available. Strong baselines explicitly exploit low-level features that are implicitly exploited by proposed systems, as well as low-level features of the dataset.

Don't remove keywords from gold annotations. Just because keywords are impossible for one system to learn, does not mean they are impossible for all systems to learn. Removing evaluation data artificially inflates system scores and limits comparison to related work.

If a proposed system is to learn associations between visual and textual features, then it is necessary to **use larger datasets**. In general, global annotations, such as scenes, is easiest; identifying specific objects is more difficult; and identification of events, activities, and other abstract qualities has a very low success rate (Fluhr et al., 2006). Alternately, **use simpler image features** that are known to have a high sucess rate. For example, Deschacht and Moens (2007) used a face detector to determine the number of faces in an image, and then used NLP to determine the names of those people from associated text.

References

K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D.M. Blei, and M.I. Jordan. 2003. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135.

Tamara L. Berg and David A. Forsyth. 2006. Animals on the web. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 1463–1470, Washington, DC, USA. IEEE Computer Society.

T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Yee-Whye Teh, E. Learned-Miller, and D.A. Forsyth. 2004. Names and faces in the news. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–848 – II–854 Vol.2, june-2 july.

E. Boiy, K. Deschacht, and M.-F. Moens. 2008. Learning visual entities and their visual attributes from text corpora. In *Database and Expert Systems Application, 2008. DEXA '08. 19th International Workshop on*, pages 48 –53, sept.

Koen Deschacht and Marie-Francine Moens. 2007. Text analysis for automatic image annotation. In *ACL*, volume 45, page 1000.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 12 - Volume 12*, CVPRW '04, pages 178–, Washington, DC, USA. IEEE Computer Society.

Yansong Feng and Mirella Lapata. 2008. Automatic image annotation using auxiliary text information. *Proceedings of ACL-08: HLT*, pages 272–280.

Yansong Feng and Mirella Lapata. 2010a. How many words is a picture worth? automatic caption generation for news images. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1239–1249, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yansong Feng and Mirella Lapata. 2010b. Topic models for image annotation and text illustration. In *HLT-NAACL*, pages 831–839.

Yansong Feng. 2011. *Automatic caption generation for news images*. Ph.D. thesis, University of Edinburgh.

Christian Fluhr, Pierre-Alain Mollic, and Patrick Hde. 2006. Usage-oriented multimedia information retrieval technological evaluation. In James Ze Wang, Nozha Boujemaa, and Yixin Chen, editors, *Multimedia Information Retrieval*, pages 301–306. ACM.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, Boulder, Colorado, June. Association for Computational Linguistics.

Allan Hanbury. 2008. A survey of methods for image annotation. *J. Vis. Lang. Comput.*, 19:617–627, October.

Nicolas Hervé and Nozha Boujemaa. 2007. Image annotation: which approach for realistic databases? In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 170–177, New York, NY, USA. ACM.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, pages 1601–1608.

Chee Wee Leong, Rada Mihalcea, and Samer Hassan. 2010. Text mining for automatic image tagging. In *COLING*, pages 647–655.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 577–584, New York, NY, USA. ACM.

D.G. Lowe. 1999. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150 –1157 vol.2.

Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. 2008. A new baseline for image annotation. In *Proceedings of the 10th European Conference on Computer Vision: Part III*, ECCV '08, pages 316–329, Berlin, Heidelberg. Springer-Verlag.

A. Makadia, V. Pavlovic, and S. Kumar. 2010. Baselines for image annotation. *International Journal of Computer Vision*, 90(1):88–105.

D. McCarthy and R. Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53.

Florent Monay and Daniel Gatica-Perez. 2003. On image auto-annotation with latent space models. In *Proceedings of the eleventh ACM international conference on Multimedia*, Multimedia '03, pages 275–278, New York, NY, USA. ACM.

S. Moran and V. Lavrenko. 2011. Optimal tag sets for automatic image.

Henning Müller, Stéphane Marchand-Maillet, and Thierry Pun. 2002. The truth about corel - evaluation in image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, CIVR '02, pages 38–49, London, UK, UK. Springer-Verlag.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. Technical report, Microsoft Research.

J. Ponce, T. Berg, M. Everingham, D. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. Russell, A. Torralba, C. Williams, J. Zhang, and A. Zisserman. 2006. Dataset issues in object recognition. In Jean Ponce, Martial Hebert, Cordelia Schmid, and Andrew Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *Lecture Notes in Computer Science*, pages 29–48. Springer Berlin / Heidelberg. 10.1007/11957959_2.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.

Jiayu Tang and Paul Lewis. 2007. A study of quality issues for image auto-annotation with the corel dataset. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 1(NO. 3):384–389, March.

T. Tsikrika, A. Popescu, and J. Kludas. 2011. Overview of the wikipedia image retrieval task at imageclef 2011. In *CLEF (Notebook Papers/LABs/Workshops): CLEF*.

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, pages 319–326, New York, NY, USA. ACM.

Thijs Westerveld and Arjen P. de Vries. 2003. Experimental evaluation of a generative probabilistic image retrieval model on 'easy' data. In *In Proceedings of the SIGIR Multimedia Information Retrieval Workshop 2003, Aug*.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Empirical Methods in Natural Language Processing (EMNLP)*, Edinburgh, Scotland.