Text Mining for Automatic Image Tagging

Chee Wee Leong and Rada Mihalcea and Samer Hassan
Department of Computer Science and Engineering
University of North Texas
cheeweeleong@my.unt.edu, rada@cs.unt.edu, samer@unt.edu

Abstract

This paper introduces several extractive approaches for automatic image tagging, relying exclusively on information mined from texts. Through evaluations on two datasets, we show that our methods exceed competitive baselines by a large margin, and compare favorably with the state-of-the-art that uses both textual and image features.

1 Introduction

With continuously increasing amounts of images available on the Web and elsewhere, it is important to find methods to annotate and organize image databases in meaningful ways. Tagging images with words describing their content can contribute to faster and more effective image search and classification. In fact, a large number of applications, including the image search feature of current search engines (e.g., Yahoo!, Google) or the various sites providing picture storage services (e.g., Flickr, Picasa) rely exclusively on the tags associated with an image in order to search for relevant images for a given query.

However, the task of developing accurate and robust automatic image annotation models entails daunting challenges. First, the availability of large and correctly annotated image databases is crucial for the training and testing of new annotation models. Although a number of image databases have emerged to serve as evaluation benchmarks for different applications, including image annotation (Duygulu et al., 2002), content-based image retrieval (Li and Wang, 2008) and cross language information retrieval (Grubinger et al., 2006), such databases are almost exclusively created by manual labeling of keywords, requiring significant human effort and time. The content of these image databases is often restricted only to a few domains, such as medical and natural photo scenes (Grubinger et al., 2006), and specific objects like cars, airplanes, or buildings (Fergus et al., 2003). For obvious practical reasons, it is important to develop models trained and evaluated on more realistic and diverse image collections.

The second challenge concerns the extraction of useful image and text features for the construction of reliable annotation models. Most traditional approaches relied on the extraction of image colors and textures (Li and Wang, 2008), or the identification of similar image regions clustered as blobs (Duygulu et al., 2002) to derive correlations between image features and annotation keywords. In comparison, there are only a few efforts that leverage on the multitude of resources available for natural language processing to derive robust linguistic-based image annotation models. One of the earliest efforts involved the use of captions for face recognition in photographs through the construction of a specific lexicon that integrates linguistic and photographic information (Srihari and Burhans, 1994). More recently, several approaches have proposed the use of WordNet as a knowledge-base to improve content-based image annotation models, either by removing noisy keywords through semantic clustering (Jin et al., 2005) or by inducing a hierarchical classification of candidate labels (Srikanth et al., 2005).

In this paper, we explore the use of several natural language resources to construct image annotation models that are capable of automatically tagging images from unrestricted domains with good accuracy. Unlike traditional image annotation methodologies that generate tags using image-based features, we propose to extract them in a manner analogous to keyword extraction. Given a target image and its surrounding text, we extract those words and phrases that are most likely to represent meaningful tags. More importantly, we

are interested to investigate the potential of such linguistic-based models on image annotation accuracy and reliability. Our work is motivated by the need for annotation models that can be efficiently applied on a very large scale (e.g. harvesting images from the web), which are required in applications that cannot afford the complexity and time associated with current image processing techniques.

The paper makes the following contributions. We first propose a new evaluation framework for image tagging, which is based on an analogy drawn between the tasks of image labeling and lexical substitution. Next, we present three extractive approaches for the task of image annotation. The methods proposed are based only on the text surrounding an image, without the use of image features. Finally, by combining several orthogonal methods through machine learning, we show that it is possible to achieve a performance that is competitive to a state-of-the-art image annotation system that relies on visual and textual features, thus demonstrating the effectiveness of text-based extractive annotation models.

2 Related Work

Several online systems have sprung into existence to achieve annotation of real world images through human collaborative efforts (Flickr) and stimulating competition (von Ahn and Dabbish, 2004). Although a large number of image tags can be generated in short time, these approaches depend on the availability of human annotators and are far from being automatic. Similarly, research in the other direction via text-to-image synthesis (Li and Fei-Fei, 2008; Collins et al., 2008; Mihalcea and Leong, 2009) has also helped to harvest images, mostly for concrete words, by refining image search engines.

Most approaches to automatic image annotation have focused on the generation of image labels using annotation models trained with image features and human annotated keywords (Barnard and Forsyth, 2001; Jeon et al., 2003; Makadia et al., 2008; Wang et al., 2009). Instead of predicting specific words, these methods generally target the generation of semantic classes (e.g. vegetation, animal, building, places etc), which they can achieve with a reasonable amount of success. Recent work has also considered the generation of labels for real-world images (Li and Wang, 2008; Feng and Lapata, 2008). To our knowledge, we are unaware of any other work that performs extractive annotation for images from unrestricted domains through the exclusive use of textual features.

3 Dataset

As the methods we propose are extractive, standard image databases with no surrounding text such as Corel (Duygulu et al., 2002) are not suitable, nor are they representative for the challenges associated with raw data from unrestricted domains. We thus create our own dataset using images randomly extracted from the Web.

To avoid sparse searches, we use a list of the most frequent words in the British National Corpus as seed words, and query the web using the Google Image API. A webpage is randomly selected from the query results if it contains a single image in the specified size range (width and height of 275 to 1000 pixels[1]) and its text contains more than 10 words. Next, we use a Document Object Model (DOM) HTML parser[2] to extract the content of the webpage. Note that we do not perform manual filtering of our images except where they contain undesirable qualities (e.g. porn, corrupted or blank images).

In total, we collected 300 image-text pairs from the web. The average image size is 496 pixels width and 461 pixels height. The average text length is 278 tokens and the average document title length is 6 tokens. In total, there are 83,522 words and the total vocabulary is 8,409 words.

For each image, we also create a gold standard of manually assigned tags, by using the labels assigned by five human annotators. The image annotation is conducted via Amazon Mechanical Turk, which was shown in the past to produce reliable annotations (Snow et al., 2008). For increased annotation reliability, we only accept annotators with an approval rating of 98%.

Given an image, an annotator extracts from the associated text a minimum of five words or collocations.Annotators can choose words freely from the text, while collocation candidates are restricted to a fixed set obtained from the n-grams (n $\leq$ 7) in the text that also appear as article names or surface forms in Wikipedia. Moreover, when interpreting the image, the annotators are instructed to focus on both the denotational and conotational attributes present in the image[3].

---

[1] Empirically determined to filter advertisements, banners and undersized images.

[2] http://search.cpan.org/dist/HTML-ContentExtractor/

[3] Annotation instructions, dataset and gold standard can

	Normal Image	**Mode** Image
Gold standard	czech (5), festival (5), oklahoma (4), yukon (4), october (4), web page (2), the first (2), event (2), success (1), every (1), year (1)	train (5), station (4), steam (4), trans siberian (4), steam train (4), travel (3), park (3), siberian (3), old (3), photo (1), trans (2), yekaterinburg (2), the web (2), photo host (1)

Table 1: Two sample images. The number besides each label indicates the number of human annotators agreeing on that label. Note that the mode image has a tag (i.e."train") in the gold standard set most frequently selected by the annotators

4 A New Evaluation Framework : Image Tagging as Lexical Substitution

While evaluations of previous work in image annotation were often based on labels provided with the images, such as tags or image captions, in our dataset such annotations are either missing or unreliable. We rely instead on human-produced extractive annotations (as described in the previous section), and formulate a new evaluation framework based on the intuition that an image can be substituted with one or more tags that convey the same meaning as the image itself. Ideally, there is a single tag that "best" describes the image overall (i.e. the gold standard tag agreed by the majority of human annotators), but there are also multiple tags that describe the fine-grained concepts present in the image. Our evaluation framework is inspired by the lexical substitution task (McCarthy and Navigli, 2007), where a system attempts to generate a word (or a set of words) to replace a target word, such that the meaning of the sentence is preserved.

Given this analogy, the evaluation metrics used for lexical substitution can be adapted to the evaluation of image tagging. Specifically, we measure the precision and the recall of a tagging method using four subtasks: **best normal**: provides precision and recall for the top-ranked tag returned by a method; **best mode**: provides precision and recall only if the top-ranked tag by a method matches the tag in the gold standard that was most frequently selected by the annotators; **out of ten (oot) nor-**

mal: provides precision and recall for the top ten tags by the system; and **out of ten (oot) mode**: similar to best mode, but it considers the top ten tags returned by the system instead of one. Table 1 show examples of a normal and a mode image.

Formally, let us assume that H is the set of annotators, namely $\{h_1, h_2, h_3, ...\}$, and I, $\{i_1, i_2, i_3, ...\}$ is the set of images for which each human annotator provide at least five tags. For each i_j, we calculate m_j, which is the most frequent tag for that image, if available. We also collect all r_j^k, which is the set of tags for the image i_j from the annotator h_k.

Let the set of those images where there is a tag agreed upon by the most annotators (i.e. the images with a mode) be denoted by IM, such that $IM \subseteq I$. Also, let $A \subseteq I$ be the set of images for which the system provides more than one tag. Let the corresponding set for the images with modes be denoted by AM, such that $AM \subseteq IM$. Let $a_j \in A$ be the set of system's extracted tags for the image i_j.

Thus, for each image i_j, we have the set of tags extracted by the system, and the set of tags from the human annotators. As the next step, the multiset union of the human tags is calculated, and the frequencies of the unique tags is noted. Therefore, for image i_j, we calculate R_j, which is $\sum r_j^k$, and the individual unique tag in R_j, say res, will have a frequency associated with it, namely $freq_{res}$.

Given this setting, the precision (P) and recall (R) metrics we use are defined below.

be downloaded at
http://lit.csci.unt.edu/index.php/Downloads

Best measures:

$$P = \frac{\sum_{a_j : i_j \in A} \frac{\sum_{res \in a_j} freq_{res}}{|a_j|}}{|A|}$$

$$R = \frac{\sum_{a_j : i_j \in I} \frac{\sum_{res \in a_j} freq_{res}}{|a_j|}}{|I|}$$

$$modeP = \frac{\sum_{bestguess_j \in AM}(1 if\_best\_guess = m_j)}{|AM|}$$

$$modeR = \frac{\sum_{bestguess_j \in IM}(1 if\_best\_guess = m_j)}{|IM|}$$

Out of ten (oot) measures:

$$P = \frac{\sum_{a_j : i_j \in A} \frac{\sum_{res \in a_j} freq_{res}}{|R_j|}}{|A|}$$

$$R = \frac{\sum_{a_j : i_j \in I} \frac{\sum_{res \in a_j} freq_{res}}{|R_j|}}{|I|}$$

$$modeP = \frac{\sum_{a_j : i_j \in AM}(1 if\_any\_guess \in a_j = m_j)}{|AM|}$$

$$modeR = \frac{\sum_{a_j : i_j \in IM}(1 if\_any\_guess \in a_j = m_j)}{|IM|}$$

As a simplified example (with less tags), consider i_j showing a picture of a Chihuahua being labeled by five annotators with the following tags :

Annotator	Tags
1	dog,pet
2	chihuahua
3	animal,dog
4	dog,chihuahua
5	dog

In this case, $r_j^1 = \{dog,pet\}$, $r_j^2 = \{chihuahua\}$, $r_j^3 = \{animal,dog\}$ and so on. The tag "dog" appears the most frequent among the five annotators, hence $m_j = \{dog\}$. $R_j = \{dog, dog, dog, dog, chihuahua, chihuahua, animal, pet\}$. The res with associated frequencies would be dog 4, chihuahua 2, animal 1, pet 1. If the system's proposed tag for i_j is $\{dog, animal\}$, then the numerator of P and R for best subtask would be $\frac{\frac{4+1}{2}}{8} = 0.313$. Similarly, the numerator of P and R for oot subtask is $\frac{4+1}{8} = 0.625$.

5 Extractive Image Annotation

The main idea underlying our work is that we can perform effective image annotation using information drawn from the associated text. Following (Feng and Lapata, 2008), we propose that an image can be annotated with keywords capturing the denotative (entities or objects depicted) and connotative (semantics or ideologies interpreted) attributes in the image. For instance, a picture showing a group of athletes and a ball may also be tagged with words like "soccer," or "sports activity." Specifically, we use a combination of knowledge sources to model the denotative quality of a word as its picturability, and the connotative attribute as its saliency. The idea of visualness and salience as textual features for discovering named entities in an image was first pursued by (Deschacht and Moens, 2007), using data from the news domain. In contrast, we are able to perform annotation of images from unrestricted domains using content words (nouns, verbs and adjectives). In the following, we first describe three unsupervised extractive approaches for image annotation, followed by a supervised method using a re-ranking hypothesis that combines all the methods.

5.1 Flickr Picturability

Featuring a repository of four billion images, Flickr (http://www.flickr.com) is one of the most comprehensive image resources on the web. As a photo management and sharing application, it provides users with the ability to tag, organize, and share their photos online. Interestingly, an inspection of Flickr tags for randomly selected images reveal that users tend to describe the denotational attributes of images, using concrete and picturable words such as *cat*, *bug*, *car* etc. This observation lends evidence to Flickr's suitability as a resource to model the picturability of words.

Given the text (T) of an image, we can use the $getRelatedTags$ API to retrieve the most frequent Flickr tags associated with a given word, and use them as corpus evidence to filter or promote words in the text. In the filtering phase we ignore any words that return an empty list of Flickr's related tags, based on the assumption that these words are not used in the Flickr tags repository. We also discard words with a length that is less than three characters ($\alpha=3$). In the promotion phase, we reward any retrieved tags that appear as surface forms in the text. This reward is proportional to the term frequency of these tags in the

Algorithm 1 Flickr Picturability Algorithm

Start : L[]=ϕ , TF[]=tf of each word in T
for each word in T **do**
  if $length(word) \geq \alpha$ **then**
    RelatedTags=getRelatedTags(word);
    if $size(RelatedTags) > 0$ **then**
      L[word]+=β*TF[word]
      for each tag in RelatedTags **do**
        if $exists\ TF[tag]$ **then**
          L[tag]+=TF[tag]
        end if
      end for
    end if
  end if
end for

text. Additionally, we also include in the final label set any word that returns a non-empty related tags set with a discounted weight (β=0.5) of its term frequency, to the end of enriching our labels set while assuring more credit are given to the picturable words.

To extract multiword labels, we locate all n-grams formed exclusively from our extracted set of possible labels. The subsequent score for each of these n-grams is:

$$L[w_i..w_{i+k}] = (\sum_{j=i}^{j=i+k} L[wj])/k$$

By reverse sorting the associative array in L, we can retrieve the top K words to label the image. For illustration, let us consider the following text snippet.

> *On the Origin of Species, published by Charles Darwin in 1859, is considered to be the foundation of evolutionary biology.*

After removing stopwords, we consider the remaining words as candidate labels. For each of these candidates w_i (i.e. $origin$, $species$, $published$, $charles$, $darwin$, $foundation$, $evolutionary$, and $biology$), we query Flickr and obtain their related tag set R_i. $origin$, $published$, and $foundation$ return an empty set of related tags and hence are removed from our set of candidate labels, leaving $species$, $charles$, $darwin$, $evolutionary$, and $biology$ as possible annotation keywords with the initial score of 0.5. In the promotion phase, we score each w_i based on the number of votes it receives from the remaining w_j

(Figure 1). Each vote represents an occurrence of the candidate tag w_i in the related tag set R_j of the candidate tag w_j. For example, $darwin$ appeared in the Flickr related tags for $charles$, $evolutionary$, and $biology$, hence it has a weight of 3.5. The final list of candidate labels are shown in Table 2.
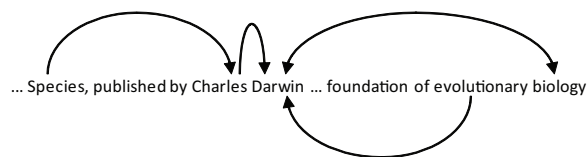

... Species, published by Charles Darwin ... foundation of evolutionary biology

Figure 1: Flickr Picturability Labels

Label	$S(w_i)$
darwin	3.5
charles darwin	2.5
charles	1.5
biology	1.5
evolutionary biology	1.0
evolutionary	0.5
species	0.5

Table 2: Candidate labels obtained for a sample text using the Flickr model

5.2 Wikipedia Salience

We hypothesize that an image often describes the most important concepts in the associated text. Thus, the keywords selected from a text could be used as candidate labels for the image. We use a graph-based keyword extraction method similar to (Mihalcea and Tarau, 2004), enhanced with a semantic similarity measure. Starting with a text, we extract all the candidate labels and add them as vertices in the graph. A measure of word similarity is then used to draw weighted edges between the nodes. Using the PageRank algorithm, the words are assigned with a score indicating their salience within the given text.

To determine the similarity between words, we use a directed measure of similarity. Most word similarity metrics provide a single-valued score between a pair of words w_1 and w_2 to indicate their semantic similarity. Intuitively, this is not always the case, as w_1 may be represented by concepts that are entirely embedded in other concepts, represented by w_2. In psycholinguistics terms, uttering w_1 may bring to mind w_2, while the appearance of w_2 without any contextual clues may not associate with w_1. For example, $Obama$ brings to mind the concept of $president$, but $president$

may trigger other concepts such as $Washington$, $Lincoln$, $Ford$ etc., depending on the existing contextual clues. Thus, the degree of similarity of w_1 with respect to w_2 should be separated from that of w_2 with respect to w_1. Specifically, we use the following measure of similarity, based on the Explicit Semantic Analysis (ESA) vectors derived from Wikipedia (Gabrilovich and Markovitch, 2007):

$$DSim(w_i, w_j) = \frac{C_{ij}}{C_i} * Sim(w_i, w_j)$$

where C_{ij} is the count of articles in Wikipedia containing words w_i and w_j, C_i is the count of articles containing words w_i, and $Sim(w_i, w_j)$ is the cosine similarity of the ESA vectors representing the input words. The *directional weight* (C_{ij}/C_i) amounts to the degree of association of w_i with respect to w_j. Using the directional inferential similarity scores as directed edges and distinct words as vertices, we obtain a graph for each text. The directed edges denotes the idea of "recommendation" where we say w_1 recommends w_2 if and only if there is a directed edge from w_1 to w_2, with the weight of the recommendation being the directional similarity score. Starting with this graph, we use the graph iteration algorithm from (Mihalcea and Tarau, 2004) to calculate a score for each vertex in the graph. The output is a sorted list of words in decreasing order of their ranks, which are used as candidate labels to annotate the image. This is achieved by using C_j instead of C_i for the denominator in the directional weight. As an example, consider the text snippet :

> *Microsoft Corporation is a multinational computer technology corporation that develops, manufactures, licenses, and supports a wide range of software products for computing devices*

after stopword removal, the list of nouns extracted is *Microsoft, computer, corporation, devices, products, technology, software*. Note that the top-ranked word must infer some or all of the words in the text. In this case, the word *Microsoft* infers the terms *computer*, *technology* and *software*.

To calculate the semantic relatedness between two collocations, we use a simplified version of the text-to-text relatedness technique proposed by and (Mihalcea et al., 2006) that incorporate the directional inferential similarity as an underlying semantic metric.

5.3 Topical Modeling

Intuitively, every text is written with a topic in mind, and the associated image serves as an illustration of the text meaning. In this paper, we investigate the effect of topical modeling on image annotation accuracy directly. We use the Pachinko Allocation Model (PAM) (Li and McCallum, 2006) to model the topics in a text, where keywords forming the dominant topic are assumed as our set of annotation keywords. Compared with previous topic modeling approaches, such as Latent Dirichlet allocation (LDA) or its improved variant Correlated Topic Model (CTM) (Blei and Lafferty, 2007), PAM captures correlations between all the topic pairs using a directed acyclic graph (DAG). It also supports finer-grained topic modeling, and has state-of-the-art performance on the tasks of document classification and topical keyword coherence. Given a text, we use the PAM model to infer a list of *super-topics* and *sub-topics* together with words weighted according to the likelihood that they belong to each of these topics. For each text, we retrieve the top words belonging to the dominant super-topic and sub-topic. We use 50 super-topics and 100 sub-topics as operating parameters for PAM, since these values were found to provide good results in previous work on topic modeling. Default values are used for other parameters in the model.

5.4 Supervised Learning

The three tagging methods target different aspects of what constitutes a good label for an image. We use them as features in a machine learning framework, and introduce a final rank attribute $S(t_j)$, which is a linear combination of the reciprocals of the rank of each tag as given by each method,

$$S(t_j) = \sum_{m \in methods} \lambda_m \frac{1}{r^m_{t_j}}$$

where $r^m_{t_j}$ is the rank for tag t_j given by method m. The weight of each method λ_m is estimated from the training set using information gain values. Since our predicted variable (*mode* precision or recall) is continuous, we use the Support Vector Algorithm (nu-SVR) implementation of SVM (Chang and Lin, 2001) to perform regression analysis on the weights for each method via a radial basis function kernel. A ten-fold cross-validation is applied on the entire dataset of 300 images.

652

Models	Best				out-of-ten (oot)			
	Normal		Mode		Normal		Mode	
	P	R	P	R	P	R	P	R
Flickr picturability	6.32	6.32	78.57	78.57	35.61	35.61	92.86	92.86
Wikipedia Salience	6.40	6.40	7.14	7.14	35.19	35.19	92.86	92.86
Topic modeling	5.99	5.99	42.86	42.86	37.13	37.13	85.71	85.71
Combined (SVM)	6.87	6.87	67.49	67.49	37.85	37.85	100.00	100.00
Doc Title	6.40	6.40	75.00	75.00	18.97	18.97	82.14	82.14
*tf * idf*	5.94	5.94	14.29	14.29	38.40	38.40	78.57	78.57
Random	3.76	3.76	3.57	3.57	30.20	30.20	50.00	50.00
Upper bound (human)	12.23	12.07	81.48	81.48	82.44	81.55	100.00	100.00

Table 3: Results obtained on the Web dataset

6 Experiments and Evaluations

We evaluate the performance of each of the three tagging methods separately, followed by an evaluation of the combined method. Each system produces a ranked list of K words or collocations as tags assigned to a given image. A system can discretionary generate less (but not more) than K tags, depending on its confidence level.

For comparison, we implement three baselines: *tf*idf*, *Doc Title* and *Random*. For *tf*idf*, we use the British National Corpus to calculate the *idf* scores, while the frequency of a term is calculated from the entire text associated with an image. The *Doc Title* baseline is similar, except that the term frequency is calculated based on the title of the document. The *Random* baseline randomly selects words from a co-occurrence window of size K before and after an image as its annotation. Following other tagging methods, we apply a pre-processing stage, where we part-of-speech tag the text (to retain only nouns), followed by stemming. We also determine an upper bound, which is calculated as follows. For each image, the labels assigned by each of the five annotators are in turn evaluated against a gold standard consisting of the annotations of the other four annotators. The best performing annotator is then recorded. This process is repeated for each of the 300 images, and the average precision and recall are calculated. This represents an upper bound, as it is the best performance that a human can achieve on this dataset. Table 3 shows our experimental results.

Among the individual methods, the method implementing Flickr picturability has the highest individual score for *best* and *oot* modes, yielding a precision and recall of 78.57% and 92.86% respectively. The Wikipedia Saliency method also scores the highest (jointly with Flickr) in the *oot* mode, but for the *best* mode achieves a score only marginally better than the random baseline. A plausible explanation is that it tends to favor "all-

inferring" over-specific labels, while the most frequently selected tags in mode pictures are typically more "picturable" than being specific (e.g. "train" for the mode picture in Table 1). The topic modeling method has mixed results: its scores for *oot* normal and mode are somewhat competitive with *tf*idf*, but it scores consistently lower than the DocTitle in the *best* subtask, possibly due to the absence of a more sophisticated re-ranking algorithm tailored for the image annotation task other than the intrinsic ranking mechanism in PAM. It is worth noting that the combined supervised system provides the overall best results (6.87%) on the *best* normal, and achieves a perfect precision and recall (100%) for *oot* mode, which means perfect agreement with the human tagging.

7 Comparison with Related Work

We also compare our work against (Feng and Lapata, 2008) as it allows for a direct comparison with models using both image and textual features under a standard evaluation framework. We obtained the BBC dataset used in their experiments, which consists of 3121 training and 240 testing images. In this dataset, images are implicitly tagged with captions by the author of the corresponding BBC article. The evaluations are run against these captions.

In their experiments, Feng and Lapata created four annotation models. The first two (tf*idf and Document Title) are the same as used in our baseline experiments. The third model (Lavrenko03) is an application of the continuous relevance model in (Jeon et al., 2003), trained with the BBC image features and captions. Finally, the forth (ExtModel) is an extension of the relevance model using additional information in auxiliary texts. Briefly, the model assumes a multiple Bernoulli distribution for words in a caption, and generates tags for a test image using a weighted combination of the accompanying document, caption and image features learned during training.

Models	Top 10			Top 15			Top 20		
	P	R	F1	P	R	F1	P	R	F1
*tf*idf*	4.37	7.09	5.41	3.57	8.12	4.86	2.65	8.89	4.00
DocTitle	9.22	7.03	7.20	9.22	7.03	7.20	9.22	7.03	7.20
Lavrenko03	9.05	16.01	11.81	7.73	17.87	10.71	6.55	19.38	9.79
ExtModel	14.72	27.95	19.82	11.62	32.99	17.18	9.72	36.77	15.39
Flickr picturability	12.13	22.82	15.84	9.52	26.82	14.05	8.23	29.80	12.90
Wikipedia Salience	11.63	21.89	15.18	9.28	26.20	13.70	7.81	29.41	12.35
Topic Modeling	11.42	21.49	14.91	9.28	26.20	13.70	7.86	29.57	12.42
Combined (SVM)	13.38	25.17	17.47	11.08	31.29	16.37	9.50	35.76	15.01

Table 4: Results obtained on the BBC dataset used in (Feng and Lapata, 2008)

The experimental setup is similar to the earlier section, but a few modifications are made for a fair and direct comparison. First, we extend our models coverage to include content words (i.e. nouns, verbs, adjectives) determined using the Tree Tagger (Schmid, 1994). Second, no collocations are used. Third, we adopt the evaluation framework used by Feng and Lapata to extract the top 10, 15 and 20 tags. Note that in our methods, the extraction of tags for a test image is only done on the document surrounding the image, after excluding the caption. As the number of negative examples (words not present in the caption) greatly outnumber the positive instances, we employ an undersampling method (Kubat and Matwin, 1997) to balance the dataset for training.

The results are shown in Table 4. Interestingly, all our unsupervised extraction-based models perform consistently above the supervised Lavrenko03 model, indicating that textual features are more informative than captions and image features taken together. Comparing with models using significantly less document information (tf*idf and Doc title), our models gain even greater advantage. Note that the title of any BBC article does not exceed 10 words, hence comparison is only meaningful given the top 10 tags retrieved.

Feng and Lapata used LDA to perform reranking of final candidates in their ExtModel. However, when used as a model alone, the PAM topic model achieved promising scores in all the categories, performing best for top 10 keywords (F1 of 14.91%). Flickr picturability stands out as the best performing unsupervised method, scoring the highest precision (12.13%, top 10), recall (29.80%, top 20) and F1 (15.84%, top 10).

Overall, this comparative evaluation yields some important insights. First, our combined model using SVM is statistically better (p<0.1 for top 10, 15, 20) than the Laverenko03 model, but not statistically different from the ExtModel. This demonstrates the effectiveness of textual-based

models over traditional models trained with image features and captions. While it is intuitively clear that image features help in improving tagging performance, we show that mining only the text surrounding an image, where it exists, can yield a performance that is comparable to a state-of-the-art system that uses both textual and visual features. Moreover, an increase in complexity of a model by using more features may hinder its applicability to large datasets, but not necessarily improving annotation performance (Makadia et al., 2008). On this, text-based annotation models can provide a desirable compromise. For instance, our unsupervised models implementing Flickr picturability and Wikipedia Salience are able to extract annotations from a BBC article (average 133.85 tokens) in approximately 1 second and 20 seconds respectively.

8 Conclusions and Future Work

In this paper, we introduced several text-based extractive approaches for automatic image annotation and showed that they compare favorably with the state-of-the-art in image annotation using both text and image features. We believe our work has practical applications in mining and annotating images over the Web, where texts are naturally associated with images, and scalability is important. Our next direction seeks to derive robust annotation models using additional ontological knowledge-bases. We would also like to advance the the state-of-the-art by augmenting current textual models with image features.

Acknowledgments

References

Kobus Barnard and David Forsyth. 2001. Learning the semantics of words and pictures. In *Proceedings of International Conference on Computer Vision*.

David Blei and John Lafferty. 2007. A correlated topic model of science. In *Annals of Applied Statistics*, volume 1, pages 17–35.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*.

Brendan Collins, Jia Deng, Kai Li, and Li Fei-Fei. 2008. Towards scalable dataset construction: An active learning approach. In *Proceedings of European Conference on Computer Vision*.

Koen Deschacht and Marie-Francine Moens. 2007. Text analysis for automatic image annotation. In *Proceedings of the Association for Computational Linguistics*.

Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David Forsyth. 2002. Object recognition as machine translation:learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision*.

Yansong Feng and Mirella Lapata. 2008. Automatic image annotation using auxiliary text information. In *Proceedings of the Association for Computational Linguistics*.

Rob Fergus, Pietro Perona, and Andrew Zisserman. 2003. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *International Joint Conferences on Artificial Intelligence*.

Michael Grubinger, Clough Paul, Mller Henning, and Deselaers Thomas. 2006. The iapr benchmark: A new evaluation resource for visual information systems. In *International Conference on Language Resources and Evaluation*.

Jiwoon Jeon, Victor Lavrenko, and R Manmatha. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yohan Jin, Latifur Khan, Lei Wang, and Mamoun Awad. 2005. Image annotations by combining multiple evidence & wordnet. In *Proceedings of Annual ACM Multimedia*.

Miroslav Kubat and Stan Matwin. 1997. Addressing the curse of imbalanced training sets: one-sided selection. In *Proceedings of International Conference on Machine Learning*.

Li-Jia Li and Li Fei-Fei. 2008. Optimol: automatic online picture collection via incremental model learning. In *International Journal of Computer Vision*.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the International Conference on Machine learning*.

Jia Li and James Wang. 2008. Real-time computerized annotation of pictures. In *Proceedings of International Conference on Computer Vision*.

Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. 2008. A new baseline for image annotation. In *Proceedings of European Conference on Computer Vision*.

Diana McCarthy and Roberto Navigli. 2007. The semeval English lexical substitution task. In *Proceedings of the ACL Semeval workshop*.

Rada Mihalcea and Chee Wee Leong. 2009. Towards communicating simple sentences using pictorial representations. In *Machine Translation*, volume 22, pages 153–173.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of Empirical Methods in Natural Language Processing*.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of Association for the Advancement of Artificial Intelligence*, pages 775–780.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of Empirical Methods in Natural Language Processing*.

Srihari and Burhans. 1994. Visual semantics: Extracting visual information from text accompanying pictures. In *Proceedings of the American Association for Artificial Intelligence*.

Munirathnam Srikanth, Joshua Varner, Mitchell Bowden, and Dan Moldovan. 2005. Exploiting ontologies for automatic image annotation. In *Proceedings of the ACM Special Interest Group on Research and Development in Information Retrieval*.

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the ACM Special Interest Group on Computer Human Interaction*.

Chong Wang, David Blei, and Li Fei-Fei. 2009. Simultaneous image classification and annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.