Automatic Image Annotation Using Auxiliary Text Information

Yansong Feng and **Mirella Lapata**
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, UK
Y.Feng-4@sms.ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

The availability of databases of images labeled with keywords is necessary for developing and evaluating image annotation models. Dataset collection is however a costly and time consuming task. In this paper we exploit the vast resource of images available on the web. We create a database of pictures that are naturally embedded into news articles and propose to use their captions as a proxy for annotation keywords. Experimental results show that an image annotation model can be developed on this dataset alone without the overhead of manual annotation. We also demonstrate that the news article associated with the picture can be used to boost image annotation performance.

1 Introduction

As the number of image collections is rapidly growing, so does the need to browse and search them. Recent years have witnessed significant progress in developing methods for image retrieval[1], many of which are query-based. Given a database of images, each annotated with keywords, the query is used to retrieve relevant pictures under the assumption that the annotations can essentially capture their semantics.

One stumbling block to the widespread use of query-based image retrieval systems is obtaining the keywords for the images. Since manual annotation is expensive, time-consuming and practically infeasible for large databases, there has been great in-

terest in automating the image annotation process (see references). More formally, given an image I with visual features $V_i = \{v_1, v_2, \ldots, v_N\}$ and a set of keywords $W = \{w_1, w_2, \ldots, w_M\}$, the task consists in finding *automatically* the keyword subset $W_I \subset W$, which can appropriately describe the image I. Indeed, several approaches have been proposed to solve this problem under a variety of learning paradigms. These range from supervised classification (Vailaya et al., 2001; Smeulders et al., 2000) to instantiations of the noisy-channel model (Duygulu et al., 2002), to clustering (Barnard et al., 2002), and methods inspired by information retrieval (Lavrenko et al., 2003; Feng et al., 2004).

Obviously in order to develop accurate image annotation models, some manually labeled data is required. Previous approaches have been developed and tested almost exclusively on the Corel database. The latter contains 600 CD-ROMs, each containing about 100 images representing the same topic or concept, e.g., people, landscape, male. Each topic is associated with keywords and these are assumed to also describe the images under this topic. As an example consider the pictures in Figure 1 which are classified under the topic *male* and have the description keywords *man, male, people, cloth*, and *face*.

Current image annotation methods work well when large amounts of labeled images are available but can run into severe difficulties when the number of images and keywords for a given topic is relatively small. Unfortunately, databases like Corel are few and far between and somewhat idealized. Corel contains clusters of many closely related images which in turn share keyword descriptions, thus allowing models to learn image-keyword associations

---

[1]The approaches are too numerous to list; we refer the interested reader to Datta et al. (2005) for an overview.

Figure 1: Images from the Corel database, exemplifying the concept *male* with keyword descriptions *man, male, people, cloth*, and *face.*

reliably (Tang and Lewis, 2007). It is unlikely that models trained on this database will perform well out-of-domain on other image collections which are more noisy and do not share these characteristics. Furthermore, in order to develop robust image annotation models, it is crucial to have large and diverse datasets both for training and evaluation.

In this work, we aim to relieve the data acquisition bottleneck associated with automatic image annotation by taking advantage of resources where images and their annotations co-occur naturally. News articles associated with images and their captions spring readily to mind (e.g., BBC News, Yahoo News). So, rather than laboriously annotating images with their keywords, we simply treat captions as labels. These annotations are admittedly noisy and far from ideal. Captions can be denotative (describing the objects the image depicts) but also connotative (describing sociological, political, or economic attitudes reflected in the image). Importantly, our images are not standalone, they come with news articles whose content is shared with the image. So, by processing the accompanying document, we can effectively learn about the image and reduce the effect of noise due to the approximate nature of the caption labels. To give a simple example, if two words appear both in the caption and the document, it is more likely that the annotation is genuine.

In what follows, we present a new database consisting of articles, images, and their captions which we collected from an on-line news source. We then propose an image annotation model which can learn from our noisy annotations and the auxiliary documents. Specifically, we extend and modify Lavrenko's (2003) continuous relevance model to suit our task. Our experimental results show that this model can successfully scale to our database, without making use of explicit human annotations in any way. We also show that the auxiliary document contains important information for generating more accurate image descriptions.

2 Related Work

Automatic image annotation is a popular task in computer vision. The earliest approaches are closely related to image classification (Vailaya et al., 2001; Smeulders et al., 2000), where pictures are assigned a set of simple descriptions such as indoor, outdoor, landscape, people, animal. A binary classifier is trained for each concept, sometimes in a "one vs all" setting. The focus here is mostly on image processing and good feature selection (e.g., colour, texture, contours) rather than the annotation task itself.

Recently, much progress has been made on the image annotation task thanks to three factors. The availability of the Corel database, the use of unsupervised methods and new insights from the related fields of natural language processing and information retrieval. The co-occurrence model (Mori et al., 1999) collects co-occurrence counts between words and image features and uses them to predict annotations for new images. Duygulu et al. (2002) improve on this model by treating image regions and keywords as a bi-text and using the EM algorithm to construct an image region-word dictionary.

Another way of capturing co-occurrence information is to introduce latent variables linking image features with words. Standard latent semantic analysis (LSA) and its probabilistic variant (PLSA) have been applied to this task (Hofmann, 1998). Barnard et al. (2002) propose a hierarchical latent model in order to account for the fact that some words are more general than others. More sophisticated graphical models (Blei and Jordan, 2003) have also been employed including Gaussian Mixture Models (GMM) and Latent Dirichlet Allocation (LDA).

Finally, relevance models originally developed for information retrieval, have been successfully applied to image annotation (Lavrenko et al., 2003; Feng et al., 2004). A key idea behind these models is to find the images most similar to the test image and then use their shared keywords for annotation.

Our approach differs from previous work in two

important respects. Firstly, our ultimate goal is to develop an image annotation model that can cope with real-world images and noisy data sets. To this end we are faced with the challenge of building an appropriate database for testing and training purposes. Our solution is to leverage the vast resource of images available on the web but also the fact that many of these images are implicitly annotated. For example, news articles often contain images whose captions can be thought of as annotations. Secondly, we allow our image annotation model access to knowledge sources other than the image and its keywords. This is relatively straightforward in our case; an image and its accompanying document have shared content, and we can use the latter to glean information about the former. But we hope to illustrate the more general point that auxiliary linguistic information can indeed bring performance improvements on the image annotation task.

3 BBC News Database

Our database consists of news images which are abundant. Many on-line news providers supply pictures with news articles, some even classify news into broad topic categories (e.g., business, world, sports, entertainment). Importantly, news images often display several objects and complex scenes and are usually associated with captions describing their contents. The captions are image specific and use a rich vocabulary. This is in marked contrast to the Corel database whose images contain one or two salient objects and a limited vocabulary (typically around 300 words).

We downloaded 3,361 news articles from the BBC News website.[2] Each article was accompanied with an image and its caption. We thus created a database of image-caption-document tuples. The documents cover a wide range of topics including national and international politics, advanced technology, sports, education, etc. An example of an entry in our database is illustrated in Figure 2. Here, the image caption is *Marcin and Florent face intense competition from outside Europe* and the accompanying article discusses EU subsidies to farmers. The images are usually 203 pixels wide and 152 pixels high. The average caption length is 5.35 tokens, and the average document length 133.85 tokens. Our

[2]http://news.bbc.co.uk/

Figure 2: A sample from our BBC News database. Each entry contains an image, a caption for the image, and the accompanying document with its title.

captions have a vocabulary of 2,167 words and our documents 6,253. The vocabulary shared between captions and documents is 2,056 words.

4 Extending the Continuous Relevance Annotation Model

Our work is an extension of the continuous relevance annotation model put forward in Lavrenko et al. (2003). Unlike other unsupervised approaches where a set of latent variables is introduced, each defining a joint distribution on the space of keywords and image features, the relevance model captures the joint probability of images and annotated words *directly*, without requiring an intermediate clustering stage. This model is a good point of departure for our task for several reasons, both theoretical and empirical. Firstly, expectations are computed over every single point in the training set and

therefore parameters can be estimated without EM. Indeed, Lavrenko et al. achieve competitive performance with latent variable models. Secondly, the generation of feature vectors is modeled directly, so there is no need for quantization. Thirdly, as we show below the model can be easily extended to incorporate information outside the image and its keywords.

In the following we first lay out the assumptions underlying our model. We next describe the continuous relevance model in more detail and present our extensions and modifications.

Assumptions   Since we are using a non-standard database, namely images embedded in documents, it is important to clarify what we mean by image annotation, and how the precise nature of our data impacts the task. We thus make the following assumptions:

1. The caption describes the content of the image directly or indirectly. Unlike traditional image annotation where keywords describe salient objects, captions supply more detailed information, not only about objects, and their attributes, but also events. In Figure 2 the caption mentions Marcin and Florent the two individuals shown in the picture but also the fact that they face competition from outside Europe.
2. Since our images are implicitly rather than explicitly labeled, we do not assume that we can annotate *all* objects present in the image. Instead, we hope to be able to model event-related information such as "what happened", "who did it", "when" and "where". Our annotation task is therefore more semantic in nature than traditionally assumed.
3. The accompanying document describes the content of the image. This is trivially true for news documents where the images conventionally depict events, objects or people mentioned in the article.

To validate these assumptions, we performed the following experiment on our BBC News dataset. We randomly selected 240 image-caption pairs and manually assessed whether the caption content words (i.e., nouns, verbs, and adjectives) could describe the image. We found out that the captions express the picture's content 90% of the time. Furthermore, approximately 88% of the nouns in sub-

ject or object position directly denote salient picture objects. We thus conclude that the captions contain useful information about the picture and can be used for annotation purposes.

Model Description   The continuous relevance image annotation model (Lavrenko et al., 2003) generatively learns the joint probability distribution $P(V, W)$ of words W and image regions V. The key assumption here is that the process of generating images is conditionally independent from the process of generating words. Each annotated image in the training set is treated as a latent variable. Then for an unknown image I, we estimate:

$$P(V_I, W_I) = \sum_{s \in D} P(V_I|s)P(W_I|s)P(s), \qquad (1)$$

where D is the number of images in the training database, V_I are visual features of the image regions representing I, W_I are the keywords of I, s is a latent variable (i.e., an image-annotation pair), and $P(s)$ the prior probability of s. The latter is drawn from a uniform distribution:

$$P(s) = \frac{1}{N_D} \qquad (2)$$

where N_D is number of the latent variables in the training database D.

When estimating $P(V_I|s)$, the probability of image regions and words, Lavrenko et al. (2003) reasonably assume a generative Gaussian kernel distribution for the image regions:

$$P(V_I|s) = \prod_{r=1}^{N_{V_I}} P_g(v_r|s) \qquad (3)$$

$$= \prod_{r=1}^{N_{V_I}} \frac{1}{n_{s_v}} \sum_{i=1}^{n_{s_v}} \frac{\exp\left\{(v_r - v_i)^T \Sigma^{-1}(v_r - v_i)\right\}}{\sqrt{2^k \pi^k |\Sigma|}}$$

where N_{V_I} is the number of regions in image I, v_r the feature vector for region r in image I, n_{s_v} the number of regions in the image of latent variable s, v_i the feature vector for region i in s's image, k the dimension of the image feature vectors and Σ the feature covariance matrix. According to equation (3), a Gaussian kernel is fit to every feature vector v_i corresponding to region i in the image of the latent variable s. Each kernel here is determined by the feature covariance matrix Σ, and for simplicity, Σ is assumed to be a diagonal matrix: $\Sigma = \beta I$, where I is the identity matrix; and β is a scalar modulating the bandwidth of

the kernel whose value is optimized on the development set.

Lavrenko et al. (2003) estimate the word probabilities $P(W_I|s)$ using a multinomial distribution. This is a reasonable assumption in the Corel dataset, where the annotations have similar lengths and the words reflect the salience of objects in the image (the multinomial model tends to favor words that appear multiple times in the annotation). However, in our dataset the annotations have varying lengths, and do not necessarily reflect object salience. We are more interested in modeling the *presence* or *absence* of words in the annotation and thus use the multiple-Bernoulli distribution to generate words (Feng et al., 2004). And rather than relying solely on annotations in the training database, we can also take the accompanying document into account using a weighted combination.

The probability of sampling a set of words W given a latent variable s from the underlying multiple Bernoulli distribution that has generated the training set D is:

$$P(W|s) = \prod_{w \in W} P(w|s) \prod_{w \notin W} (1 - P(w|s)) \quad (4)$$

where $P(w|s)$ denotes the probability of the w'th component of the multiple Bernoulli distribution. Now, in estimating $P(w|s)$ we can include the document as:

$$P_{est}(w|s) = \alpha P_{est}(w|s_a) + (1-\alpha)P_{est}(w|s_d) \quad (5)$$

where α is a smoothing parameter tuned on the development set, s_a is the annotation for the latent variable s and s_d its corresponding document.

Equation (5) smooths the influence of the annotation words and allows to offset the negative effect of the noise inherent in our dataset. Since our images are implicitly annotated, there is no guarantee that the annotations are all appropriate. By taking into account $P_{est}(w|s_d)$, it is possible to annotate an image with a word that appears in the document but is not included in the caption.

We use a Bayesian framework for estimating $P_{est}(w|s_a)$. Specifically, we assume a beta prior (conjugate to the Bernoulli distribution) for each word:

$$P_{est}(w|s_a) = \frac{\mu\, b_{w,s_a} + N_w}{\mu + D} \quad (6)$$

where μ is a smoothing parameter estimated on the development set, b_{w,s_a} is a Boolean variable denoting whether w appears in the annotation s_a, and N_w is the number of latent variables that contain w in their annotations.

We estimate $P_{est}(w|s_d)$ using maximum likelihood estimation (Ponte and Croft, 1998):

$$P_{est}(w|s_d) = \frac{num_{w,s_d}}{num_{s_d}} \quad (7)$$

where num_{w,s_d} denotes the frequency of w in the accompanying document of latent variable s and num_{s_d} the number of all tokens in the document. Note that we purposely leave P_{est} unsmoothed, since it is used as a means of balancing the weight of word frequencies in annotations. So, if a word does not appear in the document, the possibility of selecting it will not be greater than α (see Equation (5)).

Unfortunately, including the document in the estimation of $P_{est}(w|s)$ increases the vocabulary which in turn increases computation time. Given a test image-document pair, we must evaluate $P(w|V_I)$ for every w in our vocabulary which is the union of the caption and document words. We reduce the search space, by scoring each document word with its $tf * idf$ weight (Salton and McGill, 1983) and adding the n-best candidates to our caption vocabulary. This way the vocabulary is not fixed in advance for all images but changes dynamically depending on the document at hand.

Re-ranking the Annotation Hypotheses It is easy to see that the output of our model is a ranked word list. Typically, the k-best words are taken to be the automatic annotations for a test image I (Duygulu et al., 2002; Lavrenko et al., 2003; Jeon and Manmatha, 2004) where k is a small number and the same for all images.

So far we have taken account of the auxiliary document rather naively, by considering its vocabulary in the estimation of $P(W|s)$. Crucially, documents are written with one or more topics in mind. The image (and its annotations) are likely to represent these topics, so ideally our model should prefer words that are strong topic indicators. A simple way to implement this idea is by re-ranking our k-best list according to a topic model estimated from the entire document collection.

Specifically, we use Latent Dirichlet Allocation (LDA) as our topic model (Blei et al., 2003). LDA

represents documents as a mixture of topics and has been previously used to perform document classification (Blei et al., 2003) and ad-hoc information retrieval (Wei and Croft, 2006) with good results. Given a collection of documents and a set of latent variables (i.e., the number of topics), the LDA model estimates the probability of topics per document and the probability of words per topic. The topic mixture is drawn from a conjugate Dirichlet prior that remains the same for all documents.

For our re-ranking task, we use the LDA model to infer the m-best topics in the accompanying document. We then select from the output of our model those words that are most likely according to these topics. To give a concrete example, let us assume that for a given image our model has produced five annotations, w_1, w_2, w_3, w_4, and w_5. However, according to the LDA model neither w_2 nor w_5 are likely topic indicators. We therefore remove w_2 and w_5 and substitute them with words further down the ranked list that are topical (e.g., w_6 and w_7). An advantage of using LDA is that at test time we can perform inference without retraining the topic model.

5 Experimental Setup

In this section we discuss our experimental design for assessing the performance of the model presented above. We give details on our training procedure and parameter estimation, describe our features, and present the baseline methods used for comparison with our approach.

Data Our model was trained and tested on the database introduced in Section 3. We used 2,881 image-caption-document tuples for training, 240 tuples for development and 240 for testing. The documents and captions were part-of-speech tagged and lemmatized with Tree Tagger (Schmid, 1994).Words other than nouns, verbs, and adjectives were discarded. Words that were attested less than five times in the training set were also removed to avoid unreliable estimation. In total, our vocabulary consisted of 8,309 words.

Model Parameters Images are typically segmented into regions prior to training. We impose a fixed-size rectangular grid on each image rather than attempting segmentation using a general purpose algorithm such as normalized cuts (Shi and Malik,

Color
average of RGB components, standard deviation
average of LUV components, standard deviation
average of LAB components, standard deviation
Texture
output of DCT transformation
output of Gabor filtering (4 directions, 3 scales)
Shape
oriented edge (4 directions)
ratio of edge to non-edge

Table 2: Set of image features used in our experiments.

2000). Using a grid avoids unnecessary errors from image segmentation algorithms, reduces computation time, and simplifies parameter estimation (Feng et al., 2004). Taking the small size and low resolution of the BBC News images into account, we averagely divide each image into 6×5 rectangles and extract features for each region. We use 46 features based on color, texture, and shape. They are summarized in Table 2.

The model presented in Section 4 has a few parameters that must be selected empirically on the development set. These include the vocabulary size, which is dependent on the n words with the highest $tf * idf$ scores in each document, and the number of topics for the LDA-based re-ranker. We obtained best performance with n set to 100 (no cutoff was applied in cases where the vocabulary was less than 100). We trained an LDA model with 20 topics on our document collection using David Blei's implementation.[3] We used this model to re-rank the output of our annotation model according to the three most likely topics in each document.

Baselines We compared our model against three baselines. The first baseline is based on $tf * idf$ (Salton and McGill, 1983). We rank the document's content words (i.e., nouns, verbs, and adjectives) according to their $tf * idf$ weight and select the top k to be the final annotations. Our second baseline simply annotates the image with the document's title. Again we only use content words (the average title length in the training set was 4.0 words). Our third baseline is Lavrenko et al.'s (2003) continuous relevance model. It is trained solely on image-caption

---

[3]Available from http://www.cs.princeton.edu/~blei/lda-c/index.html.

Model	Top 10			Top 15			Top 20		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	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

Table 1: Automatic image annotation results on the BBC News database.

pairs, uses a vocabulary of 2,167 words and the same features as our extended model.

Evaluation   Our evaluation follows the experimental methodology proposed in Duygulu et al. (2002). We are given an un-annotated image I and are asked to automatically produce suitable annotations for I. Given a set of image regions V_I, we use equation (1) to derive the conditional distribution $P(w|V_I)$. We consider the k-best words as the annotations for I. We present results using the top 10, 15, and 20 annotation words. We assess our model's performance using precision/recall and F1. In our task, precision is the percentage of correctly annotated words over all annotations that the system suggested. Recall, is the percentage of correctly annotated words over the number of genuine annotations in the test data. F1 is the harmonic mean of precision and recall. These measures are averaged over the set of test words.

6   Results

Our experiments were driven by three questions: (1) Is it possible to create an annotation model from noisy data that has not been explicitly hand labeled for this task? (2) What is the contribution of the auxiliary document? As mentioned earlier, considering the document increases the model's computational complexity, which can be justified as long as we demonstrate a substantial increase in performance. (3) What is the contribution of the image? Here, we are trying to assess if the image features matter. For instance, we could simply generate annotation words by processing the document alone.

Our results are summarized in Table 1. We compare the annotation performance of the model proposed in this paper (ExtModel) with Lavrenko et al.'s (2003) original continuous relevance model (Lavrenko03) and two other simpler models which

do not take the image into account (*tf ∗ idf* and DocTitle). First, note that the original relevance model performs best when the annotation output is restricted to 10 words with an F1 of 11.81% (recall is 9.05 and precision 16.01). F1 is marginally worse with 15 output words and decreases by 2% with 20. This model does not take any document-based information into account, it is trained solely on image-caption pairs. On the Corel test set the same model obtains a precision of 19.0% and a recall of 16.0% with a vocabulary of 260 words. Although these results are not strictly comparable with ours due to the different nature of the training data (in addition, we output 10 annotation words, whereas Lavrenko et al. (2003) output 5), they give some indication of the decrease in performance incurred when using a more challenging dataset. Unlike Corel, our images have greater variety, non-overlapping content and employ a larger vocabulary (2,167 vs. 260 words).

When the document is taken into account (see ExtModel in Table 1), F1 improves by 8.01% (recall is 14.72% and precision 27.95%). Increasing the size of the output annotations to 15 or 20 yields better recall, at the expense of precision. Eliminating the LDA reranker from the extended model decreases F1 by 0.62%. Incidentally, LDA can be also used to rerank the output of Lavrenko et al.'s (2003) model. LDA also increases the performance of this model by 0.41%.

Finally, considering the document alone, without the image yields inferior performance. This is true for the *tf ∗ idf* model and the model based on the document titles.[4] Interestingly, the latter yields precision similar to Lavrenko et al. (2003). This is probably due to the fact that the document's title is in a sense similar to a caption. It often contains words that describe the document's gist and expectedly

---

[4]Reranking the output of these models with LDA slightly decreases performance (approximately by 0.2%).

tf ∗ idf	**breastfeed**, medical, intelligent, health, child	culturalism, faith, Muslim, separateness, ethnic	**ceasefire**, Lebanese, disarm, cabinet, Haaretz
DocTitle	Breast milk does not boost IQ	UK must tackle ethnic tensions	Mid-East hope as ceasefire begins
Lavrenko03	woman, **baby**, hospital, new, day, lead, good, England, look, family	bomb, city, want, day, fight, child, attack, face, help, government	war, carry, city, security, **Israeli**, attack, minister, force, government, leader
ExtModel	**breastfeed**, intelligent, **baby**, mother, **tend**, child, study, woman, sibling, advantage	aim, Kelly, faith, culturalism, community, Ms, tension, commission, multi, tackle, school	**Lebanon**, **Israeli**, Lebanese, aeroplane, **troop**, Hezbollah, Israel, force, **ceasefire**, grey
Caption	Breastfed babies tend to be brighter	Segregation problems were blamed for 2001's Bradford riots	Thousands of Israeli troops are in Lebanon as the ceasefire begins

Figure 3: Examples of annotations generated by our model (ExtModel), the continuous relevance model (Lavrenko03), and the two baselines based on *tf ∗ idf* and the document title (DocTitle). Words in bold face indicate exact matches, underlined words are semantically compatible. The original captions are in the last row.

some of these words will be also appropriate for the image. In fact, in our dataset, the title words are a subset of those found in the captions.

Examples of the annotations generated by our model are shown in Figure 3. We also include the annotations produced by Lavrenko et. al's (2003) model and the two baselines. As we can see our model annotates the image with words that are not always included in the caption. Some of these are synonyms of the caption words (e.g., *child* and *intelligent* in left image of Figure 3), whereas others express additional information (e.g., *mother*, *woman*). Also note that complex scene images remain challenging (see the center image in Figure 3). Such images are better analyzed at a higher resolution and probably require more training examples.

7 Conclusions and Future Work

In this paper, we describe a new approach for the collection of image annotation datasets. Specifically, we leverage the vast resource of images available on the Internet while exploiting the fact that many of them are labeled with captions. Our experiments show that it is possible to learn an image annotation model from caption-picture pairs even if these are not explicitly annotated in any way. We also show that the annotation model benefits substantially from

additional information, beyond the caption or image. In our case this information is provided by the news documents associated with the pictures. But more generally our results indicate that further linguistic knowledge is needed to improve performance on the image annotation task. For instance, resources like WordNet (Fellbaum, 1998) can be used to expand the annotations by exploiting information about is-a relationships.

The uses of the database discussed in this article are many and varied. An interesting future direction concerns the application of the proposed model in a semi-supervised setting where the annotation output is iteratively refined with some manual intervention. Another possibility would be to use the document to increase the annotation keywords by identifying synonyms or even sentences that are similar to the image caption. Also note that our analysis of the accompanying document was rather shallow, limited to part of speech tagging. It is reasonable to assume that results would improve with more sophisticated preprocessing (i.e., named entity recognition, parsing, word sense disambiguation). Finally, we also believe that the model proposed here can be usefully employed in an information retrieval setting, where the goal is to find the image most relevant for a given query or document.

References

K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. 2002. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.

D. Blei and M. Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference*, pages 127–134, Toronto, ON.

D. Blei, A. Ng, and M. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

R. Datta, J. Li, and J. Z. Wang. 2005. Content-based image retrieval – approaches and trends of the new age. In *Proceedings of the International Workshop on Multimedia Information Retrieval*, pages 253–262, Singapore.

P. Duygulu, K. Barnard, J. de Freitas, and D. 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*, pages 97–112, Copenhagen, Danemark.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

S. Feng, V. Lavrenko, and R. Manmatha. 2004. Multiple Bernoulli relevance models for image and video annotation. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 1002–1009, Washington, DC.

T. Hofmann. 1998. Learning and representing topic. A hierarchical mixture model for word occurrences in document databases. In *Proceedings of the Conference for Automated Learning and Discovery*, pages 408–415, Pittsburgh, PA.

J. Jeon and R. Manmatha. 2004. Using maximum entropy for automatic image annotation. In *Proceedings of the 3rd International Conference on Image and Video Retrieval*, pages 24–32, Dublin City, Ireland.

V. Lavrenko, R. Manmatha, and J. Jeon. 2003. A model for learning the semantics of pictures. In *Proceedings of the 16th Conference on Advances in Neural Information Processing Systems*, Vancouver, BC.

Y. Mori, H. Takahashi, and R. Oka. 1999. Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of the 1st International Workshop on Multimedia Intelligent Storage and Retrieval Management*, Orlando, FL.

J. M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference*, pages 275–281, New York, NY.

G. Salton and M.J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

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

J. Shi and J. Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

J. Tang and P. H. Lewis. 2007. A study of quality issues for image auto-annotation with the Corel data-set. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):384–389.

A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang. 2001. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10:117–130.

X. Wei and B. W. Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proeedings of the 29th Annual International ACM SIGIR Conference*, pages 178–185, Seattle, WA.