Research Article Impact Factor: 4.226 ISSN: 2319-507X
CP Ronald Reagan, IJPRET, 2014; Volume 3 (4): 365-379 IJPRET

INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

A MODULAR APPROACH FOR AUTOMATIC IMAGE ANNOTATION AND TAG SELECTION FOR NEWS IMAGE

CP RONALD REAGAN, P. VELAVAN

Assistant Professor, Department of Engineering, Apollo Engineering College, Chennai.

Accepted Date: 22/11/2014; Published Date: 01/12/2014

Abstract: Automatic tag selection for image is an important for many image related applications. The proposed approach explore the feasibility of automatic tag selection for News image in a knowledge-lean way and it consists of two components, namely extracting image content and rendering it in natural language. By using MixLDA represent the visual and textual modalities jointly as a probabilistic distribution over a set of topics. Automatic image annotation model take these distributions into account and finding the most likely keywords for an image and its associated documents. For caption generation, extractive and abstractive caption generation models are used to render the extracted image contents in natural language without rely on rich knowledge resources or sentence templates. Experimental results shown that the generated keywords and captions are relevant to the specific content of an image and its associated articles.

Keywords: Image Annotation, Caption Generation, Topic Models, Stemming, Stop Words, Summarization

Corresponding Author: MR. C. P. RONALD REAGAN

Access Online On:

www.ijpret.com

How to Cite This Article:

CP Ronald Reagan, IJPRET, 2014; Volume 3 (4): 365-379

PAPER-QR CODE

INTRODUCTION

The rapid growth of image collections on the internet indicates the increasing demand for searching and browsing. A large number of image search engines mainly employ the surrounding texts around the images and the image names to index the images. However, this limits the capability of the search engines in retrieving the semantically related images using a given query.

Although image understanding is a popular topic within computer vision, relatively little work has focused on caption generation for image. Caption generation for image similar to image description generation. The standard image description generation consists of two stage content selection and surface realization. Content selection analyzes the image content and it use dictionaries that specify a mapping between words and image regions or features [2] [3] [12]. Surface realization render the image content into natural language and it uses human written templates[4]. These approaches are difficult and time consuming task. So the proposed system adopts content selection and surface realization without requiring expensive manual annotation. It has two task automatic image annotation and caption generation.

Automatic image annotation [5] is the process of assigning keywords to digital images depending on the content information. Automatically assigning keywords to images is of great interest as it allows one to index, retrieve, and understand large collections of image data. Many techniques have been proposed for image annotation in the last decade that gives reasonable performance on standard datasets. The main idea of Automatic image annotation (AIA) techniques is to identify the concepts from large number of image samples, and use the concept models to label new images. Once images are annotated with semantic labels, they can be retrieved by keywords, which is similar to text document retrieval. The key characteristic of automatic image annotation is that it offers keyword searching based on image content and it employs the advantages of both the text based annotation and CBIR [6].

Automatic image annotation usually consist of two main steps i.e.; feature extraction and annotation. An image is an unstructured array of pixels represented using some features like color, texture, shape etc. For annotation, there is a need for appropriate feature selection from these pixels. Based on the different features selected and different algorithm techniques used, the performance of the semantic learning techniques vary. Automatic annotation detects and labels semantic content of images with a set of keywords automatically [1].

For generating caption for image, extractive and abstractive caption generation models [1] are used. Extractive model use the keywords generated by automatic caption generation model

and identify the sentences in the documents that share these keywords. Abstractive model operates over the description keyword and document phrases. The combination of this model gives caption for the image.

I. Proposed System

In the proposed method of automatic tag selection or caption generation for image consists of automatic image annotation and caption generation. In automatic image annotation SIFT algorithm is used for feature extraction and MixLDA is used for generating keywords. After generating keywords extractive and abstractive generation models are used for caption generation. Extractive caption generation model extract the sentence by maximally similar keywords generated in the automatic image annotation from the documents. With the help of phrase dependency predicted by the Standford parser, abstractive caption generation generated phrases for the keywords. The combination of these two models generate caption for the image.

II. Literature survey

In computer vision, there is an increasing demand for describing images or video frames more linguistically, e.g., with description sentences rather than isolated keywords lists. More emphasis has been placed on extracting content from images or video key frames, e.g, by recognizing objects or even interpreting human actions. Generally, the content is first extracted and represented as a keyword list or concept entries in a dictionary, and next a natural language generation module arranges this content into human-readable sentences, often using sentence-templates or a functional grammar-based surface realizer.

For example, H´ede et al. [2] attempt to generate descriptions for images of objects shot in a uniform background. Their work highlights the importance of a content representation with semantic relations in image database related applications, e.g., a phrase “an orange ball” explicitly indicates the modified relationship between orange and ball while isolated words “orange, ball” might describe two objects, an orange and a ball, rather than one (an orange ball). Their system relies on a manually created dictionary of objects, each entry is indexed by an image signature (e.g., raw image features, such as color and texture, and two keywords, the object’s name and category). The model first segments images into regions, retrieves corresponding signatures from the database by comparing the region features with entries in the dictionary, and produces a description sentence using the retrieved signature keywords and selected sentence templates.

Duygulu et al.[7]segment images into regions and cluster the latter using K-means into 500 classes which they call blobs. They assume that blobs correspond to objects in images. Next, they learn the correspondence between the blobs and words, using the IBM machine translation model. Using the Expectation Maximization(EM) algorithm, they learn the translation probabilities from blobs to words, which makes it easier to compute the probabilities of keywords given a test image. In this work, the size of the training data was 4500 images, with less than 5 annotated words and 10 blobs per image. This dataset is substantially smaller and thus not enough to reliably capture the correspondence between regions and keywords in IBM model and the quality of the training data which should be ideally strongly labeled, since the modeling procedure operates on the region level.

Feng et al.[8]the word generative distribution is estimated using a multiple Bernoulli model instead of a multinomial one, which places more emphasis on the presence of a word rather than its prominence. For example, given an image of two persons, the multiple Bernoulli model will focus on whether the word “people” has appeared in the annotation, while the multinomial model will care about how many times “people” has appeared in the annotation. In practice, the presence of a word is more useful than its prominence since it can make the estimated probability distribution concentrate on the concepts or objects and partially avoid the negative effect of weak labeling.

Monay and Gatica-Perez [9] propose several image annotation models based on the PLSA model. They first render images into word-like visual terms by clustering algorithm, and derive three annotation models, PLSA-Words, PLSA-Mixed and PLSA-Features, from the original PLSA structure. PLSA-Words uses annotation words alone to obtain the topic proportions as well as the textual part of the topic representation, and then folds the image features into the model in order to obtain the visual aspect. PLSA-Features relies solely on the images to determine the topic proportions and further folds-in the annotation keywords into the model. PLSA-Mixed, uses both images and annotation words to infer the topic space, and follows normal PLSA procedures.

Correspondence LDA (CorrLDA), proposed by Blei and Jordan [10]**,** has been successfully employed for modeling annotated images in the Corel domain. In CorrLDA, images are segmented into regions which are modeled by multiple Gaussian distributions and further drive the construction of the latent space to which the textual modality is linked. There is a strong assumption here: annotation keywords strictly correspond to topics which have been used to generate the regions of the current image. This assumption is based on the observation in Corel

that only salient objects in the images are labeled and the prerequisite that image regions are accurately segmented.

SumBasic, proposed by Nenkova and Vanderwende [11], is a simple but effective extractive model for multi-document summarization. The algorithm is based on the observation that content words used frequently in a document set are likely to appear in human-written summaries. Therefore a sentence is scored according to how many frequently appeared content words it has, and the summary is built up by these highest scored sentences, together with measures to penalize repeated content. This algorithm utilizes the unigram frequencies over the document set, but ignores the fact that a word appearing many times in only one document could have totally different status in a summary compared to a word appearing the same number of times evenly across the whole document set.

From the above survey the existing systems are rely on the background knowledge bases containing correspondences that help interpret visual information into textual information and fine grained sentences templates. The development of such knowledge bases usually requires significant human involvement. The proposed work is a knowledge lean way and utilize resources where images and their captions co-occur naturally.

III. SYSTEM DESIGN

The system design for automatic tag selection(caption generation) for News image using Latent Dirichlet Allocation(LDA) algorithm consists of automatic caption generation task and image and information retrieval.

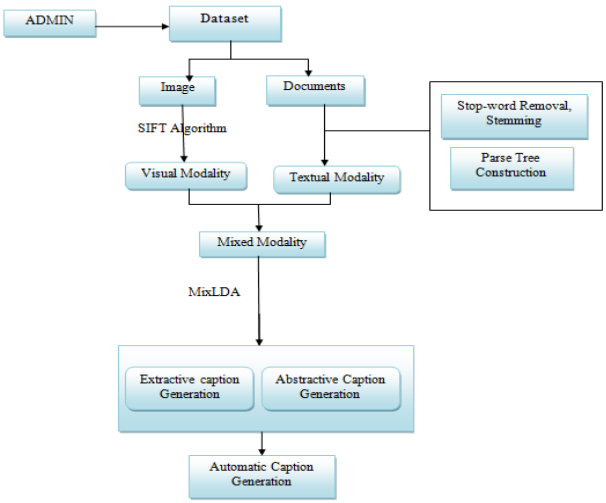

Fig 1. Automatic caption generation

Dataset

The dataset consists of 1000 BBC News articles and its associated document. The dataset covers a wide range of topics including national and international politics, technology, sports, education, and so on. News articles normally use color images which are around 200 pixels wide and 150 pixels high. The average sentence length is 20.5 words and the average document length 421.5 words. The document vocabulary is 26,795 words.

Parents warned over computer use

A third of children in the UK use blogs and social network websites but two thirds of parents do not even know what they are, a survey suggests.

The children's charity NCH said there was "an alarming gap" in technological knowledge between generations.

Even when parents had put controls on what youngsters could access, almost half the 1,003 children aged 11 to 16 surveyed said they could disable them.

The NCH said families had to learn more about technology to protect children.

'Worldly wisdom'

A tenth of the 11-year-olds who took part in the survey said their parents did not know about the people with whom they communicated online.

And 13% revealed they were never supervised while using computers at home.

John Carr, the NCH's new technology adviser, said: "Children are pretty clued up when it comes to technology but they often lack the worldly wisdom to steer them away from its potential hazards.

"That's where parents come in. But our research shows they need to increase their knowledge if they want to protect their children."

The survey also found that 69% of parents thought they knew less than their children about mobile phones.

The NCH and the supermarket chain Tesco are launching a parents' technology guide, called IT? Got IT! Good!, which is being distributed at stores.

The chief executive of Tesco Telecoms, Andy Dewhurst, said: "Young people are often in the driving seat when it comes to new technology, and mobile phones and internet use can be of huge benefit for families.

Fig 2.Example for BBC News Dataset consists of image and its associated documents

Visual modality

Visual modality means content of an image or visual terms. For getting the content of an image, Scale invariant Feature Transformation algorithm is used. Scale-invariant feature transform or SIFT is an algorithm in computer vision to detect and describe local features in images.

Textual modality

Textual modality means individual text words. Stop word, stemming process, Standford parser are used for getting textual information. The combination of visual and textual modality is called mixed modality.

MixLDA

LDA is an algorithm for generating keywords for the particular image and its associated documents. In the proposed method, visual terms and textual terms are treated as same in the document so this LDA is represented as MixLDA. MixLDA represents documents as mixtures of topics that spit out words with certain probabilities. It assumes that documents are produced in the following fashion: when writing each document,

1. Decide on the number of words N the document will have (say, according to a Poisson distribution).

2. Choose a topic mixture for the document (according to a Dirichlet distribution over a fixed set of K topics). For example, assuming that we have the two food and cute animal topics above, you might choose the document to consist of 1/3 food and 2/3 cute animals.

3. Generate each word w_i in the document by:

- First picking a topic (according to the multinomial distribution that you sampled above; for example, you might pick the food topic with 1/3 probability and the cute animals topic with 2/3 probability).
- Using the topic to generate the word itself (according to the topic's multinomial distribution). For example, if we selected the food topic, we might generate the word "broccoli" with 30% probability, "bananas" with 15% probability, and so on.

Assuming this generative model for a collection of documents, LDA then tries to backtrack from the documents to find a set of topics that are likely to have generated the collection.

Caption generation

In caption generation, caption is generated automatically for the particular image. For this, there are two kinds of caption generation. They are abstractive and extractive caption generation. Extractive caption generation focus on sentence extraction. It extracts the sentence by maximally similar keywords from the documents.

With the help of phrase dependency predicted by the Standford parser, abstractive caption generation generated phrases for the keywords. After generating caption automatically, the captions and the particular image are stored in the database.

Image and information retrieval

In Image and Information retrieval system, the generated keywords, captions and associated image are stored in the database. The user can enter the keyword for a particular image into the search engine and it can retrieve the correct image from the database. The image can be displayed with their caption.

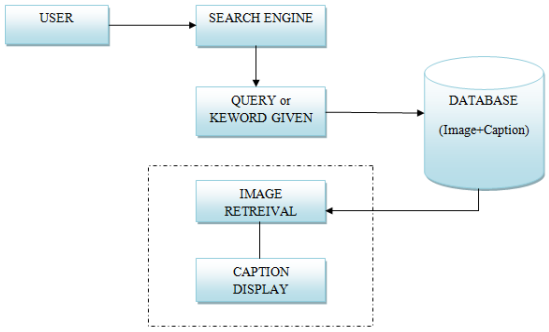

Fig 3.Image and information retrieval

IV. methodology

A. Image annotation

In image annotation module, load the dataset with images and their corresponding documents. For applying MixLDA, treat the image as documents. For this feature is detected by SIFT Detector. After detecting feature each image is abstracted by local patches. SIFT descriptor represent the patches as numerical vectors and then convert this vectors to visual terms. Now visual terms and textual terms have the same status.

For applying SIFT algorithm, first images are divided into 3 octaves and then blur the each octaves using Gaussian blur operator.

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y,) \qquad (1)$$

The symbols:

- L is a blurred image
- G is the Gaussian Blur operator
- I is an image

- x,y are the location coordinates
- σ is the amount of blur. Greater the value, greater the blur.
- The * is the convolution operation in x and y. It "applies" Gaussian blur G onto the image I.

After blurring an image, sift detector use difference of Gaussian approximation to find blob like feature points. For this, Two consecutive images in an octave are picked and one is subtracted from the other. Then the next consecutive pair is taken, and the process repeats. This is done for all octaves. The resulting images are an approximation of scale invariant laplacian of Gaussian (which is good for detecting key points). Difference of Gaussians is one such technique, locating scale-space extrema, D(*x*, *y*, σ) by computing the difference between two images, one with scale *k* times the other. D(*x*, *y*, σ) is then given by:

$$D(x, y, \sigma)=L(x, y, k\sigma) - L(x, y, \sigma) \quad (2)$$

Then we get the blob key points. After finding key points, take 16 x 16 window around the key point and broken down it into 4 x 4 window. Within each 4x4 window, gradient magnitudes and orientation are calculated. The Gaussian-smoothed image L(x,y,σ) at the key point's scale σ is taken so that all computations are performed in a scale-invariant manner. For an image sample L(x,y,σ) at scale , the gradient magnitude m(x,y), and orientation are precomputed.

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2} + \sqrt{L(x,y+1) - L(x,y-1))^2} \quad (3)$$

$$\theta(x,y) = \tan^{-1}\left(\frac{L(x,y-1)}{L(x+1,y)-L(x-1,y)}\right) \quad (4)$$

Then orientation are put into 8 bin histogram. Any gradient orientation in the range 0-44 degrees add to the first bin. 45-89 add to the next bin. And so on. The amount added to the bin depends on the magnitude of the gradient. Doing this for all 16 pixels, you would've "compiled" 16 totally random orientations into 8 predetermined bins. You do this for all sixteen 4×4 regions. So you end up with 4x4x8 = 128 numbers. These 128 numbers form the "feature vector". Typically, the SIFT descriptor is used to convert extracted image patches into visual words.

B. Parse Tree construction

In parse tree construction, stop word process is used to split the sentence into individual words and then apply stemming process. Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form-generally a written word form. For example, words such as running, runs, runnly, runned are stemmed to run. Then construct

parse tree by using Standford parser. Parse tree contains root, sibling, leaf node and it also find the dependency between the phrases. Parse tree is an ordered, rooted tree that represents the syntactic structure of a string according to some formal grammar. *John hit the ball*:

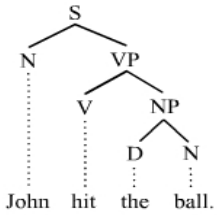

Fig 4. Example for phrase dependencies

The parse tree is the entire structure, starting from S and ending in each of the leaf nodes (*John*, *hit*, *the*, *ball*). The following abbreviations are used in the tree:

- S for sentence, the top-level structure in this example
- NP for noun phrase. The first (leftmost) NP, a single noun "John", serves as the subject of the sentence. The second one is the object of the sentence.
- VP for verb phrase, which serves as the predicate
- V for verb. In this case, it's a transitive verb *hit*.
- D for determiner, in this instance the definite article "the"

C. Keyword generation

After Parse tree construction, then generate keywords for the image and the document using Latent dirichlet allocation. LDA represents documents as mixtures of topics that spit out words with certain probabilities. Given a corpus consisting of D documents and each documents is modeled using a mixture over K topics. The words in the documents are generated

1. Choose a topic mixture for the document according to a Dirichlet distribution .

$$\text{Choose } \theta|\alpha \sim \text{Dir}(\alpha), \qquad (5)$$

$$\text{for } n \in 1,2,\ldots,N$$

where α is the parameter of the Dirichlet prior on the per-document topic distributions, $^{\theta}$ is the topic distribution for document and Dir ($^{(\alpha)}$is the dirichlet distribution.

2. Generate each word in the document by picking a topic and using the topic to generate the word itself using multinominal distribution

Choose topic $z_n|^{\theta \sim \text{Mult}(\theta),}$ (6)

Choose a word $w_n \mid z_n, \beta_{1:k} \sim$ Mult (βz_n) (7)

where z_n is the topic in the document, $^{\text{Mult}(\theta)}$ is the multinominal distribution, w_n is the words in the specified topics, $\beta_{1:k}$ is the distribution over words in the topic and(βz_n) is the multinominal distribution over words in the topic

3. Assuming this generative model for a collection of document and repeating this step for several number of times until it reached a steady state and it can be obtained by the following equation

$$P(d|\alpha,\beta) = \int P(\theta|\alpha)(\prod_{n=1}^{N} \sum_{z_k} P(z_k|\theta) P(w_n|z_k,\beta))d\theta \quad (8)$$

where d is the document in a corpus, α is the parameter of the Dirichlet prior on the per-document topic distributions, $^{\beta}$ is the parameter of the Dirichlet prior on the per-topic word distribution, $^{\theta}$ is the topic distribution for document , ^{z_k} is the topic proportions and w_n is the words in the specified topics

4. After generating keywords using LDA algorithm, then find the maximal probability keywords and is given by the following equation

$$W_I^* = \arg\max \prod P(w_t|I,D) \quad (9)$$

Where ^{w_t} is a set of textual words and I, D are represented jointly as the concatenation of textual and visual terms

5. Since I and D represents the mixed proportion of textual word and visual word and it can be represented as jointly by $d_{mix.}$ It can be given by the following equation

$$W_I^* = \arg\max \prod P(w_t|d_{mix}) \quad (10)$$

Now the maximal keywords are generated.

D. Caption Generation

In caption generation, caption is generated automatically for the particular image. For this, there are two kinds of caption generation. They are abstractive and extractive caption generation. Extractive caption generation focus on sentence extraction. It extracts the sentence by maximally similar keywords from the documents and it is given by the following equation

Overlap (11)

With the help of phrase dependency predicted by the Standford parser, abstractive caption generation generated phrases for the keywords and is given by the following equation

$$P(\rho_j \in C|\rho_j \in D) = \prod P(w_j \in$$ (12)

Where is the probability of adding phrases in the caption C generated by extractive caption generation model and in the document D.

After generating caption automatically, the captions and the particular image are stored in the database.

E. Image retrieval

The final process is to search the particular images from the search engine. It's only used by the user only. Here, the user can enter the keyword of a particular image into the search engine; it can retrieve the correct image from the database. The image can be displayed with their caption.

V. experimental results

In order to evaluate the performance of automatic tag selection for News image three measures are used precision, recall and F1 measure.

Table 1: Results of Automatic Caption Generation

Model	Precision	Recall	F1 measure
TxtLDA	7.30	16.90	10.20
ImgLDA	7.92	17.40	10.60
ContRel	14.70	27.90	19.80
MixLDA	15.30	32.10	21.58

Precision is also called positive predictive value and it is the fraction of retrieved words that are relevant to the search. Precision takes all retrieved words into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision at n or P@n.For example for a text search on a set of documents precision is the number of correct results divided by the number of all returned results. Precision is also used with recall, the percent of all relevant words that is returned by the search.

$$\text{Precision} = \frac{\text{number of relevant iter}}{\text{total number of items}} \qquad (13)$$

Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved. For example for text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned. In binary classification, recall is called sensitivity. So it can be looked at as the probability that a relevant document is retrieved by the query. It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by computing the precision

$$\text{recall} = \frac{\text{number of relevant items reti}}{\text{total number of relevant items in}} \qquad (14)$$

The F_1 score (also F-score or F1-measure) is a measure of a test's accuracy. It considers both the precision *p* and the recall *r* of the test to compute the score: *p* is the number of correct results divided by the number of all returned results and *r* is the number of correct results divided by the number of results that should have been returned. The F_1 score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst score at 0.The traditional F-measure or balanced F-score (**F_1 score**) is the harmonic mean of precision and recall and is given by

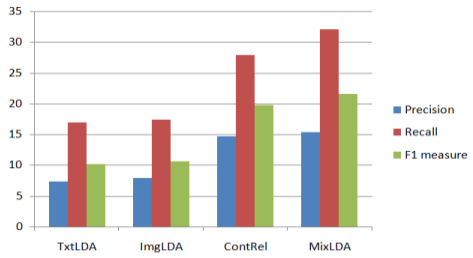

Fig 4. Performance graph for automatic caption generation

VI. CONCLUSION

The proposed system focused on the task of automatically generating captions for News image in aknowledge –lean way. The model consists of two components, namely extracting image content and rendering it in natural language. For extracting image content automatic image annotation model is used and for rendering it into natural language extractive and abstractive model is used. So the proposed system does not rely on manual annotation for image annotation and does not need manually created sentence templates or dictionaries to rendering it into natural language.

REFERENCES

1. Yansong Feng, Mirella Lapata "Automatic caption generation for News image" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, No. 4, April 2013.

2. P. He´de, P.A. Moe¨llic, J. Bourgeoys, M. Joint, and C. Thomas,"Automatic Generation of Natural Language Descriptions for Images," Proc. Recherche d'Information Assiste´e par Ordinateur, 2004.

3. B. Yao, X. Yang, L. Lin, M.W. Lee, and S. Chun Zhu, "I2T: Image Parsing to Text Description," Proc. IEEE, vol. 98, no. 8, pp. 1485-1508, 2009.

A. Farhadi, M. Hejrati, A. Sadeghi, P. Yong, C. Rashtchian, J.Hockenmaier, and D. Forsyth, "Every Picture Tells a Story: Generating Sentences from Images," Proc. 11th European Conf. Computer Vision, pp. 15-29, 2010.

A. Makadia, V. Pavlovic, and S. Kumar, "Baselines for Image Annotation," Int'l J. Computer Vision, vol. 90, no. 1, pp. 88-105, 2010.

4. A.W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1349-1380, Dec. 2000.

5. Duygulu, P., Barnard, K., de Freitas, J., and Forsyth, D. (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.

6. S. Feng, V. Lavrenko, and R. Manmatha, "Multiple Bernoulli Relevance Models for Image and Video Annotation," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1002-1009, 2004.

7. F. Monay and D. Gatica-Perez, "Modeling Semantic Aspects for Cross-Media Image Indexing," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 10, pp. 1802-1817, Oct. 2007.

8. D. Blei and M. Jordan, "Modeling Annotated Data," Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 127-134, 2003

9. Nenkova, A. and Vanderwende, L. (2005). The impact of frequency on summarization. Technical report, Microsoft Research.

10. V. Lavrenko, R. Manmatha, and J. Jeon, "A Model for Learning the Semantics of Pictures," Proc. 16th Conf. Advances in Neural Information Processing Systems, 2003.