Probabilistic Models for Topic Learning from Images and Captions in Online Biomedical Literatures

Xin Chen Caimei Lu Yuan An Palakorn Achananuparp
College of Information Science & Technology, Drexel University,
Philadelphia, PA, 19104
bruce.chen@drexel.edu, cl389@drexel.edu, yuan.an@ischool.drexel.edu, pa442@drexel.edu

ABSTRACT
Biomedical images and captions are one of the major sources of information in online biomedical publications. They often contain the most important results to be reported, and provide rich information about the main themes in published papers. In the data mining and information retrieval community, there are a lot of research works on using text mining and language modeling algorithms to extract knowledge from the text content of online biomedical publications; however, the problem of knowledge extraction from biomedical images and captions has not been fully studied yet. In this paper, a hierarchical probabilistic topic model with background distribution (HPB) is introduced to uncover the latent semantic topics from the co-occurrence patterns of caption words, visual words and biomedical concepts. With downloaded biomedical figures, restricted captions are extracted with regard to each individual image panel. During the indexing stage, the 'bag-of-words' representation of caption words is supplemented by an ontology-based concept indexing to alleviate the synonym and polysemy problems. As the visual counterpart of text words, the visual words are extracted and indexed from corresponding image panels. The model is estimated via collapsed Gibbs sampling algorithm. We compare the performance of our model with the extension of the Correspondence LDA (Corr-LDA) model under the same biomedical image annotation scenario using cross-validation. Experimental results demonstrate that our model is able to accurately extract latent patterns from complicated biomedical image-caption pairs and facilitate knowledge organization and understanding in online biomedical literatures.

Categories and Subject Descriptors
H.2.8 [**DATABASE MANAGEMENT**]: Database applications – Data mining; Image databases; I.2.6 **[ARTIFICIAL INTELLIGENCE]**: Learning

General Terms
Algorithms, Experimentation, Theory.

Keywords
Probabilistic models, topic learning, bioinformatics, Gibbs sampling, visual words, automatic image annotation.

CIKM'09, November 2–6, 2009, Hong Kong, China.

1. INTRODUCTION
Scientific research activities in biomedical and life science produce hundreds of thousands of digital publications each year. Although there are several public available digital databases such as PubMed Central, which provide users immediate access to full-text biomedical and life science journal articles, users are still facing a difficult task of organizing the massive information from the digital repositories. In particular, it is extremely difficult for users to handle the highly complicated process of mapping the visual content in biomedical images to various domain-specific terms and concepts in corresponding captions.

Biomedical images and captions are one of the major information sources in online biomedical literatures; they contain the most important results to be reported and provide rich information about the main themes in the published papers. Compared to free-form image captions (such as that from social network data source, like *Flickr.com* and *Facebook.com*), which are characterized by user-sensitive descriptions, the image captions in biomedical literatures have relatively standard representation with restricted terms used and always highly conform to the image content. In extracting biomedical concepts from captions, polysemies and synonyms are the major barrier. Biomedical ontologies (such as UMLS) provide the ability to overcome the polysemy and synonym problems. Therefore, if we can uncover the latent themes from the co-occurrence patterns of image content, caption words and biomedical concepts, we will be able to help biologists to find, understand and organize complicate knowledge from biomedical figures and satisfy their information needs.

In order to achieve that aim, the first issue is to bridge over the 'semantic gap' between image features and the user [2], which is to identify a set of image features that well preserve the semantic consistency of image content. Recently, the 'bag-of-visual-words' [6] approach exhibits very good performance in image categorization and semantic image retrieval across several well-known databases such as the *LabelMe*, the *TRECVID* and the *Visual Object Classes (VOC)* datasets [4, 8, 10, 16]. The underlying assumption of this approach is that, the patterns of different image categories can be represented by different distributions of microstructures (key-points). As an image document can be constantly represented as an unordered collection of key-points which carry rich local information, it can to some extent be regarded as a 'bag of visual words'. In practice, image patches containing key-points are quantified based on affine invariant local descriptors [9, 11, 13]. Sivic et al. further proposed the idea of assigning all the patch descriptors into clusters to build a

'vocabulary' of 'visual words' for a specific image set [6]. As a visual counterpart of the 'bag-of-word' model, the 'bag-of-visual-words' approach usually represents each image as a vector of visual words based on the visual term frequency [4, 6].

After representing image content as 'bag-of-visual words', the second issue is to uncover latent semantic themes from the co-occurrence patterns of image content (i.e. the extracted 'bag-of-visual words'), text captions and ontology-based concepts. In the data mining and information retrieval community, there are many research works on using probabilistic models to learn latent topics from text content (such as the abstract) in online publications. Several effective probabilistic models such as the Naïve Bayesian model, the Probabilistic Latent Semantic Indexing (pLSI) model [1] and the Latent Dirichlet Allocation (LDA) model [19] are proposed. Particularly, the LDA model has been very popular with the text mining community due to its solid theoretical foundation and promising performance. Despite the success of these models in text mining, however, the problem of topic learning from both images and captions has not been fully studied yet. Although there are some approaches toward modeling latent topics from visual words, such as directly using LDA [17] and using Spatial Latent Dirichlet Allocation [18]. However, to the best of our knowledge, there has not been any study combining visual words, text captions and ontology-based concepts in one single probabilistic model.

The Correspondence LDA (CorrLDA) model [7], initially proposed by Blei et al. for automatic image annotation, provides a natural way to learn the correlation between text words and other entities. In this model, topic generated from text words are used to generate other entities (such as image features). By extending the entities in the CorrLDA model to visual words and ontology-based biomedical concepts, it's not difficult to establish a probabilistic model that uncovers latent themes from the co-occurrence patterns of caption words, visual words and biomedical concepts.

Although the CorrLDA model is able to learn latent topics from the image-caption pairs, however, as indicated in our study, the discovered topics can be overwhelmed by several background words that frequently appear in the database. With this consideration, a hierarchical probabilistic topic model with background distribution is presented in this paper. With downloaded biomedical figures, restricted captions are extracted with regard to each individual image panel. During the indexing stage, the 'bag-of-words' representation of caption words is supplemented by an ontology-based concept indexing to alleviate the synonym and polysemy problems. As the visual counterpart of text words, the visual words are extracted and indexed from corresponding image panels. The model is estimated via collapsed Gibbs sampling algorithm, while the parameter selection is achieved by studying the likelihood and perplexity. We compare the performance of our model with the extension of the Correspondence LDA (Corr-LDA) model under the same biomedical image annotation scenario using cross-validation. Experimental results demonstrate that our model is able to accurately extract latent patterns from highly complicated biomedical image-caption pairs, facilitate knowledge organization and understanding in online biomedical literatures.

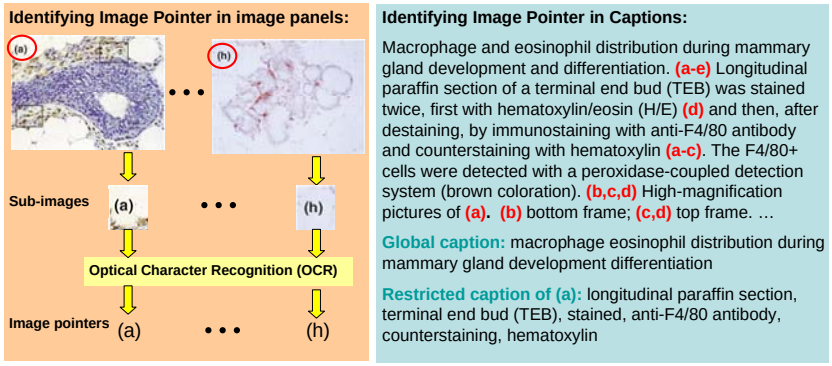
The remainder of this paper is organized as follows. In Section 2, we describe the procedure of preprocessing and indexing of biomedical figures. In Section 3, we present the extension of CorrLDA model and our hierarchical probabilistic topic model with background distribution. Section 4 provides the collapsed Gibbs sampling algorithms for inference and learning the proposed probabilistic models. Section 5 reports the experimental results of the proposed method and compares our approach to the extension of CorrLDA model. We conclude the paper in Section 6.

2. PREPROCESSING AND INDEXING OF BIOMEDICAL FIGURES

2.1 Figure Preprocessing

In our research, we deal with biomedical figures downloaded from the PubMed Central web pages. Generally, a biomedical figure involves two parts, that is, a single image composed with one or multiple image panels (sub-images) and the corresponding captions. Therefore, the preprocessing section of biomedical figures has two parts, the image processing part and the caption processing part.

Within the downloaded biomedical figures, images are segmented into several individual image panels. It should be pointed out that there are image panels which contain flow charts or diagrams. These image panels do not carry substantial visual content. Therefore, they are filtered out using basic region segmentation method.



(a) Image preprocessing **(b) caption preprocessing**

Fig. 1 Biomedical figures preprocessing

In captions texts, there are some parenthesized expressions refer to specific image panels. Most of them are simply composed of single letter such as (A), (b) or letters connected by conjunction, such as (a and b), (b,c) and (a-c). We refer to these parenthesized expressions as image pointers (as marked by red color in Fig. 1b). We develop a set of rules to extract these regular image pointers in captions, which is similar to the *HANDCODE2* method in [5].

Image pointers are commonly placed in some important positions (such as upper left and lower left corner) of image panels. Therefore, we apply the Asprise OCR Java SDK toolkit[1] for optical character recognition (OCR) in sub-images of image corners (Fig. 1a). The OCR toolkit achieved a moderate precision in our image pointer extraction, which is sufficient for our research. We check the image pointer extraction results and make necessary manual corrections.

In a figure with multiple image panels, instead of replicating the entire caption to each image panel, we develop a restricted caption scanner to identify *restricted captions* (Fig. 1b) with regard to the image pointer of each image panel. The association of texts and image pointers are determined according to different cases, such as image pointers locate at the beginning of a sentence, preceded by preposition and noun phrases, followed by a clauses, etc. Generally, the undergoing image pointer(s) for captions are disabled when the scanner meets another image pointer or reaches the end of a clause or a sentence. All the texts that don't have any assigned image pointers are regarded as *global captions* (Fig. 1b).

The image panel and captions associated with the same image pointer are named as an *image-caption pair*. In an image-caption pair, the final caption words are generated via a linear combination of restricted captions and global captions, which avoids the over-representation problem and preserves the uniqueness of each individual image panel. Each image-caption pair is assigned a unique ID like 'bcr1011-1_a', in which 'bcr1011' is the PubMed Central article ID, '1' is the number of figure in the article, while 'a' is the name of image pointer of a given image panel.

2.2 Image-Caption Pairs Indexing

During the indexing stage, we choose to represent the image content in each image-caption pair as a 'bag-of-visual-word'. Firstly, we adopt the Difference-of-Gaussian (DoG) salient point detector [13] to detect salient points from images. The detection is achieved by locating scale-space extreme points in the difference-of-Gaussian images. The main orientations of salient points are determined by image gradient. Image patches containing the salient points are then rotated to a canonical orientation and divided into 4×4 cells. In each cell, the gradient magnitudes at 8 different orientations are calculated. Consequently, each salient point is described by a 128-dimensional SIFT descriptor. Compared to other local descriptors, the SIFT descriptor is more robust and invariable to rotation and scale/luminance changes [11]. The SIFT descriptors extracted from training images are clustered into 1000 clusters using k-mean clustering to establish a codebook of 'visual words', with each cluster center as a 'visual word'. As shown in Fig. 2, the image indexing is achieved by computing the term frequency and building index of visual words for each image panel.

The indexing of captions results in two parts, the term index and the concept index (Fig. 2). The term index is simply obtained by calculating the term frequency of caption words after lemmatizing and stop-word removal. In our approach, the Van Rijsbergen's stop-word lists [14] and the UMLS biomedical stop-word list [15] are used to remove non-content-bearing terms.

The concept index is achieved by calculating the term frequency of concepts according to the results of concept extraction. In biomedical ontology, a concept carries a unique meaning and represents a set of synonymous terms. For example, *C0006149* is a concept about the benign or malignant neoplasm of the breast parenchyma in Unified Medical Language System (UMLS) [15]. It represents a set of synonyms including Breast Neoplasm, Breast Tumor, tumor of the Breast and Neoplasm of the Breast. Compared to individual words and multiple word phrases, a concept is more meaningful, therefore, used as indexing terms in large-scale biomedical literatures. In our approach, we adopt MaxMatcher [12], a dictionary-based biological concept extraction tool, to extract UMLS concept from captions.

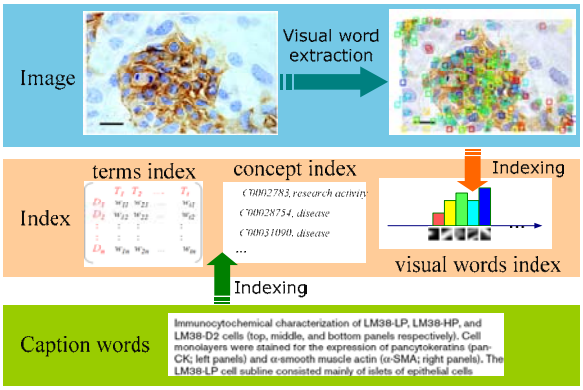


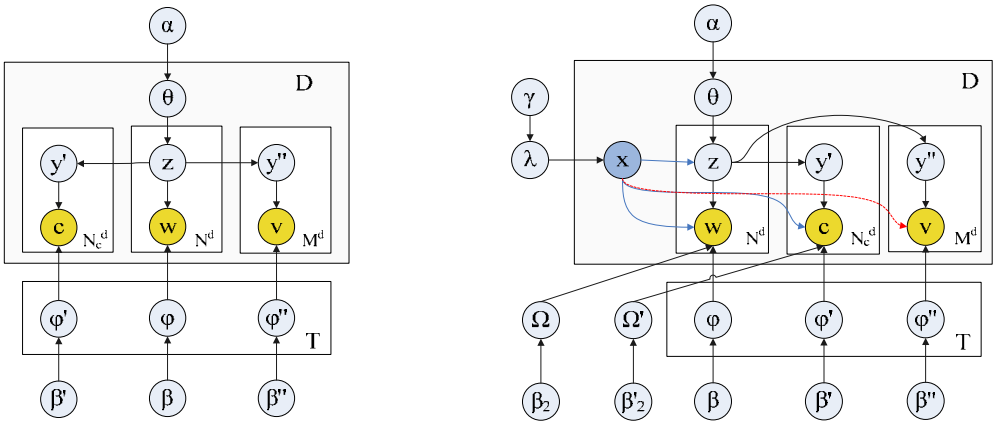
Fig. 2 The workflow for image-caption pair indexing

3. PROBABILISTIC MODELS FOR TOPIC LEARNING

In this paper, we mainly focus on learning latent semantic topics from biomedical image and captions. The underlying philosophy is that, an image-caption pair may deal with multiple topics; and the co-occurrence patterns of caption words, visual words and biomedical concepts in this image-caption pair are related to some unseen latent semantic variables, which indicate the presence/absence of specific topics.

In this section, we will present two probabilistic models, one is the extended Correspondence LDA (CorrLDA) model and the other is our proposed hierarchical probabilistic topic model with background distribution (HPB). For clarity of the notations, we name each image-caption pair as a *document*. Some notations to be used in the two probabilistic models are list as follows: D is the number of documents, T is the anticipated number of latent topics, N_d is the total number of text words in document d, N_c^d denotes the total number of extracted biomedical concepts in document d, while M_d represents the total number of extracted visual words in document d.

[1] The toolkit is downloaded from the home page of LAB Asprise! (http://asprise.com/product/ocr/selector.php) on Dec. 2008



(a) Extension of correspondence LDA (CorrLDA) model **(b) proposed HPB model**

Fig. 3 The extension of CorrLDA model and the hierarchical probabilistic model with background distribution (HPB), Yellow cycles represent the observation of words, concepts and visual words. The red dash line in (b) denotes a variation of HPB model.

3.1 The Extension of Correspondence LDA

CorrLDA model [7] provides a natural way to learn latent topics from text words and other entities. Therefore, our topic learning problem can be addressed by extending the entities in the CorrLDA model to visual words and ontology-based biomedical concepts. The differences between our extension and the original CorrLDA model are twofold, firstly, we combine visual words, text captions and ontology-based concepts in one single model; secondly, the original model only takes use of global image features such as color and texture, while our extension deals with visual words, which is robust than global image features and have similar statistical properties with text words (which are assumed to fit multinomial distributions).

The generative process for the extend CorrLDA model is:

1. For the d^{th} (d=1…D) documents, sample $\theta_d \sim Dir(\alpha)$
2. For the t^{th} (t=1…T) topic, sample $\varphi_t \sim Dir(\beta)$, $\varphi'_t \sim Dir(\beta')$ and $\varphi''_t \sim Dir(\beta'')$.
3. For each of the N_d words w_i in document d:
   a) Sample a topic $z_i \sim Mult(\theta_d)$
   b) Sample $w_i \mid z_i \sim Mult(\varphi_{z_i})$
4. For each of the N_c^d concepts c_i in document d:
   a) Sample a topic $y'_i \sim Uniform(z_{w_1},...,z_{w_{N^d}})$
   b) Sample $c_i \mid y'_i \sim Mult(\varphi'_{y'_i})$
5. For each of the M_d visual words v_i in document d:
   a) Sample a topic $y''_i \sim Uniform(z_{w_1},...,z_{w_{N^d}})$
   b) Sample $v_i \mid y''_i \sim Mult(\varphi''_{y''_i})$

In the first step, a T-dimensional topic-prior vector $\boldsymbol{\theta_d}$ is sampled for each document d, with the t^{th} dimension of the vector represents the prior probability of the t^{th} topic in d. For each document d, the generative process of the N_d words is achieved by sampling topics from the document-topic multinomial distribution (with Dirichlet prior $\boldsymbol{\theta_d}$) and sampling words from the topic-word multinomial distribution (with Dirichlet prior $\boldsymbol{\varphi_t}$). The generative process of the N_c^d concepts and M_d visual words are similar with that of the N_d words; the only difference is that only the topics that associated with the N_d words in document d are used to generate concepts and visual words. Parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\beta'}$ and $\boldsymbol{\beta''}$ are hyperparameters for the Dirichlet priors. In our approach, we assume symmetric Dirichlet priors, with α, β, β' and β'' being scalar parameters.

3.2 Hierarchical Probabilistic Model with Background Distribution (HPB)

Although the CorrLDA model is able to learn latent topics from the image-caption pairs and establish direct correlation among words, visual words and concepts, however, after looking into the discovered topics from the data collection, we found several background words appear at the top ranked terms of most discovered topics due to their high frequency. For example, when we use image-caption pairs from online journal: 'Breast Cancer Research' as training data and learn topics using the CorrLDA model, we found 'breast', 'cancer', 'mammary' are among the top-tanked words of many topics. These words, which we named as 'background words', appear frequently in many topics and take the places of the topic-specific key words. It's necessary to discover these 'background words' from the dataset, otherwise, the topic learning would be less effective.

It should be note that during the caption indexing stage, we have removed the non-content-bearing stopwords according to the Van Rijsbergen's stopword lists [14] and the UMLS stopword list [15]. Obviously, the 'background words' do not belong to regular stopwords. As we have seen, these words carry some contextual information which is shared by most image captions in a biomedical journal. As such 'background words' turn to be different from one journal to another, it's better to discover them automatically rather than manually specifying them for each journal.

In [20], Newman et al. proposed the 'SwitchLDA' model, in which a switch variable is introduced to control the fraction of entities in topics. With similar consideration, we develop a hierarchical probabilistic model with background distribution (HPB model) to capture the background topic z_0. In this model, an additional Binomial distribution λ (with a Beta prior of γ_1 and γ_2)

was incorporated to control the switch variable x (Fig. 3b), which decides whether a term should be drawn from a background topic z_0 or a regular latent topic z_i.

It's not clear whether the background words and concepts (Fig. 4) are related to certain image content, as image content may always be dramatically different from one to another. Therefore, in our research, we test this issue by present a variation of the HPB model. The generative process is as following:

1. For the d^{th} (d=1…D) documents, sample $\theta_d \sim Dir(\alpha)$ and $\lambda_d \sim Beta(\gamma_1, \gamma_2)$
2. For the t^{th} (t=1…T) topic, sample $\varphi_t \sim Dir(\beta)$, $\varphi'_t \sim Dir(\beta')$ and $\varphi''_t \sim Dir(\beta'')$; for background topic, sample $\Omega \sim Dir(\beta_2)$ and $\Omega' \sim Dir(\beta_2')$.

Variation (for HPB2 model):

For background topic, sample $\Omega'' \sim Dir(\beta_2'')$

3. For each of the N_d words w_i in document d:
   a) Sample a switch $x_i \sim Bernoulli(\lambda_d)$
   b) If $x_i = 0$, sample $w_i \mid z_0 \sim Mult(\Omega)$
   c) If $x_i = 1$, sample a topic $z_i \sim Mult(\theta_d)$ and sample $w_i \mid z_i \sim Mult(\varphi_{z_i})$
4. For each of the N_c^d concepts c_i in document d:
   a) Sample a topic $y'_i \sim Uniform(z_{w_1}, ..., z_{w_{N^d}})$
   b) If $y'_i = z_0$, sample $c_i \mid y'_i \sim Mult(\Omega')$
   c) If $y'_i = z_i \ (i = 1...T)$, sample $c_i \mid y'_i \sim Mult(\varphi'_{y'_i})$
5. For each of the M_d visual words v_i in document d:
   a) Sample a topic $y''_i \sim Uniform(z_{w_1}, ..., z_{w_{N^d}})$
   b) If $y''_i = z_0$, repeat (a)
   c) If $y''_i = z_i \ (i = 1...T)$, sample $v_i \mid y''_i \sim Mult(\varphi''_{y''_i})$

Variation (for HPB2 model):

   a) Sample a topic $y''_i \sim Uniform(z_{w_1}, ..., z_{w_{N^d}})$
   b) If $y''_i = z_0$, sample $v_i \mid y''_i \sim Mult(\Omega'')$
   c) If $y''_i = z_i \ (i = 1...T)$, sample $v_i \mid y''_i \sim Mult(\varphi''_{y''_i})$

In the proposed model, λ is the Bernoulli parameter for switch variable x. In our experiment, we assume symmetric priors and set $\alpha = 0.1, \beta = \beta' = \beta'' = 0.01, \gamma_1 = \gamma_2 = 0.5$. For clarity, we call the variation of HPB model (in gray color) as HPB2 model. In the HPB model, visual words has nothing to do with the background topic, while in HPB2 model, the presence of background topic z_0 in the caption words of document d is used to generate visual words, which results in direct correlation between visual words and the background topic.

4. COLLAPSE GIBBS SAMPLING FOR PROPOSED MODELS

The model estimation is achieved via the Collapse Gibbs Sampling procedure [3], which iteratively estimates the posterior probability conditioned on current entity-topic assignment and adopts a Monte Carlo process to determine the assignment of entity-topic in the next iteration.

Some notations to be used in Collapse Gibbs Sampling are list as following: W accounts for the vocabulary size of indexed words in the testing dataset; N_W denotes the total number of indexed words while W', $N_{W'}$ and W'', $N_{W''}$ represent the vocabulary size and the total number of concepts and visual words, respectively.

Background Topic

Top words	Probability
Cells	0.262948
breast	0.170271
Figure	0.092214
cancer	0.092214
staining	0.043864
mammary	0.041698
epithelial	0.037985
stained	0.019882
Representative	0.019573
sections	0.017484
normal	0.016479
positive	0.015241
shown	0.012456
receptor	0.01145
negative	0.010831
panel	0.010599
Arrows	0.009207
growth	0.006576
Red	0.004642
indicated	0.003482

Top Concepts	Concept Names
C0678222	Breast Cancer
C0242821	Human Body
C0487602,	Staining
C0700320	microtome
C0336721	Arrow
C0014597	Epithelial Cell
C1317667	PT PANEL
C0597357	receptor
C0587921	Magnification device
C0205159	Positive
C0205160	Negative
C0027651	Tumors
C0014609	Epithelium
C0332583	Green
C1260957	Blue
C0334227	cancer cell
C0079603	Immunofluorescence
C0441800	Grade
C0380213	MMP14 gene product
C0057142	DAPI

Fig. 4 Illustration of top-ranked words and concepts in background topic of online journal 'Breast Cancer Research'

4.1 Sampling for the Extended CorrLDA Model

Given the generative process in Section 3.1, our objective is to compute the word-topic posterior probability, which is:

$$p(z_{wi} = j \mid w_i, \mathbf{w_{-i}}, \mathbf{z_{-wi}}) \propto p(w_i \mid z_{wi} = j, \mathbf{w_{-i}}, \mathbf{z_{-wi}}) \cdot p(z = j \mid \mathbf{w_{-i}}, \mathbf{z_{-wi}})$$

The above posterior is intractable, however, it can be approximated by integrating out (collapsing) all the latent variables φ_j and θ_d separately, which is:

$$p(w_i \mid z_{wi} = j, \mathbf{w_{-i}}, \mathbf{z_{-wi}}) = \int p(w_i \mid z = j, \varphi_j, \mathbf{w_{-i}}, \mathbf{z_{-wi}}) p(\varphi_j \mid \mathbf{w_{-i}}, \mathbf{z_{-wi}}) d\varphi_j$$

$$\propto E\left(p(\varphi_j \mid \mathbf{w_{-i}}, \mathbf{z_{-wi}}) \sim Dir(\beta + n_{-i,j}^{wi})\right) = \frac{\beta + n_{-i,j}^{wi}}{W\beta + n_{-i,j}^{\cdot}}$$

$$p(z = j \mid \mathbf{w_{-i}}, \mathbf{z_{-wi}}) = \int p(z = j \mid \theta_d) \cdot p(\theta_d \mid \mathbf{w_{-i}}, \mathbf{z_{-wi}}) d\theta_d \propto \frac{\alpha + n_{-i,j}^{d}}{T\alpha + n_{-i,\cdot}^{d}}$$

Therefore, posterior probability for current word w_i is:

$$p(z_{wi} = j \mid w_i, \mathbf{w_{-i}}, \mathbf{z_{-wi}}) \propto \frac{\beta + n_{-i,j}^{wi}}{W\beta + n_{-i,j}^{\cdot}} \cdot \frac{\alpha + n_{-i,j}^{d}}{T\alpha + n_{-i,\cdot}^{d}} \qquad (1)$$

In which $n_{-i,j}^{wi}$ ($-i$ denotes that current word w_i is removed) is the total number of times word w_i being assigned to topic j except for

current one, $n^{\cdot}_{-i,j}$ is the summation of $n^{wi}_{-i,j}$, and $n^{d}_{-i,j}$ is the total number of words in document *d* assigned to topic *j* except for current word.

Having obtained the word-topic posterior probability, the Monte Carlo process is then straightforward - it is similar to throwing dice (based on the posterior probability) to determine the topic assignment for current word w_i in the next iteration.

Based on sampled topic variables for each word w_i, the posterior probabilities for visual word-topic and concept-topic can be approximated in similar formations. For simplicity, we give their posterior probabilities in a uniform expression, which is:

$$p(\tilde{z}_i = j \mid \tilde{w}_i = v, \tilde{\mathbf{z}}_{\mathbf{-i}}, \tilde{\mathbf{w}}_{\mathbf{-i}}, \mathbf{z}, \tilde{\beta}) \propto \frac{n_j}{N_d} \cdot \frac{\tilde{\beta} + n^{\tilde{w}_i}_{-i,j}}{\tilde{W}\tilde{\beta} + n^{\cdot}_{-i,j}} \quad (2)$$

In which n_j is the total number of words in document *d* assigned to topic *j*; N_d is the total number of words in document *d*; $n^{\tilde{w}_i}_{-i,j}$ is the total number of entities (concepts /visual words) assigned to topic *j* except for current entity: $\tilde{w}_i$. For concepts, we have $\tilde{W} = W'$ and $\tilde{\beta} = \beta'$; while for visual words, $\tilde{W} = W''$, $\tilde{\beta} = \beta''$.

4.2 Sampling for Proposed HPB Model

Similar to the Gibbs sampling procedure in Section 4.1, we derive the sampling equation for proposed HPB model as follows, which allow for joint sampling of the topic variables and the switch variable *x* for each word w_i :

$$p(x_{w_i} = 0, z_{w_i} = 0 \mid w_i, \mathbf{w}_{\mathbf{-i}}, \mathbf{z}_{\mathbf{-wi}}, \mathbf{x}_{\mathbf{-wi}}) \propto \frac{N^0_{d,-i} + \gamma}{N_{d,-i} + 2\gamma} \cdot \frac{\beta_2 + n^{w_i}_{-i,0}}{W\beta_2 + n^{\cdot}_{-i,0}} \quad (3)$$

$$p(x_{w_i} = 1, z_{w_i} = j \mid w_i, \mathbf{w}_{\mathbf{-i}}, \mathbf{z}_{\mathbf{-wi}}, \mathbf{x}_{\mathbf{-wi}}) \propto \frac{N^1_{d,-i} + \gamma}{N_{d,-i} + 2\gamma} \cdot \frac{\beta + n^{w_i}_{-i,j}}{W\beta + n^{\cdot}_{-i,j}} \cdot \frac{\alpha + n^{d}_{-i,j}}{T\alpha + n^{d}_{-i,\cdot}} \quad (4)$$

In which $N^0_{d,-i}$ and $N^1_{d,-i}$ are the total number of words (except for current word w_i) assigned to background topic and regular latent topics in document *d*. In equation (3), $n^{wi}_{-i,0}$ denotes the number of times word w_i being assigned to background topic except for current one, while $n^{\cdot}_{-i,0}$ is the summation of $n^{wi}_{-i,0}$. In (4), $n^{wi}_{-i,j}$ is the total number of times word w_i being assigned to topic *j* except for current one, $n^{\cdot}_{-i,j}$ is the summation of $n^{wi}_{-i,j}$,and $n^{d}_{-i,j}$ is the total number of words in document *d* assigned to topic *j* except for current word.

The sampling equations or concept and visual words have two different cases. For the **HPB model**, we have:

$$p(x_i = 0, y'_i = 0 \mid c_i, \mathbf{c}_{\mathbf{-i}}, \mathbf{y'}_{\mathbf{-i}}, \mathbf{w}, \mathbf{z}) \propto \frac{N^0_d}{N_d} \cdot \frac{\beta_2' + n^{c_i}_{-i,0}}{W'\beta_2' + n^{\cdot}_{-i,0}} \quad (5)$$

$$p(x_i = 1, y'_i = j \mid c_i, \mathbf{c}_{\mathbf{-i}}, \mathbf{y'}_{\mathbf{-i}}, \mathbf{w}, \mathbf{z}) \propto \frac{N^1_d}{N_d} \cdot \frac{n_j}{N^1_d} \cdot \frac{\beta' + n^{c_i}_{-i,j}}{W'\beta' + n^{\cdot}_{-i,j}} \quad (6)$$

$$p(y''_i = j \mid v_i, \mathbf{v}_{\mathbf{-i}}, \mathbf{y''}_{\mathbf{-i}}, \mathbf{w}, \mathbf{z}) \propto \frac{n_j}{N^1_d} \cdot \frac{\beta'' + n^{v_i}_{-i,j}}{W''\beta'' + n^{\cdot}_{-i,j}} \quad (7)$$

In which N^0_d and N^1_d are the total number of words assigned to background topic and regular latent topics in document *d*. $n^{c_i}_{-i,j}$ is the total number of times concept c_i being assigned to topic *j* except for current one, while $n^{v_i}_{-i,j}$ is the total number of times visual word v_i being assigned to topic *j* except for current one.

For the **variation** of HPB model (i.e. the **HPB2 model**), we have a uniform expression of posterior probabilities for both concept and visual words:

$$p(x_i = 0, \tilde{z}_i = 0 \mid \tilde{w}_i, \tilde{\mathbf{z}}_{\mathbf{-i}}, \tilde{\mathbf{w}}_{\mathbf{-i}}, \mathbf{w}, \mathbf{z}, \tilde{\beta}) \propto \frac{N^0_d}{N_d} \cdot \frac{\tilde{\beta}_2 + n^{\tilde{w}_i}_{-i,0}}{\tilde{W}\tilde{\beta}_2 + n^{\cdot}_{-i,0}} \quad (8)$$

$$p(x_i = 1, \tilde{z}_i = j \mid \tilde{w}_i, \tilde{\mathbf{z}}_{\mathbf{-i}}, \tilde{\mathbf{w}}_{\mathbf{-i}}, \mathbf{w}, \mathbf{z}, \tilde{\beta}) \propto \frac{N^1_d}{N_d} \cdot \frac{n_j}{N^1_d} \cdot \frac{\tilde{\beta} + n^{\tilde{w}_i}_{-i,j}}{\tilde{W}\tilde{\beta} + n^{\cdot}_{-i,j}} \quad (9)$$

The nomination in (8) and (9) is the same as that in (2).

5. EXPERIMENTAL RESULTS

In this section, we apply the proposed HPB model to topic learning and compare the performance of HPB model with that of the extended Correspondence LDA (Corr-LDA) model under the same biomedical image annotation scenario using cross-validation. For topic learning, we look into the average log-likelihood of two models and visualize the discovered latent themes. The performance of automatic image annotation is evaluated by perplexity and annotation accuracy.

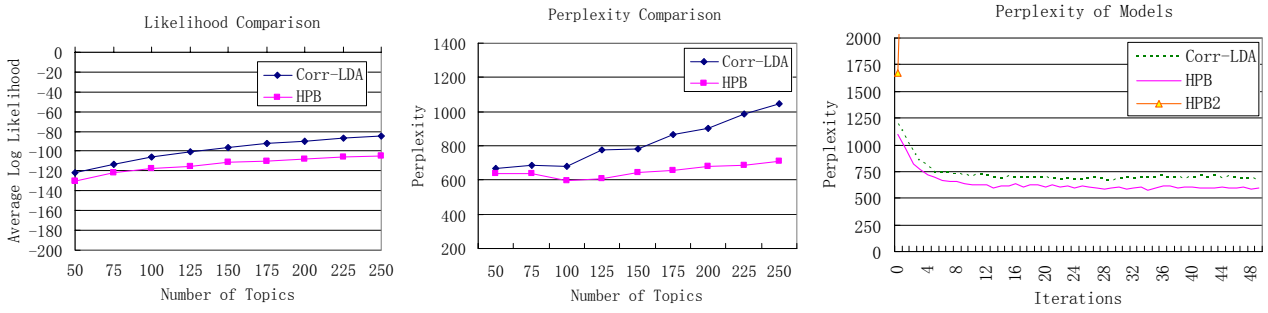
5.1 Data Collection and Settings

The data used in our experiment is from the online journal 'Breast Cancer Research' in the publicly available PubMed Central database (http://www.pubmedcentral.nih.gov/). In this journal, all the research articles (in digital version) between year 2002 and 2008 are downloaded and parsed. After that, a total of 2320 image-caption pairs are extracted from the original biomedical literatures and makeup the dataset for experiment. As introduced in Section 2, words, visual words and ontology-based biomedical concepts are indexed from image-caption pairs. In total, we indexed 132,978 text tokens which belong to 4113 unique words, 379,526 visual words from a vocabulary size of 1000, and 53,825 concepts, with 1938 unique concepts appear.

The original dataset is divided into 5 subsets with equal size. Of the 5 subsets, one subset (20%) is retained as the validation data for testing the model, and the remaining 4 subsets (80%) are used as training data. For image annotation evaluation, the cross-validation process repeats 5 times, with each of the 5 subsets used once as the validation data. After that, we take the average results for evaluation.

5.2 Topic Learning and Representation

The topic learning process of the proposed probabilistic model is achieved by running the collapse Gibbs sampling process over training dataset until converge (basically, it takes less than 100 iterations to converge in model estimation). When the topic model is estimated from the training dataset, we will be able to visualize the uncovered latent themes and tell the correlation among words, visual words and biomedical concepts.



(a) Likelihood comparison (after convergence) **b) Perplexity comparison** **(c) Perplexity over the iterations (# of topics = 100)**

Fig. 5 The likelihood and perplexity comparison of the extend Corr-LDA model and the HPB model

5.2.1 Likelihood Comparison

Log-likelihood is a standard criterion for generative models. It can be calculated by integrating out the topic variables after the convergence of Gibbs sampling. Generally, the higher log-likelihood the model assigned to the data, the better predictive power and generalization ability the model has.

The average word likelihood of the extend Corr-LDA model and the HPB model is compared. The marginal likelihood p(**w**|**z**) of the extend Corr-LDA model can be calculated by integrating out latent variables $\boldsymbol{\varphi}$:

$$p(\mathbf{w}\mid\mathbf{z}) = \prod_{t=1}^{T}\left[\int_{\varphi_{z_t}} p(\mathbf{w}\mid z_t,\varphi_{z_t})\,p(\varphi_{z_t}\mid z_t)\,d\varphi_{z_t}\right]$$
$$= \prod_{t=1}^{T}\left[\frac{\Gamma(W\beta)}{\Gamma(\beta)^{W}}\int_{\varphi_{z_t}}\prod_{i=1}^{W} p_{w_i}^{n_t^{(w_i)}+\beta-1}\,d\varphi_{z_t}\right]$$
$$= \left[\frac{\Gamma(W\beta)}{\Gamma(\beta)^{W}}\right]^{T}\cdot\prod_{t=1}^{T}\frac{\prod_{w_i}\Gamma(n_t^{(w_i)}+\beta)}{\Gamma(n_t^{(\cdot)}+W\beta)}$$

The average word likelihood can be obtained by taking the logarithm of p(**w**|**z**) and averaging the resulting summation by W.

For the HPB model, the marginal likelihood p(**w**|**z**) is:

$$p(\mathbf{w}\mid\mathbf{z}) = \left[\frac{\Gamma(W\beta)}{\Gamma(\beta)^{W}}\right]^{T}\cdot\prod_{t=1}^{T}\frac{\prod_{w_i}\Gamma(n_t^{(w_i)}+\beta)}{\Gamma(n_t^{(\cdot)}+W\beta)}\cdot\frac{\Gamma(W\beta_2)}{\Gamma(\beta_2)^{W}}\cdot\frac{\prod_{w_i}\Gamma(n_0^{(w_i)}+\beta_2)}{\Gamma(n_0^{(\cdot)}+W\beta_2)}$$

The average word likelihood of the HPB2 model is the same as the HPB model.

As illustrated in Fig. 5a, for both models, the likelihood increase as the number of topic increase, which means that a relatively larger topic numbers may potentially result in better modeling of testing data. However, it should be noted that there is a trade-off between topic numbers and convergence time of models. And, as we would see in Section 5.3, the increase of topic number does not always lead to the improvement of predictive results.

In general, the log-likelihood of the extended Corr-LDA model and the HPB model are close, the difference between two models can be explained by the introduction of background topic in the HPB model.

5.2.2 Illustration of Discovered Latent Themes

One major objective of the proposed models is to uncover the latent topics from image-caption pairs and facilitate knowledge organization and understanding in online biomedical literatures. With this consideration, we visualize the discovered latent topics by providing the top-ranked words, top-ranked concepts (Fig. 4 and 6) and most related images (Fig. 6, with probability under each image). For this example, the latent topics are learnt by the HPB model, in which the topic number is 125.

Topic28

Top words	Probability
P53	0.184625
mammary	0.096425
heterozygous	0.06976
adenocarcinoma	0.051299
carcinomas	0.038992
panels	0.036941
enhances	0.036941
deficiency	0.032839
knockout	0.032839
irradiated	0.032839
early	0.018481
developing	0.018481
tumours	0.01643
Balb	0.01643
genotypes	0.01643
Spontaneously	0.01643
MSM	0.01643
cHeA	0.014379
spares	0.012327
adjuvant	0.012327

Top Concepts	Concept Names
C0001418	Adenocarcinoma
C0858252	Breast Adenocarcinoma
C0007097	Carcinomas
C0206745	Knockout Mice
C0599772	knockout gene
C0025919	BALB C Mice
C1446974	cheA protein, E coli
C1307090	F1-20 protein, mouse
C0598034	BRCA2 Gene
C0677850	adjuvant therapy
C0001551	Immunoadjuvants
C0029463	Osteosarcoma
C1336745	Thymic Lymphoma
C0009085	Clustering
C0349966	Figs
C0315310	Salmonella uppsala
C0439828	Variable
C0557775	Van
C0439234	yr
C0906368	MPR1 transport protein

Topic70

Top words	Probability
immunostaining	0.057799
Receptor	0.040886
Intense	0.031021
Chains	0.029611
Infiltrative	0.028202
Complex	0.026792
Immunohistochemical	0.025383
Lobular	0.023974
comparative	0.022564
Interleukin	0.022564
Infiltrating	0.019745
corresponding	0.018336
Androgen	0.018336
Thymus	0.016927
Labeled	0.015517
IDC	0.014108
magnification	0.012699
Observed	0.012699
comparison	0.011289
Distinct	0.011289

Top Concepts	Concept Names
C0597357	Receptor
C0444498	In situ
C0439855	Complex
C0021764	Interleukin
C0009491	Comparative Study
C0291890	NR0B1
C0154073	Stage 0 Skin Cancer
C0002844	Androgenic Agents
C0205417	Lobular
C0034804	Receptors, Estrogen
C0264793	DCM
C0206692	Carcinoma, Lobular
C0796396	I-125
C0022262	Isotopes
C0123356	ILS
C0021010	Gm Allotype
C1101536	IL-2Ralpha
C0079004	B-Cell Subset
C0010834	Cytoplasm
C0456981	Specific antigen

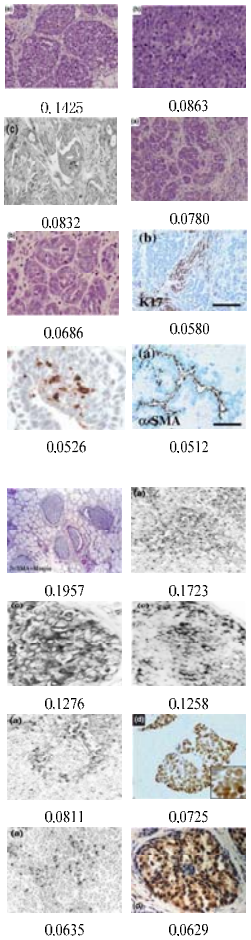


Fig. 6 Illustration of discovered latent themes by HPB model

As illustrated in Fig. 4, the background topic depicts the contextual information of the biomedical journal, such as breast cancer, human body and tumor. The regular latent topics, on the other hand, reveal some domain specific knowledge. As

illustrated in Fig. 6, the top-ranked words, concepts and images of the uncovered latent topics have high semantic consistency. The top ranked words and concepts not only contain domain specific terms such as receptor, carcinomas, breast adenocarcinoma and Immunohistochemical, which help user to interpret the topics, but also provide many protein names and gene names that are related to the uncovered latent topic.

5.3 Image Annotation and Evaluation

The proposed probabilistic models are able to establish direct correlation among caption words, visual words and biomedical concept in biomedical image-caption pairs. Therefore, given the image content, a good model should be able to predict the missing captions. Next we automatically annotate caption words and concepts for images in the testing dataset based on the uncovered latent topics from training dataset, with both captions and concepts in testing dataset regarded as unknown (missing). The performance of automatic annotation is evaluated by perplexity and annotation accuracy using cross-validation.

5.3.1 Perplexity Comparison

In our experiment, we resort to the word caption perplexity as standard criteria of the annotation performance.

The perplexity of a set of testing image-caption pairs (for all $d \in D_{test}$) is defined as the exponential of the negative normalized predictive log-likelihood using the training model, in which the topic-word conditional probability: $p(w_i \mid z_{wi} = t)$ is obtained from the Gibbs sampling process of training dataset in Section 4.

$$ppx = \exp\left\{-\frac{1}{N_W}\sum_{j=1}^{D}\sum_{i=1}^{W}\log\left[\sum_{t=1}^{T}E\left(p(w=w_{j,i}\mid z=t)\right).E\left(p(z=t\mid d=j)\right)\right]\right\}$$

With uncovered latent topics from training image-caption pairs, the estimation of prior probability of topic in a testing image can be approximated by running collapse Gibbs sampling over all the extracted visual words (no words or concepts used) in testing dataset (eq. 10) using fixed visual word-topic conditional probability $p(v_i \mid y''_i = j)$ (which is obtained from the Gibbs sampling process of training dataset in Section 4).

$$p(y''_i = t \mid v_i, \mathbf{v}_{-\mathbf{i}}, \mathbf{y''}_{\mathbf{-i}}) \propto p(v_i \mid y''_i = t)\cdot\frac{\alpha + n^d_{-i,t}}{T\alpha + n^d_{-i,\cdot}} \quad (10)$$

After the convergence of the Gibbs sampling process, the probability for the 'missing' caption words and concepts of an image can be calculated via the production of topic-word/concept conditional probability and the prior probability for each topic.

Recall that for HPB model, we assume no background topic for visual words, the prior for background topic in a document is approximated by average probability over the training dataset.

Fig. 5b represents the perplexity of CorrLDA and HPB model over different topic numbers. The perplexity of HPB model is lower than that of the CorrLDA model, which indicates that HPB model generated from training data set is 'less surprised' by the testing data, thus, it demonstrates better ability in annotation. What's more, as the topic number increases, the perplexities of both models decrease first, and then increase, with 100 topics have the lowest perplexity. It appears that the increase of topic number does not always lead to persistent improvement of predictive ability.

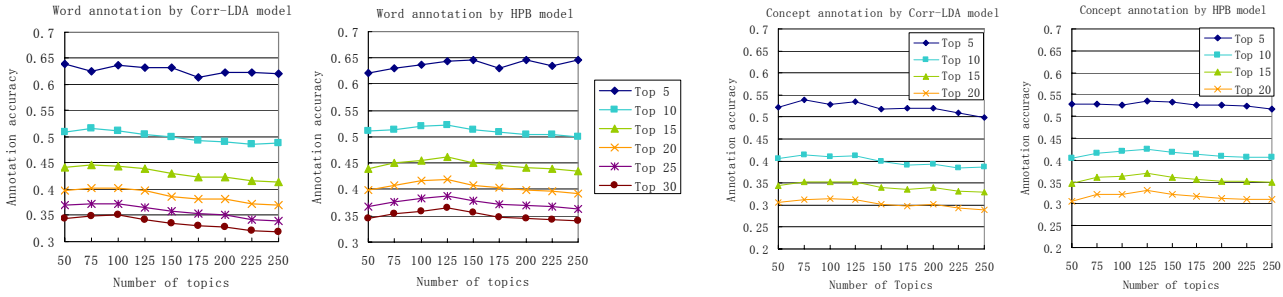
Fig. 5c illustrates the perplexities over the iterations when the topic number is 100. Although the HPB model appears to be more sophisticated than the Corr-LDA model, they converged in similar number of iterations. Recall that we have a variation of HPB model (named as the HPB2 model), which assumes that background words and concepts are related to certain image content (visual words). As in Fig. 5c, the perplexity of HPB2 increases sharply and quickly exceeds 10000, which indicates that the Gibbs sampling process for this model fails to converge. Finally, over 90% of the entities in documents are assigned to the background topic (as a comparison, only about 1/10 of the words will be assigned to background topic when the Gibbs sampling process of HPB model converges). According to the perplexity results, there is no evidence that there exist a direct correlation between image content and background information in the caption.

5.3.2 Annotation Accuracy Comparison

When the prior probability of topics in a testing image is estimated (eq. 10), the word and caption annotation for each document can be achieved by ranking words and concepts with regard to the following probability.

$$\begin{cases} p(w_i \mid d_j) = \sum_{t=1}^{T} p(w = w_i \mid z = t).p(z = t \mid d = j) \\ p(c_i \mid d_j) = \sum_{t=1}^{T} p(c = c_i \mid z = t).p(z = t \mid d = j) \end{cases} \quad (11)$$

The words and concepts that achieve highest probability value in (11) are used as the annotation of images. After that, the image annotations are compared to the original words and concepts in testing image-caption pairs for validation. During annotation evaluation, the cross-validation process repeats 5 times, and the results are averaged to produce the final annotation accuracy.



(a) Word annotation accuracy comparison **(b) concept annotation accuracy comparison**

Fig. 7 Annotation accuracy comparison over different topic numbers

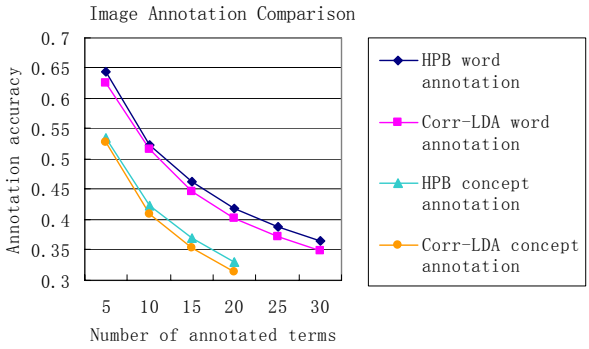


Fig. 8 Image annotation comparison

The accuracy of word and concepts annotation over different topic numbers is illustrated in Fig. 7. Specifically, Fig. 7a represents the annotation accuracy from top 5 annotation words to top 30, while Fig. 7b provides the annotation accuracy from top 5 concepts to top 20. According to the experiment results, the HPB achieves best annotation performance when topic number is 150, while the Corr-LDA model achieves best performance with 100 topics. As the topic number increases, the annotation accuracy of both models increase first, and then decrease, which is consistent with the results in perplexity comparison.

The annotation accuracy of extended Corr-LDA model and the proposed HPB model is compared using their best annotation performance (i.e. 100 topics for Corr-LDA model, and 125 topics for HPB model). As illustrated in Fig. 8, the HPB model is consistently better than the extended Corr-LDA model in both word annotation and concept annotation tasks, which is consistent with the perplexity comparison results in Section 5.3.1. What's more, according to Fig. 7 and Fig. 8, the performance of HPB model drop slower than the Corr-LDA model when considering the annotation accuracy of large number of annotation terms, which indicates that HPB model is more robust and is able to achieve better performance in annotating less frequent terms

6. CONCLUSION AND FUTURE WORK

The contribution of this paper is twofold. First, we proposed a novel HPB model to integrate background information in topic learning, incorporating contextual information to interpret the uncovered latent topic and improve the image annotation accuracy. Second, in our experiments, we discovered that there is no direct correlation between image content and the background information in the captions. In other word, the extracted visual words from images have nothing to do with the background topic. It is unnecessary to incorporate contextual information when modeling the image contents.

For future work, we plan to incorporate other kinds of knowledge (such as protein entities, gene names and concept relations) in our model to enrich the discovered latent semantic topics and facilitate knowledge organization and understanding in online biomedical literatures.

7. REFERENCES

[1] T. Hofmann. Probabilistic Latent Semantic Indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference, 1999.

[2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, Content-based image rerieval at the end of the early years, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[3] T. L. Griffiths, M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101:5228-5235, 2004.

[4] J. Yang, Y. G. Jiang, A. G. Hauptmann, C. W. Ngo, Evaluating Bag-of-Visual-Words Representations in Scene Classification. ACM SIGMM Int'l Workshop on Multimedia Information Retrieval (MIR'07), Augsburg, Germany, Sep. 2007.

[5] W. W. Cohen, R. Wang, and R. F. Murphy. Understanding captions in biological publications. ACM KDD, 2005.

[6] Sivic, J., Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos. International Conference on Computer Vision. (2003) 1470– 1477

[7] D. Blei and M. Jordan, *Modeling Annotated Data*, Proc. ACM SIGIR Conf. Research and Development in Information Retrieval, 2003.

[8] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. International Journal of Computer Vision, vol. 73, no. 2, June 2007, pp. 213-238

[9] T. Kadir and M. Brady. Scale, Saliency and Image Description. International Journal of Computer Vision. 45 (2):83-105, November 2001

[10] O. Yakhnenko, V. Honavar, Annotating images and image objects using a hierarchical Dirichlet process model, proceedings of the 9th International Workshop on Multimedia Data Mining, pp. 1-7, 2008.

[11] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, vol.2 pp. 257-264, 2003

[12] Zhou, X., Zhang, X., and Hu, X., "MaxMatcher: Biological Concept Extraction Using Approximate Dictionary Lookup," the 9th biennial The Pacific Rim International Conference on Artificial Intelligence (PRICAI 2006), Aug 9-11, 2006, Guilin, Guangxi, China, Page 1145-1149

[13] Lowe, D. Distinctive Image Features from Scale-Invariant Key Points. International Journal of Computer Vision, 60(2): 91–110, 2004.

[14] Van Rijsbergen, C.J., Information Retrieval, Butterworths, 1975.

[15] Humphreys B. and Lindberg D. – The UMLS project: making the conceptual connection between users and the information they need – Bulletin of the Medical Library Association 81(2): 170, 1993.

[16] Yu-Gang Jiang, Chong-Wah Ngo, Jun Yang: Towards optimal bag-of-features for object categorization and semantic video retrieval. CIVR 2007: 494-501

[17] L. Fei-Fei and P. Perona, A Bayesian hierarchical model for learning natural scene categories. In CVPR, volume 2, pages. 524–531, 2005

[18] X. Wang and E. Grimson, Spatial Latent Dirichlet Allocation, in Proceedings of Neural Information Processing Systems Conference (NIPS) 2007

[19] D. Blei, A. Ng. and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993-1022,2003.

[20] Newman, D., Chemudugunta, C., Smyth, P., Steyvers, M.: Statistical entity-topic models. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, Pennsylvania, USA, pp. 680–686 (2006)