Simultaneous Joint and Conditional Modeling of Documents Tagged from Two Perspectives

Pradipto Das  
CSE Dept.  
SUNY Buffalo  
Buffalo, 14260  
pdas3@buffalo.edu

Rohini Srihari  
CSE Dept.  
SUNY Buffalo  
Buffalo, 14260  
rohini@cedar.buffalo.edu

Yun Fu  
CSE Dept.  
SUNY Buffalo  
Buffalo, 14260  
yunfu@buffalo.edu

ABSTRACT
This paper explores correspondence and mixture topic modeling of documents tagged from two different perspectives. There has been ongoing work in topic modeling of documents with tags (tag-topic models) where words and tags typically reflect a single perspective, namely document content. However, words in documents can also be tagged from different perspectives, for example, syntactic perspective as in part-of-speech tagging or an opinion perspective as in sentiment tagging. The models proposed in this paper are novel in: (i) the consideration of two different tag perspectives - a document level tag perspective that is relevant to the document as a whole and a word level tag perspective pertaining to each word in the document; (ii) the attribution of latent topics with word level tags and labeling latent topics with images in case of multimedia documents; and (iii) discovering the possible correspondence of the words to document level tags. The proposed correspondence tag-topic model shows better predictive power i.e. higher likelihood on held-out test data than all existing tag topic models and even a supervised topic model. To evaluate the models in practical scenarios, quantitative measures between the outputs of the proposed models and the ground truth domain knowledge have been explored. Manually assigned (gold standard) document category labels in Wikipedia pages are used to validate model-generated tag suggestions using a measure of pairwise concept similarity within an ontological hierarchy like WordNet. Using a news corpus, automatic relationship discovery between person names was performed and compared to a robust baseline.

Categories and Subject Descriptors
I.2.7 [Computing Methodologies]: Artificial Intelligence—Natural Language Processing

General Terms
Algorithm, Experimentation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM ’11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

Keywords
Text mining, tag topic models, social media, classification and clustering

1. INTRODUCTION
This paper lays down a robust topic modeling framework to solve the problem of discovering latent topics from documents tagged from two different perspectives. Documents usually consist of at least two perspectives - a document level perspective and a word level perspective. The document level perspective often tries to summarize the contents as a small bag-of-words, while the other perspective tries to annotate the content in different ways. Tables 1, 2 and 3 show different examples of perspectives. In this paper, it is assumed that tags are non-hierarchical concepts. These concepts could be represented by words or by some other higher order representation that eventually denotes a concept. Many times, for e.g., in plain text documents, these two perspectives are not explicitly shown. However, documents hosted by today’s interactive websites are a rich source of document tagging from at least two perspectives.

<table>
<thead>
<tr>
<th>Word level tags</th>
<th>Image captions</th>
<th>Category labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syringa/0 (Lilac)/0 is/0 a/0 genus/0 of/0 about/0 ... Lilac/1 bushes/1 can/1 be/1 prone/1 to/1 powdery/1 mildew/1 disease/1 ... RHS/2 Dictionary/2 of/2 Gardening/2 /2 Macmillan/2 ISBN/2 0-333-47494-5/2</td>
<td>Syringa, jasikaea, Syringa vulgaris shrub in flower, etc.</td>
<td>Syringa, Garden plants, Flowers, Shrubs</td>
</tr>
</tbody>
</table>

Table 1: Document with word level “Position” tags and document level image caption word tags [source: http://en.wikipedia.org/wiki/Syringa]

Table 1 shows an article on lilac flower in Wikipedia. Words in the document body annotated with a slash '/' denotes a word level perspective which in this case is position of the section in which the word appears. The section offsets are binned into three positions relative to the beginning of the document - begin(0), middle(1) and end(2). Positions signify the importance of the choice of words that constitute sections in a document. The document level perspective is assumed to be captured by the images which are described by the corresponding captions. In table 1, the third column represents the “ground truth” category labels which are a set of manually edited tags that summarize the Wikipedia article. With this structure of multimedia documents, several questions come to the forefront: ‘Could there be some way of using image captions to automatically suggest specific category labels for new articles? If so how good will
those suggestions be? Further, can we discover latent topics or themes and label each theme with a multimedia object? The generative process of word generation in these kinds of documents hinges on the following intuitions: Documents are distributions over latent topics. Latent topics, in turn, are distributions over observed document level tags and main content words such that the most probable observed variable ensembles for a topic are related through the assumptions of the generative process. For one particular assumption, a word conditioned on the word level tag observed at the word’s position is sampled independently of the document level tags from a topic. The topic proportions for the document only depends on the expected number of document level tags and words being assigned to each topic through co-occurrence phenomenon. In another assumption, topics generate the document level tags first. Then a document level tag position is chosen and a word conditioned on the word level tag observed at the word’s position is sampled from the corresponding topic in the position of the document level tag. For this second assumption, there is a conditioning of words to document level tags and is thus more intuitive from the document generation point of view. For example, a writer often “thinks” of a mental image/concept and then writes words that elaborate that image/concept. Although modeling multimedia Wikipedia articles served as the primary motivation for developing the proposed models, the models are extremely generic and had been applied to a variety of other datasets for different tasks. For brevity, the document level tags are dubbed DL tags and the word level tags are dubbed WL tags.

Table 2 shows an example where DL tags are abstracted at a level higher than words. In this example, the sample sentence in the table is an excerpt from a newswire article in the dataset used in the DUC2005 Summarization track[10]. The WL tags denote a particular named entity class like PERSON, LOCATION, ORGANIZATION, NUMBER, etc. being ascribed to each particular word or not. The DL tags, however, are indicative of discourse coherence markers. These markers were constructed following the technique used in [2]. Each word in the document is associated with a grammatical or semantic role (GSR in short) like named entities (ne), nouns (nn), adjectives (adj), verbs (vb), subjects (subj), objects (obj), etc. A GSR transition (GSRt in short) is a relation between the same normal form of a word that is either present in two contextual sentences or in a single sentence, e.g. a GSRt for the word “car” can be (car, subj, obj) leading to (car, subj→obj) or (car, subj→obj→obj) if the word “car” is not seen in the succeeding sentence. It is reported in [2] that a set of sentences with the same entities in roles like “subj,” “obj,” etc. are indicative of coherent passages. Although in [2], only entities are involved in GSRts, however, in this paper, words that are not entities are also considered since in quite a few cases, the focus of attention is based not just on entities. Thus, the document level perspective for DUC2005 newswire data is that of syntactic coherence. The proposed models are general enough to model documents arising out of such perspectives also.

### 1.1 Representing Tags in Datasets

In all the experiments performed in the paper, three major datasets were considered. Also, all tags have a bag-of-words representation. For Wikipedia, only documents with images were collected. The category labels were not used as DL tags, rather the image caption words were used. Generally, if captions are not available, an initial preprocessing can be done using the work in [11]. The article title was also added to the DL tags. Each word in the main body of the article was tagged with “position” information of the sections they appeared in and were labeled as {Begin, Begin_Middle, Middle, Middle_End, End}.

Unprocessed Amazon product reviews from the dataset used in [7] (henceforth the AR dataset) was also used in the experiments. The words in each review were tagged with affect labels using a simple lexicon lookup from the dataset created in [9]. The lexicon consists of 2476 words that elicit human emotions in some form. The emotions were labeled with {Unhappy, Unsatisfactory, Melancholic, Despair, Hopeful, Contended, Satisfied, Pleased, Happy, Untagged} tags based on the maximum valence values of the affect words. Non-affect words were tagged as “Untagged”. The AR dataset did not have product tags and hence the product name and the review title were used as “captions” for reviews which served as DL tags. Finally note that for WL tagging, a word can be conditioned on only 1 tag. Table 3 shows an example for the AR dataset used. The

<table>
<thead>
<tr>
<th>Word level tags</th>
<th>Document level tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Some 167/NE-NUM people were arrested in the US/NE-LOC, including a senior executive of Columbia/NE-LOC’s national bank</td>
<td>ne, ne→ne, subj→subj, subj→obj, nn→vb, vb→vb, →adj, adj→adj, etc.</td>
</tr>
</tbody>
</table>

Table 2: Document with word level “Named Entity” tags and document level “syntactic role transition” tags

Table 3: Document with word level emotion tags and document level product feature tags

DUC (Document Understanding Conference) 2005 dataset consists of newswire articles organized in 50 folders or document sets (docsets) with each folder consisting of at least 25 sizable news reports. The documents were processed to extract named entities and roles of the words using the Stanford CoreNLP toolkit. The GSRs were obtained using the dependency parse information. However, co-reference resolution was not performed due to unsatisfactory results. An example of this kind of tagging is shown in Table 2. Normal forms of the words were used e.g., “arrested” (verb) and “arrests” (noun) have the normal form “arrest”. A total of 9 GSRs were chosen (Named Entities or ne, Subjects or subj, Objects or obj, Nouns or nn, Verb or vb, Adjective or adj, Adverb or adv, Other as ow and Null as –) resulting in a total of 81 GSRs. Note that if a word has several GSRs associated with it, only one is chosen using the priority rule: ne>subj>obj>nn>adj>vb>adv>ow. The task in the DUC2005 Summarization task was the creation of 250 word multi-document summaries for each of the docsets in response to the corresponding information needs. However, in this paper the docsets from DUC2005 dataset were used to validate entity-pair relationship discovery and not for summarization.

### 1.2 Improving Existing Tag Topic Models

Tag-topic models have been explored recently [16, 17, 1, 20] as ways of improving word-based topic models with ad-
ditional information in the form of tags, usually arising out of a single perspective. Existing mixture tag topic models [16, 17] (c.f. fig. 1b) can fit a number of latent topics to the documents without words and DL tags having direct correspondence with each other. However, there could be additional word level annotation information which are implicitly attributed to the words. 

Thus existing tag topic models of documents focus either on document level tags[16, 17] or on word level tags[20] (c.f. fig. 1c). For models like those in [16, 17] (which are referred to as MMLDA - short for Multi(nomial) Multinomial LDA) each content word and DL tag is generated independently by choosing a topic and then choosing a content or DL tag. The expected number of words in the document’s topic thus depends on the counts of both the content and the DL words ascribed to that topic. On the other hand, in the TagLDA model in [20] the content words are generated by choosing a topic and drawing a word from the topic’s distribution but conditioned on the WL tag associated with that content word. This conditional aspect allows one to explore related words sharing the same semantic relatedness but along that particular condition or facet - for e.g. all “PERSON” named entities in a dataset that are semantically related through some hidden topic. Figures 1b and 1c show the existing tag topic models - MMLDA[16] and TagLDA[20]. The model in figure 1d was implemented in this study as an improvement over MMLDA following [4] for the text domain and is referred to as corrMMLDA. However, none of MMLDA, TagLDA or corrMMLDA addresses a tag space that is split across two different perspectives. The proposed TagSquaredLDA (abbreviated as Tag^2LDA) models: METag^2LDA model (ME is abbreviated form of Multinomial Exponential) (fig. 1e) and corrMETag^2LDA model (fig. 1f) allows topic modeling of documents with both DL and WL tags. Table 4 shows the relative merits and de-merits of each model discussed in this paper. Fig. 1a shows a supervised LDA topic model (sLDA)[5] that is only used for predictive power comparison on the AR dataset. Experiments reveal the improvements of the Tag^2LDA models over current tag topic models through reduced perplexity, or more predictive power for topical inference. Also an HMM type of model is not suitable for positional WL tag assignment, since, then during inference, there is nothing to infer on position “states” - they are implicit in any document.

1.3 Applications and Quantitative Measures

Measuring model perplexity[6] is an established way of showing how good a model explains the observed data. Due to intractability issues, the lower bounds to the true log likelihoods on test data are also used which are directly proportional to perplexity measurements. However, while applying the models for a specific task, the goal is not only to measure held-out test data likelihood for a model. For example, for the Wikipedia data, it was important to have a quantitative measure of confidence between probable document tags from image captions and ground truth category labels. For this, a measure of semantic relatedness using path separation between concept pairs[14] in WordNet ontology was chosen as an evaluation tool. As an example, the connection between “fire engines” and “fire_extinguisher” can be described by a shortest path linking these two concepts in WordNet as “fire_extinguisher ⇔ device ⇔ instrumentality ⇔ container ⇔ wheeled_vehicle ⇔ self-propelled_ vehicle ⇔ motor_vehicle ⇔ truck ⇔ fire_engine” with a path length of 9 and a simple “inverse of path length” similarity score of 0.11. Under this measure, a value of 1 indicates exact match or parent/child relationship. Using this evaluation, users can be explained a “chain of reasoning” that relates a probable DL tag to a ground truth category label for a new document. For N suggested DL tags and C category word labels, scores for all possible N × C pairs P were obtained. The highest score served as a measure of DL tag suggestions. If a model captures caption words that happens
to have shorter path distances to ground truth labels, then the model is scored higher. Note that WordNet is chosen since it is widely accepted - any other customized ontology can easily be plugged in depending upon the application. Also note that Newman et. al. [13] attempted to measure cohesiveness of topic "labels" consisting of top 10 high probability words, where the "results over WordNet are patchy at best." The WordNet evaluation presented here is not to measure topic cohesiveness but to measure and explain the goodness of a probable DL tag. Although measuring topic coherence is equivalent to measuring topic intrusion [8], the notion of a topic is only a mathematical convenience for a low-dimensional subspace that tries to capture the assumptions of the statistical generative model. So qualitatively, a topic is best interpreted by the task on which the model is adapted and its corresponding assumptions.

For the DUC2005 dataset where the WL tags came from named entity classes, all pairs of PERSON named entities from documents in each of the 50 documents were collected. For a particular proposed Tag^2LDA model, hidden topics were inferred for these documents and pairs from top N entities from the "PERSON" facet of the topic were collected. Entity pairs that co-occur in a sentence were chosen as ground truth pairs that were strongly related - this is the baseline. The average of ratio of the counts of the PERSON entity pairs from topics to those from the baseline served as a quantitative measure of improvement over the latter in the entity relationship discovery application. Some qualitative results are shown in Table 8. Note that the same named entity can occur across multiple documents. This is particularly true of NUMBER and LOCATION classes and entities related to governments.

2. RELATED WORK

Joint topic and tag analysis has been used in a some recent works including [16, 17, 20] which has culminated in the creation of variants of topic models like LDA[6]. Due to space constraints, the reader needs to be directed to [6] for a full description of the basic LDA model. The principle shortcoming of these papers are the use of a single tagging perspective - either document level tags or word level tags. While the models in [16] and [17] are essentially the same, the purposes of the models had been a little different. Both use generative models of words and document tags to discover latent topics. In [16] the topic-tag and the topic-word features were used to better cluster tagged documents. In [17] new documents were "folded-in" in the latent topic space and tags were predicted based on the inferred topic tag distribution. The work in [20] is useful in the sense that the topics are discovered w.r.t to words being conditioned on WL tags. A recent work on topic-perspective modeling has been done in [12] where the authors have tried to use perspectives as hidden states that represent a discrete distribution over tags. It is important to note that the "corrLDA" model referred to in [12] is not a true correspondence model as there happens to be no direct correspondence between words and tags. Further, although there is a connection between the user-perspective and perspective-tag distributions, the connections between the perspective-tag distribution and the topic-word/topic-tag distributions weakly depend on a binary switching variable. The sLDA[5] model discovers topics based on the ensembles of document contents and the response variables. The values of the response variables are explained by the frequency counts of the words in the corresponding documents only. The labeledLDA[15] model establishes a one-to-one correspondence to the latent topics and the actual document tags. This is done in a manner similar to imposing a non-uniform prior on the latent topic proportions per document [19]. Further the words in text are corresponded to topic labels which precludes any possibility of using WL conditional tags. The proposed Tag^2LDA models use both joint (words and DL tags) and conditional (words and WL tags) modeling, thereby allowing a richer document structure to be captured.

3. THE PROPOSED MODELS

This section introduces the model description and the model parameters for the Tag^2LDA models. In all model figures in fig. 1, the symbol notations and their meanings given in Table 5 are adhered to. Note that in the Tag^2LDA models, p(w_m|C = i, β, π, t_m), where C = y_m or C = z_m, is not a simple topic multinomial anymore, but is distributed as

\[ p(w_m|C = i, β, π, t_m) = \frac{\exp(\log β_{i,m} + \log π_{i,m} w_m)}{\sum_{v=1}^{V} \exp(\log β_{v,m} + \log π_{v,m} w_m)} \]

(1)

Note that each π_i is a distribution over V, Simplified Gibbs sampling using Multinomial-Dirichlet conjugacy cannot be applied in this setting since the distribution is not a multinomial anymore. Also note that for the correspondence models y_m ~ Unif(N), short for Uniform distribution, as in [4]. The generative processes for the proposed Tag^2LDA models are illustrated below:

Table 5: Symbols used in this paper and their meaning

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning (r.v. = random variable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>total number of documents</td>
</tr>
<tr>
<td>N</td>
<td>total number of unique document level tags per document d ∈ D</td>
</tr>
<tr>
<td>M</td>
<td>total number of unique words per document d ∈ D</td>
</tr>
<tr>
<td>α</td>
<td>r.v. for Dirichlet prior for the document level topic proportions</td>
</tr>
<tr>
<td>θ_d</td>
<td>r.v. for document level latent topic proportions</td>
</tr>
<tr>
<td>ρ</td>
<td>r.v. for corpus level topic-DL tag multinomial</td>
</tr>
<tr>
<td>β</td>
<td>r.v. for corpus level (marginal) topic-word distribution</td>
</tr>
<tr>
<td>π</td>
<td>r.v. for corpus level marginal tag-word distribution</td>
</tr>
<tr>
<td>s_n</td>
<td>indicator variable for DL topic proportion</td>
</tr>
<tr>
<td>y_m</td>
<td>indicator variable for DL topic proportion</td>
</tr>
<tr>
<td>y_m</td>
<td>indicator variable for DL tag correspondence</td>
</tr>
<tr>
<td>w_m</td>
<td>r.v. for DL tag at position n; vocabulary size corrV</td>
</tr>
<tr>
<td>w_m</td>
<td>r.v. for word at position m; vocabulary size V</td>
</tr>
<tr>
<td>t_m</td>
<td>r.v. denoting tag at position m, on which word w_m is conditioned; vocabulary size T</td>
</tr>
<tr>
<td>t_d</td>
<td>r.v. denoting observed response for document d</td>
</tr>
<tr>
<td>μ, σ</td>
<td>r.v.s denoting mean and standard deviation for the observed response for document d</td>
</tr>
</tbody>
</table>

For each document d ∈ 1, ..., D, Choose a topic proportion θ_n ~ Dir(α) For each “document level” position n in document d Choose topic indicator s_n ~ Mult(θ) Choose a “document level” tag w_n|s_n = k, ρ ~ Mult(ρ_k) For each “word level” position m in document d Choose y_m ~ Unif(1, ..., N) (for corrMETagLDA - fig. 1f) or Choose y_m ~ Mult(θ) (for METag^2LDA - fig. 1c) Choose a word w_m|y_m, t, β, π ~ p(w_m|y_m, β, π, t_m) (fig. 1f) or Choose a word w_m|y_m, t, β, π ~ p(w_m|y_m, β, π, t_m) (fig. 1c)
In the correspondence models, the DL perspective plays a significant role in the quality of topic coherence. For example, when the GSRt perspective (c.f. table 2) is chosen as a DL perspective, the topics will capture words that are both co-occurring and generated from similar roles. There could be a docket on “Global warning” which will tie together words like planet and ice based on co-occurrence alone. Consider another docket concerning discovery of ice on Pluto’s surface. Thus, if the GSR of “ice” is taken to be a subject, then a “Global warning” topic can include Pluto as a probable word because of the GSRt for the word “ice” that involve “subj”. This type of tagging is beneficial where the task is to optimize can be shown to be (as in [3])

\[ \text{log } F \]

\[ \text{parameters of the variational word-topic or word-DL perspective, the topics will capture words that are both co-occurring and generated from similar roles. There could, when the GSRt perspective (c.f. table 2) is chosen as a DL perspective, the topics will capture words that are both co-occurring and generated from similar roles. Consider another docket concerning discovery of ice on Pluto’s surface. Thus, if the GSR of “ice” is taken to be a subject, then a “Global warning” topic can include Pluto as a probable word because of the GSRt for the word “ice” that involve “subj”. This type of tagging is beneficial where the task is to optimize can be shown to be (as in [3]).} \]

3.1 Latent variable inference

The Variational Bayesian Expectation Maximization algorithm [6, 3] has been used to maximize the lower bound to the true intractable likelihood of the data w.r.t. the model parameters. This section outlines the various updates of the latent variables and the parameters and subsection 3.3 outlines a general plan of implementation. To find as tight as possible an approximation to the log likelihood of the data (the joint and conditional distribution of the observed variables given the parameters), the KL divergence of an approximated factorized mean field distribution is minimized to the true posterior distribution of the latent variables given the data. A fully factorized q distribution with “free” variational parameters \( \gamma, \phi, \lambda \) is imposed as

\[ q(\theta, z, y | \gamma, \phi, \lambda) = \prod_{d=1}^{D} \prod_{n=1}^{N_d} q(\theta_{d,n} | \phi_{d,n}) \prod_{n=1}^{N_d} q(y_{d,n} | \lambda_{d,n}) \]

and then optimal values of free variables and parameters are found by optimizing the lower bound on

\[ \log p(w_m, w_n | \alpha, \beta, \rho, \pi, t) \].

The variational functional to optimize can be shown to be as in [3]

\[ F = E_q[\log p(M, w_n, \theta, z, y | \alpha, \beta, \rho, \pi, t)] - E_q[\log q(\theta, z, y | \gamma, \phi, \lambda)] \]

where \( E_q[f(.)] \) is the expectation of \( f(.) \) over the q distribution and \( F \) is the Evidence Lower BOund (ELBO) to true likelihood. This ELBO is directly related to measuring perplexity[6]. In the following subsections, it is assumed that \( K \) is the number of topics, \( \phi \) be free parameters of the variational DL-tag-topic distribution and \( \lambda \) to be the free parameters of the variational word-topic or word-DL-tag distributions. These free parameters are defined for every document \( d \in D \). As in [6], the key inferential problem that is solved here is the learning of the posterior distribution of the latent variables given the observations and parameters of the models on data that are new on count proportions. Following the inequality, \( \log(x) \leq \frac{1}{2} \log((x + \frac{1}{2} - 1) x) \), the ELBO \( F \) is changed to further lower bounds \( L \) for the two models.

For the METagD^{2}LDA model:

\[ L_{MM} = E_q[\log p(\theta | \alpha)] + \sum_{m=1}^{M_d} \sum_{i=1}^{K} \lambda_{m,i} \log(\beta_{i,w_d,m} + \log(\pi_{d,m,w_d,m})) - \sum_{m=1}^{M_d} \sum_{i=1}^{K} \sum_{n=1}^{N_d} \lambda_{n,i} \log(\beta_{i,w_d,m} + \log(\pi_{d,m,w_d,m})) + \log(\gamma_{d,m}) - 1 \]

(4)

Using the new lower bound, the maximum likelihood estimations of the hidden variables in document \( d \) are as follows:

\[ \gamma_{d,m} = \frac{1}{V} \sum_{v=1}^{V} \exp(log(\beta_{i,w_d,m} + \log(\pi_{d,m,w_d,m})) - \log(\gamma_{d,m}) - 1) \]

(7)

The expression for \( E_q[\log p(w_m | y_m, \beta, \pi, t)] \) can be written as:

\[ E_q[\log p(w_m | y_m, \beta, \pi, t)] \geq \sum_{m=1}^{M_d} \sum_{i=1}^{K} \lambda_{m,i} \log(\beta_{i,w_d,m} + \log(\pi_{d,m,w_d,m})) - \sum_{m=1}^{M_d} \sum_{i=1}^{K} \sum_{n=1}^{N_d} \lambda_{n,i} \log(\beta_{i,w_d,m} + \log(\pi_{d,m,w_d,m})) + \log(\gamma_{d,m}) - 1 \]

(10)

Using these lower bounds and the maximum likelihood estimations of the hidden variables in document \( d \) are as follows:

\[ \gamma_{d,m} = \frac{1}{V} \sum_{v=1}^{V} \exp(log(\beta_{i,w_d,m} + \log(\pi_{d,m,w_d,m})) - \log(\gamma_{d,m}) - 1) \]

(11)

The expression for \( E_q[\log p(w_m | y_m, \beta, \pi, t)] \) can be written as:

\[ E_q[\log p(w_m | y_m, \beta, \pi, t)] = \sum_{m=1}^{M_d} \sum_{i=1}^{K} \sum_{n=1}^{N_d} \lambda_{n,i} \log(\beta_{i,w_d,m} + \log(\pi_{d,m,w_d,m})) - \sum_{m=1}^{M_d} \sum_{i=1}^{K} \sum_{n=1}^{N_d} \lambda_{n,i} \log(\beta_{i,w_d,m} + \log(\pi_{d,m,w_d,m})) + \log(\gamma_{d,m}) - 1 \]

(12)
\[ \lambda_{m,n} \propto \exp \left( \frac{1}{N_d} \sum_{i=1}^{K} \phi_{n,i} \left( \log \beta_{i,w_d,m} + \log \pi_{t,d,m,w_d,m} \right) \right) \]
\[ - \zeta_m^{-1} \sum_{i=1}^{V} \sum_{v=1}^{K} \phi_{n,i} \exp \left( \log \beta_{i,w_d,m} + \log \pi_{t,d,m,w_d,m} \right) \right) \}
\[ \gamma_i = \alpha_i + \sum_{n=1}^{N_d} \phi_{n,i} \] (13)

\[ \zeta_m \geq 0 \] is an additional free variable used in the Taylor expansion of \( \log(x) \) to obtain a tractable second lower bound on the probability of word generation given the topic and tag parameters of the model. Note that \( \zeta_m \) is defined for each document \( d \in D \) and does not need to be initialized in the routines described in subsection 3.3.

### 3.2 Maximum Likelihood Parameter estimation

The expressions for the maximum likelihood of the parameters of the original graphical model using derivatives w.r.t the parameters of the functional \( \mathcal{L} \) are obtained as follows:

For the METag\(^2\)LDA model:

\[ \rho_{ij} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{d,n,i} \delta(w_{d,n}, j) \] (15)
\[ \log \beta_{i,v} = \log \left( \sum_{d=1}^{D} \sum_{m=1}^{M_d} \lambda_{d,m,i} \delta(w_{d,m}, v) \right) - \log \left( \sum_{d=1}^{D} \sum_{m=1}^{M_d} \zeta_{d,m} \lambda_{d,m,i} \exp(\log \pi_{t,d,m,v}) \delta(w_{d,m}) \right) = \log(\text{term}_1^\beta) - \log(\text{term}_2^\beta) \] (16)
\[ \log \pi_{t,v} = \log \left( \sum_{d=1}^{D} \sum_{m=1}^{M_d} \sum_{n=1}^{N_d} \lambda_{d,m,i} \delta(w_{d,m}, v) \delta(t_{d,m}) \right) - \log \left( \sum_{d=1}^{D} \sum_{m=1}^{M_d} \sum_{n=1}^{N_d} \zeta_{d,m} \lambda_{d,m,i} \exp(\log \pi_{t,d,m,v}) \delta(w_{d,m}) \delta(t_{d,m}) \right) = \log(\text{term}_1^\pi) - \log(\text{term}_2^\pi) \] (17)

For the corrMETag\(^2\)LDA model:

\[ \rho_{ij} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{d,n,i} \delta(w_{d,n}', j) \] (18)
\[ \log \beta_{i,v} = \log \left( \sum_{d=1}^{D} \sum_{m=1}^{M_d} \sum_{n=1}^{N_d} \lambda_{d,m,i} \phi_{n,i} \delta(w_{d,m}', v) \right) - \log \left( \sum_{d=1}^{D} \sum_{m=1}^{M_d} \sum_{n=1}^{N_d} \zeta_{d,m} \lambda_{d,m,i} \exp(\log \pi_{t,d,m,v}) \delta(w_{d,m}') \delta(t_{d,m}') \right) = \log(\text{term}_1^\beta) - \log(\text{term}_2^\beta) \] (19)
\[ \log \pi_{t,v} = \log \left( \sum_{d=1}^{D} \sum_{m=1}^{M_d} \sum_{n=1}^{N_d} \lambda_{d,m,i} \phi_{n,i} \delta(w_{d,m}', v) \delta(t_{d,m}') \right) - \log \left( \sum_{d=1}^{D} \sum_{m=1}^{M_d} \sum_{n=1}^{N_d} \zeta_{d,m} \lambda_{d,m,i} \exp(\log \pi_{t,d,m,v}) \delta(w_{d,m}') \delta(t_{d,m}') \right) = \log(\text{term}_1^\pi) - \log(\text{term}_2^\pi) \] (20)

where \( \delta(x) = 1 \) iff \( x = y \) and 0 otherwise and \( t' \in \{1, \ldots, T\} \).

Since the updates for \( \beta \) and \( \pi \) are unconstrained, a Gaussian regularizer with 0 mean and constant standard deviation (set to 2 in this paper) is used for every \( \beta_{i,v} \) and \( \pi_{t,v} \). If \( \beta \) and \( \pi \) are in log space as \( \beta' \) and \( \pi' \), then \( \hat{L}_{i,t} \) is transformed to

\[ \hat{L}_{i,t} = \hat{L}_{i,t} - \frac{1}{2\sigma} \left( \sum_{v=1}^{V} \left( \exp(\beta'_{i,v}) \right)^2 - \frac{1}{2\sigma} \sum_{v=1}^{V} \left( \exp(\pi'_{t,v}) \right)^2 \right) \] (21)

So, in the derivative of \( \hat{L} \) w.r.t \( \log \beta \) or \( \log \pi \) results in a quadratic in \( \exp(\beta'_{i,v}) \) or \( \exp(\pi'_{t,v}) \) as \( \text{term}_1^\beta - \text{term}_2^\beta \exp(\beta') \) \(- \frac{1}{2\sigma} \left( 2 \times \exp(\beta') \right)^2 = 0 \) as a necessary condition for extrema, where \( (\cdot) \) is \( \beta'_{i,v} \) or \( \pi'_{t,v} \). For \( \exp(\beta') \) to be \geq 0, the positive root is taken as the only solution. So the solution becomes (letting \( A = \exp(\beta') \))

\[ 2A = -\sigma^2 \text{term}_1^\beta + \sigma \sqrt{\sigma^2 (\text{term}_1^\beta)^2 + 4 \text{term}_2^\beta} \] (22)

which is \( \geq 0 \). In the derivative of \( \hat{L} \) w.r.t the log of \( \beta \) or \( \pi \), if the regularizer is not used then convergence is not achieved\(^2\). This derivation is different from that used in [20]. Further, while initializing marginal statistics for \( \beta \) and \( \pi \), random initialization works best. A complete derivation of the extrema expressions for the hidden variables and model parameters cannot be shown due to space constraints. Note that number of Lagrange multipliers used in the optimization for \( \phi_{d,n,i} \) is \( N_d \), that for \( \lambda_{d,m,i} \) or \( \lambda_{d,m,n} \) is \( M_d \) and that for \( \rho \) is \( K \). These free (\( \phi, \lambda, \pi \)) and model (\( \rho \)) parameters follow multinomial distributions and hence sum to one.

### 3.3 Algorithms for Implementation

Algorithms 1, 2, 3 and 4 outline some computational procedures for implementing the model and corresponding time complexities (given as \( \mathcal{O}(\cdot) \)). If a procedure is not defined, comments in \{\} explain the functionality of the procedures.

**Algorithm 1 VB EM**

1: if \( \text{algorithm_mode} == \text{“training”} \) then
2: \( \text{initialize\_statistics}() \); \{use seeded initialization for} \( \rho \) and random initialization for \( \beta \) and \( \pi \} \)
3: \( \text{vb\_in\_step}() \);
4: end
5: \( \text{elbo\_prev} \leftarrow 0 \);
6: \( \text{elbo\_current} \leftarrow 0 \); \( \text{iters} \leftarrow 0 \);
7: while \( \text{converged} \geq \text{EM\_CONVERGED} \) do
8: \( \text{elbo\_current} \leftarrow \text{vb\_in\_step}() \) \{update hidden variables\}
9: \( \text{vb\_in\_step}() \) \{update model parameters\}
10: \( \text{converged} \leftarrow (\text{elbo\_prev} - \text{elbo\_current})/\text{elbo\_prev} \)
11: \( \text{elbo\_prev} \leftarrow \text{elbo\_current}; \text{iters} \leftarrow \text{iters} + 1 \)
12: end while \{\( \text{O}(\text{iters} \times \text{vb\_in\_step} + \text{vb\_in\_step}) \} \}

### 4. RESULTS AND DISCUSSIONS

This section shows the relative performances of the proposed models on DUC2005, Wikipedia and the Amazon Review (AR)\(^7\) datasets (see subsection 1.1). The AR dataset was further processed to extract not more than 400 reviews per category. The reviews belong to 25 category labels including {appliance, software, magazine, food, etc.}. The Wikipedia documents were crawling using the special export url\(^3\) mostly along the categories of {food, animal, countries,cities,}\(\text{Authors thank Jordan Boyd-Graber for the hint on using regularizers\}}\)

\[^3\text{http://en.wikipedia.org/wiki/Special:Export\#/Wiki_article_name}\)
The TagLDA model [20] was not compared for this dataset since the concept of multiple GSRts at the word level breaks down for TagLDA. However, empirically it is seen that the nature of DL tags influences the predictive power of the proposed TagLDA models vs. TagLDA. For the DUC05 dataset, the DL tags were represented by coherence markers like “subj” etc. as in [2]. Typically this type of marker groups variations like “landslide:subj,” “car:subj” etc. under a common “subj” abstraction. On the other hand, words like landslide and car signify concepts that allow for identifying specific centers in coherent sentences. In this respect, the WL perspective is more important (primary) over the more abstract DL perspective, the latter capturing a coarser notion of document level coherence. The counts of document level GSRs in the form “GSR→GSR” do not allow for much variance to be exhibited by the documents at the DL perspective. This fits TagLDA better to the dataset at the cost of either ignoring abstract coherence markers altogether or discarding WL perspective and choosing only one coherence marker per word at WL annotation. However, if the document level GSRs are in the form “word:GSR→GSR”, then the proposed models fit the data much better than TagLDA owing to the variance in the DL observations that are captured nicely in the topics along with the WL variations. The GSRs in the latter case cannot be considered as secondary to the WL perspective for document representation. Figures 3a, 3b, 4a and 4b also show that in the presence of decent variations in DL tags, the corrMETag  model performs the best in terms of both
training ELBO and test ELBO. The meaning of correspondence in terms of the bag-of-words model is to find important associations where the first word comes from a document and the second from DL tags in the same document. Table 6 shows some word correspondences that were obtained on test documents from Wikipedia (see rows with $\lambda_{m,n}$). For the Wikipedia dataset, the mixture Multi-Multinomial (Exponential) class of models: MMLDA[16, 17] and METag$^2$LDA (fig. 1e) perform worst. TagLDA[20] performs a little better. This trend is seen on both the training and test sets. Note however that METag$^2$LDA does a simultaneous joint and conditional modeling of DL and WL tags w.r.t. the document’s words. Thus it, along with corrMMLDA, captures what MMLDA, corrMMLDA and TagLDA individually misses out. The ELBO trends of the correspondence class of LDAs are quite similar, with corrMETag$^2$LDA beating corrMMLDA. Again, this trend is seen on both the training and test sets. For the AR dataset, the corrMETag$^2$LDA model beats all other models convincingly in both the training and test set perplexities. The perplexity of MMLDA is the highest during training, followed by (supervised) sLDA[5] and the predictive power of sLDA decreases even further on the test set. In the AR dataset, TagLDA performs much better due to less variability in DL tags. The proposed corrMETag$^2$LDA combines the best of TagLDA and corrMMLDA to achieve the best predictive power on the AR dataset consisting mostly of very short review documents.

Table 6 shows some topics from Wikipedia dataset corresponding to the best performing corrMETag$^2$LDA model. Note that “positional facets” of topics 175 and 196 have been collapsed for space limitations. The test documents for these collapsed topics were the Wikipedia articles on “galaxy” and “fog”. The learned $\beta$ parameters contributing marginally to word generation are listed for the collapsed topics. Top suggested tags from image captions for the test documents also appear as the re-weighted $\rho$ topic multinomial over all DL tags after document inference. Top correspondence tuples are listed as $\lambda_{m,n}$. For topics 54 and 76, notice how there is a “drift” from the beginning sections of the Wikipedia articles to the end sections. Words like “University Press, ISBN” have high mass on the “Middle_to_end” and “End” facets of the topic 76. The image labels for topics are obtained from an inverted index of DL tags to thumbnailed image files. From figures 3 and 4, 200 topics are good fits to both the Tag$^2$ topic models for both Wikipedia and AR datasets. Qualitative conditional topic distributions from the AR dataset (like the ones in table 6) are not shown in the paper due to space constraints. The nature of WL affect tags, though, needs some mention. The assignment of affect tags to review words based on maximum valence score do indeed make them orthogonal and one might choose not to use them at all in the modeling process. However, including such orthogonality do have some good uses. The conditioning of topics on the WL tags allows us to discover terms that might be related to tagged words through shared topics. For example, words like “advertisements,” “ads,” “listing,” etc. that do not appear in [9] could receive higher probability mass for some topic (e.g. magazines) conditioned on “MELANCHOLIC” affect while it could receive higher probability mass for other topic (e.g. software) conditioned on “CONTENDED” affect thereby introducing a relaxation over orthogonal WL tagging constraints.

4.2 Automatically Evaluating Suggested Tags From Image Captions
For each of the Wikipedia test documents, top 5 predicted tags (coming from image captions and article names) were
chosen. Following [14], the method described in section 1.3 was chosen as a quantitative measure of tag suggestion success. Figure 5a shows the relative values of the proposed TagLDA models for macro averages of maximum of best path distance scores for all test documents. Fig. 5a suggests that people ignore the specific contents of the documents while assigning a category label. The METagLDA model, in spite of higher perplexity, performs a little better here because of the lack of specificity of suggested DL tags to the document contents. This shows that humans assign DL tags that belong to higher levels of abstraction. Nevertheless, the best DL tags suggested by both METagLDA and corrMETagLDA are only within 1 to 2 hops away from the ground truth tags based on a chosen WordNet ontology. Thus image captions in Wikipedia articles provide powerful clues for suggesting document tags. Table 7 shows “explanations” of the suggested DL tags to the ground truth category labels which is a desirable output of this type of evaluation. One could also use cross-document evidence trails[18] to measure semantic relatedness. It is to be noted here that the ontology chosen must be specific to the type of evaluation. One could also use cross-document evidence trails while assigning a category label. The METagLDA suggests that people ignore the specific contents of the documents in docsets. To validate the discoveries the following experiment was devised: For each docset in the DUC05 data, all entity pairs that were co-occurring in a sentence were counted and was treated to be a baseline measure of coverage for entity pairs that are related. Then a set of best topics were inferred for the documents in docsets by the TagLDA models. For each topic set, 2450 (=50x50-50) PERSON Named Entity-pair coverage ratio to baseline from DUC05 Newswire data.

### 4.3 Automatically Evaluating Named Entity Relationship Discovery

For the DUC2005 data, the second column in the first, third and fifth rows of table 8 show selected PERSON entity pairs that were discovered to be related through some latent topics. The WL and DL tags for this purpose were named entity classes and GSRts. The first column in the first, third and fifth rows shows queries that serve as gists of the three docsets. To validate the discoveries the following experiment was devised: For each docset in the DUC2005 data, all entity pairs were inferred for the documents in docsets by the TagLDA class of models. For each topic set, 2450 (=50x50-50) PERSON named entity-pair coverage ratio to baseline from DUC05 Newswire data.

<table>
<thead>
<tr>
<th>Table 6: Topics and correspondences from the corrMETagLDA for the Wikipedia data for ( K = 200 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
</tr>
<tr>
<td>(b)</td>
</tr>
</tbody>
</table>

**Figure 5:** (5a) The best ontological inverse path length measure between suggested DL tags from image captions and ground truth Wikipedia categories for the test set in fig. 3b and (5b) PERSON Named Entity-pair coverage ratio to baseline from DUC05 Newswire data.
that John_Nash, Nash and Dr._Nash were treated as three separate entities. The graph in fig. 5b shows that the correspondence model is 3 times better than the robust baseline at the right number of fitted topics and using the abstract CSIR DL perspective. The second, fourth and sixth rows of table 8 show how topical context ties two entities together even though they do not occur in the same sentence. Rows 7, 8 and 9 show ORGANIZATION facet, LOCATION facet and marginal topic corresponding to the best topic for docset "VW/GM Industrial Espionage".

5. CONCLUSION

This paper proposes novel extensions to joint models of text and document level tags that makes use of conditional word level tags to better capture annotated document structure. The proposed Multinomial-Exponential TagLDA models capture semantics of documents with domain knowledge coming from two different perspectives. The correspondence models also show impressive predictive power for inferring topics. Further, usefulness of the models have been explored with applications that provide deep insights into the data. Overall, it is possible to add domain knowledge from different perspectives, into topic models without sacrificing predictive power. Thus supervised models of data annotation can be made better without any intervention from topic models and still aid the latter in improving posterior inference. Adding more than one word level tag to each word or phrase, each being generated from a different perspective, is an important direction of research.

6. REFERENCES