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

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

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.

Word level tags	**Image captions**	**Category labels**
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 Macmillan/2 ISBN/2 0-333-47494-5/2	Syringa josikaea, Syringa vulgaris shrub in flower, etc.	Syringa, Garden plants, Flowers, Shrubs

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*→–) 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 foci of attentions are based not just on entities. *Thus, the document level perspective for DUC05 newswire data is that of syntactic coherence*. The proposed models are general enough to model documents arising out of such perspectives also.

Word level tags	**Document level tags**
Some 167/NE-NUM people were arrested in the US/NE-LOC , including a senior executive of Columbia/NE-LOC 's national bank .	–→ne, ne→–, –→subj, subj→subj, nn→vb, vb→vb, –→adj, adj→–, etc.

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

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

Word level tags	**DL tags**
What I like/CONTENDED is the exceptional zooming without loss/MELANCHOLIC in clarity.	compact camera, Ikon 550, 18X zoom {*Rating: 4.0*}

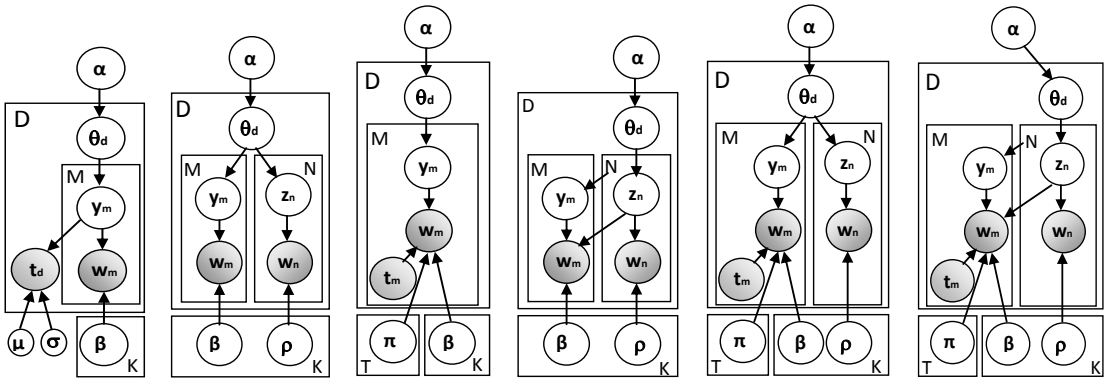
Table 3: Document with word level emotion tags and document level product feature tags

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[1]. 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 GSRts. 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-

[1] http://nlp.stanford.edu/software/corenlp.shtml



(a) supervised LDA model (b) MMLDA model (c) TagLDA model (d) corrMMLDA model (e) METag^2 LDA model (f) corrMETag^2 LDA model

Figure 1: Graphical model representations of one supervised topic model, two existing tag topic models, one extended tag topic model and two new tag squared topic models

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^{2}LDA) models: METag^{2}LDA model (ME is abbreviated form of Multinomial Exponential) (fig. 1e) and corrMETag^{2}LDA 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^{2}LDA 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 tagging, since, then during inference, there is nothing to *infer* on position "states" - they are implicit in any document.

Model Highlights	sLDA	MM LDA	Tag LDA	corr MM LDA	ME Tag^2 LDA	corr ME Tag^2 LDA
Generate words and DL tags from same topic?	×	√	×	√	√	√
Suggest related DL tags?	×	√	×	√	√	√
Associate words to DL tags probabilistically?	×	×	×	√	×	√
Decompose topics along WL tag dimensions?	×	×	√	×	√	√
Find topical-WLtag orientation of new documents ?	×	×	√	×	√	√
Document "label" prediction?	√	×	×	×	×	×

Table 4: Model features and their comparison

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 $N \times C$ 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 pugged 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 docsets were collected. For a particular proposed Tag^{2}LDA 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 docsets. 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 discreet 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^{2}LDA 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^{2}LDA 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^{2}LDA

Symbol	**Meaning** (*r.v. = random variable*)
D	total number of documents
N	total number of unique document level tags per document $d \in D$
M	total number of unique words per document $d \in D$
$\boldsymbol{\alpha}$	r.v. for Dirichlet prior for the document level topic proportions
$\boldsymbol{\theta_d}$	r.v. for document level latent topic proportions
$\boldsymbol{\rho}$	r.v. for corpus level topic-DL_tag multinomial
$\boldsymbol{\beta}$	r.v. for corpus level (marginal in figs. 1c, 1e and 1f) topic-word distribution
$\boldsymbol{\pi}$	r.v. for corpus level marginal tag-word distribution
$\mathbf{z_n}$	indicator variable for DL topic proportion
$\mathbf{y_m}$ in figs. 1b, 1c and 1e	indicator variable for DL topic proportion
$\mathbf{y_m}$ in figs. 1d and 1f	indicator variable for DL tag correspondence
$\mathbf{w_n}$	r.v. for DL tag at position n; vocabulary size $corrV$
$\mathbf{w_m}$	r.v. for word at position m; vocabulary size V
t_m in figs. 1c, 1e and 1f	r.v. denoting tag at position m, on which word w_m is conditioned; vocabulary size T
t_d in fig. 1a	r.v. denoting observed response for document d
μ, σ in fig. 1a	r.v.s denoting mean and standard deviation for the observed response for document d[5]

Table 5: Symbols used in this paper and their meaning

models, $p(w_m|C = i, \boldsymbol{\beta}, \boldsymbol{\pi}, t_m)$, where $C = y_m$ or $C = z_{y_m}$, is not a simple topic multinomial anymore, but is distributed as

$$p(w_m|C = i, \boldsymbol{\beta}, \boldsymbol{\pi}, t_m) = \frac{\exp(\log \beta_{i,w_m} + \log \pi_{t_m,w_m})}{\sum_{v=1}^{V} \exp(\log \beta_{i,w_m} + \log \pi_{t_m,w_m})} \quad (1)$$

Note that each $\boldsymbol{\pi_t}$ 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 $y_m \sim Unif(N_d)$, short for Uniform distribution, as in [4]. The generative processes for the proposed Tag^{2}LDA models are illustrated below:

For each document $d \in 1, ..., D$
  Choose a topic proportion $\boldsymbol{\theta}|\boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha})$
  For each "document level" position n in document d
    Choose topic indicator $z_n|\theta \sim Mult(\boldsymbol{\theta})$
    Choose a "document level" tag $w_n|z_n = k, \boldsymbol{\rho} \sim Mult(\boldsymbol{\rho}_{z_n})$
  For each "word level" position m in document d
    Choose $y_m \sim Unif(\mathbf{1}, ..., \mathbf{N})$ (for corrMETag^{2}LDA - fig. 1f)
    or Choose $y_m|\theta \sim Mult(\boldsymbol{\theta})$ (for METag^{2}LDA - fig. 1e)
    Choose a word $w_m|y_m, \mathbf{z}, \mathbf{t}, \boldsymbol{\beta}, \boldsymbol{\pi} \sim p(w_m|z_{y_m}, \boldsymbol{\beta}, \boldsymbol{\pi}, t_m)$ (fig. 1f)
    or Choose a word $w_m|y_m, \mathbf{t}, \boldsymbol{\beta}, \boldsymbol{\pi} \sim p(w_m|y_m, \boldsymbol{\beta}, \boldsymbol{\pi}, t_m)$ (fig. 1e)

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 docset on "Global warming" which will tie together words like planet and ice based on co-occurrence alone. Consider another docset concerning discovery of ice on Pluto's surface. Thus, if the GSR of "ice" is taken to be a subject, then a "Global warming" topic can include Pluto as a probable word because of the GSRts for the word "ice" that involve "subj". This type of tagging is beneficial where the task is not to replicate the docset structure based on co-occurrence but to provide deeper insights into data for tasks such as summarization, relationship extraction, etc. This effect is hardly observed when the DL tags tersely summarize the main contents of the documents.

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 approximate 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 $\boldsymbol{\gamma}$, $\boldsymbol{\phi}$ and $\boldsymbol{\lambda}$ is imposed as

$$q(\boldsymbol{\theta},\mathbf{z},\mathbf{y}|\boldsymbol{\gamma},\boldsymbol{\phi},\boldsymbol{\lambda}) = \prod_{d=1}^{D} q(\boldsymbol{\theta_d}|\boldsymbol{\gamma_d})[\prod_{n=1}^{N_d} q(z_{d,n}|\boldsymbol{\phi_{d,n}}) \prod_{m=1}^{M_d} q(y_{d,m}|\boldsymbol{\lambda_{d,m}})]$$

and then optimal values of free variables and parameters are found by optimizing the lower bound on $\log p(\mathbf{w_m}, \mathbf{w_n}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\pi}, \mathbf{t})$. The variational functional to optimize can be shown to be (as in [3])

$$\begin{aligned}\mathcal{F} &= E_q[\log p(\mathbf{w}_M, \mathbf{w}_N, \boldsymbol{\theta}, \mathbf{z}, \mathbf{y}|\boldsymbol{\alpha}, \beta, \boldsymbol{\rho}, \boldsymbol{\pi}, \mathbf{t})] \\ &\quad - E_q[\log q(\boldsymbol{\theta}, \mathbf{z}, \mathbf{y}, |\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\lambda})] \end{aligned} \tag{2}$$

where $E_q[f(.)]$ is the expectation of $f(.)$ over the q distribution and $\mathcal{F}$ 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, $\boldsymbol{\phi}$ to be free parameters of the variational DL_tag-topic distribution and $\boldsymbol{\lambda}$ 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, $\log(x) \leq \zeta^{-1}(x) + log(\zeta) - 1, \forall \zeta > 0$, the ELBO $\mathcal{F}$ is changed to further lower bounds $\mathcal{L}$ for the two models.

For the METag^{2}LDA model:

$$\begin{aligned}\mathcal{L}_{MM} &= E_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + E_q[\log p(\mathbf{z_n}|\boldsymbol{\theta})] + E_q[\log p(\mathbf{w_n}|\mathbf{z_n}, \rho)] \\ &\quad + E_q[\log p(\mathbf{y_m}|\boldsymbol{\theta})] + E_q[\log p(\mathbf{w_m}|\mathbf{y_m}, \beta, \boldsymbol{\pi}, \mathbf{t})] \\ &\quad - E_q[\log q(\boldsymbol{\theta}, \mathbf{z}, \mathbf{y}, |\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\lambda})] \end{aligned} \tag{3}$$

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

$$\begin{aligned}E_q[\log &p(\mathbf{w_m}|\mathbf{y_m}, \beta, \boldsymbol{\pi}, \mathbf{t})] \geq \\ &\sum_{m=1}^{M_d}\sum_{i=1}^{K} \lambda_{m,i}(\log \beta_{i,w_{d,m}} + \log \pi_{t_{d,m},w_{d,m}}) \\ &- \sum_{m=1}^{M_d}\{\zeta_{d,m}^{-1}(\sum_{i=1}^{K}\sum_{v=1}^{V} \lambda_{m,i} \exp(\log \beta_{i,w_{d,m}} \\ &+ \log \pi_{t_{d,m},w_{d,m}})) + \log \zeta_{d,m} - 1\} \end{aligned} \tag{4}$$

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

$$\zeta_m = \sum_{v=1}^{V}\sum_{i=1}^{K} \lambda_{d,m,i} \exp\left\{\log \beta_{i,v} + \log \pi_{t_{d,m},v}\right\} \tag{5}$$

$$\phi_{n,i} \propto \exp\left\{\psi(\gamma_i) - \psi(\sum_{j=1}^{K}\gamma_j) + \log \rho_{i,w_{d,n}}\right\} \tag{6}$$

$$\begin{aligned}\lambda_{m,i} \propto &\exp\{\psi(\gamma_i) - \psi(\sum_{j=1}^{K}\gamma_j) + (\log \beta_{i,w_{d,m}} + \log \pi_{t_{d,m},w_{d,m}}) \\ &- \zeta_m^{-1}\sum_{v=1}^{V} \exp(\log \beta_{i,w_{d,m}} + \log \pi_{t_{d,m},w_{d,m}})\} \end{aligned} \tag{7}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N_d}\phi_{n,i} + \sum_{m=1}^{M_d}\lambda_{m,i} \tag{8}$$

For the corrMETag^{2}LDA model:

$$\begin{aligned}\mathcal{L}_{corrMM} &= E_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + E_q[\log p(\mathbf{z_n}|\boldsymbol{\theta})] \\ &+ E_q[\log p(\mathbf{w_n}|\mathbf{z_n}, \rho)] + E_q[\log p(\mathbf{y_m}|N)] \\ &+ E_q[\log p(\mathbf{w_m}|\mathbf{y_m}, \beta, \boldsymbol{\pi}, \mathbf{t})] - E_q[\log q(\boldsymbol{\theta}, \mathbf{z}, \mathbf{y}, |\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\lambda})] \end{aligned} \tag{9}$$

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

$$\begin{aligned}E_q[\log &p(\mathbf{w_m}|\mathbf{y_m}, \beta, \boldsymbol{\pi}, \mathbf{t})] \geq \\ &\sum_{m=1}^{M_d}\sum_{i=1}^{K}(\sum_{n=1}^{N_d} \lambda_{m,n}\phi_{n,i})(\log \beta_{i,w_{d,m}} + \log \pi_{t_{d,m},w_{d,m}}) \\ &- \sum_{m=1}^{M_d}\{\zeta_{d,m}^{-1}(\sum_{i=1}^{K}\sum_{v=1}^{V}(\sum_{n=1}^{N_d} \lambda_{m,n}\phi_{n,i}) \exp(\log \beta_{i,w_{d,m}} \\ &+ \log \pi_{t_{d,m},w_{d,m}})) + \log \zeta_{d,m} - 1\} \end{aligned} \tag{10}$$

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

$$\zeta_m = \sum_{v=1}^{V}\sum_{i=1}^{K}(\sum_{n=1}^{N_d} \lambda_{m,n}\phi_{n,i}) \exp\left\{\log \beta_{i,v} + \log \pi_{t_{d,m},v}\right\} \tag{11}$$

$$\begin{aligned}\phi_{n,i} \propto &\exp\{\psi(\gamma_i) - \psi(\sum_{j=1}^{K}\gamma_j) + \log \rho_{i,w_{d,n}} \\ &+ \sum_{m=1}^{M_d} \lambda_{m,n}(\log \beta_{i,w_{d,m}} + \log \pi_{t_{d,m},w_{d,m}}) \\ &- \sum_{m=1}^{M_d} \zeta_m^{-1}\lambda_{m,n}[\sum_{v=1}^{V} \exp(\log \beta_{i,w_{d,m}} + \log \pi_{t_{d,m},w_{d,m}})]\} \end{aligned} \tag{12}$$

$$\lambda_{m,n} \propto \exp\{\frac{1}{N_d} + \sum_{i=1}^{K} \phi_{n,i}(\log \beta_{i,w_{d,m}} + \log \pi_{t_{d,m},w_{d,m}}) - \zeta_m^{-1}(\sum_{v=1}^{V}\sum_{i=1}^{K} \phi_{n,i} \exp(\log \beta_{i,w_{d,m}} + \log \pi_{t_{d,m},w_{d,m}}))\} \quad (13)$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N_d} \phi_{n,i} \quad (14)$$

$\zeta_m \geq 0$ 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 ζ_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:

$$\rho_{i,j} \propto \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{j=1}^{corrV} \phi_{d,n,i}\delta(w_{d_n}, j) \quad (15)$$

$$\log\beta_{i,v} = \log\left(\sum_{d=1}^{D}\sum_{m=1}^{M_d} \lambda_{d,m,i}\delta(w_{d_m}, v)\right) - \log\left(\sum_{d=1}^{D}\sum_{m=1}^{M_d} \zeta_{d,m}^{-1}\lambda_{d,m,i} \exp(\log \pi_{t_{d,m},v})\delta(w_{d_m}^v)\right) = \log(term_1^\beta) - \log(term_2^\beta) \quad (16)$$

$$\log\pi_{t,v} = \log\left(\sum_{d=1}^{D}\sum_{m=1}^{M_d}\sum_{i=1}^{K} \lambda_{d,m,i}\delta(w_{d_m}^v)\delta(t_{d,m}^{t'})\right) - \log\left(\sum_{d=1}^{D}\sum_{m=1}^{M_d} \zeta_{d,m}^{-1}\sum_{i=1}^{K}\lambda_{d,m,i} \exp(\log \beta_{i,v})\delta(w_{d_m}^v)\delta(t_{d,m}^{t'})\right) = \log(term_1^\pi) - \log(term_2^\pi) \quad (17)$$

For the corrMETag^{2}LDA model:

$$\rho_{i,j} \propto \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{j=1}^{corrV} \phi_{d,n,i}\delta(w_{d,n}^j) \quad (18)$$

$$\log\beta_{i,v} = \log\left(\sum_{d=1}^{D}\sum_{m=1}^{M_d}(\sum_{n=1}^{N_d} \lambda_{m,n}\phi_{n,i})\delta(w_{d_m}^v)\right) - \log\left(\sum_{d=1}^{D}\sum_{m=1}^{M_d} \zeta_{d,m}^{-1}(\sum_{n=1}^{N_d}\lambda_{m,n}\phi_{n,i}) \exp(\log \pi_{t_{d,m},v})\delta(w_{d_m}^v)\right) = \log(term_1^\beta) - \log(term_2^\beta) \quad (19)$$

$$\log\pi_{t,v} = \log\left(\sum_{d=1}^{D}\sum_{m=1}^{M_d}\sum_{i=1}^{K}(\sum_{n=1}^{N_d} \lambda_{m,n}\phi_{n,i})\delta(w_{d_m}^v)\delta(t_{d,m}^{t'})\right) - \log\left(\sum_{d=1}^{D}\sum_{m=1}^{M_d} \zeta_{d,m}^{-1}\sum_{i=1}^{K}(\sum_{n=1}^{N_d}\lambda_{m,n}\phi_{n,i}) \exp(\log \beta_{i,v})\delta(w_{d_m}^v)\delta(t_{d,m}^{t'})\right) = \log(term_1^\pi) - \log(term_2^\pi) \quad (20)$$

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

Since the updates for β and π 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 $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ are in log space as $\boldsymbol{\beta}^\ell$ and $\boldsymbol{\pi}^\ell$, then $\mathcal{L}_{i,t}$ is transformed to

$$\widehat{\mathcal{L}}_{i,t} = \mathcal{L}_{i,t} - \frac{1}{2\sigma^2}(\sum_{v=1}^{V}\left(\exp(\beta_{i,v}^\ell)\right)^2) - \frac{1}{2\sigma^2}(\sum_{v=1}^{V}\left(\exp(\pi_{t,v}^\ell)\right)^2) \quad (21)$$

So, in the derivative of $\widehat{\mathcal{L}}$ w.r.t $\log\boldsymbol{\beta}$ or $\log\boldsymbol{\pi}$ results in a quadratic in $e^{\beta_{i,v}^\ell}$ or $e^{\pi_{t,v}^\ell}$ as $term_1^{(.)} - term_2^{(.)}\exp((.)) - \frac{1}{2\sigma^2}(2\times[\exp((.))]^2) = 0$ as a necessary condition for extrema, where (.) is $\beta_{i,v}^\ell$ or $\pi_{t,v}^\ell$. For $\exp((.))$ to be ≥ 0, the positive root is taken as the only solution. So the solution becomes (letting $A = \exp((.))$),

$$2A = -\sigma^2 term_2^{(.)} + \sigma\sqrt{\sigma^2(term_2^{(.)})^2 + 4term_1^{(.)}} \quad (22)$$

which is ≥ 0. In the derivative of $\widehat{\mathcal{L}}$ w.r.t the log of β or π, 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 $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$, 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 $\boldsymbol{\rho}$ is K. These free($\boldsymbol{\phi}$,$\boldsymbol{\lambda}$) and model($\boldsymbol{\rho}$) 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 $[\mathcal{O}(.)]$). If a procedure is not defined, comments in {.} explain the functionality of the procedures.

Algorithm 1 VB EM

```
if algorithm_mode == "training" then
    initialize_statistics(); {use seeded initialization for ρ and random initialization for β and π}
    vb_m_step();
end if
elbo_prev ← 0
elbo_current ← 0; iters ← 0
while converged ≥ EM_CONVERGED do
    elbo_current ← vb_e_step() {update hidden variables}
    vb_m_step() {update model parameters}
    converged ← (elbo_prev − elbo_current)/(elbo_prev)
    elbo_prev ← elbo_current; iters ←iters+1
end while [O(iters×(vb_e_step+vb_m_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 {apparel, software, magazine, food, etc.}. The Wikipedia documents were crawled using the special export url[3] mostly along the categories of {food, animal, countries,

[2]Authors thank Jordan Boyd-Graber for the hint on using regularizers

[3]`http://en.wikipedia.org/wiki/Special:Export/<Wiki_article_name>`

Algorithm 2 vb_e_step

zero_initialize_statistics();[$\mathcal{O}$(K.corrV+K.V+T.V)]
precompute_beta_and_pi_row_sums() {precompute $\sum_{v=1}^{V} \exp\{\log \beta_{i,v} + \log \pi_{t,v}\} \forall i \in \{1, ..., K\}$ and $\forall t \in \{1, ...T\}$ in an KxT matrix }[$\mathcal{O}$(K.T.V)]
elbo_current $\leftarrow 0$
for $d = 0$ to D **do**
  doc $\leftarrow$ *corpus* $\rightarrow$ *document_vec* $\rightarrow$ *at*(d)
  elbo_current += doc_e_step(d, doc) {also accumulate $term_1^{\beta}$, $term_2^{\beta}$, $term_1^{\pi}$ and $term_1^{\pi}$ of the marginal statistics for β and $\pi \forall d, w_{d,m}$ and $t_{d,m}$ c.f. eqs. 16, 19, 17 and 20}
end for [$\mathcal{O}$(D(doc_e_step))]
return elbo_current;

Algorithm 3 doc_e_step

$\gamma_{d,i} = \alpha + \frac{(\text{doc}\rightarrow total\_num\_words + \text{doc}\rightarrow total\_num\_corr\_words)}{K}$
$\phi_{n,i} = \frac{1.0}{K}$
$\lambda_{m,i} = \frac{1.0}{K}$ {If model is METag^{2}LDA} {OR} $\lambda_{m,n} = \frac{1.0}{doc\rightarrow unique\_num\_corr\_words}$ {If model is corrMETag^{2}LDA}
elbo_current $\leftarrow 0$; *v_iter* $\leftarrow 0$
while not converged **do**
  update $\zeta_{d,m}$
  update $\phi_{d,n,i}$
  update $\lambda_{d,m,i}$ {If model is METag^{2}LDA} {OR} update $\lambda_{d,m,n}$ {If model is corrMETag^{2}LDA}
  update $\gamma_{d,i}$
  elbo_current $\leftarrow$ *compute_likelihood*() {To compute likelihoods c.f. equations 3 for METag^{2}LDA and 9 for corrMETag^{2}LDA}
  v_iter $\leftarrow$ *v_iter* + 1
end while
return elbo_current; [$\mathcal{O}$(K+KN+KM+v_iter(MK+NK+MK+KN))] for METag^{2}LDA or [$\mathcal{O}$(K+NK+MN+v_iter(MKN+NKM+MKN+KN))] for corrMETag^{2}LDA

sport, war, transportation, natural, weapon, universe and ethnic groups}. The relative positions of the sections were binned into 5 categories which served as WL tags. Standard English stopwords were removed for the Wikipedia data and after processing, it contained $33,261$ unique words and $6,902$ unique DL tags (bag-of-words from image captions and Wikipedia article names). The AR dataset contained 6017 unique words and 4271 unique DL tags from product names and review titles after processing. Tags from the affect lexicon were used as WL tags. For both datasets, words occurring once or more than a thousand times across the entire corpus were also removed. Also note that the main document word vocabulary V and the document level tag vocabulary $corrV$ were processed independently using the same token processing rules but without any correspondence. To compare the proposed models with sLDA[5] on the AR data, all DL and WL tags were discarded for sLDA. Instead, the ratings served as values of the response random variables. For both the datasets the number of topics K were set to $\{20, 50, 100, 200\}$. For the DUC2005 dataset, the DL tags were the GSRts found in respective documents and their counts (see subsection 1.1 for GSRts). Two types of WL tags were considered: word position bins like WL tags

Algorithm 4 vb_m_step

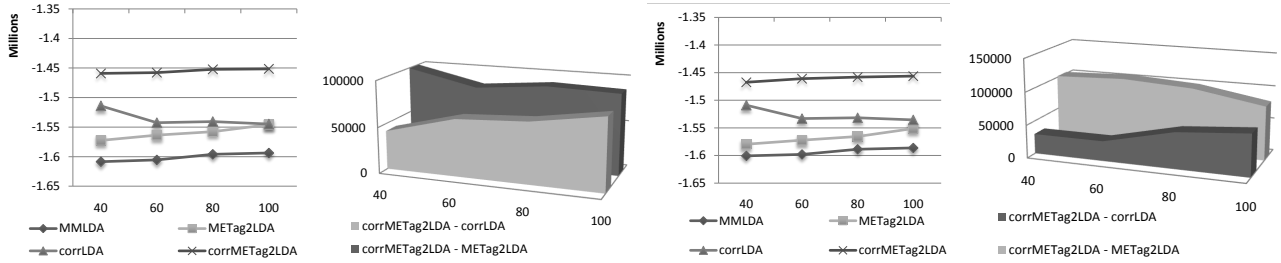
for all $i \in 1, .., K$, $v \in 1, .., V$ and $corr\_v \in 1, ..., corrV$ **do**
  update $\rho_{i,corr\_v}$ from sufficient statistics
  update $\beta_{i,v}$ from marginal statistics
  update $\pi_{t,v}$ from marginal statistics
  update α {Follow the Newton-Raphson method in [6]}
end for [$\mathcal{O}$(K.corrV+K.V+T.V)]

for Wikipedia dataset and named entity classes - PERSON, ORGANIZATION, LOCATION, NUMBER and MISC. Together, DL+WL tagging for DUC2005 data is named as GSRTPos and GSRTNe respectively (see fig. 2). Altogether, there were 36725 unique words for the DUC2005 dataset and 81 corresponding terms which were just the GSRts. K was set to $\{40, 60, 80, 100\}$ for the DUC2005 data based on human intuitions.

4.1 Model Perplexity

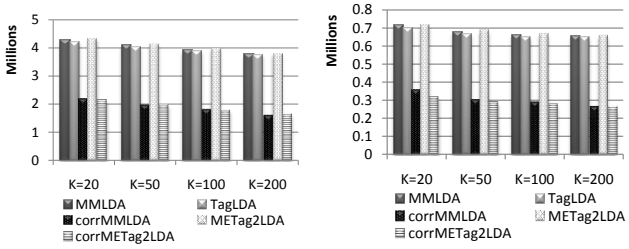
To measure predictive power of METag^{2}LDA and corrMETag^{2}LDA, a 10-fold cross validation was performed on the DUC2005 dataset. Figures 2a and 2c show the ELBO's (higher is better) on validation sets averaged over all folds. Figures 2b and 2d show the **minimum** of the differences in the ELBO's per topic across all folds for the corrMETag^{2}LDA model vs. corrMMLDA and METag^{2}LDA models. Clearly the differences prove that the improved performance of correspondence Tag^2 model is statistically significant. This is also intuitive since words in a document are always generated from a corresponding process, like visualizing an image or action role for a concept. There is a very slightly improved performance when the WL tags are chosen to be named entity classes.

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 Tag^{2}LDA models vs. TagLDA. For the DUC05 dataset, the DL tags were represented by coherence markers like "subj→subj" etc. as in [2]. Typically this type of marker groups variations like "landslide:subj→subj," "car:subj→subj" etc. under a common "subj→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 GSRts in the form of "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 GSRts 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 GSRts 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^{2}LDA model performs the best in terms of both



(a) (b) (c) (d)

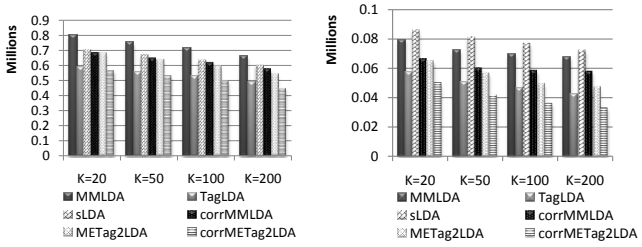
Figure 2: Cross-Validation results on DUC05 newswire data (higher is better in 2a and 2c): (2a) ELBO-Validation DUC05 GSRTNe; (2b) Minimum of differences in ELBO across topics of corrMETag^{2}LDA to corrLDA and METag^{2}LDA for GSRTNe tagging; (2c) ELBO-Validation DUC05 GSRTPos; (2d) Minimum of differences in ELBO across topics of corrMETag^{2}LDA to corrLDA and METag^{2}LDA for GSRTPos tagging



(a) (-)ELBO-Train Wiki (b) (-)ELBO-Test Wiki

Figure 3: Training and test negative ELBO plots of tag topic models on the Wiki data(Lower is better). In each K-group, the models from left to right are MMLDA, TagLDA, corrMMLDA, METag^{2}LDA and corrMETag^{2}LDA

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 $\lambda^{(\cdot)}_{m,n}$s). For the



(a) (-)ELBO-Train AR (b) (-)ELBO-Test AR

Figure 4: Training and test negative ELBO plots of tag topic models on the Amazon Review(AR) data(Lower is better). In each K-group, the models from left to right are MMLDA, TagLDA, sLDA, corrMMLDA, METag^{2}LDA and corrMETag^{2}LDA

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 corrMETag^{2}LDA, 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 $\boldsymbol{\beta}$ 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 $\boldsymbol{\rho}$ 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 the "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 another 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

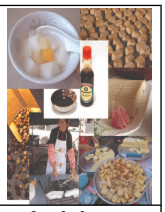
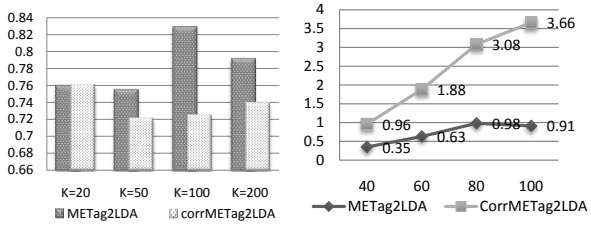
	Beginning →	Beginning To Middle →	Middle →	Middle To End →	End	Tag Suggestions	Correspondence	Image Labels
Topic54 (Artillery:Wars)	Firing, United, artillery, Army, century, guns, support, forced, targets, modern	Firing, artillery, United, guns. Army, targets, forced, century, operated, design, weapons, control	Firing, United, artillery, guns, targets, century, Army, modern, weapons, traditional, field, Battle	United, Firing, Army, weapons, guns, field, power, team, History, rockets, artillery, effective, numbers	targets, attacks, aircraft, design, combatants, radar, modern, small, capable, enemy, Electronic	artillery, guns, howitzer, French, century, Cannon,field, trajectories, PzH, self-propelled	(treating, soldiers) (construction, German) (construction, forge-welded) (postmodernism, Franco-Prussian) (demolitions, captured)	
	$\beta_{54}^{learned}$: Firing artillery guns. targets. United fuzes Army projectile mortars ammunition weapons shells. battery cannon modern							
Topic76 (Tofu:Food)	tofu, Chinese, Japanese, waters, China, Japan, century, origins, similar	tofu, Chinese, Japanese, Soy, Western, products, Asian, meat, important, milk, traditional, flavor	tofu, Chinese, Japanese, food, traditional, Western, cultures, meat, Asian, Korean, Dishes, Japan, soy	tofu, Press, Asian, fries, food, ancient, Western, play, Japanese, Ice, oil, cuisine, fresh, main, America, winter	Press, popular, cultures, China, deep, study, research, traditional, University, America, ISBN	tofu, sliced, China, water, soy, fresh, Provinces, Kong, milked, dishes, press, Hong, solid, soft, island, curds	(beans, dried) (beans, tofu) (suspended, solid) (palm, island) (feeling, sweet) (thinly, sliced) (stir-fries, hot) (canned, soya)	
	$\beta_{76}^{learned}$: tofu soy production Chinese milk firm texture flavor coagulated sauces soft Japanese “dÃšufu” fries Protein cooking fresh beans							
Topic175	$\beta_{175}^{learned}$: galaxy Star spiraled milky matter cluster Hubble gas Universe structure Formation elliptical active galactic nebula dwarf							
	ρ_{175}^{inf} : Galaxy, spiral, stars, Hubble, classification, Andromeda, rings, core, Great, compared							
	$\lambda_{m,n}^{(galaxy)}$: (Planet, Hubble) (Planet, object) (Planet, galaxy) (Herschel, Hubble) (ring, galaxy) (Heat, galaxy) (discoveries, Hubble)							
Topic196	$\beta_{196}^{learned}$: fog air Shadow Ice condensation light vapor Humidity layer temperature freeze particle cool waters moisture evaporation salt							
	ρ_{196}^{inf} : fog, Francisco, San, visible, high, temperature, streets, photo, Bai, lake, California, bridge, air							
	$\lambda_{m,n}^{(fog)}$: (dimensions, high) (beam, visible) (parallel, bridge) (droplets, fog) (combustion, temparature) (invisible, visible) (absorbed, air)							

Table 6: Topics and correspondences from the corrMETag^{2}LDA for the Wikipedia data for $K = 200$

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 Tag^{2}LDA 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 METag^{2}LDA 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 METag^{2}LDA and corrMETag^{2}LDA 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 nature of the task. For example, in medical domain, WordNet is a poor choice for providing explanations to DL tag suggestions.

Some concept pair evidence chains from domain ontology
spiral ↔ curve ↔ line ↔ shape ↔ attribute ↔ abstraction ↔ group ↔ collection ↔ ***galaxy***
weapon ↔ persuasion ↔ communication ↔ act ↔ activity ↔ ***occupation***
french ↔ sculptor ↔ artist ↔ creator ↔ person ↔ ***modern***
soy ↔ legume ↔ herb ↔ vascular_plant ↔ plant ↔ organism ↔ person ↔ inhabitant ↔ Asian ↔ ***Vietnamese***

Table 7: Sample evidence Chains for **DL Tag Suggestions** from image captions to ***actual category labels*** from the Tag^2 topic models



(a) (b)

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

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 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 Tag^{2}LDA class of models. For each topic set, 2450 (=50x50-50)PERSON entity pairs were created out of the highly probable entities appearing in the PERSON facet of the conditional topics. Note that for all entities A and B, two entity pairs (A,B) and (B,A) were created. Each docset had on average 2449 PERSON entity pairs and hence the number 2450. Note

Nobel Prize Winners in Science & Economics	(John_Harsanyi, John_Nash) (Von_Neumann, John_Harsanyi) (Von_Neumann, John_Nash)
Last week the Nobel Prize for Economics was awarded to three 'game theorists': John Harsanyi, John Nash and Rheinhard Selten. What is game theory? Game theory is still a relatively young field. Von Neumann and Oskar Morganstern introduced many of the central ideas in a book published in 1944.	
Women in Parliaments	(Mrs_Margaret_Beckett, Ms_Ann_Taylor) (Mrs_Margaret_Beckett, Ms_Clare_Short) (Mrs_Margaret_Beckett, Ms_Harriet_Harman) (Mrs_Margaret_Beckett, Ms_Hilary_Armstrong) (Mrs_Margaret_Beckett, Ms_Jo_Richardson)
There are at present just four women occupants - Mrs Margaret Beckett, Ms Ann Clwyd, Ms Ann Taylor and Ms Jo Richardson - of the 18 shadow cabinet seats elected each year. The plan now being discussed by the group is to create a 'recommended' list of women candidates. Women would be asked to ensure that they included more than three votes for group members. Beneficiaries might include Ms Harriet Harman, Ms Clare Short, Ms Marjorie Mowlam and Ms Hilary Armstrong.	
VW/GM Industrial Espionage	(Bill_Clinton, Mr_Lopez) (Dorothea_Holland, Bill_Clinton) (Bill_Clinton, Ms_Holland)
It is believed US investigators have asked for, but have been so far refused access to, evidence accumulated by German prosecutors probing allegations that former GM director, Mr Lopez, stole industrial secrets from the US group and took them with him when he joined VW last year. This investigation was launched by US President Bill Clinton and is in principle a far more simple or at least more single-minded pursuit than that of Ms Holland. Dorothea Holland, until four months ago was the only prosecuting lawyer on the German case.	
Topic34(ORG): GM Opel EC General_Motors Harvard volkswagen Justice_Department World_Bank Volkswagen the_Times FBI	
Topic34(LOC): Germany Los_Angeles California UK german Washington Europe Brazil London Slovakia european U.S. New_York	
Topic34: GM Mr_Lopez group yesterday company german week official Mr_Piech work production charge car investigation prosecutor	

Table 8: DUC2005 dataset: Related PERSON named entity pairs and evidence from documents

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 GSRt 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 Tag^{2}LDA 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

[1] Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu. Joint emotion-topic modeling for social affective text mining. *IEEE International Conference on Data Mining*, 0:699–704, 2009.

[2] Regina Barzilay and Mirella Lapata. Modeling local coherence: an entity-based approach. In *Proceedings of the 43rd ACL conference*, pages 141–148, 2005.

[3] Matthew J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[4] David M. Blei and Michael I. Jordan. Modeling annotated data. In *Proceedings of the 26th ACM SIGIR conf.*, pages 127–134, New York, NY, USA, 2003.

[5] David M. Blei and Jon D. Mcauliffe. Supervised topic models. In *NIPS*, volume 21, 2007.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[7] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th ACL conf.*, pages 440–447, Prague, CZ, 2007.

[8] Jordan Boyd-Graber, Jonathan Chang, Sean Gerrish, Chong Wang, and David Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, 2009.

[9] M.M. Bradley and P.J. Lang. Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. 1999.

[10] Hoa Trang Dang. Duc 2005: evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, SumQA '06, pages 48–55. ACL, 2006.

[11] Yansong Feng and Mirella Lapata. How many words is a picture worth? automatic caption generation for news images. In *Proceedings of the 48th ACL conf.*, pages 1239–1249, 2010.

[12] Caimei Lu, Xiaohua Hu, Xin Chen, Jung R. Park, TingTing He, and Zhoujun Li. The topic-perspective model for social tagging systems. In *Proceedings of the 16th ACM SIGKDD conf.*, pages 683–692, 2010.

[13] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *In proceedings of the NAACL conf.*, pages 100–108, Los Angeles, California, 2010.

[14] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. Wordnet::similarity - measuring the relatedness of concepts. In *AAAI*, pages 1024–1025, 2004.

[15] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 EMNLP conf.*, pages 248–256, Singapore, 2009.

[16] Daniel Ramage, Paul Heymann, Christopher D. Manning, and Hector Garcia-Molina. Clustering the tagged web. In *Proceedings of the 2nd ACM WSDM conf.*, pages 54–63, 2009.

[17] Xiance Si and Maosong Sun. Tag-LDA for Scalable Real-time Tag Recommendation. *Journal of Computational Information Systems*, 2009.

[18] Rohini Srihari, Li Xu, and Tushar Saxena. Use of ranked cross document evidence trails for hypothesis generation. In *Proceedings of the 13th KDD conference*, pages 677–686, San Jose, CA, 2007.

[19] Hanna M. Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In *NIPS*, 2009.

[20] Xiaojin Zhu, David Blei, and John Lafferty. Taglda: Bringing document structure knowledge into topic models. *UWisc Technical Report TR-1533*, 2006.