Statistical Entity-Topic Models

David Newman*
Dept. of Computer Science
Univ. of California, Irvine
Irvine, CA 92697
newman@uci.edu

Chaitanya Chemudugunta
Dept. of Computer Science
Univ. of California, Irvine
Irvine, CA 92697
chandra@uci.edu

Padhraic Smyth
Dept. of Computer Science
Univ. of California, Irvine
Irvine, CA 92697
smyth@uci.edu

Mark Steyvers
Cog. Sci.
UC Irvine
Irvine, CA
msteyver@uci

ABSTRACT

The primary purpose of news articles is to convey information about who, what, when and where. But learning and summarizing these relationships for collections of thousands to millions of articles is difficult. While statistical topic models have been highly successful at topically summarizing huge collections of text documents, they do not explicitly address the textual interactions between who/where, i.e. named entities (persons, organizations, locations) and what, i.e. the topics. We present new graphical models that directly learn the relationship between topics discussed in news articles and entities mentioned in each article. We show how these entity-topic models, through a better understanding of the entity-topic relationships, are better at making predictions about entities.

1. INTRODUCTION

News articles aim to convey information about who, what, when and where. Statistical topic models can not distinguish between these different categories and produce topical descriptions that are mixtures of whos, whats, whens and wheres. But in many applications it is important for these different concepts to be explicitly modeled. Thus in this paper we consider the problem of modeling text corpora where documents contain, in addition to ordinary words, additional classes of words (referred to as entities). For our study of news articles, the entities can be persons (e.g. "George Bush"), organizations (e.g. "NFL"), and locations (e.g. "London"). Our focus is on modeling entities and making predictions about entities based on learning that uses entities *and* words.

Many statistical topic models are based on the simple idea that individual documents are made up of one or more topics, where each topic is a distribution over words (e.g. Blei et al.'s Latent Dirichlet Allocation (LDA) model [5]). These models can efficiently describe a collection by a set of topics, retrieve information, classify documents and be used in prediction tasks. There have been several applications showing the power of these models on a wide variety of text collections (e.g. Enron emails, CiteSeer abstracts, Web pages).

In language modeling and information extraction, there is growing interest in finding and analyzing entities mentioned in text. These entities are usually proper nouns, i.e. persons, organizations or locations. We take advantage of recent developments in named entity recognition systems to identify and extract entities mentioned in news articles. Rather than build on research in named entity recognition and entity resolution, we take this line of work in a different direction. We are primarily interested in modeling and making predictions on entities, once they have been identified in the text.

In this paper, we review a series of graphical models that extend LDA to explicitly treat and model entities mentioned in text. We introduce two new models, and compare a total of five different entity-topic models that have fundamentally different generative processes. We demonstrate two primary results: (i) our proposed CorrLDA2 model has better ability to predict entities than LDA, and (ii) words can be leveraged to improve predictions about entities by up to 30%.

2. RELATED WORK

Researchers have furthered Blei et al.'s original LDA model [5] in the directions of algorithm development, applications, and model extensions. In algorithms, Griffiths and Steyvers [10] proposed the now popular Gibbs sampling method for inference; while Minka and Lafferty [15] proposed Expectation Propagation. There has been a wide variety of extensions to the original LDA model including Steyver et al.'s author-topic model [17, 16]; McCallum et al.'s author-topic-recipient model and author-role-topic model [13]; Griffith et al.'s hidden-Markov topic models for separating semantic and syntactic topics [11]; Blei's correlated and hierarchical topic models [4, 2]; and Buntine's PCA models [7].

In this paper we are specifically interested in the intersection of topic modeling and entity modeling, and in this context there are two branches of closely related work. In topic modeling, several researchers have extended basic topic models to include other information (beyond the words) contained in individual text documents. Steyvers et al.'s author-topic model uses a document's authorship information together

*Corresponding author

with the words to learn models that relate authors, topics and documents [17]. Using Gibbs sampling, they applied their Author-Topic model to a collection of CiteSeer abstracts to infer relations between authors, topics and words. Blei and Jordan [3] modeled collections of images and their captions. In this case the caption contained the words, and the image (represented as a set of real-valued features) was the other information. Using the Corel database of captioned images, they used variational EM to estimate parameters, and learn relations between images and text. Erosheva et al. [9] combined paper abstracts with bibliography information to create a mixed-membership model that identifies categories of publications. Also using variational EM, they identified categories for a collection of PNAS biological sciences publications. While this work models words and other objects, it does not specifically address entities.

Different aspects of entity analysis include named entity recognition; entity resolution; and social networks built on entities. Probabilistic modeling approaches have been applied to all of these. For example, McCallum et al. [14] use conditional random fields for noun co-references, and Bhattacharya and Getoor use a modified LDA model for entity resolution [1]. Zhu et al. use non-probabilistic latent semantic indexing to recognize named entities and find relationships between named entities in Web pages [18]. Our focus in the present work is to use simple and effective named entity recognition techniques to extract entities as a preprocessing step prior to our primary goal or relating entities, topics and words.

3. DATA SETS

To analyze entities and topics, we require text datasets that are rich in entities including persons, organizations and locations. News articles are ideal because they convey information about who, what, when and where. Our first data set is a collection of New York Times news articles taken from the Linguistic Data Consortium's English Gigaword Second Edition corpus (www.ldc.upenn.edu). We used all articles of type "story" from 2000 through 2002, resulting in 330,000 separate articles spanning three years. These include articles from the NY Times daily newspaper publication as well as a sample of news from other urban and regional US newspapers. Our second data set are articles from the Foreign Broadcast Information Service (FBIS), which comes from www.fbis.gov or wnc.dialog.com. FBIS articles come from around the globe, and include English translations of a variety of foreign (and open source) news. We used a set of 53,000 FBIS articles spanning Feb 1999 to Nov 2000.

We automatically extracted named entities (i.e. proper nouns) from each article using existing named entity recognition tools. We evaluated two tools including GATE's Information Extraction system ANNIE (gate.ac.uk), and Coburn's Perl Tagger (search.cpan.org/~acoburn/Lingua-EN-Tagger). ANNIE is rules-based and makes extensive use of gazetteers, while Coburn's tagger is based on Brill's HMM part-of-speech tagger [6]. ANNIE tends to be more conservative in identifying proper nouns.

For the NY Times 1 data, entities were extracted using Coburn's tagger. For this 2000-2002 period, the most frequently mentioned people were: George Bush; Al Gore; Bill Clinton; Yasser Arafat; Dick Cheney and John McCain. For the FBIS data, only people entities were extracted using GATE's ANNIE. Locations were omitted because long lists of countries were included at the top of every FBIS article. For this 1999-2000 period, the most frequently mentioned people were: Bill Clinton, Vladimir Putin, Jiang Zemin, Boris Yeltsin, Slobodan Milosevic, Keizo Obuchi, Boris Berezovskiy, and Ehud Barak.

Table 1: Statistics for data sets.

	NY Times 1	NY Times 2	FBIS
Train Docs	330,000	29,000	48,000
Test Docs	-	11,000	5,000
Unique Words	44,000	31,000	38,000
Unique Entities	59,000	10,000	5,700
Total Words	100,000,000	8,000,000	11,000,000
Total Entities	12,000,000	1,000,000	300,000

After tokenization and removal of stopwords, the vocabulary of unique words was filtered by requiring that a word occur in at least ten different news articles. We produced a large NY Times data set containing 330,000 documents, a vocabulary of 44,000 unique words, a list of 59,000 entities, and a total of 112 million word and entity tokens. After this processing, entities occur at the rate of 1 in 8 words (not counting stopwords). For experiments, we used 29,000 docs from July-Sept in 2000 and 2001 for training, and 11,000 docs from July-Sept 2002 for testing, creating an challenging temporal gap between training and testing. The final FBIS dataset was similarly processed, resulting in a total of 11 million word tokens. In FBIS, entities are far sparser, only occurring at the rate of 1 in 40 words. Statistics for the data sets are summarized in Table 1.

4. MODELS

In this section we describe five graphical models for entity-topic modeling. We start with LDA, and follow with more complex models that aim to better fit our multi-class text data of words and entities. Three of the models have been proposed by other researchers; we introduce two new models, namely SwitchLDA and CorrLDA2. Here we introduce some notation used in the graphical models and Gibbs sampling equations listed in the Appendix: D is the number of documents, T is the number of topics, N_d is the total number of words in document d (with $N_d = N_{w_d} + N_{\tilde{w}_d}$, the sum of all the words plus entities), α and β are Dirichlet smoothing parameters, θ is the topic-document distribution, ϕ is the word-topic distribution, z_i is a topic, v_i is a word or entity, while w_i is a word and $\tilde{w}_i$ is an entity. A tilde is used to denote the entity version of that variable.

4.1 LDA

To introduce the notation and explain the differences in the five graphical models, let us start with the LDA model shown in Figure 1(a). LDA's generative process is:

1. For all d docs sample $\theta_d \sim \text{Dir}(\alpha)$
2. For all t topics sample $\phi_t \sim \text{Dir}(\beta)$
3. For each of the N_d words v_i in document d:

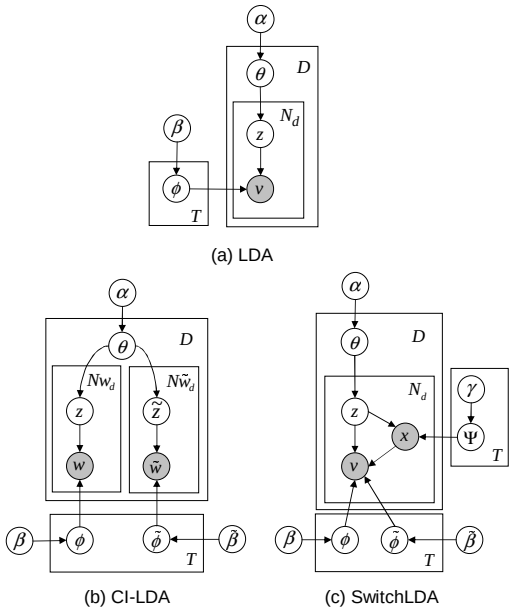

Figure 1: Graphical models for (a) LDA, (b) CI-LDA and (c) SwitchLDA

(a) Sample a topic $z_i \sim \text{Mult}(\theta_d)$

(b) Sample a word $v_i \sim \text{Mult}(\phi_{z_i})$

The learning algorithm for LDA follows the Gibbs sampling approach described in [10]. The learning algorithm for the other four models uses an analogous Gibbs sampling approach. We provide the Gibbs sampling equations for all five models in the Appendix. Note that LDA does not distinguish between words and entities, this distinction is made post-hoc (i.e. not during learning), when we make predictions about entities.

4.2 CI-LDA

The conditionally-independent LDA model (CI-LDA, Figure 1(b)) explicitly makes an a priori distinction between words and entities during learning. It is an obvious modification to the LDA model to handle multiple classes of word tokens, in our case *words* and *entities*. Cohn and Hofmann described a similar model to relate Web pages and their links [8]. The generative process is very similar to LDA's, except that we generate N_{w_d} words and $N_{\tilde{w}_d}$ entities in document d. CI-LDA's generative process is:

1. For all d docs sample $\theta_d \sim \text{Dir}(\alpha)$
2. For all t topics sample $\phi_t \sim \text{Dir}(\beta)$ and $\tilde{\phi}_t \sim \text{Dir}(\tilde{\beta})$
3. For each of the N_{w_d} words w_i in document d:
   (a) Sample a topic $z_i \sim \text{Mult}(\theta_d)$
   (b) Sample a word $w_i \sim \text{Mult}(\phi_{z_i})$
4. For each of the $N_{\tilde{w}_d}$ entities $\tilde{w}_i$ in document d:
   (a) Sample a topic $\tilde{z}_i \sim \text{Mult}(\theta_d)$
   (b) Sample an entity $\tilde{w}_i \sim \text{Mult}(\tilde{\phi}_{\tilde{z}_i})$

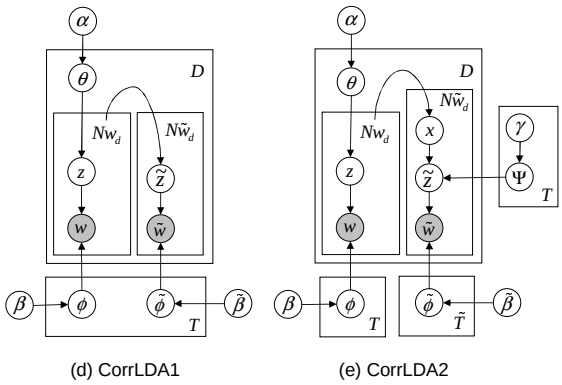

Figure 2: Graphical models for (d) CorrLDA1 and (e) CorrLDA2

4.3 SwitchLDA

One issue with CI-LDA is that it is not truly generative; every document d contains some arbitrary number of words N_{w_d} and entities $N_{\tilde{w}_d}$. It is more satisfying to have this distribution of entities in a document be part of the process itself. In our proposed SwitchLDA model (Figure 1(c)) we include an additional Binomial distribution ψ (with a Beta prior of γ) which controls the fraction of entities in topics. The generative process for SwitchLDA is:

1. For all d docs sample $\theta_d \sim \text{Dir}(\alpha)$
2. For all t topics sample $\phi_t \sim \text{Dir}(\beta)$, $\tilde{\phi}_t \sim \text{Dir}(\tilde{\beta})$ and $\psi_t \sim \text{Beta}(\gamma)$
3. For each of the N_d words v_i in document d
   (a) Sample a topic $z_i \sim \text{Mult}(\theta_d)$
   (b) Sample a flag $x_i \sim \text{Binomial}(\psi_{z_i})$
   (c) If (x_i=0) sample a word $v_i \sim \text{Mult}(\phi_{z_i})$
   (d) If (x_i=1) sample an entity $v_i \sim \text{Mult}(\tilde{\phi}_{z_i})$

The Gibbs sampling equations for CI-LDA and SwitchLDA are given in appendices A.2 and A.3. Note that the Gibbs sampling for SwitchLDA is very similar to that for LDA; in fact as the Beta prior $\gamma \to \infty$ the sampling equations become identical. One appealing feature about CI-LDA and SwitchLDA is that they are independent of ordering of the word-token classes. They are also easily generalized to handle n-classes of word tokens (for example, we may explicitly model words, people, organizations and locations).

4.4 CorrLDA1

We have found in practice that CI-LDA's word topics and entity topics can be too decoupled. To force a greater degree of correspondence between word and entity topics we use the CorrLDA1 model in Figure 2(d). This model first generates word topics for a document. Then only the topics associated with the words in the document are used to generate entities, resulting in a more direct correlation between entities and words. This CorrLDA1 model is essentially the same as Blei and Jordan's Corr-LDA model used for Image/Caption data [3]. The generative process for CorrLDA1 is:

1. For all d docs sample $\theta_d \sim \text{Dir}(\alpha)$
2. For all t topics sample $\phi_t \sim \text{Dir}(\beta)$ and $\tilde{\phi}_t \sim \text{Dir}(\tilde{\beta})$
3. For each of the N_{w_d} words w_i in document d:
   (a) Sample a topic $z_i \sim \text{Mult}(\theta_d)$
   (b) Sample a word $w_i \sim \text{Mult}(\phi_{z_i})$
4. For each of the $N_{\tilde{w}_d}$ entities $\tilde{w}_i$ in document d:
   (a) Sample a topic $\tilde{z}_i \sim \text{Unif}(z_{w_1} \ldots z_{w_{N w_d}})$
   (b) Sample an entity $\tilde{w}_i \sim \text{Mult}(\tilde{\phi}_{\tilde{z}_i})$

4.5 CorrLDA2

Finally we introduce the CorrLDA2 model, which is like CorrLDA1 but with word topics including a mixture of entity *topics* (not individual entities), as shown in Figure 2(e). The intuition is that word topics often relate to different groups of entities; say a word topic of sports may contain entity topics of NFL teams, NBA teams, and Baseball teams. Another key difference is that CorrLDA2 allows different numbers of word topics, T, and entity topics, $\tilde{T}$. The generative process for CorrLDA2 is:

1. For all d docs sample $\theta_d \sim \text{Dir}(\alpha)$
2. For all $t = 1 \ldots T$ word topics sample $\phi_t \sim \text{Dir}(\beta)$ and $\psi_t \sim \text{Dir}(\gamma)$
3. For all $t = 1 \ldots \tilde{T}$ entity topics sample $\tilde{\phi}_t \sim \text{Dir}(\tilde{\beta})$
4. For each of the N_{w_d} words w_i in document d:
   (a) Sample a topic $z_i \sim \text{Mult}(\theta_d)$
   (b) Sample a word $w_i \sim \text{Mult}(\phi_{z_i})$
5. For each of the $N_{\tilde{w}_d}$ entities $\tilde{w}_i$ in document d:
   (a) Sample a supertopic $x_i \sim \text{Unif}(z_{w_1} \ldots z_{w_{N w_d}})$
   (b) Sample a topic $\tilde{z}_i \sim \text{Mult}(\psi_{x_i})$
   (c) Sample an entity $\tilde{w}_i \sim \text{Mult}(\tilde{\phi}_{\tilde{z}_i})$

The Gibbs sampling equations for CorrLDA1 and CorrLDA2 are given in appendices A.4 and A.5. Note that the word part of the sampling equations is straightforward LDA. We also point out that for CorrLDA1 and CorrLDA2 there *is* an ordering to the classes of word tokens (for our applications using news articles we treat words as the primary class of tokens). Again, both CorrLDA1 and CorrLDA2 can be generalized to handle more than two classes of tokens.

5. RESULTS

5.1 Experimental Setup

We can use entity-topic models for multiple purposes. Beyond learning topics, they can infer and explain (latent) relationships between entities mentioned throughout a collection of text documents. They can also make predictions about entities outside a collection, based on varying amounts of additional information. Figure 3 shows a schematic of training data Wo and Eo, and test data W and E. In the entity prediction task, the models are first trained on Wo+Eo. The models then make predictions about E, entities in the test set using W, the words in the test set. In the entity-pair classification task, a set of models are trained on Wo+Eo or just Eo. The models then make predictions about whether an entity pair is in E.

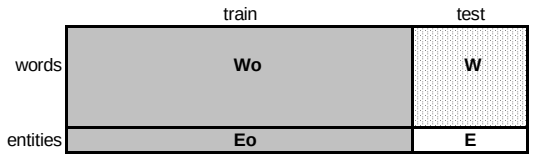

Figure 3: Schematic showing data for experiments

For all of these tasks, single samples were taken – after 400 iterations – from 10 randomly-seeded runs. We found 400 iterations to be sufficient by monitoring in-sample perplexity every 10 iterations, and observing some degree of convergence (i.e. flattening of the log likelihood). Samples were averaged before predicting entities or classifying pairs. We ran T=200 topics for all experiments, and included T=100 topic runs for entity pair classification, and a T=400 topic run for the 330,000-document NY Times 1 data. We determined that Dirichlet priors $\alpha = 0.1$ and $\beta = 0.01$ maximized test set likelihood on the NY Times 2 data, and fixed these values for all experiments.

The time complexity for the Gibbs sampling is $O(N_{tot} \cdot T \cdot Iter)$ and the space complexity is $O(3N_{tot} + T(D+W+E))$, where N_{tot} is the total number of all classes of tokens in the corpus, T is the number of topics, $Iter$ is the number of Gibbs sampler iterations, D is the number of documents, W the number of unique words and E the number of unique entities. The time and space complexity for the first four models is approximately equivalent. The time complexity for CorrLDA2 is slightly larger due to the hierarchy of entity topics.

5.2 Topics

Topics from LDA, CI-LDA, SwitchLDA and CorrLDA1 contain distributions over words, and over entities, while topics from CorrLDA2 contain distributions over words, and over entity *topics*. Three topics from the 3-year NY Times 1 data are shown in Figure 4. The Sept. 11 topic is clearly about the breaking news describing what and where, but not who (i.e. no mention of Bin Laden). The FBI Investigation topic lists 9/11 hijackers Mohamed Atta and Hani Hanjour, while the Harry Potter/Lord of the Rings topic combines these same-genre runaway successes. Selected FBIS topics (Figure 5) include a topic about foreign aid which – because of its general nature – has negligible probability assigned to entities. The N. Ireland and Russia topics have a very specific list of entities below the words. Due to imperfect named entity recognition we see entities such as "Sinn Fein" appear as words. This slippage is not critical. The entity-topic models still learn from these words even when they are not correctly tagged as entities.

Recall that an advantage of CorrLDA2 is that it can *group* related entities, and assign these entity groups to word topics. We illustrate this ability by showing CorrLDA2 topics from the NY Times 2 data in Figures 6 and 7. Five loosely-

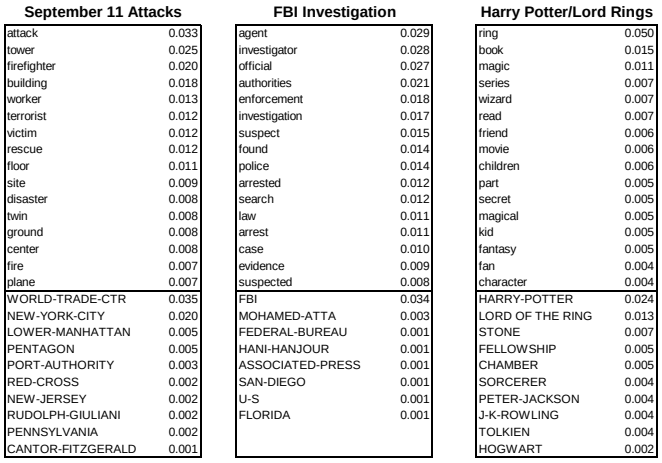

Figure 4: Selected topics from a 400-topic LDA run of the 3-year NY Times data. In each topic we list the most likely words in the topic with their probability, and below that the most likely entities. The title above each box is a human-assigned topic label.

related word topics about Sept. 11 and Washington contain mixtures of five entity topics that span different groups of entities, from ones specifically related to Sept. 11 (World-Trade-Center), to ones related to US security (NSC, CIA). The Computers topic contains just a single entity topic of computer manufacturers. The Arts topic neatly includes two separate groups of entities; one relating to theater, and one relating to music. This feature of CorrLDA2 is nice; why combine theatre and music entities into a single entity topic when they naturally belong in two separate groups?

5.3 Prediction Results

Since we are primarily interested in modeling and making predictions about entities, we evaluated all five models on a specific entity prediction task. In this entity prediction task, the models were first trained on Wo+Eo (Figure 3). The models then make predictions about E, entities in each test set document using some or all of W, the words in each test set document. The likelihood of an entity in an unseen test document is $p(e|d) = \sum_t p(e|t)p(t|d)$, where $p(e|t)$ is learned during training, and the topic mix in the test document $p(t|d)$ is estimated by resampling some or all of the test document words using the saved $p(w|t)$ word distribution. We illustrate this process in Tables 2 and 3 using two examples from the NY Times 2 data. The first table shows an excerpt from a 7/2/02 news article about the Sept 11. attacks. The top box shows an excerpt from the article, with redacted entities indicated by XXXX. Using the model parameters learned in training, the models compute, using all or some of the words in this top box, the likelihood of every possible entity (10,000 entities for NY Times, 5,700 entities for FBIS). The bottom box lists these predicted entities in order of likelihood, and matches with actual entities are underlined. We then determine from the list of actual entities, the highest (best) ranked, lowest ranked and median rank. For this example, the top predicted entity ("fbi") is an actual entity, so the best rank is 1. These best and median ranks are averaged over all the documents in the training set (11,000 docs for NY Times, 5,000 docs for FBIS).

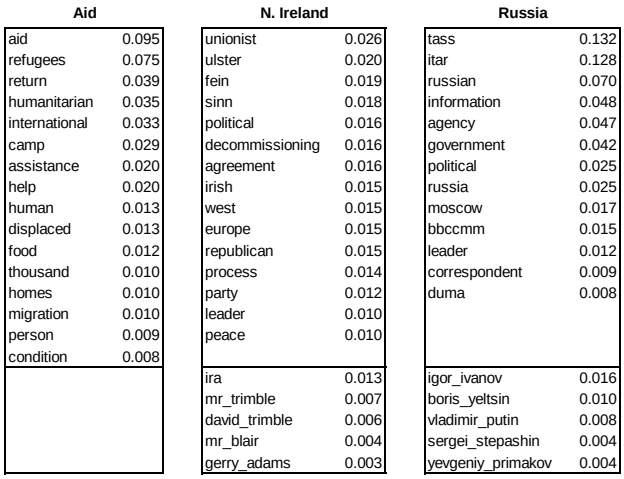

Figure 5: Selected topics from a 200-topic LDA run of the FBIS data. The Aid topic has negligible probability assigned to entities.

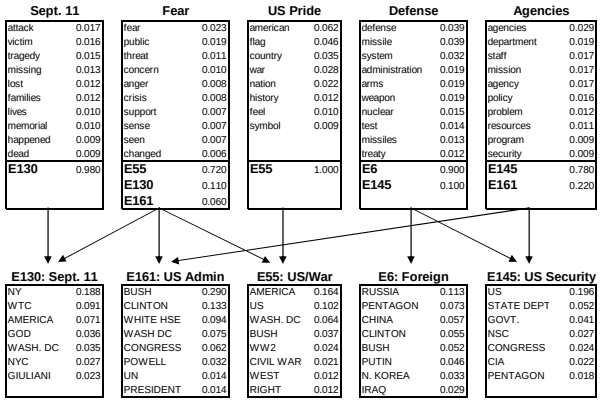

Figure 6: Sept. 11 and Washington-related word topics and entity topics from a 200-topic CorrLDA2 run of the 6-month NY Times 2 data. The word topics include a mix of entity *topics* (not entities). The lower level shows the entities in each entity topic.

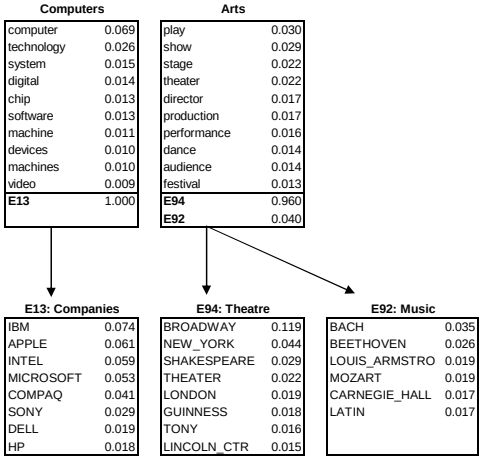

Figure 7: Computer and Arts-related word topics and entity topics from a 200-topic CorrLDA2 run of the 6-month NY Times 2 data.

Table 2: Predicting entities in a 7/2/02 Cox Newspapers article (NY Times News Service). The top box shows an excerpt from the article, with redacted entities indicated by XXXX. Below that we alphabetically list (for evaluation purposes) all the entities that are mentioned in the article. The bottom shows the most likely entities predicted by the model, with matches underlined.

words: XXXX, charged as a conspirator in the Sept. 11 terrorist attacks, called XXXX "my brother in Islam" and "my father in XXXX," in hand-written motions released this week. But the 34-year-old French citizen, who could be sentenced to death, also claimed innocence in the terrorist plot. "I am a mujahideen, if XXXX accept me. I am a terrorist in your eyes. But it does not mean that I took part in Sept. 11 ...
actual entities: afghanistan allah britain darwin eunice-moscoso europe fbi federal france germany god hamburg jihad minnesota osama-bin-laden pennsylvania pentagon u-s u-s-district-court united-flight united-states virginia world-trade-center zacarias-moussaoui
predicted entities: fbi united-states afghanistan u-s taliban pakistan washington osama-bin-laden america federal new-york pentagon

Table 3: Predicting entities in a 7/2/02 Boston Globe article (NY Times News Service). The three boxes are described in Table 2.

words: Shares of XXXX slid 8 percent, or $1.10, to $12.65 Tuesday, as major credit agencies said the conglomerate would still be challenged in repaying its debts, despite raising $4.6 billion Monday in taking its finance group public. Analysts at XXXX Investors service in XXXX said they were keeping XXXX and its subsidiaries under review for a possible debt downgrade, saying the company "will continue to face a significant debt burden," with large slices of debt coming due, over the next 18 months. XXXX said ...
actual entities: fitch goldman-sachs lehman-brother moody morgan-stanley new-york-stock-exchange standard-and-poor tyco tyco-international wall-street worldcom
predicted entities: wall-street new-york nasdaq securities-exchange-commission sec merrill-lynch new-york-stock-exchange goldman-sachs standard-and-poor

Table 4: Entity prediction results for NY Times.

model	avg best rank	avg median rank
LDA	19.4	435.2
CI-LDA	19.4	433.5
SwitchLDA	18.3	433.7
CorrLDA1	18.6	419.5
CorrLDA2	18.1	417.5

Table 5: Entity prediction results for FBIS.

model	avg best rank	avg median rank
LDA	136.9	220.1
CI-LDA	173.7	291.1
SwitchLDA	135.0	221.8
CorrLDA1	163.3	278.9
CorrLDA2	153.1	262.3

Our proposed CorrLDA2 model gives a 7% improvement in average best rank, and a 4% improvement in average median rank over the standard LDA model for the NY Times 2 data. This average is computed over 11,000 test documents. The remaining three models fall in between LDA and CorrLDA2 (Table 4). Note that random guessing would produce an average median rank of 5000 (since there are a total of 10,000 unique entities). Articles contain on average 18 different entities, so a median rank around 400 seems reasonable (relative to an average best rank of 20). This slight improvement of CorrLDA2 over LDA is seen even when we only partially observe the words in the test document. Figure 8 shows that when just 8, 32 and 128 randomly selected words are chosen from each document, CorrLDA2 consistently produces better entity rankings than LDA (note that documents contain on average 300 words).

The results using the FBIS data in Table 5 give a different result. SwitchLDA and LDA are very similar and produce the best ranking results. The CI-LDA model does the worst. CorrLDA2's ability to group entities into a two-level hierarchy is perhaps overkill given the relative sparsity of entities in the FBIS data (1 entity for every 40 words)

5.4 Classification of Entity Pairs

It is useful to be able to compute the likelihood of a pair of entities co-occurring in future documents, particularly if they have previously never been seen together [12]. Our entity-topic models can infer relationships between entities, even when those entities never appear together in any document. We measure this relationship using the entity-entity affinity, defined as $p(e_i|e_j)/2 + p(e_j|e_i)/2$, where $p(e_i|e_j) = \sum_t p(e_i|t)p(t|e_j)$ is computed from the learned model parameters.

For this experiment, we generate two sets of entity pairs. The first set ("true pairs") contains pairs that were never seen in any training document, but were seen in test documents. Some examples include breaking news stories in 2002 about Martha Stewart & ImClone; and Jack Grubman & WordCom (Table 6); both these pairs relate to events that occurred after the 2000-2001 period containing the training documents. The second set ("false pairs") contains pairs

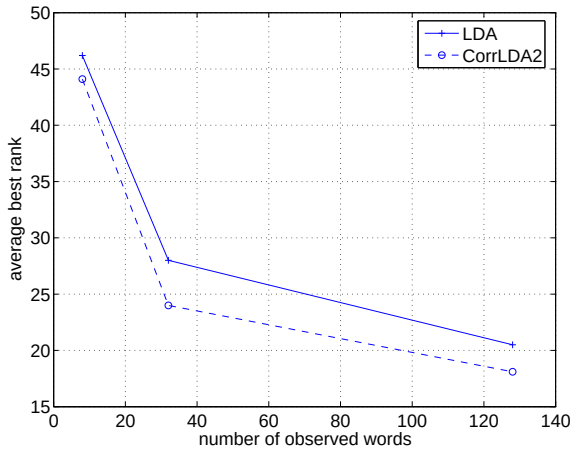

Figure 8: Average best rank versus number of observed words for NY Times 2 data.

that were never seen in any training or test document. To control for entity frequency, the false pairs set is generated from the true pairs set, by a random permutation of the second entity in each pair. So while the pair Tony Blair & Johnny Adair (loyalist leader), is seen in the FBIS test documents, the pair Tony Blair & Francoise de Panafieu (city councilor in Paris) never appear together in any of our FBIS documents.

The N true pairs and N false pairs are combined into one list, and we compute the (e_1, e_2) affinity for each of the $2N$ pairs, and save the median value. Given the equal numbers of true and false, we classify as true, pairs whose affinity is above the median, and classify as false, pairs whose affinity is below the median. For the NY Times 2 data, $N = 2000$ and for the FBIS data, $N = 200$ (the lower FBIS number is due to the much greater sparsity of entities in FBIS which create fewer pairing possibilities).

An important question is whether the text (i.e. the non-entity words) available during training improves classification accuracy. For both the NY Times and FBIS data, we trained the LDA model using (i) both words and entities, and (ii) just the entities. The prediction accuracy for T=100 and T=200 topics (Table 7) show several results: all models and data sets do better than random guessing; using words & entities improves accuracy by 3% to 30% over using just entities; and 100 topics gives uniformly better accuracy than 200 topics. The dependency on number of topics is a simple case of overfitting. The accuracy improvement (by up to 30%) by training on words and entities is a powerful result – it tells us that words that tend to co-occur with entities (and therefore topics) help us better understand and model these entities. In the FBIS dataset, where entities made up only 2.5% of the written words this improvement was the most dramatic. This can be explained by the fact that there is relatively limited information in the entities alone; the 40-fold boost in data from including words – even ones that are not relevant – clearly drives the ability to topically characterize an entity and therefore make better predictions about the connection of that entity to other entities. As a footnote, we mention that we chose LDA for this classification task as a matter of convenience and because we did not expect much difference by using the other models.

Table 6: Examples from data generated for classification of entity pairs.

Not observed in train (Eo) Was observed in test (E)	Never observed i.e. fabricated pair
NY Times:	NY Times:
(Martha Stewart, ImClone)	(Barry Bonds, Commonwealth)
(Jack Grubman, WordCom)	(Baylor University, Wrigley Field)
FBIS:	FBIS:
(Tony Blair, Johnny Adair)	(Tony Blair, Francoise de Panafieu)
(Jacques Chirac, Jean Tiberi)	(Silvio Berlusconi, Bal Thackeray)
(Madeline Albright, Y. Arafat)	(Gerhard Schroeder, Paul Kagame)

Table 7: Classification accuracy of entity pairs.

	words & entities Wo+Eo	just entities Eo
random	0.50	0.50
NYT (T=100)	0.66	0.64
NYT (T=200)	0.64	0.61
FBIS (T=100)	0.80	0.61
FBIS (T=200)	0.75	0.60

5.5 Entity-Entity Relationships

Given a collection of news articles, we can create a social network by aggregating, for each article, the co-mentions of entities. For example George Bush and Saddam Hussein co-appeared in over 400 news articles in the NY Times 2 training data of 29,000 articles. But can we infer latent entity-entity relationships purely based on topical information associated with each entity? And what if our named entity recognition and entity resolution system misses entities or produces slight variations in the entity string? As described in the Section on classification of entity pairs, we can measure entity-entity affinity from our trained entity-topic models.

A latent entity-entity network based on the 400-topic LDA run of 3-year NY Times 1 data shows associations between pairs of entities that *never* appeared in any single news article (Figure 9). We see Edmond Pope (American convicted of espionage in Russia) connected to Boris Berezovsky (a Russian businessman). While these two people never appeared together in any one of our 330,000 news articles, a Google search for this pair indicates a close connection. Ayman al-Zawahiri (Al-Qaeda, Bin Laden's physician) and Wadih el-Hage (Al-Qaeda, 1998 US embassy bombings) are also never co-referenced, but have an obvious association.

Not only can our entity-topic models identify connections in these latent social networks, they can also topically describe the nature of the connection between any two entities, and ultimately provide the evidence (as a list of most relevant documents) that supports the connection. The statistical nature of these models alleviates the problem of imperfect entity resolution, and, by leveraging word information, can discover relationships even when entity mentions are relatively scarce.

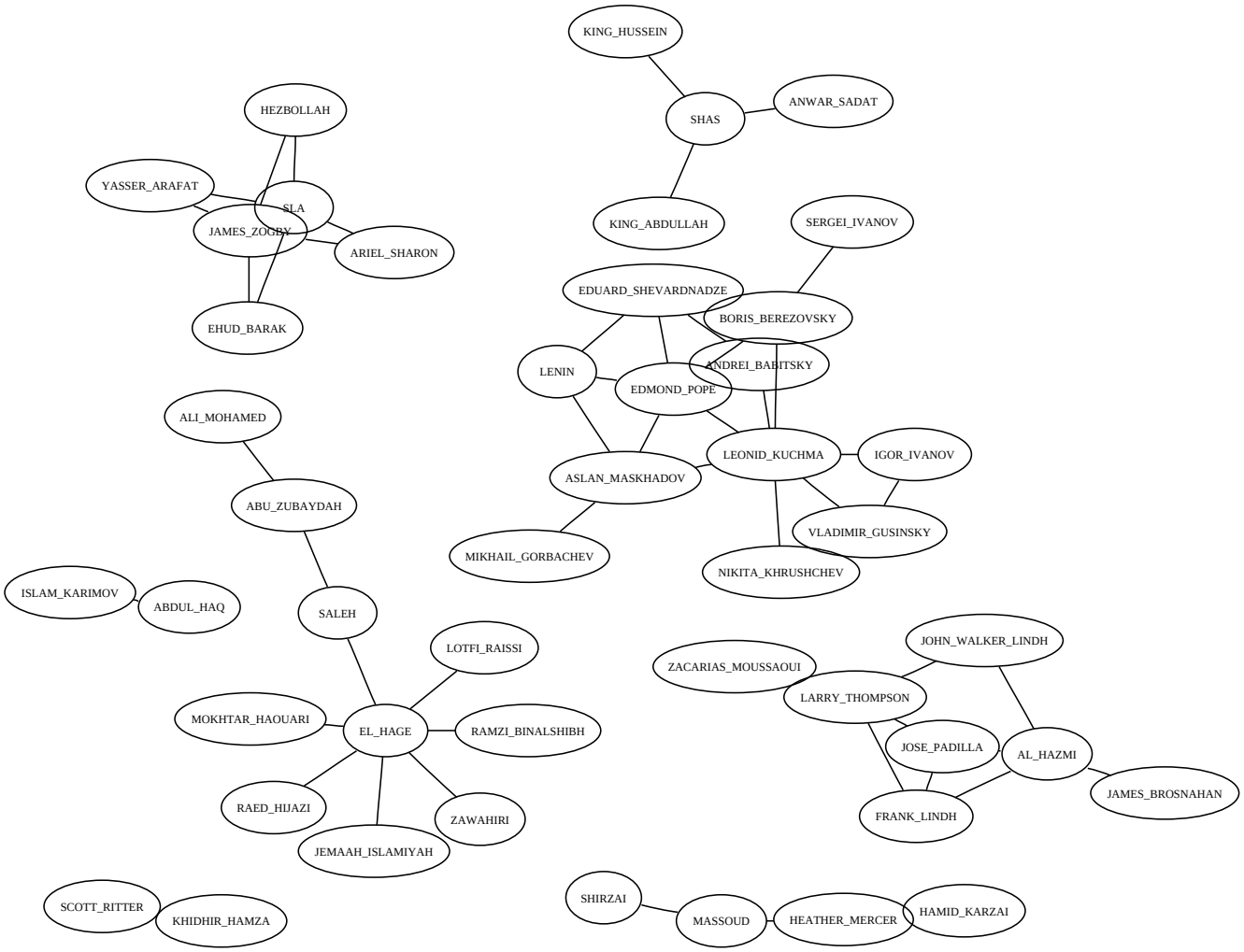

Figure 9: Latent entity-entity network based on 400-topic LDA run of 3-year NY Times 1 data. Links exist where the model-based entity-entity affinity $p(e_i|e_j)/2 + p(e_j|e_i)/2$ is above a certain threshold but where the (e_i, e_j) pair does not co-exist in any single news article. We see Edmond Pope (American convicted of espionage in Russia) connected to Boris Berezovsky (a Russian businessman). While these two people never appeared together in any one of our 330,000 news articles, a Google search for this pair indicates a close connection. Ayman al-Zawahiri (Al-Qaeda, Bin Laden's physician) and Wadih el-Hage (Al-Qaeda, 1998 US embassy bombings) are also never co-referenced, but have an obvious association.

6. CONCLUSIONS

We have developed two new graphical models – CorrLDA2 and SwitchLDA – specifically designed for text data that contains words and entities (e.g. persons, organizations, locations). We compare these two plus three other models on various entity prediction tasks using two large collections of news articles. We show how one can leverage the latent structure in text to make up to a 30% better prediction about entities by learning the relationships between entities mentioned in the text and topics learned from word co-occurrences. For one data set that is rich with entities, our CorrLDA2 model shows an improved ability to predict unseen entities in test documents.

Finally we give some examples of how this type of entity-topic modeling can be applied to construct social networks of entities based on latent information, showing links between people who never co-appear in any document. Gaining extra knowledge – through text – about entity-entity relationships is especially useful when entity mentions are sparse.

The models discussed in this paper are all generalizable to handle multiple classes of word tokens in data. For example one could model the interrelationships between words, people, organizations and locations mentioned in, say, a series of news articles. Other application areas of this entity-topic modeling include medical literature (e.g. PubMed) where one could create entity-topic models where the entities are genes and proteins mentioned in the text.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under award number IIS-0083489 (as part of the Knowledge Discovery and Dissemination Program) and under award number ITR-0331707.

8. ADDITIONAL AUTHORS

Additional author: Mark Steyvers, Dept of Cognitive Sciences, University of California, Irvine (`msteyver@uci.edu`).

9. REFERENCES

[1] I. Bhattacharya and L. Getoor. A latent dirichlet allocation model for entity resolution. *University of Maryland tech report*, 2005.

[2] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Neural Information Processing Systems*, volume 16, 2003.

[3] D. Blei and M. I. Jordan. Modeling annotated data. In *Proceedings of the Annual Conference on Research and Development in Information Retrieval (SIGIR03)*, 2003.

[4] D. Blei and J. Lafferty. Correlated topic models. In *Neural Information Processing Systems*, volume 18, 2006.

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

[6] E. Brill. Some advances in transformation-based part of speech tagging. *National Conference on Artificial Intelligence*, 1994.

[7] W. Buntine, J. Lofstrm, J. Perki, S. Perttu, V. Poroshin, T. Silander, H. Tirri, A. Tuominen, and V. Tuulos. A scalable topic-based open source search engine. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 228–234, 2004.

[8] D. Cohn and T. Hofmann. The missing link—a probabilistic model of document content and hypertext connectivity. In *Advances in Neural Information Processing Systems 13*, pages 430–436. MIT Press, 2001.

[9] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101:5220–5227, 2004.

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

[11] T. L. Griffiths, M. Steyvers, D. Blei, and J. B. Tenenbaum. Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, 2005.

[12] J. O. Madadhain, J. Hutchins, and P. Smyth. Prediction and ranking algorithms for event-based network data. In *ACM SIGKDD Explorations: Special Issue on Link Mining*, volume 7, pages 23–30, 2006.

[13] A. McCallum, , A. Corrada Emmanuel, and X. Wang. The author-recipient-topic model for topic and role discovery in social networks. Technical Report UM-CS-2004-096, Department of Computer Science, University of Massachusetts, 2004.

[14] A. McCallum and B. Wellner. Conditional models of identity uncertainty with applications to noun coreference. In *Neural Information Processing Systems*, 2004.

[15] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 352–359, San Francisco, CA, 2002.

[16] M. Rosen-Zvi, T. Griffith, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494, San Francisco, CA, 2004.

[17] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the Tenth ACM International Conference on Knowledge Discovery and Data Mining (ACM Press)*, pages 306–315, 2004.

[18] J. Zhu, A. Goncalves, and V. Uren. Adaptive named entity recognition for social network analysis and domain ontology maintenance. In *Proceedings of 3rd Professional Knowledge Management Conference, Springer, LNAI*, 2005.

APPENDIX

A. GIBBS SAMPLING EQUATIONS

In the following equations, α, β are Dirichlet priors and γ is a Beta prior. The notation C_{pq}^{PQ} represents counts from respective count matrices, e.g. counts of words in a topic, or counts of topics in a document. Word-topic distributions are represented by ϕ, and topic-document distributions are represented by θ. Words are either denoted by $\mathbf{w}$ or $\mathbf{v}$, and entities are denoted by $\tilde{\mathbf{w}}$. Correspondingly, topic or word-topic assignments are denoted by $\mathbf{z}$, and entity-topic assignments $\tilde{\mathbf{z}}$. The variable $\mathbf{x}$ is the switching flag in SwitchLDA, and a supertopic for entities in CorrLDA2.

A.1 LDA

The Gibbs sampling equation for LDA has two terms. The first term measures the likelihood of a topic in the document, and the second term measures the likelihood of the word in that topic. The Gibbs sampling equations and expectation equations for all other models can be viewed as respective modifications to the cannonical LDA model.

$$p(\mathbf{z}_i = t \,|\mathbf{v}_i = v, \mathbf{z}_{-i}, \mathbf{v}_{-i}, \alpha, \beta) \propto \frac{C^{TD}_{td,-i} + \alpha}{\sum_{t'} C^{TD}_{t'd,-i} + T\alpha} \frac{C^{VT}_{vt,-i} + \beta}{\sum_{v'} C^{VT}_{v't,-i} + V\beta}$$

$$\mathsf{E}\left[\phi_{vt} \,|\mathbf{z}, \mathbf{v}, \beta\right] = \frac{C^{VT}_{vt} + \beta}{\sum_{v'} C^{VT}_{v't} + V\beta}$$

$$\mathsf{E}\left[\Theta_{td} \,|\mathbf{z}, \alpha\right] = \frac{C^{TD}_{td} + \alpha}{\sum_{t'} C^{TD}_{t'd} + T\alpha}$$

A.2 CI-LDA

$$p(\mathbf{z}_i = t \,|\mathbf{w}_i = w, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \tilde{\mathbf{z}}, \alpha, \beta) \propto \frac{C^{TD}_{td,-i} + \alpha}{\sum_{t'} C^{TD}_{t'd,-i} + T\alpha} \frac{C^{WT}_{wt,-i} + \beta}{\sum_{w'} C^{WT}_{w'd,-i} + W\beta}$$

$$p\left(\tilde{\mathbf{z}}_i = t \,\middle|\tilde{\mathbf{w}}_i = e, \tilde{\mathbf{z}}_{-i}, \tilde{\mathbf{w}}_{-i}, \mathbf{z}, \alpha, \tilde{\beta}\right) \propto \frac{C^{TD}_{td,-i} + \alpha}{\sum_{t'} C^{TD}_{t'd,-i} + T\alpha} \frac{C^{ET}_{et,-i} + \tilde{\beta}}{\sum_{e'} C^{ET}_{e't,-i} + E\tilde{\beta}}$$

$$\mathsf{E}\left[\phi_{wt} \,|\mathbf{z}, \mathbf{w}, \beta\right] = \frac{C^{WT}_{wt} + \beta}{\sum_{w'} C^{WT}_{w't} + W\beta}$$

$$\mathsf{E}\left[\tilde{\phi}_{et} \,\middle|\tilde{\mathbf{z}}, \tilde{\mathbf{w}}, \tilde{\beta}\right] = \frac{C^{ET}_{et} + \tilde{\beta}}{\sum_{e'} C^{ET}_{e't} + E\tilde{\beta}}$$

$$\mathsf{E}\left[\Theta_{td} \,|\mathbf{z}, \tilde{\mathbf{z}}, \alpha\right] = \frac{C^{TD}_{td} + \alpha}{\sum_{t'} C^{TD}_{t'd} + T\alpha}$$

A.3 SwitchLDA

$$p(\mathbf{z}_i = t \,|\mathbf{v}_i = w, \mathbf{x}_i = 0, \mathbf{z}_{-i}, \mathbf{x}_{-i}, \mathbf{v}_{-i}, \alpha, \beta, \gamma) \propto \frac{C^{TD}_{td,-i} + \alpha}{\sum_{t'} C^{TD}_{t'd,-i} + T\alpha} \frac{n_{t,-i} + \gamma}{n_{t,-i} + \tilde{n}_t + 2\gamma} \frac{C^{WT}_{wt,-i} + \beta}{\sum_{w'} C^{WT}_{w't,-i} + W\beta}$$

$$p\left(\mathbf{z}_i = t \,\middle|\mathbf{w}_i = e, \mathbf{x}_i = 1, \mathbf{z}_{-i}, \mathbf{x}_{-i}, \mathbf{w}_{-i}, \alpha, \tilde{\beta}, \gamma\right) \propto \frac{C^{TD}_{td,-i} + \alpha}{\sum_{t'} C^{TD}_{t'd,-i} + T\alpha} \frac{\tilde{n}_{t,-i} + \gamma}{n_t + \tilde{n}_{t,-i} + 2\gamma} \frac{C^{ET}_{et,-i} + \tilde{\beta}}{\sum_{e'} C^{ET}_{e't,-i} + E\tilde{\beta}}$$

$$\mathsf{E}\left[\phi_{wt} \,|\mathbf{z}, \mathbf{x}, \mathbf{w}, \beta\right] = \frac{C^{WT}_{wt} + \beta}{\sum_{w'} C^{WT}_{w't} + W\beta}$$

$$\mathsf{E}\left[\tilde{\phi}_{et} \,\middle|\mathbf{z}, \mathbf{x}, \mathbf{w}, \tilde{\beta}\right] = \frac{C^{ET}_{et} + \tilde{\beta}}{\sum_{e'} C^{ET}_{e't} + E\tilde{\beta}}$$

$$\mathsf{E}\left[\Psi_t \,|\mathbf{z}, \mathbf{x}, \gamma\right] = \frac{\tilde{n}_t + \gamma}{n_t + \tilde{n}_t + 2\gamma}$$

$$\mathsf{E}\left[\Theta_{td} \,|\mathbf{z}, \alpha\right] = \frac{C^{TD}_{td} + \alpha}{\sum_{t'} C^{TD}_{t'd} + T\alpha}$$

A.4 CorrLDA1

$$p(\mathbf{z}_i = t \,|\mathbf{w}_i = w, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \alpha, \beta) \propto \frac{C^{TD}_{td,-i} + \alpha}{\sum_{t'} C^{TD}_{t'd,-i} + T\alpha} \frac{C^{WT}_{wt,-i} + \beta}{\sum_{w'} C^{WT}_{w't,-i} + W\beta}$$

$$p\left(\tilde{\mathbf{z}}_i = \tilde{z} \,\middle|\mathbf{w}_i = e, \tilde{\mathbf{z}}_i, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \beta, \tilde{\beta}\right) \propto \frac{C^{TD}_{\tilde{z}d}}{Nw_d} \frac{C^{ET}_{e'\tilde{z},-i} + \tilde{\beta}}{\sum_{e'} C^{ET}_{e'\tilde{z},-i} + E\tilde{\beta}}$$

A.5 CorrLDA2

$$p(\mathbf{z}_i = t \,|\mathbf{w}_i = w, \mathbf{z}_{-i}, \mathbf{w}_{-i}, \alpha, \beta) \propto \frac{C^{TD}_{td,-i} + \alpha}{\sum_{t'} C^{TD}_{t'd,-i} + T\alpha} \frac{C^{WT}_{wt,-i} + \beta}{\sum_{w'} C^{WT}_{w't,-i} + W\beta}$$

$$p\left(\tilde{\mathbf{z}}_i = \tilde{z}, \mathbf{x}_i = t \,\middle|\tilde{\mathbf{w}}_i = e, \tilde{\mathbf{z}}_{-i}, \mathbf{z}, \tilde{\mathbf{w}}_{-i}, \tilde{\beta}\right) \propto \frac{C^{TD}_{td}}{Nw_d} \frac{C^{\tilde{T}T}_{\tilde{z}t,-i} + \gamma}{\sum_{\tilde{z}'} C^{\tilde{T}T}_{\tilde{z}'t,-i} + \tilde{T}\gamma} \frac{C^{E\tilde{T}}_{e\tilde{z},-i} + \tilde{\beta}}{\sum_{e'} C^{E\tilde{T}}_{e'\tilde{z},-i} + E\tilde{\beta}}$$