Structural Topic Model for Latent Topical Structure Analysis

Hongning Wang, Duo Zhang, ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana IL, 61801 USA
{wang296, dzhang22, czhai}@cs.uiuc.edu

Abstract

Topic models have been successfully applied to many document analysis tasks to discover topics embedded in text. However, existing topic models generally cannot capture the latent topical structures in documents. Since languages are intrinsically cohesive and coherent, modeling and discovering latent topical transition structures within documents would be beneficial for many text analysis tasks.

In this work, we propose a new topic model, Structural Topic Model, which simultaneously discovers topics and reveals the latent topical structures in text through explicitly modeling topical transitions with a latent first-order Markov chain. Experiment results show that the proposed Structural Topic Model can effectively discover topical structures in text, and the identified structures significantly improve the performance of tasks such as sentence annotation and sentence ordering.

1 Introduction

A great amount of effort has recently been made in applying statistical topic models (Hofmann, 1999; Blei et al., 2003) to explore word co-occurrence patterns, i.e. topics, embedded in documents. Topic models have become important building blocks of many interesting applications (see e.g., (Blei and Jordan, 2003; Blei and Lafferty, 2007; Mei et al., 2007; Lu and Zhai, 2008)).

In general, topic models can discover word clustering patterns in documents and project each document to a latent topic space formed by such word clusters. However, the topical structure in a document, i.e., the internal dependency between the topics, is generally not captured due to the exchangeability assumption (Blei et al., 2003), i.e., the document generation probabilities are invariant to content permutation. In reality, natural language text rarely consists of isolated, unrelated sentences, but rather collocated, structured and coherent groups of sentences (Hovy, 1993). Ignoring such latent topical structures inside the documents means wasting valuable clues about topics and thus would lead to non-optimal topic modeling.

Taking apartment rental advertisements as an example, when people write advertisements for their apartments, it's natural to first introduce *"size"* and *"address"* of the apartment, and then *"rent"* and *"contact"*. Few people would talk about *"restriction"* first. If this kind of topical structures are captured by a topic model, it would not only improve the topic mining results, but, more importantly, also help many other document analysis tasks, such as sentence annotation and sentence ordering.

Nevertheless, very few existing topic models attempted to model such structural dependency among topics. The Aspect HMM model introduced in (Blei and Moreno, 2001) combines pLSA (Hofmann, 1999) with HMM (Rabiner, 1989) to perform document segmentation over text streams. However, Aspect HMM separately estimates the topics in the training set and depends on heuristics to infer the transitional relations between topics. The Hidden Topic Markov Model (HTMM) proposed by (Gruber et al., 2007) extends the traditional topic models by assuming words in each sentence share the same topic assignment, and topics transit between adjacent sentences. However, the transitional structures among topics, i.e., how likely one topic would follow another topic, are not captured in this model.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1526–1535,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

In this paper, we propose a new topic model, named Structural Topic Model (strTM) to model and analyze both latent topics and topical structures in text documents. To do so, strTM assumes: 1) words in a document are either drawn from a content topic or a *functional* (i.e., background) topic; 2) words in the same sentence share the same content topic; and 3) content topics in the adjacent sentences follow a topic transition that satisfies the first order Markov property. The first assumption distinguishes the semantics of the occurrence of each word in the document, the second requirement confines the unrealistic "bag-of-word" assumption into a tighter unit, and the third assumption exploits the connection between adjacent sentences.

To evaluate the usefulness of the identified topical structures by strTM, we applied strTM to the tasks of sentence annotation and sentence ordering, where correctly modeling the document structure is crucial. On the corpus of 8,031 apartment advertisements from craiglist (Grenager et al., 2005) and 1,991 movie reviews from IMDB (Zhuang et al., 2006), strTM achieved encouraging improvement in both tasks compared with the baseline methods that don't explicitly model the topical structure. The results confirm the necessity of modeling the latent topical structures inside documents, and also demonstrate the advantages of the proposed strTM over existing topic models.

2 Related Work

Topic models have been successfully applied to many problems, e.g., sentiment analysis (Mei et al., 2007), document summarization (Lu and Zhai, 2008) and image annotation (Blei and Jordan, 2003). However, in most existing work, the dependency among the topics is loosely governed by the prior topic distribution, e.g., Dirichlet distribution.

Some work has attempted to capture the interrelationship among the latent topics. Correlated Topic Model (Blei and Lafferty, 2007) replaces Dirichlet prior with logistic Normal prior for topic distribution in each document in order to capture the correlation between the topics. HMM-LDA (Griffiths et al., 2005) distinguishes the short-range syntactic dependencies from long-range semantic dependencies among the words in each document. But in HMM-LDA, only the latent variables for the syntactic classes are treated as a locally dependent sequence, while latent topics are treated the same as in other topic models. Chen et al. introduced the generalized Mallows model to constrain the latent topic assignments (Chen et al., 2009). In their model, they assume there exists a canonical order among the topics in the collection of related documents and the same topics are forced not to appear in disconnected portions of the topic sequence in one document (sampling without replacement). Our method relaxes this assumption by only postulating transitional dependency between topics in the adjacent sentences (sampling with replacement) and thus potentially allows a topic to appear multiple times in disconnected segments. As discussed in the previous section, HTMM (Gruber et al., 2007) is the most similar model to ours. HTMM models the document structure by assuming words in the same sentence share the same topic assignment and successive sentences are more likely to share the same topic. However, HTMM only loosely models the transition between topics as a binary relation: the same as the previous sentence's assignment or draw a new one with a certain probability. This simplified coarse modeling of dependency could not fully capture the complex structure across different documents. In contrast, our strTM model explicitly captures the regular topic transitions by postulating the first order Markov property over the topics.

Another line of related work is discourse analysis in natural language processing: discourse segmentation (Sun et al., 2007; Galley et al., 2003) splits a document into a linear sequence of multi-paragraph passages, where lexical cohesion is used to link together the textual units; discourse parsing (Soricut and Marcu, 2003; Marcu, 1998) tries to uncover a more sophisticated hierarchical coherence structure from text to represent the entire discourse. One work in this line that shares a similar goal as ours is the content models (Barzilay and Lee, 2004), where an HMM is defined over text spans to perform information ordering and extractive summarization. A deficiency of the content models is that the identification of clusters of text spans is done separately from transition modeling. Our strTM addresses this deficiency by defining a generative process to simultaneously capture the topics and the transitional relationship among topics: allowing topic modeling and transition modeling to reinforce each other in a principled framework.

3 Structural Topic Model

In this section, we formally define the Structural Topic Model (strTM) and discuss how it captures the latent topics and topical structures within the documents simultaneously. From the theory of linguistic analysis (Kamp, 1981), we know that document exhibits internal structures, where structural segments encapsulate semantic units that are closely related. In strTM, we treat a sentence as the basic structure unit, and assume all the words in a sentence share the same topical aspect. Besides, two adjacent segments are assumed to be highly related (capturing cohesion in text); specifically, in strTM we pose a strong transitional dependency assumption among the topics: the choice of topic for each sentence directly depends on the previous sentence's topic assignment, i.e., first order Markov property. Moveover, taking the insights from HMM-LDA that not all the words are content conveying (some of them may just be a result of syntactic requirement), we introduce a dummy *functional* topic z_B for every sentence in the document. We use this functional topic to capture the document-independent word distribution, i.e., corpus background (Zhai et al., 2004). As a result, in strTM, every sentence is treated as a mixture of content and functional topics.

Formally, we assume a corpus consists of *D* documents with a vocabulary of size *V*, and there are *k* content topics embedded in the corpus. In a given document *d*, there are *m* sentences and each sentence *i* has N_i words. We assume the topic transition probability $p(z|z')$ is drawn from a Multinomial distribution *Mul*($\alpha_{z'}$), and the word emission probability under each topic $p(w|z)$ is drawn from a Multinomial distribution *Mul*(β_z).

To get a unified description of the generation process, we add another dummy topic *T-START* in strTM, which is the initial topic with position "-1" for every document but does not emit any words. In addition, since our functional topic is assumed to occur in all the sentences, we don't need to model its transition with other content topics. We use a Binomial variable π to control the proportion between content and functional topics in each sentence. Therefore, there are *k*+1 topic transitions, one for *T-START* and others for *k* content topics; and *k* emission probabilities for the content topics, with an additional one for the functional topic z_B (in total *k*+1 emission probability distributions).

Conditioned on the model parameters $\Theta = (\alpha, \beta, \pi)$, the generative process of a document in strTM can be described as follows:

1. For each sentence s_i in document *d*:
   (a) Draw topic z_i from Multinomial distribution conditioned on the previous sentence s_{i-1}'s topic assignment z_{i-1}:
   $z_i \sim Mul(\alpha_{z_{i-1}})$
   (b) Draw each word w_{ij} in sentence s_i from the mixture of content topic z_i and functional topic z_B:
   $w_{ij} \sim \pi p(w_{ij}|\beta, z_i) + (1-\pi) p(w_{ij}|\beta, z_B)$

The joint probability of sentences and topics in one document defined by strTM is thus given by:

$$p(S_0, S_1, \ldots, S_m, \mathbf{z}|\alpha, \beta, \pi) = \prod_{i=1}^{m} p(z_i|\alpha, z_{i-1}) p(S_i|z_i) \tag{1}$$

where the topic to sentence emission probability is defined as:

$$p(S_i|z_i) = \prod_{j=0}^{N_i} \left[\pi p(w_{ij}|\beta, z_i) + (1-\pi) p(w_{ij}|\beta, z_B)\right] \tag{2}$$

This process is graphically illustrated in Figure 1.

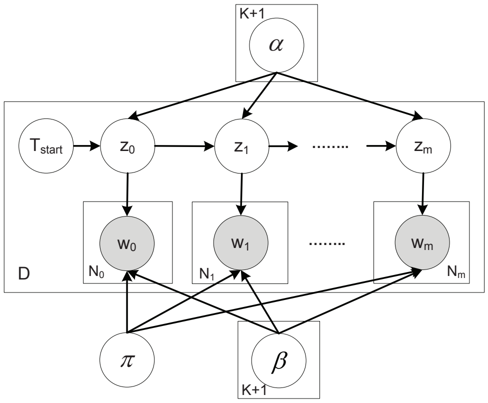


Figure 1: Graphical Representation of strTM.

From the definition of strTM, we can see that the document structure is characterized by a document-specific topic chain, and forcing the words in one sentence to share the same content topic ensures semantic cohesion of the mined topics. Although we do not directly model the topic mixture for each document as the traditional topic models do, the word co-occurrence patterns within the same document are captured by topic propagation through the transitions. This can be easily understood when we write down the posterior probability of the topic assignment for a particular sentence:

$$\begin{aligned}
&p(z_i|S_0,S_1,\ldots,S_m,\Theta)\\
=&\frac{p(S_0,S_1,\ldots,S_m|z_i,\Theta)p(z_i)}{p(S_0,S_1,\ldots,S_m)}\\
\propto&\, p(S_0,S_1,\ldots,S_i,z_i)\times p(S_{i+1},S_{i+2},\ldots,S_m|z_i)\\
=&\sum_{z_{i-1}}p(S_0,\ldots,S_{i-1},z_{i-1})p(z_i|z_{i-1})p(S_i|z_i)\\
&\times\sum_{z_{i+1}}p(S_{i+1},\ldots,S_m|z_{i+1})p(z_{i+1}|z_i) \qquad (3)
\end{aligned}$$

The first part of Eq(3) describes the recursive influence on the choice of topic for the i*th* sentence from its preceding sentences, while the second part captures how the succeeding sentences affect the current topic assignment. Intuitively, when we need to decide a sentence's topic, we will look "backward" and "forward" over all the sentences in the document to determine a "suitable" one. In addition, because of the first order Markov property, the local topical dependency gets more emphasis, i.e., they are interacting directly through the transition probabilities $p(z_i|z_{i-1})$ and $p(z_{i+1}|z_i)$. And such interaction on sentences farther away would get damped by the multiplication of such probabilities. This result is reasonable, especially in a long document, since neighboring sentences are more likely to cover similar topics than two sentences far apart.

4 Posterior Inference and Parameter Estimation

The chain structure in strTM enables us to perform exact inference: posterior distribution can be efficiently calculated by the forward-backward algorithm, the optimal topic sequence can be inferred using the Viterbi algorithm, and parameter estimation can be solved by the Expectation Maximization (EM) algorithm. More technical details can be found in (Rabiner, 1989). In this section, we only discuss strTM-specific procedures.

In the E-Step of EM algorithm, we need to collect the expected count of a sequential topic pair (z, z') and a topic-word pair (z, w) to update the model parameters α and β in the M-Step. In strTM, $E[c(z, z')]$ can be easily calculated by forward-backward algorithm. But we have to go one step further to fetch the required sufficient statistics for $E[c(z, w)]$, because our emission probabilities are defined over sentences.

Through forward-backward algorithm, we can get the posterior probability $p(s_i, z|d, \Theta)$. In strTM, words in one sentence are independently drawn from either a specific content topic z or functional topic z_B according to the mixture weight π. Therefore, we can accumulate the expected count of (z, w) over all the sentences by:

$$E[c(z,w)]=\sum_{d,s\in d}\frac{\pi p(w|z)p(s,z|d,\Theta)c(w,s)}{\pi p(w|z)+(1-\pi)p(w|z_B)} \qquad (4)$$

where $c(w, s)$ indicates the frequency of word w in sentence s.

Eq(4) can be easily explained as follows. Since we already observe topic z and sentence s co-occur with probability $p(s, z|d, \Theta)$, each word w in s should share the same probability of being observed with content topic z. Thus the expected count of $c(z, w)$ in this sentence would be $p(s, z|d, \Theta)c(w, s)$. However, since each sentence is also associated with the functional topic z_B, the word w may also be drawn from z_B. By applying the Bayes' rule, we can properly reallocate the expected count of $c(z, w)$ by Eq(4). The same strategy can be applied to obtain $E[c(z_B, w)]$.

As discussed in (Johnson, 2007), to avoid the problem that EM algorithm tends to assign a uniform word/state distribution to each hidden state, which deviates from the heavily skewed word/state distributions empirically observed, we can apply a Bayesian estimation approach for strTM. Thus we introduce prior distributions over the topic transition *Mul*($\alpha_{z'}$) and emission probabilities *Mul*(β_z), and use the Variational Bayesian (VB) (Jordan et al., 1999) estimator to obtain a model with more skewed word/state distributions.

Since both the topic transition and emission probabilities are Multinomial distributions in strTM, the conjugate Dirichlet distribution is the natural choice for imposing a prior on them (Diaconis and Ylvisaker, 1979). Thus, we further assume:

$$\alpha_z \sim Dir(\eta) \tag{5}$$

$$\beta_z \sim Dir(\gamma) \tag{6}$$

where we use exchangeable Dirichlet distributions to control the sparsity of α_z and β_z. As η and γ approach zero, the prior strongly favors the models in which each hidden state emits as few words/states as possible. In our experiments, we empirically tuned η and γ on different training corpus to optimize log-likelihood.

The resulting VB estimation only requires a minor modification to the M-Step in the original EM algorithm:

$$\bar{\alpha}_z = \frac{\Phi(E[c(z', z)] + \eta)}{\Phi(E[c(z)] + k\eta)} \tag{7}$$

$$\bar{\beta}_z = \frac{\Phi(E[c(w, z)] + \gamma)}{\Phi(E[c(z)] + V\gamma)} \tag{8}$$

where $\Phi(x)$ is the exponential of the first derivative of the log-gamma function.

The optimal setting of π for the proportion of content topics in the documents is empirically tuned by cross-validation over the training corpus to maximize the log-likelihood.

5 Experimental Results

In this section, we demonstrate the effectiveness of strTM in identifying latent topical structures from documents, and quantitatively evaluate how the mined topic transitions can help the tasks of sentence annotation and sentence ordering.

5.1 Data Set

We used two different data sets for evaluation: apartment advertisements (Ads) from (Grenager et al., 2005) and movie reviews (Review) from (Zhuang et al., 2006).

The Ads data consists of 8,767 advertisements for apartment rentals crawled from Craigslist website. 302 of them have been labeled with 11 fields, including *size*, *feature*, *address*, etc., on the sentence level. The review data contains 2,000 movie reviews discussing 11 different movies from IMDB. These reviews are manually labeled with 12 movie feature labels (We didn't use the additional opinion annotations in this data set.) , e.g., *VP* (vision effects), *MS* (music and sound effects), etc., also on the sentences, but the annotations in the review data set is much sparser than that in the Ads data set (see in Table 1). The sentence-level annotations make it possible to quantitatively evaluate the discovered topic structures.

We performed simple preprocessing on these two data sets: 1) removed a standard list of stop words, terms occurring in less than 2 documents; 2) discarded the documents with less than 2 sentences; 3) aggregated sentence-level annotations into document-level labels (binary vector) for each document. Table 1 gives a brief summary on these two data sets after the processing.

	Ads	*Review*
Document Size	8,031	1,991
Vocabulary Size	21,993	14,507
Avg Stn/Doc	8.0	13.9
Avg Labeled Stn/Doc	7.1*	5.1
Avg Token/Stn	14.1	20.0

*Only in 302 labeled ads

Table 1: Summary of evaluation data set

5.2 Topic Transition Modeling

First, we qualitatively demonstrate the topical structure identified by strTM from Ads data[1]. We trained strTM with 11 content topics in Ads data set, used word distribution under each class (estimated by maximum likelihood estimator on document-level labels) as priors to initialize the emission probability *Mul(β_z)* in Eq(6), and treated document-level labels as the prior for transition from T-START in each document, so that the mined topics can be aligned with the predefined class labels. Figure 2 shows the identified topics and the transitions among them. To get a clearer view, we discarded the transitions below a threshold of 0.1 and removed all the isolated nodes.

From Figure 2, we can find some interesting topical structures. For example, people usually start with *"size"*, *"features"* and *"address"*, and end with *"contact"* information when they post an apart-

[1] Due to the page limit, we only show the result in Ads data set.

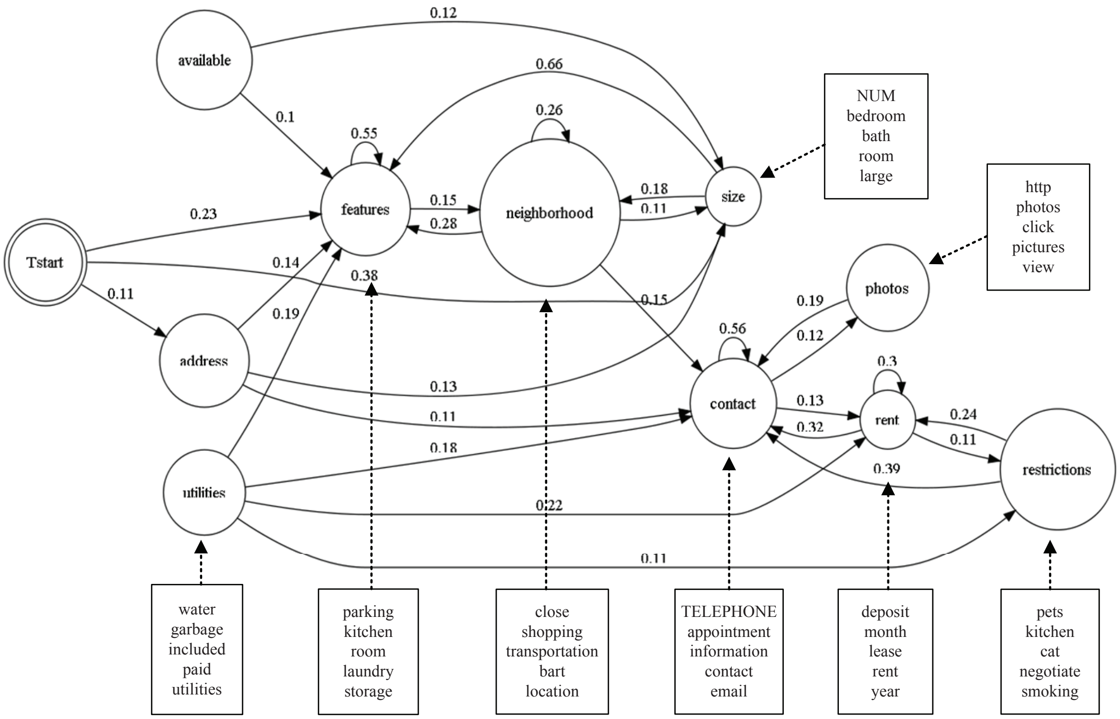


Figure 2: Estimated topics and topical transitions in Ads data set

ment ads. Also, we can discover a strong transition from *"size"* to *"features"*. This intuitively makes sense because people usually write "it's a two bedrooms apartment" first, and then describe other *"features"* about the apartment. The mined topics are also quite meaningful. For example, *"restrictions"* are usually put over pets and smoking, and parking and laundry are always the major *"features"* of an apartment.

To further quantitatively evaluate the estimated topic transitions, we used *Kullback-Leibler* (*KL*) divergency between the estimated transition matrix and the "ground-truth" transition matrix as the metric. Each element of the "ground-truth" transition matrix was calculated by Eq(9), where $c(z, z')$ denotes how many sentences annotated by z' immediately precede one annotated by z. δ is a smoothing factor, and we fixed it to 0.01 in the experiment.

$$\bar{p}(z|z') = \frac{c(z, z') + \delta}{c(z) + k\delta} \qquad (9)$$

The *KL* divergency between two transition matrices is defined in Eq(10). Because we have a $k \times k$ transition matrix (T_{start} is not included), we calculated the average *KL* divergency against the ground-truth over all the topics:

$$avgKL = \frac{\sum_{i=1}^{k} KL(p(z|z'_i)||\bar{p}(z|z'_i)) + KL(\bar{p}(z|z'_i)||p(z|z'_i))}{2k} \qquad (10)$$

where $\bar{p}(z|z')$ is the ground-truth transition probability estimated by Eq(9), and $p(z|z')$ is the transition probability given by the model.

We used pLSA (Hofmann, 1999), latent permutation model (lPerm) (Chen et al., 2009) and HTMM (Gruber et al., 2007) as the baseline methods for the comparison. Because none of these three methods can generate a topic transition matrix directly, we extended them a little bit to achieve this goal. For pLSA, we used the document-level labels as priors for the topic distribution in each document, so that the estimated topics can be aligned with the predefined class labels. After the topics were estimated, for each sentence we selected the topic that had the highest posterior probability to generate the sentence as its class label. For lPerm and HTMM, we used Kuhn-Munkres algorithm (Lovász and Plummer, 1986) to find the optimal topic-to-class alignment based on the sentence-level annotations. After the sentences were annotated with class labels, we estimated the topic transition matrices for all of these three methods by Eq(9).

Since only a small portion of sentences are annotated in the Review data set, very few neighboring sentences are annotated at the same time, which introduces many noisy transitions. As a result, we only performed the comparison on the Ads data set. The "ground-truth" transition matrix was estimated based on all the 302 annotated ads.

	pLSA+prior	lPerm	HTMM	strTM
avgKL	0.743	1.101	0.572	0.372
p-value	0.023	1e-4	0.007	–

Table 2: Comparison of estimated topic transitions on Ads data set

In Table 2, the p-value was calculated based on t-test of the *KL* divergency between each topic's transition probability against strTM. From the results, we can see that avgKL of strTM is smaller than the other three baseline methods, which means the estimated transitional relation by strTM is much closer to the ground-truth transition. This demonstrates that strTM captures the topical structure well, compared with other baseline methods.

5.3 Sentence Annotation

In this section, we demonstrate how the identified topical structure can benefit the task of sentence annotation. Sentence annotation is one step beyond the traditional document classification task: in sentence annotation, we want to predict the class label for each sentence in the document, and this will be helpful for other problems, including extractive summarization and passage retrieval. However, the lack of detailed annotations on sentences greatly limits the effectiveness of the supervised classification methods, which have been proved successful on document classifications.

In this experiment, we propose to use strTM to address this annotation task. One advantage of strTM is that it captures the topic transitions on the sentence level within documents, which provides a regularization over the adjacent predictions.

To examine the effectiveness of such structural regularization, we compared strTM with four baseline methods: pLSA, lPerm, HTMM and Naive Bayes model. The sentence labeling approaches for strTM, pLSA, lPerm and HTMM have been discussed in the previous section. As for Naive Bayes model, we used EM algorithm [2] with both labeled and unlabeled data for the training purpose (we used the same unigram features as in topics models). We set weights for the unlabeled data to be 10^{-3} in Naive Bayes with EM.

The comparison was performed on both data sets. We set the size of topics in each topic model equal to the number of classes in each data set accordingly. To tackle the situation where some sentences in the document are not strictly associated with any classes, we introduced an additional *NULL* content topic in all the topic models. During the training phase, none of the methods used the sentence-level annotations in the documents, so that we treated the whole corpus as the training and testing set.

To evaluate the prediction performance, we calculated accuracy, recall and precision based on the correct predictions over the sentences, and averaged over all the classes as the criterion.

Model	Accuracy	Recall	Precison
pLSA+prior	0.432	0.649	0.457
lPerm	0.610	0.514	0.471
HTMM	0.606	0.588	0.443
NB+EM	0.528	0.337	0.612
strTM	**0.747**	**0.674**	**0.620**

Table 3: Sentence annotation performance on Ads data set

Model	Accuracy	Recall	Precison
pLSA+prior	0.342	0.278	0.250
lPerm	0.286	0.205	0.184
HTMM	0.369	0.131	0.149
NB+EM	0.341	0.354	**0.431**
strTM	**0.541**	**0.398**	0.323

Table 4: Sentence annotation performance on Review data set

Annotation performance on the two data sets is shown in Table 3 and Table 4. We can see that strTM outperformed all the other baseline methods on most of the metrics: strTM has the best accuracy and recall on both of the two data sets. The improvement confirms our hypothesis that besides solely depending on the local word patterns to perform predic-

[2] Mallet package: `http://mallet.cs.umass.edu/`

tions, adjacent sentences provide a structural regularization in strTM (see Eq(3)). Compared with lPerm, which postulates a strong constrain over the topic assignment (sampling without replacement), strTM performed much better on both of these two data sets. This validates the benefit of modeling local transitional relation compared with the global ordering. Besides, strTM achieved over 46% accuracy improvement compared with the second best HTMM in the review data set. This result shows the advantage of explicitly modeling the topic transitions between neighbor sentences instead of using a binary relation to do so as in HTMM.

To further testify how the identified topical structure can help the sentence annotation task, we first randomly removed 100 annotated ads from the training corpus and used them as the testing set. Then, we used the ground-truth topic transition matrix estimated from the training data to order those 100 ads according to their fitness scores under the ground-truth topic transition matrix, which is defined in Eq(11). We tested the prediction accuracy of different models over two different partitions, top 50 and bottom 50, according to this order.

$$fitness(d) = \frac{1}{|d|} \sum_{i=0}^{|d|} \log \bar{p}(t_i|t_{i-1}) \qquad (11)$$

where t_i is the class label for i*th* sentence in document d, $|d|$ is the number of sentences in document d, and $\bar{p}(t_i|t_{i-1})$ is the transition probability estimated by Eq(9).

	Top 50	p-value	Bot 50	p-value
pLSA+prior	0.496	4e-12	0.542	0.004
lPerm	0.669	0.003	0.505	8e-4
HTMM	0.683	0.004	0.579	0.003
NB + EM	0.492	1e-12	0.539	0.002
strTM	0.752	–	0.644	–

Table 5: Sentence annotation performance according to structural fitness

The results are shown in Table 5. From this table, we can find that when the testing documents follow the regular patterns as in the training data, i.e., top 50 group, strTM performs significantly better than the other methods; when the testing documents don't share such structure, i.e., bottom 50 group, strTM's performance drops. This comparison confirms that when a testing document shares similar topic structure as the training data, the topical transitions captured by strTM can help the sentence annotation task a lot. In contrast, because pLSA and Naive Bayes don't depend on the document's structure, their performance does not change much over these two partitions.

5.4 Sentence Ordering

In this experiment, we illustrate how the learned topical structure can help us better arrange sentences in a document. Sentence ordering, or text planning, is essential to many text synthesis applications, including multi-document summarization (Goldstein et al., 2000) and concept-to-text generation (Barzilay and Lapata, 2005).

In strTM, we evaluate all the possible orderings of the sentences in a given document and selected the optimal one which gives the highest generation probability:

$$\bar{\sigma}(m) = \arg\max_{\sigma(m)} \sum_{\mathbf{z}} p(S_{\sigma[0]}, S_{\sigma[1]}, \ldots, S_{\sigma[m]}, \mathbf{z}|\Theta) \qquad (12)$$

where $\sigma(m)$ is a permutation of 1 to m, and $\sigma[i]$ is the i*th* element in this permutation.

To quantitatively evaluate the ordering result, we treated the original sentence order (OSO) as the perfect order and used *Kendall's* $\tau(\sigma)$ (Lapata, 2006) as the evaluation metric to compute the divergency between the optimum ordering given by the model and OSO. *Kendall's* $\tau(\sigma)$ is widely used in information retrieval domain to measure the correlation between two ranked lists and it indicates how much an ordering differs from OSO, which ranges from 1 (perfect matching) to -1 (totally mismatching).

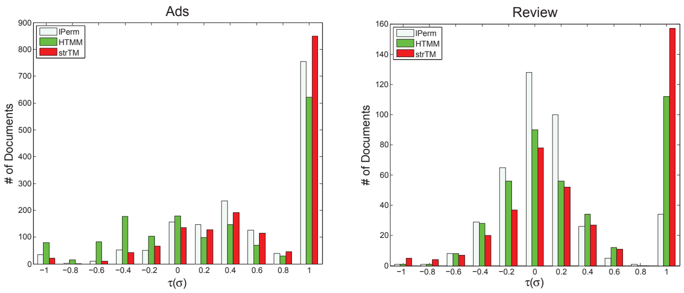
Since only the HTMM and lPerm take the order of sentences in the document into consideration, we used them as the baselines in this experiment. We ranked OSO together with candidate permutations according to the corresponding model's generation probability. However, when the size of documents becomes larger, it's infeasible to permutate all the orderings, therefore we randomly permutated 200 possible orderings of sentences as candidates when there were more than 200 possible candidates. The

2bedroom 1bath in very nice complex! Pool, carport, laundry facilities!! *Call Don (650)207-5769 to see!* Great location!! Also available, 2bed.2bath for $1275 in same complex.	$\Longrightarrow$	2bedroom 1bath in very nice complex! Pool, carport, laundry facilities!! Great location!! Also available, 2bed.2bath for $1275 in same complex. *Call Don (650)207-5769 to see!*
2 bedrooms 1 bath + a famyly room in a cul-de-sac location. *Please drive by and call Marilyn for appointment 650-652-5806.* Address: 517 Price Way, Vallejo. No Pets Please!	$\Longrightarrow$	2 bedrooms 1 bath + a famyly room in a cul-de-sac location. Address: 517 Price Way, Vallejo. No Pets Please! *Please drive by and call Marilyn for appointment 650-652-5806.*

Table 6: Sample results for document ordering by strTM

experiment was performed on both data sets with 80% data for training and the other 20% for testing.

We calculated the $\tau(\sigma)$ of all these models for each document in the two data sets and visualized the distribution of $\tau(\sigma)$ in each data set with histogram in Figure 3. From the results, we could observe that strTM's $\tau(\sigma)$ is more skewed towards the positive range (with mean 0.619 in Ads data set and 0.398 in review data set) than lPerm's results (with mean 0.566 in Ads data set and 0.08 in review data set) and HTMM's results (with mean 0.332 in Ads data set and 0.286 in review data set). This indicates that strTM better captures the internal structure within the documents.



(a) Ads (b) Review

Figure 3: Document Ordering Performance in $\tau(\sigma)$.

We see that all methods performed better on the Ads data set than the review data set, suggesting that the topical structures are more coherent in the Ads data set than the review data. Indeed, in the Ads data, strTM perfectly recovered 52.9% of the original sentence order. When examining some mismatched results, we found that some of them were due to an "outlier" order given by the original document (in comparison to the "regular" patterns in the set). In Table 6, we show two such examples where we see the learned structure "suggested" to move the contact information to the end, which intuitively gives us a more regular organization of the ads. It's hard to say that in this case, the system's ordering is inferior to that of the original; indeed, the system order is arguably more natural than the original order.

6 Conclusions

In this paper, we proposed a new structural topic model (strTM) to identify the latent topical structure in documents. Different from the traditional topic models, in which exchangeability assumption precludes them to capture the structure of a document, strTM captures the topical structure explicitly by introducing transitions among the topics. Experiment results show that both the identified topics and topical structure are intuitive and meaningful, and they are helpful for improving the performance of tasks such as sentence annotation and sentence ordering, where correctly recognizing the document structure is crucial. Besides, strTM is shown to outperform not only the baseline topic models that fail to model the dependency between the topics, but also the semi-supervised Naive Bayes model for the sentence annotation task.

Our work can be extended by incorporating richer features, such as named entity and co-reference, to enhance the model's capability of structure finding. Besides, advanced NLP techniques for document analysis, e.g., shallow parsing, may also be used to further improve structure finding.

7 Acknowledgments

We thank the anonymous reviewers for their useful comments. This material is based upon work supported by the National Science Foundation under Grant Numbers IIS-0713581 and CNS-0834709, and NASA grant NNX08AC35A.

References

R. Barzilay and M. Lapata. 2005. Collective content selection for concept-to-text generation. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 331–338.

R. Barzilay and L. Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of HLT-NAACL*, pages 113–120.

D.M. Blei and M.I. Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference*, pages 127–134.

D.M. Blei and J.D. Lafferty. 2007. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35.

D.M. Blei and P.J. Moreno. 2001. Topic segmentation with an aspect hidden Markov model. In *Proceedings of the 24th annual international ACM SIGIR conference*, page 348. ACM.

D.M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(2-3):993 – 1022.

H. Chen, SRK Branavan, R. Barzilay, and D.R. Karger. 2009. Global models of document structure using latent permutations. In *Proceedings of HLT-NAACL*, pages 371–379.

P. Diaconis and D. Ylvisaker. 1979. Conjugate priors for exponential families. *The Annals of statistics*, 7(2):269–281.

M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 562–569.

J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 Workshop on Automatic summarization*, pages 40–48.

T. Grenager, D. Klein, and C.D. Manning. 2005. Unsupervised learning of field segmentation models for information extraction. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 371–378.

T.L. Griffiths, M. Steyvers, D.M. Blei, and J.B. Tenenbaum. 2005. Integrating topics and syntax. *Advances in neural information processing systems*, 17:537–544.

Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic markov models. volume 2, pages 163–170.

T. Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

E.H. Hovy. 1993. Automated discourse generation using discourse structure relations. *Artificial intelligence*, 63(1-2):341–385.

M. Johnson. 2007. Why doesn't EM find good HMM POS-taggers. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 296–305.

M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. 1999. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

H. Kamp. 1981. A theory of truth and semantic representation. *Formal methods in the study of language*, 1:277–322.

M. Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484.

L. Lovász and M.D. Plummer. 1986. *Matching theory*. Elsevier Science Ltd.

Y. Lu and C. Zhai. 2008. Opinion integration through semi-supervised topic modeling. In *Proceeding of the 17th international conference on World Wide Web*, pages 121–130.

Daniel Marcu. 1998. The rhetorical parsing of natural language texts. In *ACL '98*, pages 96–103.

Q. Mei, X. Ling, M. Wondra, H. Su, and C.X. Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180.

L.R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

R. Soricut and D. Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the NAACL-HTC*, pages 149–156.

B. Sun, P. Mitra, C.L. Giles, J. Yen, and H. Zha. 2007. Topic segmentation with shared topic detection and alignment of multiple documents. In *Proceedings of the 30th ACM SIGIR*, pages 199–206.

ChengXiang Zhai, Atulya Velivelli, and Bei Yu. 2004. A cross-collection mixture model for comparative text minning. In *Proceeding of the 10th ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 743–748.

L. Zhuang, F. Jing, and X.Y. Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50.