An Adaptation of Topic Modeling to Sentences

Ruey-Cheng Chen†, Reid Swanson‡, and Andrew S. Gordon‡
†Dept. of Computer Science and Information, Engineering, National Taiwan University
1 Roosevelt Rd. Sec. 4, Taipei 106, Taiwan
‡Institute for Creative Technologies, University of Southern California
13274 Fiji Way, Marina del Rey, CA 90292
cobain@turing.csie.ntu.edu.tw, {swanson,gordon}@ict.usc.edu

Abstract

Advances in topic modeling have yielded effective methods for characterizing the latent semantics of textual data. However, applying standard topic modeling approaches to sentence-level tasks introduces a number of challenges. In this paper, we adapt the approach of latent-Dirichlet allocation to include an additional layer for incorporating information about the sentence boundaries in documents. We show that the addition of this minimal information of document structure improves the perplexity results of a trained model.

1 Introduction

Topic models, such as probabilistic latent-semantic analysis (Hofmann, 1999) and latent-Dirichlet allocation (Blei et al., 2003), were first introduced in the NLP community as a means of characterizing the latent semantics in textual data. A large body of research has made use of many successful variants of the original models. Generally, most of these previous efforts began with the assumption that a document is, in principle, a bag of topics, and words observed in the document are indirectly inferred from the underlying distribution of these background topics.

The generative process for realizing a collection of words from a set of topics relies on two distributions: the document-level topic distribution $\Pr(z|d)$ and the topic-level word distribution $\Pr(w|z)$, for topic z, document d and word w. This approach has garnered great success in various document-based tasks. However, there are problems in directly applying this approach in sentence-level tasks, e.g., clustering individual sentences by topic.

One way to apply LDA to sentence data is to treat individual sentences as whole documents. The problem with this approach is that it does not account for the context in which the individual sentences appear. Alternatively, the topic distribution of individual sentences could be treated as the same as that of the whole document. Here the problem is that the differences between that contributions of individual sentences to document topics is ignored. Given the large number of NLP applications where topic modeling could be applied to sentence-level tasks, a new model is needed.

This work adapts the standard LDA model to better account for the contribution of sentences to the topics of documents. Our approach is to add an additional layer in the standard LDA model that integrates information about the sentence boundaries in documents. The ensembles in this layer are indicators that point to a set of multinomial distributions over topics; each sentence in the document chooses a distribution to follow and generates the word topics accordingly. The idea is to offer a set of *switches* in between documents and topics, serving to balance the contribution of sentences and documents to the topic distribution.

In this paper, we present this new model as an adaptation of the standard LDA model. We demonstrate that our model performs better on test-set perplexity than the standard LDA model. These results suggest that this model may have applicability in future sentence-level tasks.

2 Sentence-Layered LDA

We propose a generative topic model, called *sentence-layered LDA*, that incorporates sentence boundaries into the original LDA framework. We introduce the notion of *sentence topics* by adding a set of latent variables which serve as additional sub-document constructs in between the document and the words. In this model, documents do not explicitly generate word topics, but instead guide sentences toward certain topic distributions by producing a set of sentence topics. The sentence topics are indicator scalars, pointing to specific discrete distributions over word topics. The adaptation can be seen as a number of LDA machines working as individuals at the sentence level being governed by the document node. The generative process is summarized by the following variables.

$$\begin{aligned} \theta^d &\sim \text{Dirichlet}(\alpha) \\ \tau^{(x_j)} &\sim \text{Dirichlet}(\gamma) \\ \phi^{(z_i)} &\sim \text{Dirichlet}(\beta) \\ x_j &\sim \text{Multinomial}(\theta^{(d)}) \\ z_i &\sim \text{Discrete}(\tau^{(x_j)}) \\ w_i &\sim \text{Discrete}(\phi^{(z_i)}) \end{aligned}$$

Besides the usual constructs inherited from the LDA model, the newly-introduced discrete distributions $\tau^{(x_j)}$ are governed by the Dirichlet prior γ. For simplicity, symmetric Dirichlet priors are assumed here. The complete plate notation of the model is shown in Figure 1.

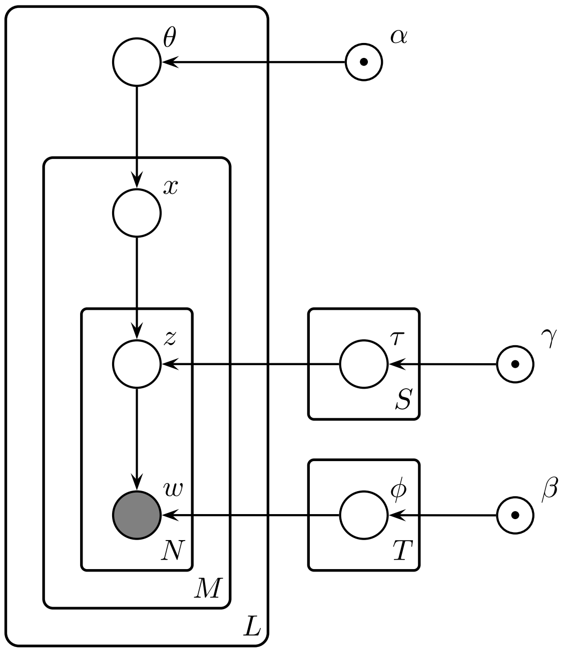

Figure 1: The sentence-layered LDA model in plate notation. S and T denotes the number of sentence and word topics; N, M, and L denotes the number of words, sentences, and documents, respectively.

Since the joint probability $\Pr(\mathbf{z}, \mathbf{x})$ is generally intractable, we break it down to two conditional probability distributions and integrate out the priors. The procedure largely follows the paradigm suggested in (Blei et al., 2003) and (Griffiths and Steyvers, 2004). The inference is done by a straightforward application of Gibbs sampler.

Conditional for $\mathbf{z_i}$. The conditional distribution $\Pr(z_i = j|\mathbf{z}_{-i}, \mathbf{w}, \mathbf{x})$ can be shown to be as follows.

$$\begin{aligned} &\Pr(z_i = j|\mathbf{z}_{-i}, \mathbf{w}, \mathbf{x}) \\ &\propto \Pr(w_i|z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) \Pr(z_i = j|\mathbf{z}_{-i}, x, \mathbf{x}') \\ &= \frac{\beta + n_{-i,j}^{(w_i)}}{V\beta + n_{-i,j}^{(\cdot)}} \frac{\gamma + n_{-i,x}^{(j)}}{T\gamma + n_{-i,x}^{(\cdot)}} \end{aligned}$$

Here, x denotes the sentence topic that induces z_i, $\mathbf{x}$ denotes the entire sentence set, and $\mathbf{x}' = \mathbf{x} - \{x\}$. The count for all the occurrences for word w that are of word topic j, excluding the current assignment of at i-th word, is represented as $n_{-i,j}^{(w)}$; the shorthand $n_{-i,j}^{(\cdot)}$ is a summation of the counts for all the words. In the second term, $n_{-i,x}^{(j)}$ denotes the number of words being assigned to topic j that are also governed by sentences assigned to the same topic as that of x, with the current assignment at i-th word left out. In other words, we consider only sentences that are of same sentence-level topic as that of x and count the number of words in these sentences being assigned to topic j.

Conditional for $\mathbf{x}$. The conditional $\Pr(x = l|\mathbf{x}', \mathbf{z})$ can be rewritten as:

$$\begin{aligned} &\Pr(x = l|\mathbf{x}', \mathbf{z}) \\ &\propto \Pr(\mathbf{z}^*|x = l, \mathbf{x}', \mathbf{z}') \Pr(x = l|\mathbf{x}', \mathbf{z}') \\ &= \frac{\mathcal{B}(\{\gamma + n_{-x,l}^{(z)} + n_x^{(z)}; \forall z\})}{\mathcal{B}(\{\gamma + n_{-x,l}^{(z)}; \forall z\})} \frac{\alpha + n_{-x,d}^{(l)}}{S\alpha + n_{-x,d}^{(\cdot)}} \end{aligned}$$

Here, $\mathbf{z}^*$ and $\mathbf{z}'$ denote all the topics governed by the current sentence and the others, respectively (i.e., $\mathbf{z} = \mathbf{z}^* \cup \mathbf{z}'$). We use $n_x^{(z)}$ to denote the count of words in the current sentences being assigned to topic z, and $n_{-x,d}^{(l)}$ to denote the count of sentences in document d that are of topic l, excluding the current assignment. Note that $\mathcal{B}(\cdot)$ is the multinomial beta function defined by

$$\mathcal{B}(a_1, \dots, a_n) = \frac{\prod_i \Gamma(a_i)}{\Gamma(\sum_i a_i)}.$$

Parameter Equation	Adjusted R^2
$\alpha = 0.6433 \times 1/S$	0.4872
$\beta = 1.46 \times 10^{-4*} \times S + 1.4528713 \times 1/T$	0.877
$\gamma = 5.276 \times 10^{-5} \times S + 0.2156 \times 1/T$	0.9135

Table 1: The linear fit for model parameters. All weight values with are significant at $p < 0.001$, except $(^*)$ significant at $p < 0.1$.

3 Parameter Estimation

The performance of LDA-based models such as this one depends largely on the parameters for the Dirichlet priors. To the best of our knowledge, there is no standard approach for optimizing these parameters; they are normally determined through experimentation. Accordingly, we used the following experimental procedure to estimate the parameters of our model.

We decided to learn the parameters on a large, multi-topic corpus. We chose to use the ICWSM 2009 Spinn3r dataset of 44 million weblog posts (Burton et al., 2009) due to its large size and broad coverage over topics. We took a sample of the corpus and divided it into the training, test, and development sets.

Our approach was to learn the optimal parameters (α, β, γ) for various numbers of word and sentence topics. To do this, we conducted a three-dimensional grid search for these parameters when given 5, 10, 50, or 100 word or sentence topics (4×4 possible combinations). For each (α, β, γ) triple, we assigned values of 0.05, 0.01, 0.005, and 0.001 to each parameter and trained the sentence-layered model. The trained models were tested against the development set to calculate the test-set perplexity. We selected the model with the lowest test-set perplexity and recorded the corresponding (α, β, γ) parameters for the given number of sentence and word topics.

Next, we formed a linear regression model based on the recorded data. We derived a linear fit function for each parameter for a given number of word/sentence topics. The resulting equations are shown in Table 1.

4 Perplexity Evaluation

One way to assess the performance of a topic model is through test-set perplexity. Test-set perplexity evaluates how well the model generalizes on a held-out data. Perplexity is measured in the amount of uncertainty, which is a positive real number. The lower perplexity we achieve for the model, the better fit the model is to the data. The test-set perplexity is calculated as:

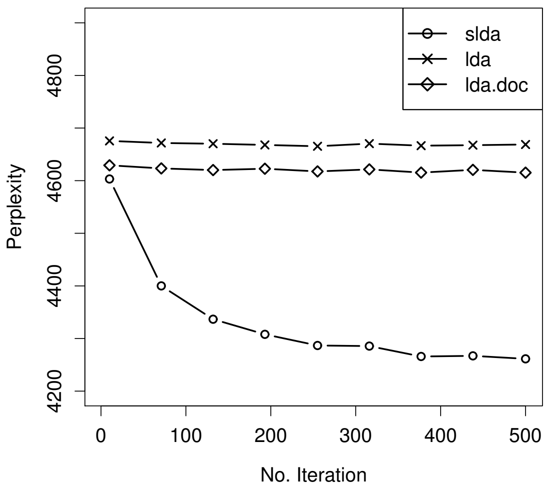

Figure 2: Performance of the three topic models in terms of test-set perplexity on the Penn Tree-Bank. The sentence-layered LDA is denoted as *slda*, the original LDA trained on sentence level as *lda*, and the original LDA trained on document level as *lda.doc*

$$\text{perplexity}_{\text{test}} = \exp(\frac{-\sum_w \log \Pr(w|d_{\text{test}})}{N_{\text{test}}})$$

The formulation of $\Pr(w_{\text{test}})$ is different for each topic model. For standard LDA, the probability is evaluated as follows.

$$\Pr(w|d_{\text{test}}) = \sum_z \Pr(w|z) \Pr(z|d)$$

The probability for the sentence LDA is a bit more complicated due to an additional layer of latent semantics.

$$\Pr(w|d_{\text{test}}) = \sum_x \sum_z \Pr(w|z) \Pr(z|x) \Pr(x|d)$$

To calculate perplexity for each model, we begin with probabilities $\Pr(w|z)$ learned from the training set. Since the training and test sets did not share documents or sentences, the probabilities related to z, x, and d needed to be re-sampled using texts from the test data. We employed the same Gibbs sampling procedure with the initialized values for $\Pr(w|z)$ and started another 500-iteration burn-in on the test sets. For simplicity, we read out only one sample for $\Pr(z|x)$, $\Pr(x|d)$, and $\Pr(z|d)$ at the end of burn-in.

As a corpus for evaluation, we chose the documents in the Penn TreeBank (Marcus et al., 1994). The Penn TreeBank is composed of 2,312 parsed documents, where each document contains 21.3 sentences on average. We further divided the TreeBank into two sets, one of 2,300 documents and the other of 154. We trained the topic models on the first set and evaluated the perplexity on the second. For both models, we had the Gibbs sampler burned in for 1,000 iterations before reading out probability estimates.

We set up the parameters of the standard LDA model as suggested in (Griffiths and Steyvers, 2004), i.e., $\alpha = 50/T$ and $\beta = 0.01$. For the sentence-layered LDA model, the parameters were determined by applying the regression results described in Section 3. In this experiment, we tested a total of three topic models: our sentence-layered LDA model and two standard LDA models, one trained at the document level and the other at the sentence level. When trained at the document level, LDA discards sentence boundaries and treats the entire document as one text unit. When trained at the sentence level, LDA treat each sentence as individual "documents". Tests were made only at the sentence level.

The number of word topics T for all models were set to 10; we figured the word topics in both models are functionally equivalent and therefore should be set to the same value. The number of sentence topics for the sentence-layered model were assigned to 20, which empirically achieved the best performance in our preliminary tests.

The experimental results showed that our sentence-layered LDA model achieved lower test-set perplexity than the standard LDA models. As shown in Figure 2, the perplexity for our model converged in roughly 300 iterations, while the LDA models were stable at an early stage. We believe that both of the standard LDA models suffered from one of two issues. When the LDA model was applied at the sentence level, the association of topic words did not span across the sentence boundaries, resulting in a loss of co-occurrence statistics. Likewise, when the LDA model was trained at the document level, it disregarded the structure of the text and treated all the word occurrences as equally important. As a result, the model overfitted the data and failed to address the diversity of co-occurred words at the sentence level. Our sentence-layered LDA model avoids these problems by adding a minimal layer of document structure directly into the model, allowing for a balance of inferred latent semantics between the document level and the sentence level.

5 Related Work

This work was inspired by a number of efforts for extending the standard LDA model. Our work differs from Pachinko allocation model (Li and McCallum, 2006), nested Chinese restaurant process (Blei et al., 2004), and mixture network (Heinrich, 2009), each of which allows an arbitrary number of sub-document units and an arbitrary number of dependency links in between to model topic correlations. Instead, our model adheres strictly to the sentence boundaries to define document structure. Our approach is most similar to that of the latent-Dirichlet co-clustering model (Shafiei and Milios, 2006). Our work differs in that it utilizes multiple sentence-level LDA machines to better account for the contribution of sentences to the topics of documents.

6 Conclusions

Although these perplexity results are encouraging, there remains difficult challenges in directly applying the topic distributions of our model in sentence-level tasks. Because there are two topic distributions in our model, some additional consideration is needed to coordinate the contribution of each to the generation of words. Additionally, good perplexity results on their own do not guarantee that the model captures the features most needed in topic modeling tasks. These challenges will be the focus of our future work.

References

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

D. Blei, T.L. Griffiths, M.I. Jordan, and J.B. Tenenbaum. 2004. Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*, 16:106.

K. Burton, A. Java, and I. Soboroff. 2009. The ICWSM 2009 Spinn3r Dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*.

T.L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5228.

G. Heinrich. 2009. A generic approach to topic models. In *ECML/PKDD*. Springer.

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. ACM New York, NY, USA.

W. Li and A. McCallum. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine learning*, page 584. ACM.

M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

M.M. Shafiei and E.E. Milios. 2006. Latent dirichlet co-clustering. In *Proceedings of the Sixth International Conference on Data Mining*, pages 542–551. IEEE Computer Society.