Syntactic Topic Models

Ke Zhai[*]
Computer Science
University of Maryland

Jordan Boyd-Graber[**]
Institute for Advanced Computer Studies
University of Maryland

David M. Blei[†]
Computer Science Department
Princeton University

The syntactic topic model (STM) is a Bayesian nonparametric model of language that discovers latent distributions of words (topics) that are both topically and syntactically coherent. The STM models dependency parsed corpora where sentences are grouped into documents. It assumes that each word is drawn from a latent topic chosen by combining document-level features and the local syntactic context. Each document has a distribution over latent topics, as in topic models, which provides the semantic consistency. Each element in the dependency parse tree also has a distribution over the topics of its children, as in latent-state syntax models, which provides the syntactic consistency. These distributions are convolved so that the topic of each word is likely under both its document and syntactic context. We derive a fast posterior inference algorithm based on variational methods. We report qualitative and quantitative studies on both synthetic data and hand-parsed documents. We show that the STM is a more predictive model of language than current models based only on syntax or only on topics.

When we read a sentence, we use two kinds of reasoning: one for understanding its syntactic structure and another for integrating its meaning into the wider context of other sentences, other paragraphs, and other documents. Both mental processes are crucial, and psychologists have found that they are distinct. A syntactically correct sentence that is semantically implausible takes longer for people to understand than its semantically plausible counterpart (Rayner et al. 1983). Furthermore, recent brain imaging experiments have localized these processes in different parts of the brain (Dapretto and Bookheimer 1999). Both of these types of reasoning should be accounted for in a probabilistic model of language.

To see how these mental processes interact, consider the following sentence from a travel brochure,

> Next weekend, you could be relaxing in ____.

How do we reason about filling in the blank? First, because the missing word is the object of a preposition, it should act like a noun, perhaps a location like "bed," "school," or "church." Second, because the document is about travel, we expect travel-related terms. This further restricts the space of possible terms, leaving alternatives like "Nepal," "Paris," or "Bermuda" as likely

[*] 3126 AV Williams, College Park, MD 20742. E-mail: zhaike@umiacs.umd.edu.
[**] 3219 AV Williams, College Park, MD 20742. E-mail: jbg@umiacs.umd.edu.
[†] 35 Olden Street, Princeton NJ, 08540. E-mail: blei@cs.princeton.edu.

possibilities. Each type of reasoning restricts the likely solution to a subset of words, but the best candidates for the missing word are in their *intersection*.

In this article we develop a probabilistic model of language that mirrors this process. Probabilistic modeling has emerged as a powerful formalism for expressing assumptions about natural language and analyzing texts under those assumptions (Manning and Schütze 1999). Current models, however, tend to focus on finding and exploiting either syntactic or thematic regularities. On one hand, *probabilistic syntax models* capture how different words are used in different parts of speech and how those parts of speech are organized into sentences (Charniak 1997; Collins 2003; Klein and Manning 2002). On the other hand, *probabilistic topic models* find patterns of words that are topically related in a large collection of documents (Blei et al. 2003; Griffiths et al. 2007).

Each type of model captures one kind of regularity in language, but ignores the other kind of regularity. Returning to the example, suppose that the correct answer is the noun "Bermuda." A syntax model would fill in the missing word with a noun, but would ignore the semantic distinction between words like "bed" and "Bermuda."[1] A topic model would consider travel words to be more likely than others, but would ignore functional differences between words like "sailed" and "Bermuda." To arrive at "Bermuda" with higher probability requires a model that simultaneously accounts for both syntax and topic.

Thus, our model assumes that language arises from an interaction between syntactic regularities at the sentence level and topical regularities at the document level. The syntactic component examines the sentence at hand and restricts attention to nouns; the topical component examines the rest of the document and restricts attention to travel words. Our model makes its ultimate prediction from the intersection of these two restrictions. As we will see, these modeling assumptions lead to a more predictive model of language.

In general, hierarchical Bayesian models of language posit that the observed words arise probabilistically via hidden structure, such as syntactic structure or topical structure. Given a collection of texts, one uses *posterior inference* to uncover the hidden structure from the observed language. In topic models, one uncovers patterns of words that evince a coherent topic; in syntax models, one uncovers patterns of words that share a similar functional roles.

Both topic models and syntax models assume that each word of the data is drawn from a mixture component, a distribution over a vocabulary that represents recurring patterns of words. The central difference between topic models and syntax models is how the component weights are shared: topic models are bag-of-words models where component weights are shared within a document; syntax models share components within a functional category (e.g. the production rules for non-terminals). Components learned from these assumptions reflect either document-level patterns of co-occurrence, which look like themes, or tree-level patterns of co-occurrence, which look like syntactic elements. In both topic models and syntax models, Bayesian nonparametric methods are used to embed the choice of the number of components into the model (Teh et al. 2006; Finkel et al. 2007). These methods further allow for new components to appear with new data.

In the *syntactic topic model* (STM), the components arise from both document-level and sentence-level distributions and therefore reflect both syntactic and thematic patterns in the texts. This captures the two types of understanding described above: the document-level distribution over components restricts attention to those that are thematically relevant; the tree-level distribution over components restricts attention to those that are syntactically appropriate. We emphasize that

[1] A proponent of lexicalized parsers might argue that conditioning on the word might be enough to answer this question completely. However, many of the most frequently used words have such broad meanings (e.g. "go") that knowledge of the broader context is necessary.

rather than choose between a thematic component or syntactic component from its appropriate context, as is done in the model of Griffiths et al (**?**), components are drawn that are consistent with both sets of weights.

This complicates posterior inference algorithms and requires developing new methodology in hierarchical Bayesian modeling of language. However, it leads to a more expressive and predictive model. In Section 1 we review latent variable models for topics and syntax and Bayesian nonparametric methods. In Section 2, building on these formalisms, we present the STM. In Section 2.2 we derive a fast approximate posterior inference algorithm based on variational methods. Finally, in Section 3 we present qualitative and quantitative results on both synthetic text and a large collection of parsed documents.

1. Background: Topics and Syntax

Our approach builds on probabilistic topic models, probabilistic syntactic models, and Bayesian nonparametric methods. We review these ideas here.

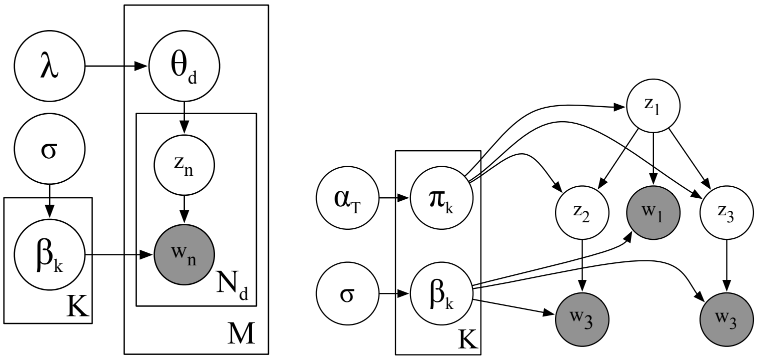
1.1 Probabilistic Topic Models

Probabilistic topic models are hierarchical Bayesian models of text that can be used to automatically discover a hidden thematic structure in a large collection of otherwise unstructured documents. Topic models have emerged as a powerful tool for unsupervised analysis of text (Blei et al. 2003) and have been extended in many ways, e.g., to authorship (Rosen-Zvi et al. 2004), citation (Mimno and McCallum 2007), sentiment analysis (Blei and McAuliffe 2007; Titov and McDonald 2008), corpus exploration (Hall et al. 2008), part-of-speech labeling (Toutanova and Johnson 2008), discourse segmentation (Purver et al. 2006), word sense induction (Brody and Lapata 2009), and word sense disambiguation (Boyd-Graber et al. 2007). Topic models have also been applied to non-language data, such as images (**?**), population genetics (Pritchard et al. 2000), and music (Hu and Saul 2009). There are several reviews of topic modeling and related literature (Blei and Lafferty 2009; Griffiths et al. 2007).

Here we will build on latent Dirichlet allocation (LDA) (Blei et al. 2003), which is often used as a building block for other topic models. The modeling assumptions behind LDA are made clear through its generative probabilistic process, the imaginary process by which a document collection is created. LDA posits that there are K topics in a collection, each of which is a distribution over terms. For each document, LDA first draws a vector of topic proportions from a Dirichlet distribution and then draws each word from a topic which is chosen from those proportions. The corpus is associated with a set of topics, and each document as associated with a random mixture of those topics. In statistics, these kinds of assumptions are called mixed-membership assumptions (Erosheva et al. 2007).

Analyzing a corpus with LDA amounts to “reversing” this process to compute the posterior distribution of the topic proportions, topic assignments, and topics conditioned on the observed documents. Of particular interest are the topics themselves, which reflect corpus-wide patterns of word co-occurrence, and the topic proportions, which describe the documents in terms of their constituent topics.[2] Notice that LDA ignores the order of words within a document but uses the document context to make inferences about the topics. For example, the term “stock” might have a high probability in both a financial topic and a culinary topic. But if “stock,” “soup,” and “broth”

[2] The topics tend to correspond to a psychologically plausible interpretation of the themes that pervade the documents (Griffiths et al. 2007). Thus, they are called topics.



(a) latent Dirichlet allocation

(b) finite tree with independent children

Figure 1
This figure introduces the graphical model notation used throughout the paper and illustrates two models: latent Dirichlet allocation (LDA) and the finite tree with independent children (FTIC). The rectangular plates denote replication, and the numbers in the lower right denote how often the variables inside the plate are replicated. Nodes represent random variables; edges indicate possible probabilistic dependence; shaded variables are observed; unshaded variables are hidden. For LDA (left), topic distributions $\boldsymbol{\beta}_k$ are drawn for each of the K topics, topic proportions $\boldsymbol{\theta}_d$ are drawn for each of each of the M documents, and topic assignments $z_{d,n}$ and words $w_{d,n}$ are drawn for each of the N_d words in a document. For FTIC (right), each state has a distribution over words, $\boldsymbol{\beta}$, and a distribution over successors, $\boldsymbol{\pi}$. Each word is associated with a hidden state z_n, which is chosen from the distribution $\boldsymbol{\pi}_{z_{p(n)}}$, the transition distribution based on the parent node's state.

also appear in the document, the posterior will likely assign appearances of "stock" to the culinary topic.

Topic models represent a fully probabilistic perspective on techniques like latent semantic analysis (LSA) (Deerwester et al. 1990) and probabilistic latent semantic analysis (pLSA) (Hofmann 1999). LSA and pLSA do not embody fully generative probabilistic processes. By adopting a fully generative model, LDA exhibits better generalization performance and is more easily used as a module in more complicated models. (Blei et al. 2003; Blei and Lafferty 2009).

1.2 Probabilistic Syntax Models

LDA is effective at capturing semantic correlations between words, but it ignores syntactic correlations and connections. The finite tree with independent children model (FTIC) can be seen as the syntactic complement to LDA (Finkel et al. 2007). As in LDA, this model assumes that observed words are generated by latent states. However, rather than considering words in the context of their shared document, the FTIC considers each word in the context of its sentence as determined by its location in a dependency parse.

The FTIC embodies a generative process over a collection of sentences with given parses. It is parameterized by a set of "syntactic states," where each state is associated with three parameters: a distribution over terms, a set of transition probabilities to other states, and a probability of being chosen as the root state. Each sentence is generated by traversing the structure of the parse tree. For each node, draw a syntactic state from the transition probabilities of its parent (or root probabilities) and draw the word from the corresponding distribution over terms. A parse of a sentence with three words is depicted as a graphical model in Figure 1.

While LDA is constructed to analyze a collection of documents, the FTIC is constructed to analyze a collection of parsed sentences. The states discovered through posterior inference correlate with part of speech labels (Finkel et al. 2007). For LDA the components respect the way words co-occur in documents. For FTIC the components respect the way words occur within parse trees.

1.3 Random Distributions and Bayesian nonparametric methods

Many recently developed probabilistic models of language, including those described above, employ distributions as random variables. These random distributions are sometimes a prior over a parameter, as in traditional Bayesian statistics, or a latent variable within the model. For example, in LDA the topic proportions and topics are random distributions; in the FTIC, the transition probabilities and term generating distributions are random.

In this section, we review the Dirichlet distribution, a commonly used distribution of multinomial parameter vectors, and the stick breaking distribution, a distribution on multinomial parameter vectors with a countably infinite number of components. We will describe the connection between the stick-breaking distribution and the Dirichlet process (DP), which is a distribution over arbitrary discrete distributions and a foundational building block of Bayesian nonparametric methods. These distributions are pivotal to the STM.

A $(K-1)$ dimensional Dirichlet distribution is a distribution over finite probability distributions of K elements. Thus its support is the simplex, non-negative vectors that sum to one.[3] It is parameterized by a mean value $\boldsymbol{\rho}$, which is a point on the $(K-1)$ simplex, and a scalar λ, which controls the variance around the mean. A random variable drawn from a Dirichlet is denoted $\theta \sim \text{Dir}(\lambda\boldsymbol{\rho})$.

In LDA, for example, the per-document topic proportions are drawn from a $(K-1)$ dimensional Dirichlet and the topics themselves are assumed drawn from a $(V-1)$ dimensional Dirichlet. (Recall that K is the number of topics and V is the number of terms in the vocabulary.) In the FTIC, the syntactic states are assumed drawn from a $(V-1)$ dimensional Dirichlet, and the transition probabilities between states are drawn from a $(K-1)$ dimensional Dirichlet.[4]

Both the FTIC and LDA assume that the number of latent components, i.e., topics or syntactic states, is fixed. Choosing this number a priori can be difficult. Recent research has extended Bayesian nonparametric methods to build more flexible models where the number of latent components is unbounded and is determined by the data (Teh et al. 2006; Liang and Klein 2007). The STM uses this methodology.

We first describe the stick breaking distribution, a distribution over the infinite simplex. The idea behind this distribution is to draw an infinite number of Beta random variables, i.e., values between zero and one, and then combine them to form a vector whose infinite sum is one. This can be understood with a stick-breaking metaphor. Consider a unit length stick that is infinitely broken into smaller and smaller pieces. The length of each successive piece is determined by taking a random proportion of the remaining stick. The random proportions are drawn from a Beta distribution,

$$\mu_k \sim \text{Beta}(1, \alpha),$$

[3] The dimensionality is $(K-1)$ rather than K because of the constraint that the vector sum to one.

[4] The Dirichlet distribution is a convenient distribution for generating multinomials, but there are other alternatives that provide different sparsity or correlation patterns. These have proved promising in limited-data frameworks; the logistic normal prior has been applied to grammar induction (Cohen et al. 2008) and integer programming has been applied to unsupervised part of speech tagging (Ravi and Knight).

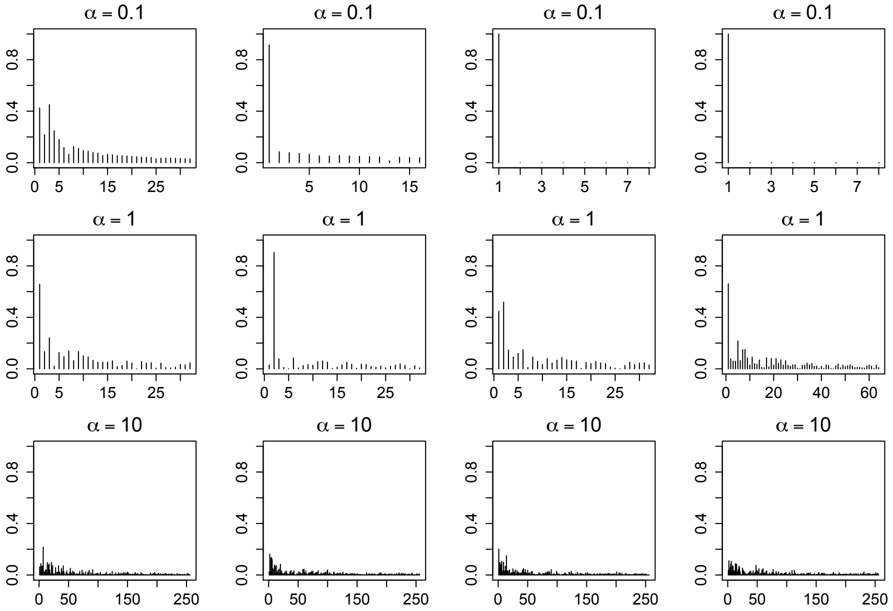


Figure 2
Draws for three settings of the parameter α of a stick-breaking distribution (enough indices are shown to account for 0.95 of the probability). When the parameter is substantially less than one (top row), very low indices are favored. When the parameter is one (middle row), the weight tapers off more slowly. Finally, if the magnitude of the parameter is larger (bottom row), weights are nearer a uniform distribution.

and the resulting stick lengths are defined from these breaking points,

$$\tau_k = \mu_k \prod_{l=1}^{k-1} (1 - \mu_l).$$

With this process, the vector τ is a point on the infinite simplex (Sethuraman 1994). This distribution is notated $\tau \sim$ GEM(α).[5]

The stick breaking distribution is a size-biased distribution—the probability tends to concentrate around the initial components. The Beta parameter α determines how many components of the probability vector will have high probability. Smaller values of α result in a peakier distributions; larger values result in distributions that are more spread out. Regardless of α, for large enough k, the value of τ_k still goes to zero because the vector must sum to one. Figure 2 illustrates draws from the stick breaking distribution for several values of α.

The stick-breaking distribution provides a constructive definition of the Dirichlet process, which is a distribution over arbitrary distributions [Ferguson 1973]. Consider a base distribution G_0, which can be any type of distribution, and the following random variables

$$\begin{aligned} \tau_i &\sim \text{GEM}(\alpha) \quad i \in \{1, 2, 3, \ldots\} \\ \mu_i &\sim G_0 \quad i \in \{1, 2, 3, \ldots\}. \end{aligned}$$

[5] GEM stands for Griffiths, Engen and McCloskey (Pitman 2002).

Now define the random distribution

$$G = \sum_{i=1}^{\infty} \tau_i \delta_{\mu_i}(\cdot)$$

which places mass τ_i on the point μ_i. This is a random distribution because its components are random variables, and note that it is a discrete distribution even if G_0 is defined on a continuous space. Marginalizing out τ_i and μ_i, the distribution of G is called a Dirichlet process (DP). It is parameterized by the base distribution G_0 and a scaling parameter α. The scaling parameter, as for the finite Dirichlet, determines how close the resulting random distribution is to G_0. Smaller α yields distributions that are further from G_0; larger α yields distributions that are closer to G_0.[6] The base distribution is also called the mean of the DP because $\mathrm{E}[G \,|\, G_0, \alpha] = G_0$. The Dirichlet process is a commonly used prior in Bayesian nonparametric statistics, where we seek a prior over arbitrary distributions (Antoniak 1974; Escobar and West 1995; Neal 2000).

In a hierarchical model, the DP can be used to define a topic model with an unbounded number of topics. In such a model, unlike LDA, the data determine the number of topics through the posterior and new documents can ignite previously unseen topics. This extension is an application of a hierarchical Dirichlet process (HDP), a model of grouped data where each group arises from a DP whose base measure is itself a draw from a DP (Teh et al. 2006). In the HDP for topic modeling, the finite dimensional Dirichlet distribution over per-document topic proportions is replaced with a draw from a DP, and the base measure of that DP is drawn once per-corpus from a stick-breaking distribution. The stick-breaking random variable describes the overall prominence of topics in a collection; the draws from the Dirichlet process describe how each document exhibits those topics.

Similarly, applying the HDP to the FTIC model of Section 1.2 results in a model where the mean of the Dirichlet process represents the overall prominence of syntactic states. This extension is described as the infinite tree with independent children (ITIC) (Finkel et al. 2007). For each syntactic state, the transition distributions drawn from the Dirichlet process allow each state to prefer certain children states in the parse tree. Other work has applied this nonparametric framework to create language models (Teh 2006), full parsers for Chomsky normal form grammars (Liang et al. 2007), models of lexical acquisition (Goldwater 2007), synchronous grammars (Blunsom et al. 2008), and adaptor grammars for morphological segmentation (Johnson et al. 2006).

2. The Syntactic Topic Model

Topic models like LDA and syntactic models like FTIC find different decompositions of language. Syntactic models ignore document boundaries but account for the order of words within each sentence–thus the components of syntactic models reflect how words are used in sentences. Topic models respect document boundaries but ignore the order of words within a document–thus the components of topic models reflect how words are used in documents. We now develop the syntactic topic model (STM), a hierarchical probabilistic model of language that finds components which reflect both the syntax of the language and the topics of the documents.

[6] The formal connection between the DP and the finite dimensional Dirichlet is that the finite dimensional distributions of the DP are finite Dirichlet, and the DP was originally defined via the Kolmogorov consistency theorem(Ferguson 1973). The infinite stick breaking distribution was developed for a more constructive definition (Sethuraman 1994). We will not be needing these mathematical details here.

For the STM, our observed data are documents, each of which is an exchangeable collection (i.e. a bag) of dependency parse trees. (Note that in LDA, the documents are simply collections of words.) The main idea is that words arise from topics, and that topic occurrence depends on both a document-level variable and parse tree-level variable. We emphasize that, unlike a parser, the STM does not model the tree structure itself and nor does it use any syntactic labeling. While it would be possible to add syntactic labeling to the model, one of our goals is to contrast the learned clusters of words with traditional syntactic categories. Thus, only the words as observed in the tree structure are modeled.

The document-level and parse tree-level variables are both distributions over topics, which we call topic weights. These distributions are never drawn from directly. Rather, they are convolved—that is, they are multiplied and renormalized—and the topic assignment for a word is drawn from the convolution. The parse-tree level topic weight enforces syntactic consistency and the document-level topic weight enforces thematic consistency. The resulting set of topics—the distributions over words that the topic weights refer to—will be those that thus reflect both thematic and syntactic constraints. Our model is a Bayesian nonparametric model, so the number of such topics is determined by the data.

We now describe this model in more mathematical detail. The STM contains topics ($\boldsymbol{\beta}$), transition distributions ($\boldsymbol{\pi}$), per-document topic weights ($\boldsymbol{\theta}$), and top level weights ($\boldsymbol{\tau}$) as hidden random variables. In the STM, *topics* are multinomial distributions over a fixed vocabulary (β_k). Each topic maintains a *transition vector* which governs the topics assigned to children of parents assigned a given topic ($\boldsymbol{\pi}_k$). *Document weights* model how much a document is about specific topics. Finally, each word has a *topic assignment* ($z_{d,n}$) that decides from which topic the word is drawn. The STM posits a joint distribution using these building blocks and, from the posterior conditioned on the observed documents, we find transitions, per-document topic distributions, and topics.

As mentioned, we use Bayesian nonparametric methods to avoid having to set the number of topics. We assume that there is a vector τ of infinite length which tells us which topics are actually in use (as discussed in Section 1.3). These top-level weights are a random probability distribution drawn from a stick-breaking distribution. Putting this all together, the generative process for the data is as follows:

1. Choose global weights $\boldsymbol{\tau} \sim$ GEM(α)
2. For each topic index $k = \{1, \dots\}$:
   (a) Choose transition distribution $\boldsymbol{\pi}_k \sim$ DP$(\alpha_T \boldsymbol{\tau})$
3. For each document $d = \{1, \dots M\}$:
   (a) Choose document weights $\boldsymbol{\theta}_d \sim$ DP$(\alpha_D \boldsymbol{\tau})$
   (b) For each sentence root node with index $(d, r) \in$ SENTENCE-ROOTS$_d$:
      i. Choose topic assignment $z_{d,r} \propto \boldsymbol{\theta}_d \boldsymbol{\pi}_{start}$
      ii. Choose root word $w_{d,r} \sim$ mult$(1, \beta_{z_r})$
   (c) For each additional child with index (d, c) and parent with index (d, p):
      i. Choose topic assignment

$$z_{d,c} \propto \boldsymbol{\theta}_d \boldsymbol{\pi}_{z_{d,p}} \tag{1}$$

      ii. Choose word $w_{d,c} \sim$ mult$(1, \boldsymbol{\beta}_{z_{d,n}})$

This process is illustrated as a probabilistic graphical model in Figure 3. Note that Equation 1 assumes that the parent's topic has already been drawn. This enforces that the topics are chosen in an order that respects the given parse tree: the topic assignments of a sentence are chosen starting at the root and then downward until the leaves of the tree are reached.

Data analysis with this model amounts to "reversing" this process to determine the posterior distribution of the latent variables. The posterior distribution is conditioned on observed words organized into parse trees and documents. It provides a distribution over all of the hidden structure—the topics, the syntactic transition probabilities, the per-document topic weights, and the corpus-wide topic weights.

Because both documents and local syntax shape the choice of possible topics for a word, the posterior distribution over topics favors topics that are consistent with *both* contexts. For example, placing all nouns in a single topic would respect the syntactic constraints but not the thematic, document-level properties, as not all nouns are equally likely to appear in a given document. Instead, the posterior prefers topics which would divide syntactically similar words into different categories based on how frequently they co-occur in documents.

In addition to determining what the topics are, i.e., which words appear in a topic with high probability, the posterior also defines a distribution over how those topics are used. It encourages topics to appear in similar documents based on the per-document topic distributions θ and encourages topics to appear in similar local syntactic contexts based on the transition distribution π. For each word, two different views of its generation are at play. On one hand, a word is part of a document and reflects that document's themes. On the other hand, a word is part of a local syntactic structure and reflects the likely type of word that is associated with a child of its parent. The posterior balances both these views to determine which topic is associated with each word.

Finally, through the stick-breaking and DP machinery, the posterior selects the number of topics that are used. This strikes a balance between explaining the data well (e.g. reflecting syntax and document-level properties) and not using too many topics, as governed by the hyperparameter α (see Section 1.3).

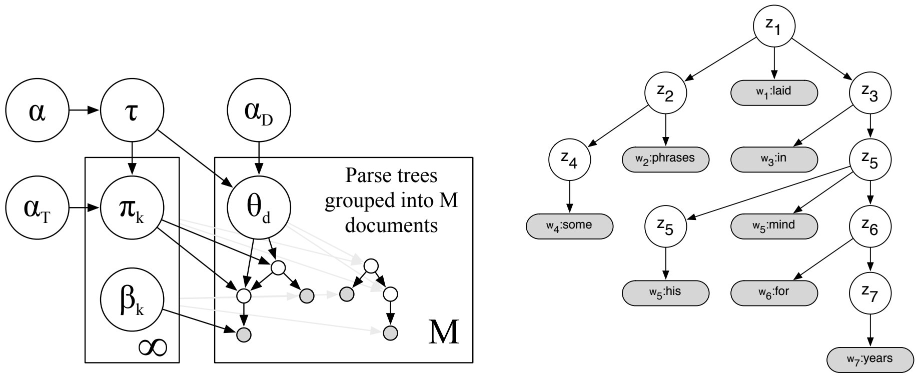
As we will see below, combining document-level properties and syntax (Equation 1) complicates posterior inference (compared to HDP or ITIC) but allows us to simultaneously capture both syntactic and semantic patterns. Under certain limiting assumptions, the STM reduces to the models discussed in Section 1 . The STM reduces to the HDP if we fix $\boldsymbol{\pi}$ to be a vector of ones, thus removing the influence of the tree structure. The STM reduces to the ITIC if we fix $\boldsymbol{\theta}$ to be a vector of ones, removing the influence of the documents.

2.1 Relationships to Other Work

The STM attempts to discover patterns of syntax and semantics simultaneously. In this section, we review previous methods to model syntax and semantics simultaneously and the statistical tools that we use to combine syntax and semantics. We also discuss other methodologies from word sense disambiguation, word clustering, and parsers that are similar to the STM.

While the STM combines topics and syntax using a single distribution (Equation 1), an alternative is, for each word, to choose one of the two distributions. In such a model, the topic assignment comes from either the parent's topic transition $\boldsymbol{\pi}_{z_{(d,p)}}$ or from the document weights $\boldsymbol{\theta}_d$, based on a binary selector variable (instead of being drawn from a product of the two distributions). Griffiths et al's topics and syntax model (**?**) did this on the linear order of words in a sentence. A mixture of topics and syntax in a similar manner over parse trees would create different types of topics, individually modeling either topics or syntax. It would not, however, enforce consistency with parent nodes *and* a document's themes. A word need only be consistent with either view.

Rather, the STM draws on the idea behind the product of experts (Hinton 1999), multiplying two vectors and renormalizing to obtain a new distribution. Taking the point-wise product can be thought of as viewing one distribution through the "lens" of another, effectively choosing only words whose appearance can be explained by both.



(a) syntactic topic model

(b) sentence graphical model

Figure 3
In this graphical model depiction of the syntactic topic model, the dependency parse representation of FTIC in Figure 1(b) are grouped into documents, as in LDA in 1(a). For each of the words in the sentence, the topics weights of a document $\boldsymbol{\theta}$ and the parent's topic transition $\boldsymbol{\pi}$ together choose the topic. (For clarity, some of the sentence node dependencies have been grayed out.) An example of the structure of a sentence is on the right, as demonstrated by an automatic parse of the sentence "Some phrases laid in his mind for years." The STM assumes that the tree structure and words are given, but the latent topics $\boldsymbol{z}$ are not.

Instead of applying the lens to the selection of the latent classes, the topics, once selected, could be altered based on syntactic features of the text. This is the approach taken by TagLDA (Zhu et al. 2006), where each word is associated with a single tag (such as a part of speech), and the model learns a weighting over the vocabulary terms for each tag. This weighting is combined with the per-topic weighting to emit the words. Unlike the STM, this model does not learn relationships between different syntactic classes and, because the tags are fixed, cannot adjust its understanding of syntax to better reflect the data.

There has also been other work that does not seek to model syntax explicitly but nevertheless seeks to use local context to influence topic selection. One example is the hidden topic Markov model (Gruber et al. 2007), which finds chains of homogeneous topics within a document. Like the STM and Griffiths et al, the HTMM sacrifices the exchangibility of a topic model to incorporate local structure. Similarly, Wallach's bigram topic model (Wallach 2006) assumes a generative model that chooses topics in a fashion identical to LDA but instead chooses words from a distribution based on per-topic bigram probabilities, thus partitioning bigram probabilities across topics.

A similar vein of research is discourse-based WSD methods. The Yarowsky algorithm, for instance, uses clusters of similar contexts to disambiguate the sense of a word in a given context (Yarowsky 1995; Abney 2004). While the result does not explicitly model syntax, it does have a notion of both document theme (as all senses in a document must have the same sense) and the local context of words (the feature vectors used for clustering mentions). However, the algorithm is only defined on a word-by-word basis and does not build a consistent picture of the corpus for all the words in a document.

Local context is better captured by explicitly syntactic models. Work such as Lin similarity (Lin 1998) and semantic space models (Padó and Lapata 2007) build sets of related terms that appear in similar syntactic contexts. However, they cannot distinguish between uses that always appear in different kinds of documents. For instance, the string “fly” is associated with both terms from baseball and entomology.

These syntactic models use the output of parsers as input. Some parsing formalisms, such as adaptor grammars (Johnson et al. 2006; Johnson and Goldwater 2009), are broad and expressive enough to also describe topic models. However, there has been no systematic attempt to combine syntax and semantic in a unified framework. The development of statistical parsers has increasingly turned to methods to refine the latent classes that generate the words and transitions present in a parser. Whether through subcategorization (Klein and Manning 2003) or lexicalization (Collins 2003; Charniak 2000), broad categories are constrained to better model idiosyncrasies of the text. While the STM is not a full parser, it offers an alternate way of constraining the latent classes of terms to be consistent across similar documents.

2.2 Posterior inference with variational methods

We have described the modeling assumptions behind the STM. As detailed, the STM assumes a decomposition of the parsed corpus by a hidden semantic and syntactic structure encoded with latent variables. Given a data set, the central computational challenge for the STM is to compute the posterior distribution of that hidden structure given the observed documents, and data analysis proceeds by examining this distribution. Computing the posterior is “learning from data” from the perspective of Bayesian statistics.

This posterior distribution, as for many hierarchical Bayesian models, is not tractable to compute exactly and we must appeal to an approximation. (Developing algorithms for approximating posterior distributions of complex hierarchical models is an active research problem in Bayesian statistics and machine learning.) One of the most widely used approximation techniques for such models is Monte Carlo Markov chain (MCMC) sampling, where one samples from a Markov chain whose limiting distribution is the posterior of interest (Neal 1993; Robert and Casella 2004). Gibbs sampling in particular, where the Markov chain is defined by the conditional distribution of each latent variable, has found widespread use in Bayesian nonparametric models and topic models (Neal 1993; Teh 2006; Griffiths and Steyvers 2004; Finkel et al. 2007).

MCMC is a powerful methodology, but it has drawbacks. Convergence of the sampler to its stationary distribution is difficult to diagnose, and sampling algorithms can be slow to converge in high dimensional models (Robert and Casella 2004). An alternative to MCMC is variational inference. Variational methods, which are based on related techniques from statistical physics, use optimization to find a distribution over the latent variables that is close to the posterior of interest (Jordan et al. 1999; Wainwright and Jordan 2008). Variational methods provide effective approximations in topic models and nonparametric Bayesian models (Blei et al. 2003; Blei and Jordan 2005; Teh et al. 2006; Liang et al. 2007; Kurihara et al. 2007).

Variational methods enjoy a clear convergence criterion and tend to be faster than MCMC in high-dimensional problems.[7] Variational methods provide particular advantages over sampling when latent variable pairs are not conjugate. Gibbs sampling requires conjugacy, and other forms of sampling that can handle non-conjugacy, such as Metropolis-Hastings, are much slower than variational methods. Non-conjugate pairs appear in the dynamic topic model (Blei and Lafferty 2006; Wang et al. 2008), correlated topic model (Blei et al. 2007), and in the STM considered here. Specifically, in the STM the topic assignment is drawn from a renormalized product of two Dirichlet-distributed vectors (Equation 1). The distribution for each word’s topic does not form a conjugate pair with the document or transition topic distributions. In this section, we develop an approximate posterior inference algorithm for the STM that is based on variational methods.

[7] Understanding the general trade-offs between variational methods and Gibbs sampling is an open research question.

Our goal is to compute the posterior of topics $\boldsymbol{\beta}$, topic transitions $\boldsymbol{\pi}$, per-document weights $\boldsymbol{\theta}$, per-word topic assignments $\boldsymbol{z}$, top-level weights $\boldsymbol{\tau}$ given a collection of documents and the model described in Section 2. The difficulty around this posterior is that the hidden variables are connected through a complex dependency pattern. With a variational method, we begin by positing a family of distributions of the same variables with a simpler dependency pattern. This distribution is called the variational distribution. Here we use the fully-factorized variational distribution,

$$q(\boldsymbol{\tau}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\pi}, \beta | \boldsymbol{\tau}^*, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\nu}) = q(\boldsymbol{\tau}|\boldsymbol{\tau}^*) \prod_k q(\boldsymbol{\pi}_k|\boldsymbol{\nu}_k) \prod_d \left[q(\boldsymbol{\theta}_d|\gamma_d) \prod_n q(z_{d,n}|\boldsymbol{\phi}_{d,n}) \right].$$

Note that the latent variables are independent and each is governed by its own parameter. The idea behind variational methods is to adjust these parameters to find the member of this family that is close to the true distribution.

Following Liang (2007), $q(\tau|\tau^*)$ is not a full distribution but is a degenerate point estimate truncated so that all weights with index greater than K are zero in the variational distribution. The variational parameters γ_d and ν_z index Dirichlet distributions, and ϕ_n is a topic multinomial for the n^{th} word.

With this variational family in hand, we optimize the *evidence lower bound* (ELBO), a lower bound on the marginal probability of the observed data,

$$\begin{aligned} &\mathcal{L}(\gamma, \nu, \phi; \tau, \theta, \pi, \beta) = \\ &\quad \mathbb{E}_q \left[\log p(\boldsymbol{\tau}|\alpha)\right] + \mathbb{E}_q \left[\log p(\boldsymbol{\theta}|\alpha_D, \boldsymbol{\tau})\right] + \mathbb{E}_q \left[\log p(\boldsymbol{\pi}|\alpha_P, \boldsymbol{\tau})\right] + \mathbb{E}_q \left[\log p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\pi})\right] \\ &\qquad + \mathbb{E}_q \left[\log p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta})\right] + \mathbb{E}_q \left[\log p(\boldsymbol{\beta}|\sigma)\right] - \mathbb{E}_q \left[\log q(\boldsymbol{\theta}) + \log q(\boldsymbol{\pi}) + \log q(\boldsymbol{z})\right]. \end{aligned} \tag{2}$$

In simpler models, it is possible to explicitly compute the values of the latent variables that maximize the likelihood of the observed data. In contrast, variational inference amounts to fitting the variational parameters to tighten this lower bound. This is equivalent to minimizing the KL divergence between the variational distribution and the posterior. Once fit, the variational distribution is used as an approximation to the posterior.

Optimization of Equation 2 proceeds by coordinate ascent, optimizing each variational parameter while holding the others fixed. Each pass through the variational parameters increases the ELBO, and we iterate this process until reaching a local optimum. When possible, we find the per-parameter maximum value in closed form. When such updates are not possible, we employ gradient-based optimization (Galassi et al. 2003).

One can divide the ELBO into document terms and global terms. The document terms reflect the variational parameters of a single document and the global terms reflect variational parameters which are shared across all documents. This can be seen in the plate notion in Figure 4; the variational parameters on the right hand side are specific to individual documents. We expand Equation 2 and divide it into a document component (Equation B.5) and a global component (Equation D.1), which contains a sum of all the document contributions, in the appendix.

In coordinate ascent, the global parameters are fixed as we optimize the document level parameters. Thus, we can optimize a single document's contribution to the ELBO ignoring all other documents. This allows us to parallelize our implementation at the document level; each parallel document-level optimization is followed by an optimization step for the global variational parameters. We iterate these steps until we find a local optimum. In practice, several random starting points are used and we select the variational parameters provide the highest ELBO.

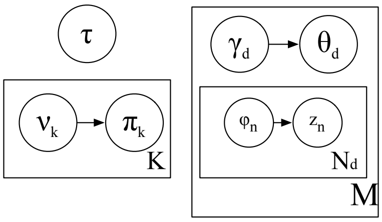


Figure 4
The truncated variational distribution removes constraints that are imposed because of the interactions of the full model and also truncates the possible number of topics (c.f. the full model in Figure 3). This family of distributions is used to approximate the log likelihood of the data and uncover the model's true parameters.

In the next sections, we outline the variational updates for the word-specific terms, document-specific terms, and corpus-wide terms. This exposition preserves the parallelization in our implementation and highlights the separate influences of topic modeling and syntactic models.

2.2.1 Document-specific Terms. We begin with $\phi_{d,n}$, the variational parameter that corresponds to the nth observed word's assignment to a topic. We can explicitly solve for the value of $\boldsymbol{\phi}_n$ which maximizes document d's contribution to the ELBO:

$$\begin{aligned}\phi_{n,i} \propto \exp \Bigg\{ & \Psi\left(\gamma_i\right) - \Psi\left(\textstyle\sum_{j=1}^K \gamma_j\right) + \sum_{j=1}^K \phi_{p(n),j}\left(\Psi\left(\nu_{j,i}\right) - \Psi\left(\textstyle\sum_{k=1}^K \nu_{j,k}\right)\right) \\ & - \sum_{c\in c(n)} \omega_c^{-1} \sum_j^K \frac{\gamma_j \nu_{i,j}}{\sum_k \gamma_k \sum_k \nu_{i,k}} \\ & + \sum_{c\in c(n)} \sum_{j=1}^K \phi_{c,j}\left(\Psi\left(\nu_{i,j}\right) - \Psi\left(\textstyle\sum_{k=1}^K \nu_{i,k}\right)\right) + \log \beta_{i,w_{d,n}} \Bigg\}. \end{aligned} \tag{3}$$

(Note that we have suppressed the document index d on ϕ and γ.)

This update reveals the influences on our estimate of the posterior of a single word's topic assignment. In the first line, the first two terms with the Dirichlet parameter $\boldsymbol{\gamma}$ show the influence of the document's distribution over topics; the term with multinomial parameter $\boldsymbol{\phi}_{p(n)}$ and Dirichlet parameter $\boldsymbol{\nu}$ reflects the interaction between the topic of the parent and transition probabilities. In the second line, the interaction between the document and transitions forces the document and syntax to be consistent (this is mediated by an additional variational parameter ω_c defined in Appendix B). In the final line, the influence of the children's' topic on the current word's topic is expressed in the first term, and the probability of a word given a topic in the second.

The other document-specific term is the per-document variational Dirichlet over topic proportions γ_d. Intuitively, topic proportions should reflect the expected number of words assigned to each topic in a document (the first two terms of equation 4), with the constraint that γ must be consistent with the syntactic transitions in the document, which is reflected by the ν term (the final term of Equation 4). This interaction prevents us from performing the update directly, so we use the gradient (derived in Appendix C)

$$\frac{\partial \mathcal{L}}{\partial \gamma_i} = \Psi'\left(\gamma_i\right)\left(\alpha_{D,i}\tau_i^* + \sum_{n=1}^{N}\phi_{n,i} - \gamma_i\right) - \Psi'\left(\textstyle\sum_{j=1}^{N}\gamma_j\right)\sum_{j=1}^{K}\left[\alpha_{D,j}\tau_j^* + \sum_{n=1}^{N}\phi_{n,j} - \gamma_j\right]$$
$$- \sum_{n=1}^{N}\omega_n^{-1}\sum_{j=1}^{K}\left[\phi_{p(n),j}\frac{\nu_{j,i}\sum_{k\neq j}^{N}\gamma_k - \sum_{k\neq j}^{N}\nu_{j,k}\gamma_k}{\left(\sum_{k=1}^{N}\gamma_k\right)^2\sum_{k=1}^{N}\nu_{j,k}}\right]. \qquad (4)$$

to compute the update for each document's per-document variational Dirichlet over topic proportions γ_d. This numerical optimization happens for each document in each iteration of the outer ELBO optimization loop and is done using Fletcher-Reeves conjugate gradient optimization (Galassi et al. 2003).

Now we turn to updates which require input from all documents and cannot be parallelized. Each document optimization, however, produces expected counts which are summed together; this is similar to the how the the E-step of EM algorithms can be parallelized and summed as input to the M-step (Wolfe et al. 2008).

2.2.2 Global Variational Terms. In this section, we consider optimizing the variational parameters for the transitions between topics and the top-level topic weights. Note that these variational parameters, in contrast with the previous section, are more concerned with the overall syntax, which is shared across all documents. Instead of optimizing a single ELBO term for each document, we now seek to maximize the entirety of Equation 2, expanded in Equation D.1 in the appendix.

The nonparametric models in Section 1.3 use a random variable τ drawn from a stick-breaking distribution to control how many components the model uses. The prior for τ attempts use as few topics as possible; the ELBO balances this desire against using more topics to better explain the data. We use numerical methods to optimize τ with respect to the gradient of the global ELBO, which is given in Equation D.2 in the appendix.

Finally, we optimize the variational distribution ν_i. If there were no interaction between θ and π, the update for $\nu_{i,j}$ would be proportional to the expected number of transitions from parents of topic i to children of topic j (this will set the first two terms of Equation 5 to zero). However, the objective function also encourages ν to be consistent with γ (the final term of Equation 5); thus, if γ excludes topics from being observed in a document, the optimization will not allow transitions to those topics. Again, this optimization is done using numerical optimization using the gradient of the ELBO,

$$\frac{\partial L}{\partial \nu_{i,j}} = \Psi'\left(\nu_{i,j}\right)\left(\alpha_{T,j}\tau_j^* + \sum_{n=1}^{N}\sum_{c\in c(n)}\phi_{n,i}\phi_{c,j} - \nu_{i,j}\right)$$
$$-\Psi'\left(\textstyle\sum_{k=1}^{K}\nu_{i,k}\right)\sum_{k=1}^{K}\left[\alpha_{T,k}\tau_k^* + \sum_{n=1}^{N}\sum_{c\in c(n)}\phi_{n,i}\phi_{c,k} - \nu_{i,k}\right]$$
$$-\sum_{n}^{N}\phi_{n,i}\sum_{c\in c(n)}\left[\omega_c^{-1}\frac{\gamma_j\sum_{k\neq j}^{N}\nu_{i,k} - \sum_{k\neq j}^{N}\nu_{i,k}\gamma_k}{\left(\sum_{k=1}^{N}\nu_{j,k}\right)^2\sum_{k=1}^{N}\gamma_k}\right]. \qquad (5)$$

Fixed Syntax		
S	→	VP
VP	→	NP V (PP) (NP)
NP	→	(Det) (Adj) N (PP)
PP	→	P NP
P	→	(“about”, “on”, “over”, “with”)
Det	→	(“a”, “that”, “the”, “this”)

Document-specific Vocabulary			
V	→	(“falls”, “runs”, “sits”)	**or**
		(“bucks”, “climbs”, “falls”, “surges”)	...
N	→	(“COW”, “PONY”, “SHEEP”)	**or**
		(“MUTUAL_FUND”, “SHARE”, “STOCK”)	...
Adj	→	(“American”, “German”, “Russian”)	**or**
		(“blue”, “purple”, “red”, “white”)	...

Table 1
The procedure for generating synthetic data. Syntax is shared across all documents, but each document chooses one of the thematic terminal distribution for verbs, nouns, and adjectives. This simulates how all documents share syntax and subsets of documents share topical themes. All expansion rules are chosen uniformly at random.

3. Experiments

We demonstrate how the STM works on data sets of increasing complexity. First, we show that the STM captures properties of a simple synthetic dataset that elude both topic and syntactic models individually. Next, we use a larger real-word dataset of hand-parsed sentences to show that both thematic and syntactic information is captured by the STM.

3.1 Topics Learned from Synthetic Data

We demonstrate the STM on synthetic data that resemble natural language. The data were generated using the grammar specified in Table 1. Each of the parts of speech except for prepositions and determiners was divided into themes, and a document contains a single theme for each part of speech. For example, a document can only contain nouns from a single “economic,” “academic,” or “livestock” theme, verbs from a possibly different theme, etc. Documents had between twenty and fifty sentences. An example of two documents is shown in Figure 5.

Using a truncation level of 16, we fit three different nonparametric Bayesian language models to the synthetic data (Figure 6).[8] Because the infinite tree model is aware of the tree structure but not documents, it is able to separate all parts of speech successfully except for adjectives and determiners (Figure 6c). However, it ignores the thematic distinctions that actually divided the terms between documents. The HDP is aware of document groupings and treats the words exchangeably within them and is thus able to recover the thematic topics, but it misses the connections between the parts of speech, and has conflated multiple parts of speech (Figure 6b).

[8] In Figure 6 and Figure 7, we mark topics which represent a single part of speech and are essentially the lone representative of that part of speech in the model. This is a subjective determination of the authors, does not reflect any specialization or special treatment of topics by the model, and is done merely for didactic purposes.

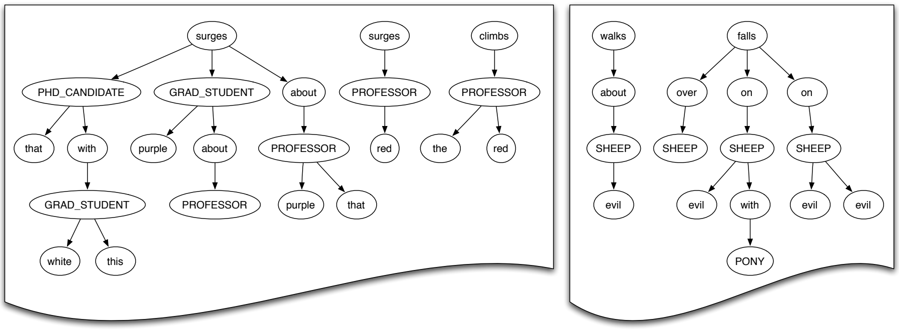


Figure 5
Two synthetic documents with multiple sentences. Nouns are shown in upper case. Each document chooses a theme for each part of speech independently; for example, the document on the left uses motion verbs, academic nouns, and color adjectives. Various models are applied to these data in Figure 6.

The STM is able to capture the the topical themes and recover parts of speech (with the exception of prepositions placed in the same topic as nouns with a self loop). Moreover, it was able to identify the same interconnections between latent classes that were apparent from the infinite tree. Nouns are dominated by verbs and prepositions, and verbs are the root (head) of sentences. Figure 6d shows the two divisions as separate axes; going form left to right, the thematic divisions that the HDP was able to uncover are clear. Going from top to bottom, the syntactic distinctions made by the infinite tree are revealed.

3.2 Qualitative Description of Topics learned by the STM from Hand-annotated Data

The same general properties, but with greater variation, are exhibited in real data. We converted the Penn Treebank (**?**), a corpus of manually curated parse trees, into a dependency parse (Johansson and Nugues 2007). The vocabulary was pruned to terms that appeared in at least ten documents; and word that did not appear in the vocabulary was replaced with an OOV character based on its part of speech (thus, if “doorhinge” was an out of vocabulary term then it was replaced by “OOV-Noun” when it appeared).

Figure 7 shows a subset of topics learned by the STM with truncation level 32. Many of the resulting topics illustrate both syntactic and thematic consistency. A few non-specific function topics emerged (pronoun, possessive pronoun, general verbs, etc.). Many of the noun categories were more specialized. For instance, Figure 7 shows clusters of nouns relating to media, individuals associated with companies (“mr,” “president,” “chairman”), and abstract nouns related to stock prices (“shares,” “quarter,” “earnings,” “interest”), all of which feed into a topic that modifies nouns (“his,” “their,” “other,” “last”).

Griffiths et al (**?**) observed that nouns, more than other parts of speech, tend to specialize into distinct topics, and this is also evident here. In Figure 7, the unspecialized syntactic categories (shaded and with rounded edges) serve to connect many different specialized thematic categories, which are predominantly nouns (although the adjectives also showed bifurcation). For example, verbs are mostly found in a single topic, but then have a large number of outgoing transitions to many noun topics. Because of this relationship, verbs look like a syntactic “source” in Figure 7. Many of these noun topics then point to thematically unified topics such as “personal pronouns,” which look like syntactic “sinks.”

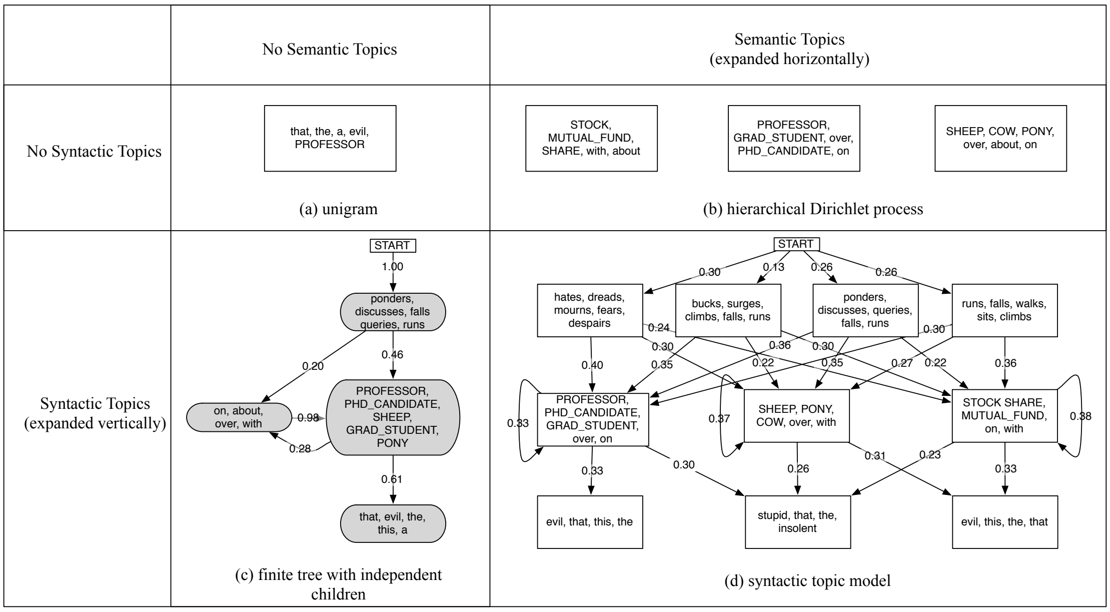

No Semantic Topics
Semantic Topics (expanded horizontally)
No Syntactic Topics
that, the, a, evil, PROFESSOR
(a) unigram
STOCK, MUTUAL_FUND, SHARE, with, about
PROFESSOR, GRAD_STUDENT, over, PHD_CANDIDATE, on
SHEEP, COW, PONY, over, about, on
(b) hierarchical Dirichlet process
Syntactic Topics (expanded vertically)
START
1.00
ponders, discusses, falls queries, runs
0.20
0.46
on, about, over, with
0.98
0.28
PROFESSOR, PHD_CANDIDATE, SHEEP, GRAD_STUDENT, PONY
0.61
that, evil, the, this, a
(c) finite tree with independent children
START
0.30
0.13
0.26
0.26
hates, dreads, mourns, fears, despairs
bucks, surges, climbs, falls, runs
ponders, discusses, queries, falls, runs
runs, falls, walks, sits, climbs
0.24
0.30
0.30
0.36
0.30
0.40
0.35
0.22
0.35
0.27
0.22
0.36
PROFESSOR, PHD_CANDIDATE, GRAD_STUDENT, over, on
SHEEP, PONY, COW, over, with
STOCK SHARE, MUTUAL_FUND, on, with
0.33
0.37
0.38
0.33
0.30
0.26
0.31
0.23
0.33
evil, that, this, the
stupid, that, the, insolent
evil, this, the, that
(d) syntactic topic model

Figure 6
We contrast the different views of data that are available by using syntactic and semantic topics based on our synthetic data. Three models were fit to the synthetic data described in Section 3. Each box illustrates the top five words of a topic; boxes that represent homogeneous parts of speech have rounded edges and are shaded; and nouns are in upper case. Edges between topics are labeled with estimates of their transition weight π. If we have neither syntactic nor semantic topics, we have a unigram (a) model that views words as coming from a single distribution over words. Adding syntactic topics allows us to recover the parts of speech (c), but this lumps all topics together. Although the HDP (b) can discover themes of recurring words, it cannot determine the interactions between topics or separate out ubiquitous words that occur in all documents. The STM (d) is able to recover both the syntax and the themes.

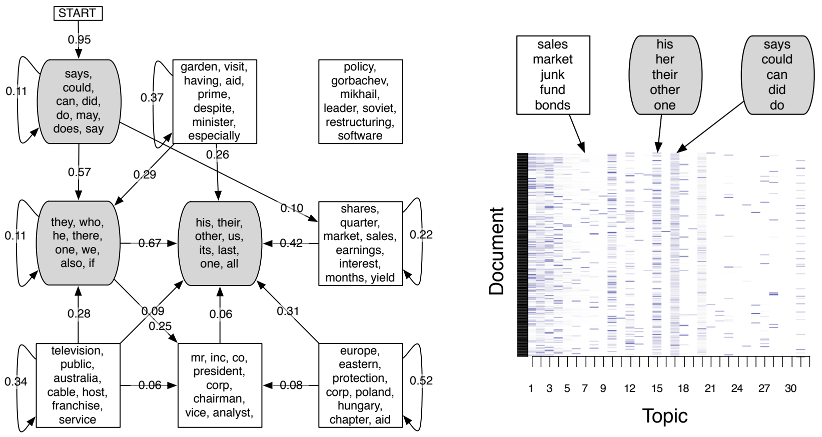
It is important to note that Figure 7 only presents half of the process of choosing a topic for a word. While the transition distribution of verb topics allows many different noun topics as possible dependents, because the topic is chosen from a product of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$, $\boldsymbol{\theta}$ can filter out the noun topics that are inconsistent with a document's theme.

This division between functional and topical uses for the latent classes can also been seen in the values for the per-document multinomial over topics. A number of topics in Figure 7(b), such as 17, 15, 10, and 3, appear to some degree in nearly every document, while other topics are used more sparingly to denote specialized content. With $\alpha = 0.1$, this plot also shows that the nonparametric Bayesian framework choosing to use a number of topics below the truncation level.

3.3 Quantitative Results on Synthetic and Hand-annotated Data

To study the performance of the STM on new data, we estimated the held out probability of previously unseen documents with an STM trained on a portion of the dataset. For each position in the parse trees, we estimate the probability of the observed word using the variational approximation of the posterior.

This is done by keeping the topics fixed, and using the inference procedure to fit the document-specific variational parameters of the held-out document to maximize the document's


START
0.95
says, could, can, did, do, may, does, say
0.11
0.37
garden, visit, having, aid, prime, despite, minister, especially
policy, gorbachev, mikhail, leader, soviet, restructuring, software
0.57
0.29
0.26
0.10
they, who, he, there, one, we, also, if
0.11
0.67
his, their, other, us, its, last, one, all
0.42
shares, quarter, market, sales, earnings, interest, months, yield
0.22
0.28
0.09
0.25
0.06
0.31
television, public, australia, cable, host, franchise, service
0.34
0.06
mr, inc, co, president, corp, chairman, vice, analyst,
0.08
europe, eastern, protection, corp, poland, hungary, chapter, aid
0.52
sales market junk fund bonds
his her their other one
says could can did do
Document
1 3 5 7 9 12 15 18 21 24 27 30
Topic

(a) learned topics

(b) topic prevalence

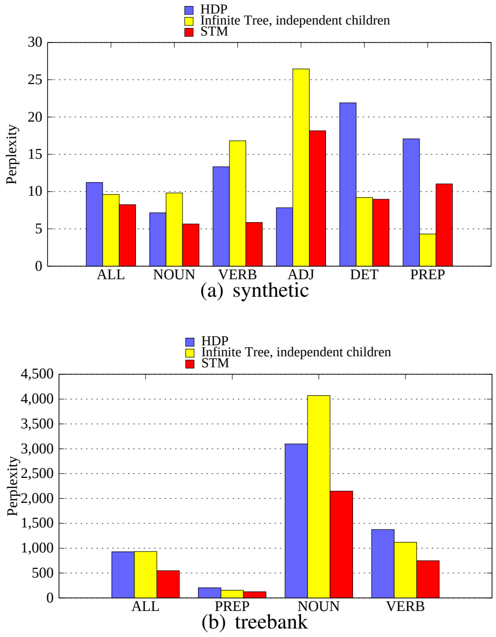
Figure 7
Topics discovered from fitting the syntactic topic model on the Treebank corpus. As in Figure 6, parts of speech that aren't subdivided across themes are indicated and edges between topics are labeled with estimates of the the transition probability π. Head words (verbs) are shared across many documents and allow many different types of nouns as possible dependents. These dependents, in turn, share topics that look like pronouns as common dependents. The specialization of topics is also visible in plots of the estimates for the per-document topic distribution θ for the first 300 documents of the Treebank (right), where three topics columns have been identified. Many topics are used to some extent in every document, showing that they are performing a functional role, while others are used more sparingly for topical content.

$\mathcal{L}$ contribution. We compute the perplexity as the exponent of the inverse of the per-word average log probability,

$$perplexity(D) = \exp\left(\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M} N_d}\right).$$

The lower the perplexity, the better the model has captured the patterns in the data. We also computed perplexity for individual parts of speech (thus, summing only over words in a particular part of speech) to study the differences in predictive power between content words, such as nouns and verbs, and function words, such as prepositions and determiners. (We use parts of speech only to differentiate how the model performs for certain types of words; we emphasize that the model does not have access to the parts of speech.) This illustrates how different algorithms better capture aspects of context. We expect function words to be dominated by local context and content words to be determined more by the themes of the document.

This trend is seen not only in the synthetic data (Figure 8(a)), where syntactic models better predict functional categories like prepositions, and document-only models fail to account for patterns of verbs and determiners, but also in real data. Figure 8(b) shows that HDP and STM both perform better than syntactic models in capturing the patterns behind nouns, while both STM and the infinite tree have lower perplexity for verbs. Like syntactic models, our model was better able to predict the appearance of prepositions but also remained competitive with HDP on content words. On the whole, STM had lower perplexity than HDP and the infinite tree.



(a) synthetic

(b) treebank

Figure 8
After fitting three models on synthetic data, the syntactic topic model has better (lower) perplexity on all word classes except for adjectives. HDP is better able to capture document-level patterns of adjectives. The infinite tree captures prepositions best, which have no cross-document variation. On real data 8(b), the syntactic topic model was able to combine the strengths of the infinite tree on functional categories like prepositions with the strengths of the HDP on content categories like nouns to attain lower overall perplexity.

4. Conclusion

In this work, we explored the common threads that link syntactic and topic models and created a model that is simultaneously aware of both thematic and syntactic influences in a document. These models are aware of more structure than either model individually.

More generally, this work serves as an example of how a mixture model can support two different, simultaneous explanations for how the latent class is chosen. Although this model used discrete observations, the variational inference setup is flexible enough to support other distributions over the output.

While this work's primary goal was to demonstrate how these two views of context could be simultaneously learned, there are a number of extensions that could lead to more accurate parsers. First, this model could be further extended by integrating a richer syntactic model that does not just model the words that appear in a given structure but one that also models the parse structure itself. This would allow the model to use large, diverse corpora without relying upon an external parser to provide the tree structure.

Removing the independence restriction between children also would allow for this model to closer approximate the state of the art syntactic models and to be better distinguish the children of parent nodes (this is especially the problem for head verbs, which often have many children). Finally, this model could also make use of labeled dependency relations and lexicalization.

With the ability to adjust to specific document or corpus-based contexts, a parser built using this framework could adapt to handle different domains while still sharing information between them. The classification and clustering implicitly provided by the topic components would allow the parser to specialize its parsing model when necessary, allowing both sentence-level and document-level information to shape the model's understanding of a document.

References

[Abney2004]Steven Abney. 2004. Understanding the Yarowsky Algorithm. Computational Linguistics, 30(3):365–395.

[Antoniak1974]Charles E. Antoniak. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics, 2(6):1152–1174.

[Blei and Jordan2005]David M. Blei and Michael I. Jordan. 2005. Variational inference for Dirichlet process mixtures. Journal of Bayesian Analysis, 1(1):121–144.

[Blei and Lafferty2006]David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In Proceedings of the International Conference of Machine Learning.

[Blei and Lafferty2009]David M. Blei and John Lafferty, 2009. Text Mining: Theory and Applications, chapter Topic Models. Taylor and Francis, London.

[Blei and McAuliffe2007]David M. Blei and Jon D. McAuliffe. 2007. Supervised topic models. In NIPS.

[Blei et al.2003]David M. Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022.

[Blei et al.2007]David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2007. The nested Chinese restaurant process and hierarchical topic models.

[Blunsom et al.2008]Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008. Bayesian synchronous grammar induction. In Proc of NIPS.

[Boyd-Graber et al.2007]Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In Proceedings of Emperical Methods in Natural Language Processing.

[Brody and Lapata2009]Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In Proceedings of the European Chapter of the Association for Computational Linguistics, Athens, Greece.

[Charniak1997]Eugene Charniak. 1997. Statistical techniques for natural language parsing. AI Magazine, 18:33–44.

[Charniak2000]Eugene Charniak. 2000. A maximum-entropy-inspired parser. In Conference of the North American Chapter of the Association for Computational Linguistics, pages 132–139.

[Cohen et al.2008]Shay B. Cohen, Kevin Gimpel, and Noah A. Smith. 2008. Logistic normal priors for unsupervised probabilistic grammar induction. In NIPS 21.

[Collins2003]Michael Collins. 2003. Head-driven statistical models for natural language parsing. Computational Linguistics, 29(4):589–637.

[Dapretto and Bookheimer1999]Mirella Dapretto and Susan Y. Bookheimer. 1999. Form and content: Dissociating syntax and semantics in sentence comprehension. Neuron, 24(2):427–432.

[Deerwester et al.1990]Scott Deerwester, Susan Dumais, Thomas Landauer, George Furnas, and Richard Harshman. 1990. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391–407.

[Erosheva et al.2007]Elena A. Erosheva, Stephen E. Fienberg, and Cyrille Joutard. 2007. Describing disability through individual-level mixture models for multivariate binary data. Annals of Applied Statistics, 1:502.

[Escobar and West1995]Michael D. Escobar and Mike West. 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association, 90:577–588.

[Ferguson1973]Thomas S. Ferguson. 1973. A Bayesian analysis of some nonparametric problems. The Annals of Statistics, 1(2).

[Finkel et al.2007]Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2007. The infinite tree. In Proceedings of the Association for Computational Linguistics.

[Galassi et al.2003]Mark Galassi, Jim Davies, James Theiler, Brian Gough, Gerard Jungman, Michael Booth, and Fabrice Rossi. 2003. Gnu Scientific Library: Reference Manual. Network Theory Ltd.

[Goldwater2007]Sharon Goldwater. 2007. Nonparametric Bayesian Models of Lexical Acquisition. Ph.D. thesis, Brown University.

[Griffiths and Steyvers2004]Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. Proceedings of the National Academy of Sciences, 101(Suppl 1):5228–5235.

[Griffiths et al.2007]Thomas L. Griffiths, Mark Steyvers, and Joshua Tenenbaum. 2007. Topics in semantic representation. Psychological Review, 114(2):211–244.

[Gruber et al.2007]Amit Gruber, Michael Rosen-Zvi, and Yair Weiss. 2007. Hidden topic Markov models. In Artificial Intelligence and Statistics.

[Hall et al.2008]David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In Proceedings of Emperical Methods in Natural Language Processing.

[Hinton1999]Geoffrey Hinton. 1999. Products of experts. In Proceedings of the Ninth International Conference on Artificial Neural Networks, pages 1–6, Edinburgh, Scotland. IEEE.

[Hofmann1999]Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In Proceedings of Uncertainty in Artificial Intelligence.

[Hu and Saul2009]Diane Hu and Lawrence K. Saul. 2009. A probabilistic model of unsupervised learning for musical-key profiles. In International Society for Music Information Retrieval Conference.

[Johansson and Nugues2007]Richard Johansson and Pierre Nugues. 2007. Extended constituent-to-dependency conversion for English. In Proceedings of the Nordic Conference on Computational Linguistics.

[Johnson and Goldwater2009]Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In Conference of the North American Chapter of the Association for Computational Linguistics, pages 317–325, Boulder, Colorado, June. Proceedings of the Association for Computational Linguistics.

[Johnson et al.2006]Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In Proceedings of Advances in Neural Information Processing Systems.

[Jordan et al.1999]Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. 1999. An introduction to variational methods for graphical models. Machine Learning, 37(2):183–233.

[Klein and Manning2002]Dan Klein and Christopher D. Manning. 2002. Fast exact inference with a factored model for natural language parsing. In Advances in Neural Information Processing Systems 15 (NIPS 2002).

[Klein and Manning2003]Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In Proceedings of the Association for Computational Linguistics, pages 423–430. Association for Computational Linguistics.

[Kurihara et al.2007]Kenichi Kurihara, Max Welling, and Yee Whye Teh. 2007. Collapsed variational Dirichlet process mixture models. In International Joint Conference on Artificial Intelligence.

[Liang and Klein2007]Percy Liang and Dan Klein. 2007. Structured Bayesian nonparametric models with variational inference (tutorial). In Proceedings of the Association for Computational Linguistics.

[Liang et al.2007]Percy Liang, Slav Petrov, Michael Jordan, and Dan Klein. 2007. The infinite PCFG using hierarchical Dirichlet processes. In Proceedings of Emperical Methods in Natural Language Processing, pages 688–697.

[Lin1998]Dekang Lin. 1998. An information-theoretic definition of similarity. In Proceedings of the International Conference of Machine Learning, pages 296–304.

[Manning and Schütze1999]Christopher D. Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA.

[Mimno and McCallum2007]David Mimno and Andrew McCallum. 2007. Mining a digital library for influential authors. In JCDL '07, New York, NY, USA. ACM.

[Neal1993]Radford M. Neal. 1993. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, University of Toronto.

[Neal2000]Radford M. Neal. 2000. Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics, 9(2):249–265.

[Padó and Lapata2007]Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. Computational Linguistics, 33(2):161–199.

[Pitman2002]Jim Pitman. 2002. Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. Combinatorics, Probability and Computing, 11:501–514.

[Pritchard et al.2000]Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics, 155:945–959.

[Purver et al.2006]Matthew Purver, Konrad Körding, Thomas L. Griffiths, and Joshua Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In Proceedings of the Association for Computational Linguistics.

[Ravi and Knight]Sujith Ravi and Kevin Knight. Minimized models for unsupervised part-of-speech tagging. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 504–512, Suntec, Singapore.

[Rayner et al.1983]Keith Rayner, Marcia Carlson, and Lyn Frazier. 1983. The interaction of syntax and semantics during sentence processing — Eye-movements in the analysis of semantically biased sentences. Journal of Verbal Learning and Verbal Behavior, 22(3):358–374.

[Robert and Casella2004]Christian Robert and George Casella. 2004. Monte Carlo Statistical Methods. Springer Texts in Statistics. Springer-Verlag, New York, NY.

[Rosen-Zvi et al.2004]Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In Proceedings of Uncertainty in Artificial Intelligence.

[Sethuraman1994]Jayaram Sethuraman. 1994. A constructive definition of Dirichlet priors. Statistica Sinica, 4:639–650.

[Teh et al.2006]Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476):1566–1581.

[Teh2006]Yee Whye Teh. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In Proceedings of the Association for Computational Linguistics.

[Titov and McDonald2008]Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In Proceedings of the Association for Computational Linguistics.

[Toutanova and Johnson2008]Kristina Toutanova and Mark Johnson. 2008. A Bayesian LDA-based model for semi-supervised part-of-speech tagging. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, Advances in Neural Information Processing Systems 20, pages 1521–1528. MIT Press, Cambridge, MA.

[Wainwright and Jordan2008]Martin J. Wainwright and Michael I. Jordan. 2008. Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning, 1(1–2):1–305.

[Wallach2006]Hanna M. Wallach. 2006. Topic modeling: Beyond bag-of-words. In Proceedings of the International Conference of Machine Learning.

[Wang et al.2008]Chong Wang, David M. Blei, and David Heckerman. 2008. Continuous time dynamic topic models. In Proceedings of Uncertainty in Artificial Intelligence.

[Winn2003]John Winn. 2003. Variational Message Passing and its Applications. PhD thesis, University of Cambridge, Cambridge, UK.

[Wolfe et al.2008]Jason Wolfe, Aria Haghighi, and Dan Klein. 2008. Fully distributed EM for very large datasets. In Proceedings of the International Conference of Machine Learning, pages 1184–1191.

[Yarowsky1995]David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the Association for Computational Linguistics, pages 189–196, Morristown, NJ, USA. Association for Computational Linguistics.

[Zhu et al.2006]Xiaojin Zhu, David M. Blei, and John Lafferty. 2006. TagLDA: Bringing document structure knowledge into topic models. Technical Report TR-1553, University of Wisconsin, Madison.

Appendix A: Dirichlet in the Exponential Family

These appendices explain the derivation of the updates for variational inference for the Syntactic Topic Model (STM). After some mathematical preliminaries in Appendix A, we expand the expectations in the variational likelihood bound in Appendix B and then, having expanded the objective function, derive the updates which optimize the bound for individual documents (Appendix C) and for the entire corpus (Appendix D).

First, we delve deeper into the Dirichlet distribution, which will allow us to expand the expectations over distributions in our objective function. We will show that the Dirichlet distribution can be expressed as a member of the exponential family of probability distributions and derive the expectation of the log of a Dirichlet distributed random variable under its variational distribution, which will appear repeatedly in the ELBO objective function.

A probability distribution is a member of the exponential family of distributions if it can be expressed using the exponential family form

$$p(x|\eta) = h(x)\text{exp}\left\{g(\eta)^T u(x) - a(\eta)\right\}, \tag{A.1}$$

where $g(\eta)$ is the natural parameter vector, $u(x)$ is the natural statistic vector, $h(x)$ is the measure of the space, and $a(\eta)$ is the normalization. We can express the Dirichlet distribution (first discussed in Section 1.3) as an exponential family distribution, rewriting the conventional density function,

$$\mathrm{Dir}(\boldsymbol{\theta}\,|\,\alpha_1,\ldots,\alpha_K)=\underbrace{\frac{\Gamma\left(\sum_k\alpha_k\right)}{\prod_k\Gamma\left(\alpha_k\right)}}_{\text{normalization}}\prod_k\theta_k^{\alpha_k-1},$$

as an exponential family distribution

$$\mathrm{Dir}(\boldsymbol{\theta}|\alpha)=\!\exp\left\{\begin{bmatrix}\alpha_1-1\\ \vdots\\ \alpha_K-1\end{bmatrix}^T\begin{bmatrix}\log\theta_1\\ \vdots\\ \log\theta_K\end{bmatrix}+\log\Gamma\left(\sum_{i=1}^K\alpha_i\right)-\sum_{i=1}^K\Gamma\left(\alpha_i\right)\right\}. \tag{A.2}$$

One property of the exponential family of distributions that we state without proof (Winn 2003) is that the expectation of the natural statistic vector is the derivative of the log normalizer, with respect to the natural parameter. For a Dirichlet distribution,

$$\mathbb{E}_\theta\left[\log\theta_i\right]=\Psi\left(\alpha_i\right)-\Psi\left(\sum_{j=1}^K\alpha_j\right). \tag{A.3}$$

Thus, if we take the expectation of a Dirichlet distribution $p(\theta|\alpha)$ with respect to a variational Dirichlet distribution over θ parameterized by γ, we have

$$\begin{aligned}\mathbb{E}_q\left[\log p(\theta|\alpha)\right]=&\mathbb{E}_q\left[\log\left(\frac{\Gamma\left(\sum_{i=1}^K\alpha_i\right)}{\sum_{i=1}^K\Gamma\left(\alpha_i\right)}\prod_{i=1}^K\theta^{\alpha_i-1}\right)\right]\\ =&\mathbb{E}_q\left[\log\Gamma\left(\sum_{i=1}^K\alpha_i\right)-\sum_{i=1}^K\log\Gamma\left(\alpha_i\right)+\sum_{i=1}^K\left(\alpha_i-1\right)\log\theta_i\right].\end{aligned}$$

Only θ is not a constant with respect to q, and $\log\theta_i$ is the natural statistic of the Dirichlet distribution when it as written as an exponential family distribution. In the variational distribution, θ comes from a Dirichlet parameterized by γ, so using Equation A.3, we have

$$\mathbb{E}_q\left[\log p(\theta|\alpha)\right]=\log\Gamma\left(\sum_{i=1}^K\alpha_i\right)-\sum_{i=1}^K\log\Gamma\left(\alpha_i\right)+\sum_{i=1}^K\left(\alpha_i-1\right)\left(\Psi\left(\gamma_i\right)-\Psi\left(\sum_{j=1}^K\gamma_j\right)\right). \tag{A.4}$$

Appendix B: Expanding the Likelihood Bound for Document-Specific Terms

Recall that the objective function for the STM is

$$\begin{aligned}&\mathcal{L}(\gamma,\nu,\phi;\tau,\theta,\pi,\beta)=\\ &\quad\mathbb{E}_q\left[\log p(\boldsymbol{\tau}|\alpha)\right]+\mathbb{E}_q\left[\log p(\boldsymbol{\theta}|\alpha_D,\boldsymbol{\tau})\right]+\mathbb{E}_q\left[\log p(\boldsymbol{\pi}|\alpha_P,\boldsymbol{\tau})\right]+\mathbb{E}_q\left[\log p(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{\pi})\right]\\ &\qquad+\mathbb{E}_q\left[\log p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta})\right]+\mathbb{E}_q\left[\log p(\boldsymbol{\beta}|\sigma)\right]-\mathbb{E}_q\left[\log q(\boldsymbol{\theta})+\log q(\boldsymbol{\pi})+\log q(\boldsymbol{z})\right].\end{aligned} \tag{B.1}$$

In this section, we expand the terms in the $\mathcal{L}$ needed to perform document-specific expectations. This will provide the information needed to optimize document-specific variational parameters in the next section. We save the expansion of the remaining terms from $\mathcal{L}$ until Section D.

LDA-like terms

The terms of equation B.2 specific to a single document are

$$\begin{aligned}\mathcal{L}_d =&\mathbb{E}_q\left[\log p(\boldsymbol{\theta}_d|\alpha_D,\boldsymbol{\tau})\right]+\mathbb{E}_q\left[\log p(\boldsymbol{z}_d|\boldsymbol{\theta}_d,\boldsymbol{\pi})\right]\\ &+\mathbb{E}_q\left[\log p(\boldsymbol{w}|\boldsymbol{z}_d,\boldsymbol{\beta})\right]-\mathbb{E}_q\left[\log q(\boldsymbol{\theta}_d)+\log q(\boldsymbol{z}_d)\right].\end{aligned} \tag{B.2}$$

We now expand each of these using the formula given in Equation A.4. First, if we consider the expectation over the document's distribution over latent topics,

$$\begin{aligned}\mathbb{E}_q\left[\log p(\boldsymbol{\theta}_d|\alpha_D,\boldsymbol{\tau})\right]=&\log\Gamma\left(\textstyle\sum_{j=1}^{K}\alpha_{D,j}\tau^*\right)-\sum_{i=1}^{K}\log\Gamma\left(\alpha_{D,i}\tau^*\right)+\\ &\sum_{i=1}^{K}\left(\alpha_{D,i}\tau^*-1\right)\left(\Psi\left(\gamma_i\right)-\Psi\left(\textstyle\sum_{j=1}^{K}\gamma_j\right)\right),\end{aligned}$$

we can treat the truncated Dirichlet process as a Dirichlet distribution with a parameter that has been scaled by α_D. We postpone expanding the expectation over topic assignments $\boldsymbol{z}_d$ until the next section. For the expectation over the words, we note that the probability of the n^{th} word in document d taking topic k under the variational distribution is $\phi_{d,n,k}$ or (suppressing the document index) $\phi_{n,k}$ and given that assignment, the probability of the corresponding token $w_{d,n}$ being produced by topic k is $\beta_{k,w_{d,n}}$. Thus,

$$\mathbb{E}_q\left[\log p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta})\right]=\sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{n,i}\log\beta_{i,w_{d,n}}.$$

We are left with the entropy terms. First, the entropy for the per-document topic distribution is

$$\begin{aligned}\mathbb{E}_q\left[\log q(\boldsymbol{\theta})\right]=&\log\Gamma\left(\textstyle\sum_{j=1}^{K}\gamma_j\right)-\sum_{i=1}^{K}\log\Gamma\left(\gamma_i\right)+\\ &\sum_{i=1}^{K}(\gamma_i-1)\left(\Psi\left(\gamma_i\right)-\Psi\left(\textstyle\sum_{j=1}^{K}\gamma_j\right)\right),\end{aligned}$$

which follows by the same reasoning used in equation A.4. The entropy of a multinomial distribution is straightforward

$$\mathbb{E}_q\left[\log q(\boldsymbol{z})\right]=\sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{n,i}\log\phi_{n,i}.$$

The Interaction of Syntax and Semantics

We now move on to expanding $\mathbb{E}_q\left[\log p(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{\pi})\right]$ from Equation B.2. Rather than drawing the topic of a word directly from a multinomial, the topic is chosen from the renormalized point-wise product of two multinomial distributions. In order to handle the expectation of the log sum introduced by the renormalization, we introduce an additional variational parameter ω_n for each word via a Taylor approximation of the logarithm

$$\log(x) \leq \frac{x}{\omega} + \log\omega - 1 \qquad\qquad \forall\omega > 0. \tag{B.3}$$

Using this approximation, we find that $\mathbb{E}_q\left[\log p(\mathbf{z}|\boldsymbol{\theta},\boldsymbol{\pi})\right] =$

$$\mathbb{E}_q\left[\log\prod_{n=1}^{N}\frac{\theta_{z_n}\pi_{z_{p(n)},z_n}}{\sum_i^K\theta_i\pi_{z_{p(n)},i}}\right] = \mathbb{E}_q\left[\sum_{n=1}^{N}\log\theta_{z_n}\pi_{z_{p(n)},z_n} - \sum_{n=1}^{N}\log\sum_{i=1}^{K}\theta_i\pi_{z_{p(n)},i}\right]$$

$$\geq\sum_{n=1}^{N}\mathbb{E}_q\left[\log\theta_{z_n}\pi_{z_{p(n)},z_n}\right] - \sum_{n=1}^{N}\mathbb{E}_q\left[\omega_n^{-1}\sum_{i=1}^{K}\theta_i\pi_{z_{p(n)},i}\right] + \log\omega_n - 1$$

$$= \sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{n,i}\left(\Psi\left(\gamma_i\right) - \Psi\left(\textstyle\sum_{j=1}^{K}\gamma_j\right)\right) + \sum_{n=1}^{N}\sum_{i=1}^{K}\sum_{j=1}^{K}\phi_{n,i}\phi_{p(n),j}\left(\Psi\left(\nu_{j,i}\right) - \Psi\left(\textstyle\sum_{k=1}^{K}\nu_{j,k}\right)\right)$$

$$-\left(\sum_{n=1}^{N}\omega_n^{-1}\sum_{i=1}\sum_{j=1}\phi_{p(n),j}\frac{\gamma_i\nu_{j,i}}{\sum_{k=1}^{K}\gamma_k\sum_{k=1}^{K}\nu_{j,k}} + \log\omega_n - 1\right). \tag{B.4}$$

Combining this with the other expansions for a document gives us an individual document's contribution to the objective function

$$\mathcal{L}_d = \log\Gamma\left(\textstyle\sum_{j=1}^{K}\alpha_{D,j}\tau^*\right) - \sum_{i=1}^{K}\log\Gamma\left(\alpha_{D,i}\tau^*\right) + \sum_{i=1}^{K}\left(\alpha_{D,i}\tau^* - 1\right)\left(\Psi\left(\gamma_i\right) - \Psi\left(\textstyle\sum_{j=1}^{K}\gamma_j\right)\right)$$

$$+\sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{n,i}\left(\Psi\left(\gamma_i\right) - \Psi\left(\textstyle\sum_{j=1}^{K}\gamma_j\right)\right) + \sum_{n=1}^{N}\sum_{i=1}^{K}\sum_{j=1}^{K}\phi_{n,i}\phi_{p(n),j}\left(\Psi\left(\nu_{j,i}\right) - \Psi\left(\textstyle\sum_{k=1}^{K}\nu_{j,k}\right)\right)$$

$$-\left(\sum_{n=1}^{N}\omega_n^{-1}\sum_{i=1}\sum_{j=1}\phi_{p(n),j}\frac{\gamma_i\nu_{j,i}}{\sum_{k=1}^{K}\gamma_k\sum_{k=1}^{K}\nu_{j,k}} + \log\omega_n - 1\right)$$

$$+\sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{n,i}\log\beta_{i,w_{d,n}}$$

$$-\log\Gamma\left(\textstyle\sum_{j=1}^{K}\gamma_j\right) + \sum_{i=1}^{K}\log\Gamma\left(\gamma_i\right) - \sum_{i=1}^{K}(\gamma_i - 1)\left(\Psi\left(\gamma_i\right) - \Psi\left(\textstyle\sum_{j=1}^{K}\gamma_j\right)\right)$$

$$-\sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{n,i}\log\phi_{n,i}. \tag{B.5}$$

Apart from the terms derived in Equation B.4, the other terms here are very similar to the objective function for LDA. The expectation of the log of $p(\boldsymbol{\theta})$, $q(\boldsymbol{\theta})$, $p(\boldsymbol{z})$, $q(\boldsymbol{z})$, and $p(\boldsymbol{w})$ all appear in the LDA likelihood bound.

Appendix C: Document-specific Variational Updates

In this section, we derive the updates for all document-specific variational parameters other than $\boldsymbol{\phi}_n$, which is updated according to Equation 3.

Because we cannot assume that the point-wise product of of π_k and θ_d sums to one, we introduced a slack term ω_n in Equation B.4; its update is

$$\omega_n = \sum_{i=1} \sum_{j=1} \phi_{p(n),j} \frac{\gamma_i \nu_{j,i}}{\sum_{k=1}^{K} \gamma_k \sum_{k=1}^{K} \nu_{j,k}}.$$

Because we couple $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$, the interaction between these terms in the normalizer prevents us from solving the optimization for $\boldsymbol{\gamma}$ and $\boldsymbol{\nu}$ explicitly. Instead, for each $\boldsymbol{\gamma}_d$ we compute the partial derivative with respect to $\gamma_{d,i}$ for each component of the vector. We then maximize the likelihood bound for each $\boldsymbol{\gamma}_d$. In deriving the gradient, the following derivative is useful:

$$\begin{aligned} f(x) &= \sum_{i=1}^{N} \alpha_i \frac{x_i}{\sum_{i=j}^{N} x_j} \\ \Rightarrow \frac{\partial f}{\partial x_i} &= \frac{\alpha_i \sum_{j \neq i}^{N} x_j - \sum_{j \neq i}^{N} \alpha_j x_j}{\left(\sum_{i=1}^{N} x_i\right)^2}. \end{aligned} \tag{C.1}$$

This allows us to more easily compute the partial derivative of Equation B.5 with respect to γ_i to be

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \gamma_i} = \Psi'\left(\gamma_i\right) \left(\alpha_{D,i} \tau_i^* + \sum_{n=1}^{N} \phi_{n,i} - \gamma_i\right) - \Psi'\left(\sum_{j=1}^{N} \gamma_j\right) \sum_{j=1}^{K} \left[\alpha_{D,j} \tau_j^* + \sum_{n=1}^{N} \phi_{n,j} - \gamma_j\right] \\ - \sum_{n=1}^{N} \omega_n^{-1} \sum_{j=1}^{K} \left[\phi_{p(n),j} \frac{\nu_{j,i} \sum_{k \neq j}^{N} \gamma_k - \sum_{k \neq j}^{N} \nu_{j,k} \gamma_k}{\left(\sum_{k=1}^{N} \gamma_k\right)^2 \sum_{k=1}^{N} \nu_{j,k}}\right]. \end{aligned}$$

A function that computs $\mathcal{L}$ and the derivative with respect to γ is sufficient to use a numerical optimization technique to find the value of γ that maximizes $\mathcal{L}$. However, the parameterization of the Dirichlet distribution (recall that γ is a variational Dirichlet parameter) only allows $\gamma_i > 0$. Thus, the actual optimization is with respect to $\log \gamma$.

Appendix D: Global Updates

In this section, we expand the terms of Equation 2 that were not expanded in Equation B.5. First, we note that $\mathbb{E}_q \left[\log \text{GEM}(\boldsymbol{\tau}; \alpha)\right]$, because the variational distribution only puts weight on $\boldsymbol{\tau}^*$, is just $\log \text{GEM}(\boldsymbol{\tau}^*; \alpha)$.

We can return to the stick-breaking weights by dividing each τ_z^* by the sum of all of the indices greater than z (recalling that τ sums to one), $T_z \equiv 1 - \sum_{i=1}^{z-1} \tau_i$. Using this reformulation, the total likelihood bound, including Equation B.5 as $\mathcal{L}_d$, is then[9]

$$
\begin{aligned}
\mathcal{L} = & \sum_{d}^{M} \mathcal{L}_d \\
& + (\alpha - 1) \log T_K - \sum_{z}^{K-1} \log T_z \\
& + \sum_{k=1}^{K+1} [\log \Gamma \left(\textstyle\sum_{j=1}^{K} \alpha_{T,j} \tau^* \right) - \sum_{i=1}^{K} \log \Gamma \left(\alpha_{T,i} \tau^* \right) + \sum_{i=1}^{K} \left(\alpha_{T,i} \tau^* - 1 \right) \left(\Psi \left(\nu_{k,i} \right) - \Psi \left(\textstyle\sum_{j=1}^{K} \nu_{k,j} \right) \right) \\
& - \log \Gamma \left(\textstyle\sum_{j=1}^{K} \nu_{k,j} \right) + \sum_{i=1}^{K} \log \Gamma \left(\nu_{k,i} \right) - \sum_{i=1}^{K} (\nu_{k,i} - 1) \left(\Psi \left(\nu_{k,i} \right) - \Psi \left(\textstyle\sum_{j=1}^{K} \nu_{k,j} \right) \right)].
\end{aligned}
\tag{D.1}
$$

Variational Dirichlet for Parent-child Transitions. Like the update for $\boldsymbol{\gamma}$, the interaction between $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ in the normalizer prevents us from solving the optimization for each of the ν_i explicitly. Differentiating the global likelihood bound, keeping in mind Equation C.1, gives

$$
\begin{aligned}
\frac{\partial L}{\partial \nu_{i,j}} = & \Psi' \left(\nu_{i,j} \right) \left(\alpha_{T,j} \tau_j^* + \sum_{n=1}^{N} \sum_{c \in c(n)} \phi_{n,i} \phi_{c,j} - \nu_{i,j} \right) \\
& - \Psi' \left(\textstyle\sum_{k=1}^{K} \nu_{i,k} \right) \sum_{k=1}^{K} \left[\alpha_{T,k} \tau_k^* + \sum_{n=1}^{N} \sum_{c \in c(n)} \phi_{n,i} \phi_{c,k} - \nu_{i,k} \right] \\
& - \sum_{n}^{N} \phi_{n,i} \sum_{c \in c(n)} \left[\omega_c^{-1} \frac{\gamma_j \sum_{k \neq j}^{N} \nu_{i,k} - \sum_{k \neq j}^{N} \nu_{i,k} \gamma_k}{\left(\sum_{k=1}^{N} \nu_{j,k} \right)^2 \sum_{k=1}^{N} \gamma_k} \right].
\end{aligned}
$$

Each of the $\boldsymbol{\nu}_i$ are then maximized individually using conjugate gradient optimization after transforming the vector into logspace to assure non-negativity.

Variational Top-level Weights. The last variational parameter is $\boldsymbol{\tau}^*$, which is the variational estimate of the top-level weights $\boldsymbol{\tau}$. Because τ_K^* is implicitly defined as $\left(1 - \sum_{i=0}^{K-1} \tau_i^* \right)$, τ_K^* appears in the partial derivative of $\boldsymbol{\tau}^*$ with respect to τ_k^* for $k < K$. Similarly, we must also use implicit differentiation with respect to the stick breaking proportions T_z, defined above. Taking

[9] For simplicity, we present inference with the per-topic distribution β as a parameter. Inference for the complete model with β from a Dirichlet distribution requires adding an additional variational parameter. This is straightforward, but would further complicate the exposition.

the derivative and implicitly differentiating τ_K gives us

$$\begin{aligned}\frac{\partial L_{\tau^*}}{\partial \tau_k^*} &= \left(\sum_{z=k+1}^{K-1} \frac{1}{T_z} \right) - \frac{\alpha - 1}{T_K} \\ &+ \alpha_D \sum_{d}^{M} \left(\Psi\left(\gamma_{d,k}\right) - \Psi\left(\textstyle\sum_{j=1}^{K} \gamma_{d,j}\right)\right) - \alpha_D \sum_{d}^{M} \left(\Psi\left(\gamma_{d,K}\right) - \Psi\left(\textstyle\sum_{j=1}^{K} \gamma_{d,j}\right)\right) \\ &+ \alpha_T \sum_{z}^{K} \left(\Psi\left(\nu_{z,k}\right) - \Psi\left(\textstyle\sum_{j=1}^{K} \nu_{z,j}\right)\right) - \alpha_T \sum_{z}^{K} \left(\Psi\left(\nu_{z,K}\right) - \Psi\left(\textstyle\sum_{j=1}^{K} \nu_{z,j}\right)\right) \\ &- K \left[\alpha_T \Psi\left(\alpha_T \tau_k^*\right) - \alpha_T \Psi\left(\alpha_T \tau_K^*\right)\right] \\ &- M \left[\alpha_D \Psi\left(\alpha_D \tau_k^*\right) - \alpha_D \Psi\left(\alpha_D \tau_K^*\right)\right] \end{aligned} \tag{D.2}$$

again, we transform the vector into logspace to ensure that τ is greater than zero.