Topic Modeling in Embedding Spaces

Adji B. Dieng
Columbia University
New York, NY, USA
abd2141@columbia.edu

Francisco J. R. Ruiz*
DeepMind
London, UK
franrruiz@google.com

David M. Blei
Columbia University
New York, NY, USA
david.blei@columbia.edu

Abstract

Topic modeling analyzes documents to learn meaningful patterns of words. However, existing topic models fail to learn interpretable topics when working with large and heavy-tailed vocabularies. To this end, we develop the *embedded topic model* (ETM), a generative model of documents that marries traditional topic models with word embeddings. More specifically, the ETM models each word with a categorical distribution whose natural parameter is the inner product between the word's embedding and an embedding of its assigned topic. To fit the ETM, we develop an efficient amortized variational inference algorithm. The ETM discovers interpretable topics even with large vocabularies that include rare words and stop words. It outperforms existing document models, such as latent Dirichlet allocation, in terms of both topic quality and predictive performance.

1 Introduction

Topic models are statistical tools for discovering the hidden semantic structure in a collection of documents (Blei et al., 2003; Blei, 2012). Topic models and their extensions have been applied to many fields, such as marketing, sociology, political science, and the digital humanities. Boyd-Graber et al. (2017) provide a review.

Most topic models build on latent Dirichlet allocation (LDA) (Blei et al., 2003). LDA is a hierarchical probabilistic model that represents each topic as a distribution over terms and represents each document as a mixture of the topics. When fit to a collection of documents, the topics summarize their contents, and the topic proportions provide a low-dimensional representation of each document. LDA can be fit to large datasets of text by using variational inference and stochastic optimization (Hoffman et al., 2010, 2013).

LDA is a powerful model and it is widely used. However, it suffers from a pervasive technical problem—it fails in the face of large vocabularies. Practitioners must severely prune their vocabularies in order to fit good topic models—namely, those that are both predictive and interpretable. This is typically done by removing the most and least frequent words. On large collections, this pruning may remove important terms and limit the scope of the models. The problem of topic modeling with large vocabularies has yet to be addressed in the research literature.

In parallel with topic modeling came the idea of word embeddings. Research in word embeddings begins with the neural language model of Bengio et al. (2003), published in the same year and journal as Blei et al. (2003). Word embeddings eschew the ''one-hot'' representation of words—a vocabulary-length vector of zeros with a single one—to learn a distributed representation, one where words with similar meanings are close in a lower-dimensional vector space (Rumelhart and Abrahamson, 1973; Bengio et al., 2006). As for topic models, researchers scaled up embedding methods to large datasets (Mikolov et al., 2013a,b; Pennington et al., 2014; Levy and Goldberg, 2014; Mnih and Kavukcuoglu, 2013). Word embeddings have been extended and developed in many ways. They have become crucial in many applications of natural language processing (Maas et al., 2011; Li and Yang, 2018), and they have also been extended to datasets beyond text (Rudolph et al., 2016).

In this paper, we develop the *embedded topic model* (ETM), a document model that marries LDA and word embeddings. The ETM enjoys the good properties of topic models and the good properties

*Work done while at Columbia University and the University of Cambridge.

Transactions of the Association for Computational Linguistics, vol. 8, pp. 439–453, 2020. https://doi.org/10.1162/tacl_a_00325
Action Editor: Doug Downey. Submission batch: 2/2020; Revision batch: 5/2020; Published 7/2020.
© 2020 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00325 by guest on 29 March 2024

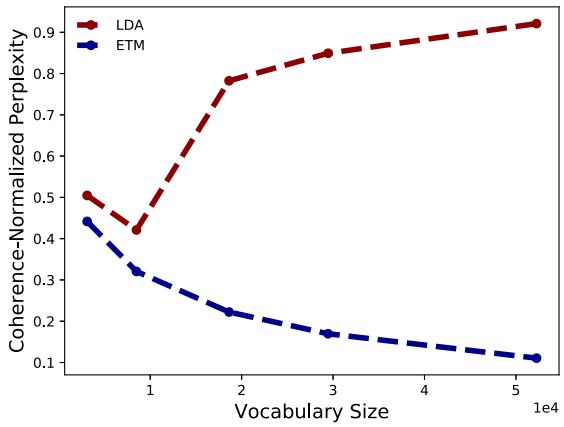


Figure 1: Ratio of the held-out perplexity on a document completion task and the topic coherence as a function of the vocabulary size for the ETM and LDA on the *20NewsGroup* corpus. The perplexity is normalized by the size of the vocabulary. While the performance of LDA deteriorates for large vocabularies, the ETM maintains good performance.

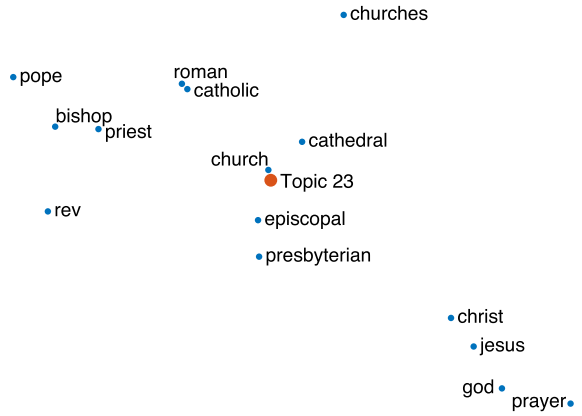


Figure 2: A topic about Christianity found by the ETM on *The New York Times*. The topic is a point in the word embedding space.

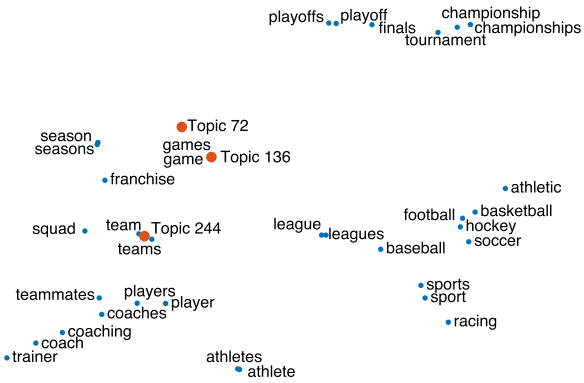


Figure 3: Topics about sports found by the ETM on *The New York Times*. Each topic is a point in the word embedding space.

of word embeddings. As a topic model, it discovers an interpretable latent semantic structure of the documents; as a word embedding model, it provides a low-dimensional representation of the meaning of words. The ETM robustly accommodates large vocabularies and the long tail of language data.

Figure 1 illustrates the advantages. This figure shows the ratio between the perplexity on held-out documents (a measure of predictive performance) and the topic coherence (a measure of the quality of the topics), as a function of the size of the vocabulary. (The perplexity has been normalized by the vocabulary size.) This is for a corpus of 11.2K articles from the *20NewsGroup* and for 100 topics. The red line is LDA; its performance deteriorates as the vocabulary size increases—the predictive performance and the quality of the topics get worse. The blue line is the ETM; it maintains good performance, even as the vocabulary size become large.

Like LDA, the ETM is a generative probabilistic model: Each document is a mixture of topics and each observed word is assigned to a particular topic. In contrast to LDA, the per-topic conditional probability of a term has a log-linear form that involves a low-dimensional representation of the vocabulary. Each term is represented by an embedding and each topic is a point in that embedding space. The topic's distribution over terms is proportional to the exponentiated inner product of the topic's embedding and each term's embedding. Figures 2 and 3 show topics from a 300-topic ETM of *The New York Times*. The figures show each topic's embedding and its closest words; these topics are about Christianity and sports.

Representing topics as points in the embedding space allows the ETM to be robust to the presence of stop words, unlike most topic models. When stop words are included in the vocabulary, the ETM assigns topics to the corresponding area of the embedding space (we demonstrate this in Section 6).

As for most topic models, the posterior of the topic proportions is intractable to compute. We derive an efficient algorithm for approximating the posterior with variational inference (Jordan et al., 1999; Hoffman et al., 2013; Blei et al., 2017) and additionally use amortized inference to efficiently approximate the topic proportions (Kingma and Welling, 2014; Rezende et al., 2014).

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00325 by guest on 29 March 2024

The resulting algorithm fits the ETM to large corpora with large vocabularies. This algorithm can either use previously fitted word embeddings, or fit them jointly with the rest of the parameters. (In particular, Figures 1 to 3 were made using the version of the ETM that uses pre-fitted skip-gram word embeddings.)

We compared the performance of the ETM to LDA, the neural variational document model (NVDM) (Miao et al., 2016), and PRODLDA (Srivastava and Sutton, 2017).[1] The NVDM is a form of multinomial matrix factorization and PRODLDA is a modern version of LDA that uses a product of experts to model the distribution over words. We also compare to a document model that combines PRODLDA with pre-fitted word embeddings. The ETM yields better predictive performance, as measured by held-out log-likelihood on a document completion task (Wallach et al., 2009b). It also discovers more meaningful topics, as measured by topic coherence (Mimno et al., 2011) and topic diversity. The latter is a metric we introduce in this paper that, together with topic coherence, gives a better indication of the quality of the topics. The ETM is especially robust to large vocabularies.

2 Related Work

This work develops a new topic model that extends LDA. LDA has been extended in many ways, and topic modeling has become a subfield of its own. For a review, see Blei (2012) and Boyd-Graber et al. (2017).

A broader set of related works are neural topic models. These mainly focus on improving topic modeling inference through deep neural networks (Srivastava and Sutton, 2017; Card et al., 2017; Cong et al., 2017; Zhang et al., 2018). Specifically, these methods reduce the dimension of the text data through amortized inference and the variational auto-encoder (Kingma and Welling, 2014; Rezende et al., 2014). To perform inference in the ETM, we also avail ourselves of amortized inference methods (Gershman and Goodman, 2014).

As a document model, the ETM also relates to works that learn per-document representations as part of an embedding model (Le and Mikolov, 2014; Moody, 2016; Miao et al., 2016; Li et al., 2016). In contrast to these works, the document variables in the ETM are part of a larger probabilistic topic model.

One of the goals in developing the ETM is to incorporate word similarity into the topic model, and there is previous research that shares this goal. These methods either modify the topic priors (Petterson et al., 2010; Zhao et al., 2017b; Shi et al., 2017; Zhao et al., 2017a) or the topic assignment priors (Xie et al., 2015). For example, Petterson et al. (2010) use a word similarity graph (as given by a thesaurus) to bias LDA towards assigning similar words to similar topics. As another example, Xie et al. (2015) model the per-word topic assignments of LDA using a Markov random field to account for both the topic proportions and the topic assignments of similar words. These methods use word similarity as a type of ''side information'' about language; in contrast, the ETM directly models the similarity (via embeddings) in its generative process of words.

However, a more closely related set of works directly combine topic modeling and word embeddings. One common strategy is to convert the discrete text into continuous observations of embeddings, and then adapt LDA to generate real-valued data (Das et al., 2015; Xun et al., 2016; Batmanghelich et al., 2016; Xun et al., 2017). With this strategy, topics are Gaussian distributions with latent means and covariances, and the likelihood over the embeddings is modeled with a Gaussian (Das et al., 2015) or a Von-Mises Fisher distribution (Batmanghelich et al., 2016). The ETM differs from these approaches in that it is a model of categorical data, one that goes through the embeddings matrix. Thus it does not require pre-fitted embeddings and, indeed, can learn embeddings as part of its inference process. The ETM also differs from these approaches in that it is amenable to large datasets with large vocabularies.

There are few other ways of combining LDA and embeddings. Nguyen et al. (2015) mix the likelihood defined by LDA with a log-linear model that uses pre-fitted word embeddings; Bunk and Krestel (2018) randomly replace words drawn from a topic with their embeddings drawn from a Gaussian; Xu et al. (2018) adopt a geometric perspective, using Wasserstein distances to learn topics and word embeddings jointly; and Keya et al. (2019) propose the neural embedding allocation (NEA), which has a similar generative process to the ETM but is fit using a pre-fitted LDA model as

[1] Code is available at https://github.com/adjidieng/ETM.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00325 by guest on 29 March 2024

a target distribution. Because it requires LDA, the NEA suffers from the same limitation as LDA. These models often lack scalability with respect to the vocabulary size and are fit using Gibbs sampling, limiting their scalability to large corpora.

3 Background

The ETM builds on two main ideas, LDA and word embeddings. Consider a corpus of D documents, where the vocabulary contains V distinct terms. Let $w_{dn} \in \{1, \dots, V\}$ denote the n^{th} word in the d^{th} document.

Latent Dirichlet Allocation. LDA is a probabilistic generative model of documents (Blei et al., 2003). It posits K topics $\beta_{1:K}$, each of which is a distribution over the vocabulary. LDA assumes each document comes from a mixture of topics, where the topics are shared across the corpus and the mixture proportions are unique for each document. The generative process for each document is the following:

1. Draw topic proportion $\theta_d \sim \text{Dirichlet}(\alpha_\theta)$.
2. For each word n in the document:
   (a) Draw topic assignment $z_{dn} \sim \text{Cat}(\theta_d)$.
   (b) Draw word $w_{dn} \sim \text{Cat}(\beta_{z_{dn}})$.

Here, $\text{Cat}(\cdot)$ denotes the categorical distribution. LDA places a Dirichlet prior on the topics,

$$\beta_k \sim \text{Dirichlet}(\alpha_\beta) \text{ for } k = 1, \dots, K.$$

The concentration parameters α_β and α_θ of the Dirichlet distributions are fixed model hyperparameters.

Word Embeddings. Word embeddings provide models of language that use vector representations of words (Rumelhart and Abrahamson, 1973; Bengio et al., 2003). The word representations are fitted to relate to meaning, in that words with similar meanings will have representations that are close. (In embeddings, the ''meaning'' of a word comes from the contexts in which it is used [Harris, 1954].)

We focus on the continuous bag-of-words (CBOW) variant of word embeddings (Mikolov et al., 2013b). In CBOW, the likelihood of each word w_{dn} is

$$w_{dn} \sim \text{softmax}(\rho^\top \alpha_{dn}). \quad (1)$$

The embedding matrix ρ is a $L \times V$ matrix whose columns contain the embedding representations of the vocabulary, $\rho_v \in \mathbb{R}^L$. The vector α_{dn} is the *context embedding*. The context embedding is the sum of the context embedding vectors (α_v for each word v) of the words surrounding w_{dn}.

4 The Embedded Topic Model

The ETM is a topic model that uses embedding representations of both words and topics. It contains two notions of latent dimension. First, it embeds the vocabulary in an L-dimensional space. These embeddings are similar in spirit to classical word embeddings. Second, it represents each document in terms of K latent topics.

In traditional topic modeling, each topic is a full distribution over the vocabulary. In the ETM, however, the k^{th} topic is a vector $\alpha_k \in \mathbb{R}^L$ in the embedding space. We call α_k a *topic embedding*—it is a distributed representation of the k^{th} topic in the semantic space of words.

In its generative process, the ETM uses the topic embedding to form a per-topic distribution over the vocabulary. Specifically, the ETM uses a log-linear model that takes the inner product of the word embedding matrix and the topic embedding. With this form, the ETM assigns high probability to a word v in topic k by measuring the agreement between the word's embedding and the topic's embedding.

Denote the $L \times V$ word embedding matrix by ρ; the column ρ_v is the embedding of term v. Under the ETM, the generative process of the d^{th} document is the following:

1. Draw topic proportions $\theta_d \sim \mathcal{LN}(0, I)$.
2. For each word n in the document:
   a. Draw topic assignment $z_{dn} \sim \text{Cat}(\theta_d)$.
   b. Draw the word $w_{dn} \sim \text{softmax}(\rho^\top \alpha_{z_{dn}})$.

In Step 1, $\mathcal{LN}(\cdot)$ denotes the logistic-normal distribution (Aitchison and Shen, 1980; Blei and Lafferty, 2007); it transforms a standard Gaussian random variable to the simplex. A draw θ_d from this distribution is obtained as

$$\delta_d \sim \mathcal{N}(0, I); \quad \theta_d = \text{softmax}(\delta_d). \quad (2)$$

(We replaced the Dirichlet with the logistic normal to easily use reparameterization in the inference algorithm; see Section 5.)

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00325 by guest on 29 March 2024

Steps 1 and 2a are standard for topic modeling: They represent documents as distributions over topics and draw a topic assignment for each observed word. Step 2b is different; it uses the embeddings of the vocabulary ρ and the assigned topic embedding $\alpha_{z_{dn}}$ to draw the observed word from the assigned topic, as given by z_{dn}.

The topic distribution in Step 2b mirrors the CBOW likelihood in Eq. 1. Recall CBOW uses the surrounding words to form the context vector α_{dn}. In contrast, the ETM uses the topic embedding $\alpha_{z_{dn}}$ as the context vector, where the assigned topic z_{dn} is drawn from the per-document variable θ_d. The ETM draws its words from a document context, rather than from a window of surrounding words.

The ETM likelihood uses a matrix of word embeddings ρ, a representation of the vocabulary in a lower dimensional space. In practice, it can either rely on previously fitted embeddings or learn them as part of its overall fitting procedure. When the ETM learns the embeddings as part of the fitting procedure, it simultaneously finds topics and an embedding space.

When the ETM uses previously fitted embeddings, it learns the topics of a corpus in a particular embedding space. This strategy is particularly useful when there are words in the embedding that are not used in the corpus. The ETM can hypothesize how those words fit in to the topics because it can calculate $\rho_v^\top \alpha_k$ even for words v that do not appear in the corpus.

5 Inference and Estimation

We are given a corpus of documents $\{\mathbf{w}_1, \ldots, \mathbf{w}_D\}$, where the d^{th} document $\mathbf{w}_d$ is a collection of N_d words. How do we fit the ETM to this corpus?

The Marginal Likelihood. The parameters of the ETM are the word embeddings $\rho_{1:V}$ and the topic embeddings $\alpha_{1:K}$; each α_k is a point in the word embedding space. We maximize the log marginal likelihood of the documents,

$$\mathcal{L}(\alpha, \rho) \;=\; \sum_{d=1}^{D} \log p(\mathbf{w}_d \,|\, \alpha, \rho). \tag{3}$$

The problem is that the marginal likelihood of each document—$p(\mathbf{w}_d \,|\, \alpha, \rho)$—is intractable to compute. It involves a difficult integral over the topic proportions, which we write in terms of the untransformed proportions δ_d in Eq. 2,

$$p(\mathbf{w}_d \,|\, \alpha, \rho) \;=\; \int p(\delta_d) \prod_{n=1}^{N_d} p(w_{dn} \,|\, \delta_d, \alpha, \rho)\, \mathrm{d}\delta_d. \tag{4}$$

The conditional distribution $p(w_{dn} \,|\, \delta_d, \alpha, \rho)$ of each word marginalizes out the topic assignment z_{dn},

$$p(w_{dn} \,|\, \delta_d, \alpha, \rho) \;=\; \sum_{k=1}^{K} \theta_{dk} \beta_{k, w_{dn}}. \tag{5}$$

Here, θ_{dk} denotes the (transformed) topic proportions (Eq. 2) and $\beta_{k,v}$ denotes a traditional ''topic,'' that is, a distribution over words, induced by the word embeddings ρ and the topic embedding α_k,

$$\beta_{kv} \;=\; \text{softmax}(\rho^\top \alpha_k)\big|_v. \tag{6}$$

Eqs. 4, 5, 6 flesh out the likelihood in Eq. 3.

Variational Inference. We sidestep the intractable integral in Eq. eq:integral with variational inference (Jordan et al., 1999; Blei et al., 2017). Variational inference optimizes a sum of per-document bounds on the log of the marginal likelihood of Eq. 4.

To begin, posit a family of distributions of the untransformed topic proportions $q(\delta_d \,;\, \mathbf{w}_d, \nu)$. This family of distributions is parameterized by ν. We use amortized inference, where $q(\delta_d \,;\, \mathbf{w}_d, \nu)$ (called a *variational distribution*) depends on both the document $\mathbf{w}_d$ and shared parameters ν. In particular, $q(\delta_d \,;\, \mathbf{w}_d, \nu)$ is a Gaussian whose mean and variance come from an ''inference network,'' a neural network parameterized by ν (Kingma and Welling, 2014). The inference network ingests a bag-of-words representation of the document $\mathbf{w}_d$ and outputs the mean and covariance of δ_d. (To accommodate documents of varying length, we form the input of the inference network by normalizing the bag-of-word representation of the document by the number of words N_d.)

We use this family of distributions to bound the log of the marginal likelihood in Eq. 4. The bound is called the evidence lower bound (ELBO)

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00325 by guest on 29 March 2024

and is a function of the model parameters and the variational parameters,

$$\mathcal{L}(\alpha, \rho, \nu) = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}q[\, \log p(w_{nd} \,|\, \delta_d, \rho, \alpha)\,] - \sum_{d=1}^{D} \mathrm{KL}(q(\delta_d; \mathbf{w}_d, \nu) \,||\, p(\delta_d)). \quad (7)$$

The first term of the ELBO (Eq. 7) encourages variational distributions $q(\delta_d\,;\, \mathbf{w}_d, \nu)$ that place mass on topic proportions δ_d that explain the observed words and the second term encourages $q(\delta_d\,;\, \mathbf{w}_d, \nu)$ to be close to the prior $p(\delta_d)$. Maximizing the ELBO with respect to the model parameters (α, ρ) is equivalent to maximizing the expected complete log-likelihood, $\sum_d \log p(\delta_d, \mathbf{w}_d \,|\, \alpha, \rho)$.

The ELBO in Eq. 7 is intractable because the expectation is intractable. However, we can form a Monte Carlo approximation of the ELBO,

$$\tilde{\mathcal{L}}(\alpha, \rho, \nu) \;=\; \frac{1}{S} \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{s=1}^{S} \log p(w_{nd} \,|\, \delta_d^{(s)}, \rho, \alpha) - \sum_{d=1}^{D} \mathrm{KL}(q(\delta_d; \mathbf{w}_d, \nu) \,||\, p(\delta_d)), \quad (8)$$

where $\delta_d^{(s)} \sim q(\delta_d; \mathbf{w}_d, \nu)$ for $s = 1 \ldots S$. To form an unbiased estimator of the ELBO and its gradients, we use the reparameterization trick when sampling the unnormalized proportions $\delta_d^{(1)}, \ldots, \delta_d^{(S)}$ (Kingma and Welling, 2014; Titsias and Lázaro-Gredilla, 2014; Rezende et al., 2014). That is, we sample $\delta_d^{(s)}$ from $q(\delta_d; \mathbf{w}_d, \nu)$ as

$$\epsilon_d^{(s)} \sim \mathcal{N}(0, I) \text{ and } \delta_d^{(s)} \;=\; \mu_d + \Sigma_d^{\frac{1}{2}} \epsilon_d^{(s)}, \quad (9)$$

where μ_d and Σ_d are the mean and covariance of $q(\delta_d; \mathbf{w}_d, \nu)$ respectively, which depend implicitly on ν and $\mathbf{w}_d$ via the inference network. We use a diagonal covariance matrix Σ_d.

We also use data subsampling to handle large collections of documents (Hoffman et al., 2013). Denote by $\mathcal{B}$ a minibatch of documents. Then the approximation of the ELBO using data subsampling is

$$\tilde{\mathcal{L}}(\alpha, \rho, \nu) \;=\; \frac{D}{|\mathcal{B}|} \sum_{d \in \mathcal{B}} \sum_{n=1}^{N_d} \sum_{s=1}^{S} \log p(w_{nd} \,|\, \delta_d^{(s)}, \rho, \alpha) - \frac{D}{|\mathcal{B}|} \sum_{d \in \mathcal{B}} \mathrm{KL}(q(\delta_d; \mathbf{w}_d, \nu) \,||\, p(\delta_d)). \quad (10)$$

Algorithm 1 Topic modeling with the ETM

```
Initialize model and variational parameters
for iteration i = 1, 2, ... do
    Compute β_k = softmax(ρ^⊤ α_k) for each topic k
    Choose a minibatch B of documents
    for each document d in B do
        Get normalized bag-of-word representat. x_d
        Compute μ_d = NN(x_d ; ν_μ)
        Compute Σ_d = NN(x_d ; ν_Σ)
        Sample θ_d using Eq. 9 and θ_d = softmax(δ_d)
        for each word in the document do
            Compute p(w_dn | θ_d, ρ, α) = θ_d^⊤ β_{·,w_dn}
        end for
    end for
    Estimate the ELBO using Eq. 10 and Eq. 11
    Take gradients of the ELBO via backpropagation
    Update model parameters α_{1:K} (ρ if necessary)
    Update variational parameters (ν_μ, ν_Σ)
end for
```

Given that the prior $p(\delta_d)$ and $q(\delta_d; \mathbf{w}_d, \nu)$ are both Gaussians, the KL admits a closed-form expression,

$$\mathrm{KL}(q(\delta_d; \mathbf{w}_d, \nu) \,||\, p(\delta_d)) \;= \frac{1}{2} \left\{ \mathrm{tr}(\Sigma_d) + \mu_d^\top \mu_d - \log \det(\Sigma_d) - K \right\}. \quad (11)$$

We optimize the stochastic ELBO in Equation 10 with respect to both the model parameters (α, ρ) and the variational parameters ν. We set the learning rate with Adam (Kingma and Ba, 2015). The procedure is shown in Algorithm 1, where we set the number of Monte Carlo samples $S = 1$ and the notation $\mathrm{NN}(\mathbf{x}\,;\, \nu)$ represents a neural network with input $\mathbf{x}$ and parameters ν.

6 Empirical Study

We study the performance of the ETM and compare it to other unsupervised document models. A good document model should provide both coherent patterns of language and an accurate distribution of words, so we measure performance in terms of both predictive accuracy and topic interpretability. We measure accuracy with log-likelihood on a document completion task (Rosen-Zvi et al., 2004; Wallach et al., 2009b); we measure topic interpretability as a blend of topic coherence and diversity. We find that, of the interpretable models, the ETM is the one that provides better predictions and topics.

In a separate analysis (Section 6.1), we study the robustness of each method in the presence

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00325 by guest on 29 March 2024

Dataset	Minimum DF	#Tokens Train	#Tokens Valid	#Tokens Test	Vocabulary
20Newsgroups	100	604.9 K	5,998	399.6 K	3,102
	30	778.0 K	7,231	512.5 K	8,496
	10	880.3 K	6,769	578.8 K	18,625
	5	922.3 K	8,494	605.9 K	29,461
	2	966.3 K	8,600	622.9 K	52,258
New York Times	5,000	226.9 M	13.4 M	26.8 M	9,842
	200	270.1 M	15.9 M	31.8 M	55,627
	100	272.3 M	16.0 M	32.1 M	74,095
	30	274.8 M	16.1 M	32.3 M	124,725
	10	276.0 M	16.1 M	32.5 M	212,237

Table 1: Statistics of the different corpora studied. DF denotes document frequency, K denotes a thousand, and M denotes a million.

of stop words. Standard topic models fail in this regime—because stop words appear in many documents, every learned topic includes some stop words, leading to poor topic interpretability. In contrast, the ETM is able to use the information from the word embeddings to provide interpretable topics.

Corpora. We study the *20Newsgroups* corpus and the *New York Times* corpus; the statistics of both corpora are summarized in Table 1.

The *20Newsgroup* corpus is a collection of newsgroup posts. We preprocess the corpus by filtering stop words, words with document frequency above 70%, and tokenizing. To form the vocabulary, we keep all words that appear in more than a certain number of documents, and we vary the threshold from 100 (a smaller vocabulary, where $V = 3{,}102$) to 2 (a larger vocabulary, where $V = 52{,}258$). After preprocessing, we further remove one-word documents from the validation and test sets. We split the corpus into a training set of 11,260 documents, a test set of 7,532 documents, and a validation set of 100 documents.

The *New York Times* corpus is a larger collection of news articles. It contains more than 1.8 million articles, spanning the years 1987–2007. We follow the same preprocessing steps as for *20Newsgroups*. We form versions of this corpus with vocabularies ranging from $V = 9{,}842$ to $V = 212{,}237$. After preprocessing, we use 85% of the documents for training, 10% for testing, and 5% for validation.

Models. We compare the performance of the ETM against several document models. We briefly describe each below.

We consider latent Dirichlet allocation (LDA) (Blei et al., 2003), a standard topic model that posits Dirichlet priors for the topics β_k and topic proportions θ_d. (We set the prior hyperparameters to 1.) It is a conditionally conjugate model, amenable to variational inference with coordinate ascent. We consider LDA because it is the most commonly used topic model, and it has a similar generative process as the ETM.

We also consider the neural variational document model (NVDM) (Miao et al., 2016). The NVDM is a multinomial factor model of documents; it posits the likelihood $w_{dn} \sim \text{softmax}\ (\beta^\top \theta_d)$, where the K-dimensional vector $\theta_d \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$ is a per-document variable, and β is a real-valued matrix of size $K \times V$. The NVDM uses a per-document real-valued latent vector θ_d to average over the embedding matrix β in the logit space. Like the ETM, the NVDM uses amortized variational inference to jointly learn the approximate posterior over the document representation θ_d and the model parameter β.

NVDM is not interpretable as a topic model; its latent variables are unconstrained. We study a more interpretable variant of the NVDM which constrains θ_d to lie in the simplex, replacing its Gaussian prior with a logistic normal (Aitchison and Shen, 1980). (This can be thought of as a semi-nonnegative matrix factorization.) We call this document model Δ-NVDM.

We also consider PRODLDA (Srivastava and Sutton, 2017). It posits the likelihood $w_{dn} \sim \text{softmax}(\beta^\top \theta_d)$ where the topic proportions θ_d are from the simplex. Contrary to LDA, the topic-matrix β s unconstrained.

PRODLDA shares the generative model with Δ-NVDM but it is fit differently. PRODLDA uses

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00325 by guest on 29 March 2024

Skip-gram embeddings

love	**family**	**woman**	**politics**
loved	families	man	political
passion	grandparents	girl	religion
loves	mother	boy	politicking
affection	friends	teenager	ideology
adore	relatives	person	partisanship

ETM embeddings

love	**family**	**woman**	**politics**
joy	children	girl	political
loves	son	boy	politician
loved	mother	mother	ideology
passion	father	daughter	speeches
wonderful	wife	pregnant	ideological

NVDM embeddings

love	**family**	**woman**	**politics**
loves	sons	girl	political
passion	life	women	politician
wonderful	brother	man	politicians
joy	son	pregnant	politically
beautiful	lived	boyfriend	democratic

Δ-NVDM embeddings

love	**family**	**woman**	**politics**
miss	home	life	political
young	father	marriage	faith
born	son	women	marriage
dream	day	read	politicians
younger	mrs	young	election

PRODLDA embeddings

love	**family**	**woman**	**politics**
loves	husband	girl	political
affection	wife	boyfriend	politician
sentimental	daughters	boy	liberal
dreams	sister	teenager	politicians
laugh	friends	ager	ideological

Table 2: Word embeddings learned by all document models (and skip-gram) on the *New York Times* with vocabulary size 118,363.

amortized variational inference with batch normalization (Ioffe and Szegedy, 2015) and dropout (Srivastava et al., 2014).

Finally, we consider a document model that combines PRODLDA with pre-fitted word embeddings ρ, by using the likelihood $w_{dn} \sim \text{softmax}(\rho^\top \theta_d)$. We call this document model PRODLDA-PWE, where PWE stands for Pre-fitted Word Embeddings.

We study two variants of the ETM, one where the word embeddings are pre-fitted and one where they are learned jointly with the rest of the parameters. The variant with pre-fitted embeddings is called the ETM-PWE.

For PRODLDA-PWE and the ETM-PWE, we first obtain the word embeddings (Mikolov et al., 2013b) by training skip-gram on each corpus. (We reuse the same embeddings across the experiments with varying vocabulary sizes.)

Algorithm Settings. Given a corpus, each model comes with an approximate posterior inference problem. We use variational inference for all of the models and employ SVI (Hoffman et al., 2013) to speed up the optimization. The minibatch size is 1,000 documents. For LDA, we set the learning rate as suggested by Hoffman et al. (2013): the delay is 10 and the forgetting factor is 0.85.

Within SVI, LDA enjoys coordinate ascent variational updates; we use five inner steps to optimize the local variables. For the other models, we use amortized inference over the local variables θ_d. We use 3-layer inference networks and we set the local learning rate to 0.002. We use ℓ_2 regularization on the variational parameters (the weight decay parameter is 1.2×10^{-6}).

Qualitative Results. We first examine the embeddings. The ETM, NVDM, Δ-NVDM, and PRODLDA all learn word embeddings. We illustrate them by fixing a set of terms and showing the closest words in the embedding space (as measured by cosine distance). For comparison, we also illustrate word embeddings learned by the skip-gram model.

Table 2 illustrates the embeddings of the different models. All the methods provide interpretable embeddings—words with related meanings are close to each other. The ETM, the NVDM, and PRODLDA learn embeddings that are similar to those from the skip-gram. The embeddings of Δ-NVDM are different; the simplex constraint on the local variable and the inference procedure change the nature of the embeddings.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00325 by guest on 29 March 2024

LDA						
time	year	officials	mr	city	percent	state
day	million	public	president	building	million	republican
back	money	department	bush	street	company	party
good	pay	report	white	park	year	bill
long	tax	state	clinton	house	billion	mr
NVDM						
scholars	japan	gansler	spratt	assn	ridership	pryce
gingrich	tokyo	wellstone	tabitha	assoc	mtv	mickens
funds	pacific	mccain	mccorkle	qtr	straphangers	mckechnie
institutions	europe	shalikashvili	cheetos	yr	freierman	mfume
endowment	zealand	coached	vols	nyse	riders	filkins
Δ-NVDM						
concerto	servings	nato	innings	treas	patients	democrats
solos	tablespoons	soviet	scored	yr	doctors	republicans
sonata	tablespoon	iraqi	inning	qtr	medicare	republican
melodies	preheat	gorbachev	shutout	outst	dr	senate
soloist	minced	arab	scoreless	telerate	physicians	dole
PRODLDA						
temptation	grasp	electron	played	amato	briefly	giant
repressed	unruly	nuclei	lou	model	precious	boarding
drowsy	choke	macal	greg	delaware	serving	bundle
addiction	drowsy	trained	bobby	morita	set	distance
conquering	drift	mediaone	steve	dual	virgin	foray
PRODLDA-PWE						
mercies	cheesecloth	scoreless	chapels	distinguishable	floured	gillers
lockbox	overcook	floured	magnolias	cocktails	impartiality	lacerated
pharm	strainer	hitless	asea	punishable	knead	polshek
shims	kirberger	asterisk	bogeyed	checkpoints	refrigerate	decimated
cp	browned	knead	birdie	disobeying	tablespoons	inhuman
ETM-PWE						
music	republican	yankees	game	wine	court	company
dance	bush	game	points	restaurant	judge	million
songs	campaign	baseball	season	food	case	stock
opera	senator	season	team	dishes	justice	shares
concert	democrats	mets	play	restaurants	trial	billion
ETM						
game	music	united	wine	company	yankees	art
team	mr	israel	food	stock	game	museum
season	dance	government	sauce	million	baseball	show
coach	opera	israeli	minutes	companies	mets	work
play	band	mr	restaurant	billion	season	artist

Table 3: Top five words of seven most used topics from different document models on 1.8M documents of the *New York Times* corpus with vocabulary size 212,237 and $K = 300$ topics.

We next look at the learned topics. Table 3 displays the seven most used topics for all methods, as given by the average of the topic proportions θ_d. LDA and both variants of the ETM provide interpretable topics. The rest of the models do not provide interpretable topics; their matrices β are unconstrained and thus are not interpretable as distributions over the vocabulary that mix to

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00325 by guest on 29 March 2024

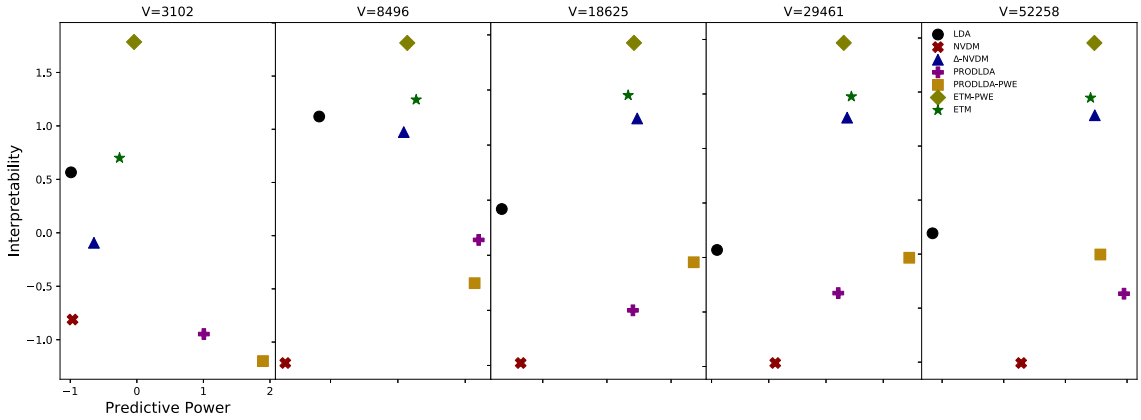


Figure 4: Interpretability as measured by the exponentiated topic quality (the higher the better) vs. predictive performance as measured by log-likelihood on document completion (the higher the better) on the *20NewsGroup* dataset. Both interpretability and predictive power metrics are normalized by subtracting the mean and dividing by the standard deviation across models. Better models are on the top right corner. Overall, the ETM is a better topic model.

form documents. Δ-NVDM also suffers from this effect although it is less apparent (see, e.g., the fifth listed topic for Δ-NVDM).

Quantitative Results. We next study the models quantitatively. We measure the quality of the topics and the predictive performance of the model. We found that among the models with interpretable topics, the ETM provides the best predictions.

We measure topic quality by blending two metrics: topic coherence and topic diversity. Topic coherence is a quantitative measure of the interpretability of a topic (Mimno et al., 2011). It is the average pointwise mutual information of two words drawn randomly from the same document,

$$\text{TC} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{45}\sum_{i=1}^{10}\sum_{j=i+1}^{10} f(w_i^{(k)}, w_j^{(k)}),$$

where $\{w_1^{(k)}, \ldots, w_{10}^{(k)}\}$ denotes the top-10 most likely words in topic k. We choose $f(\cdot,\cdot)$ as the normalized pointwise mutual information (Bouma, 2009; Lau et al., 2014),

$$f(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}.$$

Here, $P(w_i, w_j)$ is the probability of words w_i and w_j co-occurring in a document and $P(w_i)$ is the marginal probability of word w_i. We approximate these probabilities with empirical counts.

The idea behind topic coherence is that a coherent topic will display words that tend to occur in the same documents. In other words, the most likely words in a coherent topic should have high mutual information. Document models with higher topic coherence are more interpretable topic models.

We combine coherence with a second metric, topic diversity. We define topic diversity to be the percentage of unique words in the top 25 words of all topics. Diversity close to 0 indicates redundant topics; diversity close to 1 indicates more varied topics.

We define the overall quality of a model's topics as the product of its topic diversity and topic coherence.

A good topic model also provides a good distribution of language. To measure predictive power, we calculate log likelihood on a document completion task (Rosen-Zvi et al., 2004; Wallach et al., 2009b). We divide each test document into two sets of words. The first half is observed: it induces a distribution over topics which, in turn, induces a distribution over the next words in the document. We then evaluate the second half under this distribution. A good document model should provide high log-likelihood on the second half. (For all methods, we approximate the likelihood by setting θ_d to the variational mean.)

We study both corpora and with different vocabularies. Figures 4 and 5 show interpretability of the topics as a function of predictive power. (To ease visualization, we exponentiate topic quality and normalize all metrics by subtracting the mean and dividing by the standard deviation across

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00325 by guest on 29 March 2024

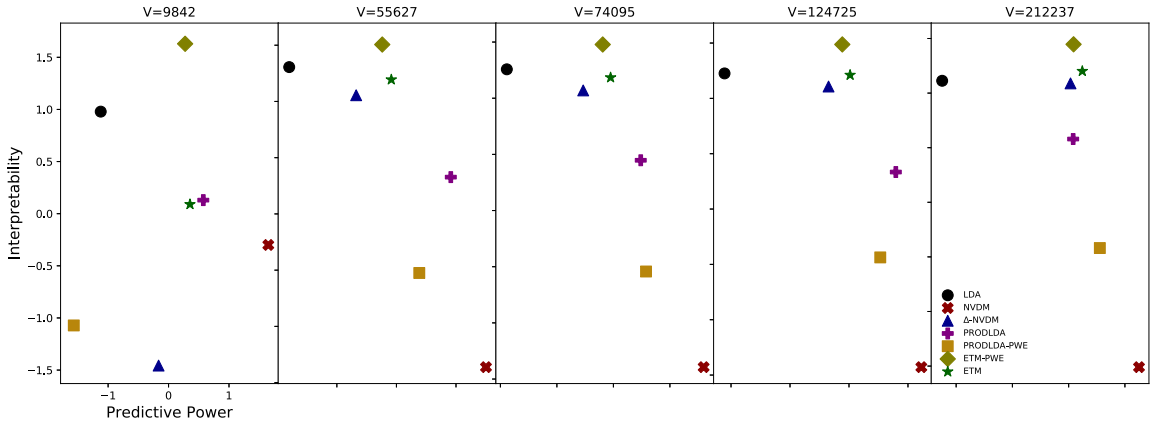


Figure 5: Interpretability as measured by the exponentiated topic quality (the higher the better) vs. predictive performance as measured by log-likelihood on document completion (the higher the better) on the *New York Times* dataset. Both interpretability and predictive power metrics are normalized by subtracting the mean and dividing by the standard deviation across models. Better models are on the top right corner. Overall, the ETM is a better topic model.

methods.) The best models are on the upper right corner.

LDA predicts worst in almost all settings. On the *20NewsGroups*, the NVDM's predictions are in general better than LDA but worse than for the other methods; on the *New York Times*, the NVDM gives the best predictions. However, topic quality for the NVDM is far below the other methods. (It does not provide ''topics'', so we assess the interpretability of its β matrix.) In prediction, both versions of the ETM are at least as good as the simplex-constrained Δ-NVDM. More importantly, both versions of the ETM outperform the PRODLDA-PWE; signaling the ETM provides a better way of integrating word embeddings into a topic model.

These figures show that, of the interpretable models, the ETM provides the best predictive performance while keeping interpretable topics. It is robust to large vocabularies.

6.1 Stop Words

We now study a version of the *New York Times* corpus that includes all stop words. We remove infrequent words to form a vocabulary of size 10,283. Our goal is to show that the ETM-PWE provides interpretable topics even in the presence of stop words, another regime where topic models typically fail. In particular, given that stop words appear in many documents, traditional topic models learn topics that contain stop words, regardless of the actual semantics of the topic. This leads to poor topic interpretability. There are extensions of topic models specifically designed to cope with stop words (Griffiths et al., 2004; Chemudugunta et al., 2006; Wallach et al., 2009a); our goal here is not to establish comparisons with these methods but to show the performance of the ETM-PWE in the presence of stop words.

	TC	TD	Quality
LDA	0.13	0.14	0.0182
Δ-NVDM	0.17	0.11	0.0187
PRODLDA-PWE	0.03	**0.53**	0.0159
ETM-PWE	**0.18**	0.22	**0.0396**

Table 4: Topic quality on the *New York Times* data in the presence of stop words. Topic quality here is given by the product of topic coherence and topic diversity (higher is better). The ETM-PWE is robust to stop words; it achieves similar topic coherence than when there are no stop words.

We fit LDA, the Δ-NVDM, the PRODLDA-PWE, and the ETM-PWE with $K = 300$ topics. (We do not report the NVDM because it does not provide interpretable topics.) Table 4 shows the topic quality (the product of topic coherence and topic diversity). Overall, the ETM-PWE gives the best performance in terms of topic quality.

While the ETM has a few ''stop topics'' that are specific for stop words (see, e.g., Figure 6), Δ-NVDM and LDA have stop words in almost every topic. (The topics are not displayed here for space constraints.) The reason is that stop words co-occur in the same documents as every other word;

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00325 by guest on 29 March 2024



Figure 6: A topic containing stop words found by the ETM-PWE on *The New York Times*. The ETM is robust even in the presence of stop words.

therefore traditional topic models have difficulties telling apart content words and stop words. The ETM-PWE recognizes the location of stop words in the embedding space; its sets them off on their own topic.

7 Conclusion

We developed the ETM, a generative model of documents that marries LDA with word embeddings. The ETM assumes that topics and words live in the same embedding space, and that words are generated from a categorical distribution whose natural parameter is the inner product of the word embeddings and the embedding of the assigned topic.

The ETM learns interpretable word embeddings and topics, even in corpora with large vocabularies. We studied the performance of the ETM against several document models. The ETM learns both coherent patterns of language and an accurate distribution of words.

Acknowledgments

DB and AD are supported by ONR N00014-17-1-2131, ONR N00014-15-1-2209, NIH 1U01MH115727-01, NSF CCF-1740833, DARPA SD2 FA8750-18-C-0130, Amazon, NVIDIA, and the Simons Foundation. FR received funding from the EU's Horizon 2020 R&I programme under the Marie Skłodowska-Curie grant agreement 706760. AD is supported by a Google PhD Fellowship.

References

John Aitchison and Shir Ming Shen. 1980. Logistic normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272.

Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Sam Gershman. 2016. Nonparametric spherical topic modeling with word embeddings. In *Association for Computational Linguistics*, volume 2016, page 537.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. 2006, Neural probabilistic language models. In *Innovations in Machine Learning*.

David M. Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

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

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *German Society for Computational Linguistics and Language Technology Conference*.

Jordan Boyd-Graber, Yuening Hu, and David Mimno. 2017. Applications of topic models. *Foundations and Trends in Information Retrieval*, 11(2–3):143–296.

Stefan Bunk and Ralf Krestel. 2018. WELDA: Enhancing topic models by incorporating local word context. In *ACM/IEEE Joint Conference on Digital Libraries*.

Dallas Card, Chenhao Tan, and Noah A. Smith. 2017. A neural framework for generalized topic models. In *arXiv:1705.09296*.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00325 by guest on 29 March 2024

Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. Modeling general and specific aspects of documents with a probabilistic topic model. In *Advances in Neural Information Processing Systems*.

Yulai Cong, Bo C. Chen, Hongwei Liu, and Mingyuan Zhou. 2017. Deep latent Dirichlet allocation with topic-layer-adaptive stochastic gradient Riemannian MCMC. In *International Conference on Machine Learning*.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Samuel J. Gershman and Noah D. Goodman. 2014. Amortized inference in probabilistic reasoning. In *Annual Meeting of the Cognitive Science Society*.

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2004. Integrating topics and syntax. In *Advances in Neural Information Processing Systems*.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2–3):146–162.

Matthew D. Hoffman, David M. Blei, and Francis Bach. 2010. Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*.

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.

Kamrun Naher Keya, Yannis Papanikolaou, and James R. Foulds. 2019. Neural embedding allocation: Distributed representations of topic models. *arXiv preprint arXiv:1909.04702*.

Diederik P. Kingma and Jimmy L. Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Jey H. Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Neural Information Processing Systems*.

Shaohua Li, Tat-Seng Chua, Jun Zhu, and Chunyan Miao. 2016. Generative topic embedding: A continuous representation of documents. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yang Li and Tao Yang. 2018. Word Embedding for Understanding Natural Language: A Survey, Springer International Publishing.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International Conference on Machine Learning*.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00325 by guest on 29 March 2024

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems*.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Conference on Empirical Methods in Natural Language Processing*.

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Neural Information Processing Systems*.

Christopher E. Moody. 2016. Mixing Dirichlet topic models and word embeddings to make LDA2vec. *arXiv:1605.02019*.

Dat Q. Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Conference on Empirical Methods on Natural Language Processing*.

James Petterson, Wray Buntine, Shravan M. Narayanamurthy, Tibério S. Caetano, and Alex J. Smola. 2010. Word features for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*.

Danilo J. Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Uncertainty in Artificial Intelligence*.

Maja Rudolph, Francisco J. R. Ruiz, Stephan Mandt, and David M. Blei. 2016. Exponential family embeddings. In *Advances in Neural Information Processing Systems*.

David E. Rumelhart and Adele A. Abrahamson. 1973. A model for analogical reasoning. *Cognitive Psychology*, 5(1):1–28.

Bei Shi, Wai Lam, Shoaib Jameel, Steven Schockaert, and Kwun P. Lai. 2017. Jointly learning word embeddings and latent topics. In *ACM SIGIR Conference on Research and Development in Information Retrieval*.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *International Conference on Learning Representations*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Michalis K. Titsias and Miguel Lázaro-Gredilla. 2014. Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*.

Hanna M. Wallach, David M. Mimno, and Andrew McCallum. 2009a. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems*.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009b. Evaluation methods for topic models. In *International Conference on Machine Learning*.

Pengtao Xie, Diyi Yang, and Eric Xing. 2015. Incorporating word correlation knowledge into topic modeling. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Hongteng Xu, Wenlin Wang, Wei Liu, and Lawrence Carin. 2018. Distilled Wasserstein learning for word embedding and topic modeling. In *Advances in Neural Information Processing Systems*.

Guangxu Xun, Vishrawas Gopalakrishnan, Fenglong Ma, Yaliang Li, Jing Gao, and Aidong Zhang. 2016. Topic discovery for short texts using word embeddings. In *IEEE International Conference on Data Mining*.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00325 by guest on 29 March 2024

Guangxu Xun, Yaliang Li, Wayne Xin Zhao, Jing Gao, and Aidong Zhang. 2017. A correlated topic model using word embeddings. In *Joint Conference on Artificial Intelligence*.

Hao Zhang, Bo Chen, Dandan Guo, and Mingyuan Zhou. 2018. WHAI: Weibull hybrid autoencoding inference for deep topic modeling. In *International Conference on Learning Representations*.

He Zhao, Lan Du, and Wray Buntine. 2017a. A word embeddings informed focused topic model. In *Asian Conference on Machine Learning*.

He Zhao, Lan Du, Wray Buntine, and Gang Liu. 2017b. MetaLDA: A topic model that efficiently incorporates meta information. In *IEEE International Conference on Data Mining*.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00325 by guest on 29 March 2024