The Dynamic Embedded Topic Model

Adji B. Dieng[*]
Columbia University
abd2141@columbia.edu

Francisco J. R. Ruiz[*]
Columbia University
Cambridge University
f.ruiz@columbia.edu

David M. Blei
Columbia University
david.blei@columbia.edu

Abstract

Topic modeling analyzes documents to learn meaningful patterns of words. Dynamic topic models capture how these patterns vary over time for a set of documents that were collected over a large time span. We develop the dynamic embedded topic model (D-ETM), a generative model of documents that combines dynamic latent Dirichlet allocation (D-LDA) and word embeddings. The D-ETM models each word with a categorical distribution whose parameter is given by the inner product between the word embedding and an embedding representation of its assigned topic at a particular time step. The word embeddings allow the D-ETM to generalize to rare words. The D-ETM learns smooth topic trajectories by defining a random walk prior over the embeddings of the topics. We fit the D-ETM using structured amortized variational inference. On a collection of United Nations debates, we find that the D-ETM learns interpretable topics and outperforms D-LDA in terms of both topic quality and predictive performance.[2]

1 Introduction

Topic models are useful tools for the statistical analysis of document collections (Blei et al., 2003; Blei, 2012). They have been applied to learn the hidden patterns in documents from many fields, including marketing, sociology, political science, and the digital humanities (Boyd-Graber et al., 2017). One of the most common topic models is latent Dirichlet allocation (LDA) (Blei et al., 2003), a probabilistic model that represents each topic as a distribution over words and each document as a mixture of the topics. LDA has been extended in different ways, for example to capture correlations among the topics (Lafferty and Blei, 2005), to classify documents (Blei and McAuliffe, 2007), or to analyze documents in different languages (Mimno et al., 2009).

In this paper, we focus on analyzing the temporal evolution of topics in large document collections. That is, given a corpus that was collected over a large number of years, our goal is to use topic modeling to find how the latent patterns of the documents change over time.

Dynamic latent Dirichlet allocation (D-LDA) (Blei and Lafferty, 2006) shares the same goal. D-LDA is an extension of LDA that uses a probabilistic time series to allow the topics to vary smoothly over time.[3] However, D-LDA suffers from the same limitations as LDA. In particular, it does not capture the distribution of rare words and the long tail of language data (Dieng et al., 2019).

Relative to classical LDA, the embedded topic model (ETM) solves these problems (Dieng et al., 2019). It improves LDA in terms of predictive performance and topic quality by using continuous representations of words (Bengio et al., 2006; Mikolov et al., 2013b). The ETM defines each topic

[*]equal contributions.

[2]Code for this work can be found at https://github.com/adjidieng/DETM

[3]Blei and Lafferty (2006) called it *dynamic topic model*, but throughout the paper we refer to it as D-LDA because it is a dynamic extension of LDA, despite the fact that it does not use Dirichlet distributions.

Preprint. Under review.

as a vector on the word embedding space. The ETM then uses the dot product between the word and topic embeddings to define the likelihood of a word in a given topic. While the ETM better fits the distribution of words, it cannot analyze a corpus whose topics shift over time.

To address this limitation we develop the dynamic embedded topic model (D-ETM), a model that extends D-LDA and the ETM. Similarly to D-LDA, the D-ETM involves a probabilistic time series to allow the topics to vary smoothly over time. However, each topic in the D-ETM is a time-varying vector on the word embedding space. Similarly to the ETM, the likelihood of a word under the D-ETM is given by a categorical distribution whose natural parameter depends on the inner product between each term's embedding and its assigned topic's embedding. In contrast to the ETM, the topic embeddings of the D-ETM vary over time.

As for most probabilistic models, the posterior of the D-ETM is intractable to compute, and we need to approximate it. We use variational inference (Jordan et al., 1999; Blei et al., 2017). To scale up the algorithm to large datasets, we use data subsampling (Hoffman et al., 2013) and amortization (Gershman and Goodman, 2014); these techniques reduce the number of variational parameters and speed up the learning procedure. Additionally, we use a structured variational approximation parameterized by a long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997).

We use the D-ETM to analyze the transcriptions of the United Nations (UN) general debates from 1970 to 2015 (Baturo et al., 2017). The D-ETM reveals the topics discussed in the political debates and their trajectories, which are aligned with historical events. In addition, we quantitatively assess the performance of the D-ETM in terms of predictive performance and topic quality. We found that the D-ETM provides better predictions and topic quality than D-LDA.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 reviews LDA, D-LDA, and the ETM. Section 4 presents the D-ETM and the corresponding inference algorithm. Finally, Section 5 contains the empirical study and Section 6 concludes the paper.

2 Related Work

The D-ETM is a topic model that incorporates continuous representations of words (Rumelhart and Abrahamson, 1973; Bengio et al., 2003, 2006; Mikolov et al., 2013a,b; Pennington et al., 2014; Levy and Goldberg, 2014). There exist other methods that achieve a similar goal by modifying the prior distributions (Petterson et al., 2010; Xie et al., 2015; Shi et al., 2017; Zhao et al., 2017a,b). These methods use word embeddings as a type of "side information;" in contrast the D-ETM directly uses the embeddings in its generative process. There are also methods that combine LDA with word embeddings by first converting the discrete text into continuous observations of embeddings (Das et al., 2015; Xun et al., 2016; Batmanghelich et al., 2016; Xun et al., 2017). These works adapt LDA for real-valued observations; for example using a Gaussian likelihood. In contrast, the D-ETM is a probabilistic model of discrete data. Other ways of combining LDA and word embeddings modify the likelihood (Nguyen et al., 2015), randomly replace words drawn from a topic with the embeddings drawn from a Gaussian (Bunk and Krestel, 2018), or use Wasserstein distances to learn topics and embeddings jointly (Xu et al., 2018).

Another line of research improves topic modeling inference through deep neural networks; these are called neural topic models (Miao et al., 2016; Srivastava and Sutton, 2017; Card et al., 2017; Cong et al., 2017; Zhang et al., 2018). Most of these works are based on the variational autoencoder (Kingma and Welling, 2014) and use amortized inference (Gershman and Goodman, 2014). Finally, the ETM (Dieng et al., 2019) is a probabilistic topic model that also makes use of word embeddings and uses amortization in its inference procedure.

Other works find embedding representations of the words that vary over time (Bamler and Mandt, 2017; Rudolph and Blei, 2018). Despite incorporating a time-varying component, these works have a different goal than the D-ETM. Rather than modeling the temporal evolution of documents, they model how the meaning of words shifts over time.

None of the methods above can model collections of documents whose topics evolve over time, which is the goal of dynamic topic models. The first and most common dynamic topic model is D-LDA (Blei and Lafferty, 2006). Bhadury et al. (2016) scale up the inference method of D-LDA using a sampling procedure. Other extensions of D-LDA use stochastic processes to introduce stronger correlations in the topic dynamics (Wang and McCallum, 2006; Wang et al., 2008; Jähnichen et al.,

2018). The D-ETM is also an extension of D-LDA, but it has a different goal than previous extensions. The D-ETM better fits the distribution of words via the use of distributed representations for both the words and the topics.

3 Background

Here we review the models on which we build the D-ETM. We start by reviewing LDA and the ETM; both are non-dynamic topic models. We then review D-LDA, the dynamic extension of LDA.

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 considers K topics $\beta_{1:K}$, each of which is a distribution over the vocabulary. It further considers a vector of topic proportions θ_d for each document d in the collection; each element θ_{dk} expresses how prevalent the k^{th} topic is in that document. In the generative process of LDA, each word is assigned to topic k with probability θ_{dk}, and the word is then drawn from the distribution β_k. The generative process for each document is as follows:

1. Draw topic proportions $\theta_d \sim \text{Dirichlet}(\alpha_\theta)$.
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 also places a Dirichlet prior on the topics, $\beta_k \sim \text{Dirichlet}(\alpha_\beta)$. The concentration parameters α_β and α_θ of the Dirichlet distributions are model hyperparameters.

Embedded topic model. The ETM uses vector representations of words (Rumelhart and Abrahamson, 1973; Bengio et al., 2003) to improve the performance of LDA in terms of topic quality and predictive accuracy, specially in the presence of large vocabularies (Dieng et al., 2019). Let ρ be an $L \times V$ matrix containing L-dimensional embeddings of the words in the vocabulary, such that each column $\rho_v \in \mathbb{R}^L$ corresponds to the embedding representation of the v^{th} term. The ETM uses the embedding matrix ρ to define each topic β_k; in particular it sets

$$\beta_k = \text{softmax}(\rho^\top \alpha_k). \tag{1}$$

Here, $\alpha_k \in \mathbb{R}^L$ is an embedding representation of the k^{th} topic, called *topic embedding*. The topic embedding is a distributed representation of the topic in the semantic space of words. The ETM uses the topic embeddings in its generative process, which is analogous to LDA:

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 word $w_{dn} \sim \text{Cat}(\text{softmax}(\rho^\top \alpha_{z_{dn}}))$.

The notation $\mathcal{LN}$ in Step 1 refers to the logistic-normal distribution (Aitchison and Shen, 1980), which transforms Gaussian random variables to the simplex.

In using the word representations $\rho_{1:V}$ in the definition of $\beta_{1:K}$, the ETM learns the topics of a corpus in a particular embedding space. The intuition behind the ETM is that semantically related words will be assigned to similar topics—since their embedding representations are close, they will interact similarly with the topic embeddings $\alpha_{1:K}$.

Dynamic latent Dirichlet allocation. D-LDA allows the topics to vary over time to analyze a corpus that spans a large period of time (Blei and Lafferty, 2006). The generative model of D-LDA differs from LDA in that the topics are time-specific, i.e., they are $\beta^{(t)}_{1:K}$, where $t \in \{1, \dots, T\}$ indexes time steps. Moreover, the prior over the topic proportions θ_d depends on the time stamp of document d, denoted $t_d \in \{1, \dots, T\}$. The generative process for each document is:

1. Draw topic proportions $\theta_d \sim \mathcal{LN}(\eta_{t_d}, a^2 I)$.
2. For each word n in the document:
   (a) Draw topic assignment $z_{dn} \sim \text{Cat}(\theta_d)$.
   (b) Draw word $w_{dn} \sim \text{Cat}(\beta^{(t_d)}_{z_{dn}})$.

Here, a is a model hyperparameter and η_t is a latent variable that controls the prior mean over the topic proportions at time t. To encourage smoothness over the topics and topic proportions, D-LDA places random walk priors over $\beta_{1:K}^{(t)}$ and η_t,

$$\beta_k^{(t)} = \text{softmax}(\widetilde{\beta}_k^{(t)}), \;\; p(\widetilde{\beta}_k^{(t)} \mid \widetilde{\beta}_k^{(t-1)}) = \mathcal{N}(\widetilde{\beta}_k^{(t-1)}, \sigma^2 I), \text{ and } p(\eta_t \mid \eta_{t-1}) = \mathcal{N}(\eta_{t-1}, \delta^2 I). \quad (2)$$

The variables $\widetilde{\beta}_k^{(t)} \in \mathbb{R}^V$ are the transformed topics; the topics $\beta_k^{(t)}$ are obtained after mapping $\widetilde{\beta}_k^{(t)}$ to the simplex. The hyperparameters σ and δ control the smoothness of the Markov chains.

4 The Dynamic Embedded Topic Model

Here we develop the D-ETM, a model that combines the advantages of D-LDA and the ETM. Like D-LDA, it allows the topics to vary smoothly over time to accommodate datasets that span a large period of time. Like the ETM, the D-ETM uses word embeddings, making it generalize better and improving the predictive performance and the topics of D-LDA. We describe the model in Section 4.1 and then we develop an efficient inference algorithm in Section 4.2.

4.1 Model Description

The D-ETM is a dynamic topic model that uses embedding representations of words and topics. For each term v, it considers an L-dimensional embedding representation ρ_v. The D-ETM posits an embedding $\alpha_k^{(t)} \in \mathbb{R}^L$ for each topic k at a given time stamp $t = 1, \ldots, T$. That is, the D-ETM represents each topic as a time-varying real-valued vector, unlike traditional topic models (where topics are distributions over the vocabulary). We refer to $\alpha_k^{(t)}$ as *topic embedding* (Dieng et al., 2019); it is a distributed representation of the k^{th} topic in the semantic space of words.

The D-ETM forms distributions over the vocabulary using the word and topic embeddings. Specifically, under the D-ETM, the probability of a word under a topic is given by the (normalized) exponentiated inner product between the embedding representation of the word and the topic's embedding at the corresponding time step,

$$p(w_{dn} = v \mid z_{dn} = k, \alpha_k^{(t_d)}) \propto \exp\{\rho_v^\top \alpha_k^{(t_d)}\}. \quad (3)$$

Due to the inner product in Eq. 3, the probability of a particular term is higher when the term's embedding and the topic's embeddings are in agreement. Therefore, semantically similar words will be assigned to similar topics, since their representations are close in the embedding space.

The D-ETM enforces smooth variations of the topics by using a Markov chain over the topic embeddings $\alpha_k^{(t)}$, in which the topic representations evolve under Gaussian noise with variance γ^2,

$$p(\alpha_k^{(t)} \mid \alpha_k^{(t-1)}) = \mathcal{N}(\alpha_k^{(t-1)}, \gamma^2 I). \quad (4)$$

Similarly to D-LDA, the D-ETM considers time-varying priors over the topic proportions θ_d. This allows the model to capture how the general topic usage evolves over time (in addition to the evolution of the topic themselves). In particular, the prior over θ_d depends on a latent variable η_{t_d} (recall that t_d is the time stamp of document d),

$$p(\theta_d \mid \eta_{t_d}) = \mathcal{LN}(\eta_{t_d}, a^2 I), \qquad \text{where} \qquad p(\eta_t \mid \eta_{t-1}) = \mathcal{N}(\eta_{t-1}, \delta^2 I). \quad (5)$$

The D-ETM's graphical model is depicted in Figure 1. The generative process is as follows:

1. Draw topic embeddings[4] $\alpha_k^{(t)} \sim \mathcal{N}(\alpha_k^{(t-1)}, \gamma^2 I)$ for $k = 1, \ldots, K$ and $t = 1, \ldots, T$.
2. Draw the latent means $\eta_t \sim \mathcal{N}(\eta_{t-1}, \delta^2 I)$ for $t = 1, \ldots, T$.
3. For each document d:
   (a) Draw topic proportions $\theta_d \sim \mathcal{LN}(\eta_{t_d}, a^2 I)$.
   (b) For each word n in the document:
      i. Draw topic assignment $z_{dn} \sim \text{Cat}(\theta_d)$.
      ii. Draw word $w_{dn} \sim \text{Cat}(\text{softmax}(\rho^\top \alpha_{z_{dn}}^{(t_d)}))$.

[4]For the first time step, we consider a standard Gaussian prior $\mathcal{N}(0, I)$ over $\alpha_k^{(1)}$ and η_1.

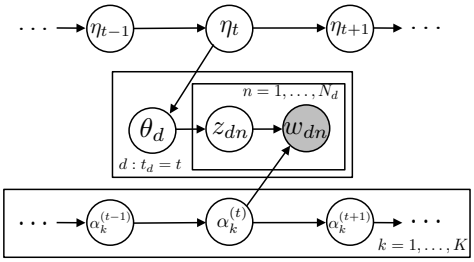

Figure 1. Graphical representation of D-ETM. The topic embeddings $\alpha_k^{(t)}$ and the latent means η_t evolve over time. For each document at a particular time step t, the prior over the topic proportions θ_d depends on η_t. The variables z_{dn} and w_{dn} denote topic assignment and observed words, respectively.

Step 1 gives the prior over the topic embeddings; it encourages smoothness on the resulting topics. Steps 2 is shared with D-LDA; it describes the evolution of the prior mean over the topic proportions. Steps 3a and 3b-i are standard for topic modeling; they represent documents as distributions over topics and draw a topic assignment for each word. Step 3b-ii is different—it uses the $L \times V$ word embedding matrix ρ and the assigned topic embedding $\alpha_{z_{dn}}^{(t_d)}$ at time instant t_d to form a categorical distribution over the vocabulary.

Since the D-ETM uses embedding representations of the words, it learns the topics in a particular embedding space. This is particularly useful when an embedding of a word that is not used in the corpus is available. Indeed, consider a term $v^\star$ that was not seen in the corpus. The D-ETM can assign it to topics by computing the inner product $\rho_{v^\star}^\top \alpha_k^{(t)}$, thus leveraging the semantic information of the word's embedding.

4.2 Inference Algorithm

Given a dataset $\mathcal{D}$ of documents $\{\mathbf{w}_1, \ldots, \mathbf{w}_D\}$ and their time stamps $\{t_1, \ldots, t_D\}$, fitting a D-ETM involves finding the posterior distribution over the model's latent variables, $p(\theta, \eta, \alpha \mid \mathcal{D})$, where we have marginalized out the topic assignments z from Eq. 3 for convenience,[5]

$$p(w_{dn} \mid \alpha_k^{(t_d)}) = \sum_{k=1}^{K} p(w_{dn} \mid z_{dn} = k, \alpha_k^{(t_d)}). \tag{6}$$

However, the posterior distribution is intractable, and therefore we approximate it using variational inference (Jordan et al., 1999; Blei et al., 2017). Variational inference approximates the posterior using a family of distributions $q_\nu(\theta, \eta, \alpha)$. The parameters ν that index this family are called variational parameters; they are optimized to minimize the Kullback-Leibler (KL) divergence between the approximation and the posterior. Solving this optimization problem is equivalent to maximizing the evidence lower bound (ELBO),

$$\mathcal{L}(\nu) = \mathbb{E}_q \left[\log p(\mathcal{D}, \theta, \eta, \alpha) - \log q_\nu(\theta, \eta, \alpha)\right]. \tag{7}$$

To reduce the number of variational parameters and speed-up the inference algorithm, we use an amortized variational distribution, i.e., we let the parameters of the approximating distributions be functions of the data (Gershman and Goodman, 2014; Kingma and Welling, 2014). Additionally, we use a structured variational family to preserve some of the conditional dependencies of the graphical model (Saul and Jordan, 1996). The specific variational family in the D-ETM takes the form

$$q(\theta, \eta, \alpha) = \prod_d q(\theta_d \mid \eta_{t_d}, \mathbf{w}_d) \times \prod_t q(\eta_t \mid \eta_{1:t-1}, \widetilde{\mathbf{w}}_t) \times \prod_k \prod_t q(\alpha_k^{(t)}), \tag{8}$$

where we have removed the dependency on the variational parameters to avoid clutter.

The distribution over the topic proportions $q(\theta_d \mid \eta_{t_d}, \mathbf{w}_d)$ is a logistic-normal whose mean and covariance parameters are functions of both the latent mean η_{t_d} and the bag-of-words representation

[5]The marginalization of z_{dn} reduces the number of variational parameters to fit and avoids discrete latent variables in the inference procedure, which is useful to form reparameterization gradients.

Algorithm 1: Dynamic topic modeling with the D-ETM

input : Documents $\{\mathbf{w}_1, \ldots, \mathbf{w}_D\}$ and their time stamps $\{t_1, \ldots, t_D\}$
Initialize all variational parameters
for iteration $1, 2, 3, \ldots$ **do**
  Sample the latent means and the topic embeddings, $\eta \sim q(\eta \mid \widetilde{\mathbf{w}})$ and $\alpha \sim q(\alpha)$
  Compute the topics $\beta_k^{(t)} = \text{softmax}(\rho^\top \alpha_k^{(t)})$ for $k = 1, \ldots, K$ and $t = 1, \ldots, T$
  Obtain a minibatch of documents
  for each document d in the minibatch **do**
    Sample the topic proportions $\theta_d \sim q(\theta_d \mid \eta_{t_d}, \mathbf{w}_d)$
    for each word n in the document **do**
      Compute $p(w_{dn} \mid \theta_d) = \sum_k \theta_{dk} \beta_{k,w_{dn}}^{(t_d)}$
    end
  end
  Estimate the ELBO in Eq. 7 and its gradient w.r.t. the variational parameters (backpropagation)
  Update the model and variational parameters using Adam
end

of document d. In particular, these functions are parameterized by feed-forward neural networks that input both η_{t_d} and the normalized bag-of-words representation. The distribution over the latent means $q(\eta_t \mid \eta_{1:t-1}, \widetilde{\mathbf{w}}_t)$ depends on all previous latent means $\eta_{1:t-1}$. We use a LSTM to capture this temporal dependency. We choose a Gaussian distribution $q(\eta_t \mid \eta_{1:t-1}, \widetilde{\mathbf{w}}_t)$ whose mean and covariance are given by the output of the LSTM. The input to the LSTM at time t is the average of the bag-of-words representation of all documents whose time stamp is t. Here, $\widetilde{\mathbf{w}}_t$ denotes the normalized bag-of-words representation of all such documents. Finally, we use the mean-field family for the approximation over the topic embeddings, $q(\alpha_k^{(t)})$.

We optimize the ELBO with respect to the variational parameters. Since the expectations in Eq. 7 are intractable, we use stochastic optimization, obtaining unbiased gradient estimators of the ELBO. In particular, we form reparameterization gradients (Kingma and Welling, 2014; Titsias and Lázaro-Gredilla, 2014; Rezende et al., 2014). To speed up the algorithm, we take a minibatch of documents at each iteration; this allows to handle large collections of documents (Hoffman et al., 2013). We set the learning rate with Adam (Kingma and Ba, 2015). Algorithm 1 summarizes the procedure.

5 Empirical Study

In this section, we use the D-ETM to analyze the transcriptions of the UN general debates from 1970 to 2015. The D-ETM reveals the temporal evolution of the topics discussed in the debates (such as climate change, war, poverty, or human rights). Besides the qualitative analysis, we quantitatively compare the D-ETM against D-LDA, obtaining that the D-ETM provides better predictive power and higher topic quality.

Dataset. The UN debates corpus spans a period of $T = 46$ years (Baturo et al., 2017). Each year, leaders and other senior officials deliver statements that present their government's perspective on the major issues in world politics. The corpus contains the transcriptions of each country's statement at the UN General Assembly.

The dataset is available online.[6] We apply standard preprocessing techniques, such as tokenization and removal of numbers and punctuation marks. In this dataset, the speeches are lengthy and each one discusses a multitude of topics; therefore a topic model trained on entire speeches as documents would not be able to distinguish those topics. Instead, we follow Lefebure (2018) and split the speeches into paragraphs, treating each paragraph as a separate document for the analysis. We additionally filter stop words, i.e., words with document frequency above 70%, as well as standard stop words from a list. We use 85% randomly chosen documents for training, 10% for testing, and

[6] See https://www.kaggle.com/unitednations/un-general-debates.

Table 1. The D-ETM outperforms D-LDA on two different vocabulary settings. Here, PPL denotes perplexity on a document completion task (lower is better), TC is topic coherence (higher is better), TD is topic diversity (higher is better), and TQ is a measure of overall topic quality (higher is better). The D-ETM performs better on both datasets and on all metrics except for TC.

	Smaller Vocabulary				Larger Vocabulary			
Method	PPL	TC	TD	TQ	PPL	TC	TD	TQ
D-LDA (Blei and Lafferty, 2006)	2053	**0.117**	0.266	0.031	2561	**0.115**	0.270	0.031
D-ETM (this paper)	**1529**	0.100	**0.679**	**0.068**	**2099**	0.083	**0.728**	**0.061**

5% for validation. We remove one-word documents from the validation and test sets. After all the preprocessing steps, the number of training documents is 196K.

We build two versions of the dataset, each with a different vocabulary size. To form each version, we keep only the words that appear in more than 100 or 30 documents, leading to vocabulary sizes of $V = 7{,}307$ and $V = 12{,}466$, respectively.

Methods. Besides the D-ETM, we fit D-LDA, as it is the dynamic topic model on which the D-ETM is built. Other dynamic topic models use stochastic processes to model the evolution of topics; we see these models as complementary to the D-ETM. In fact, the D-ETM can be built on continuous time dynamic topic models as well. For a fair comparison, we choose D-LDA as the baseline.

Settings. We fit a 50-topic dynamic topic model, and we follow Blei and Lafferty (2006) to set the hypeparameters as $\delta^2 = \sigma^2 = \gamma^2 = 0.005$ and $a^2 = 1$.

For the D-ETM, we first fit 300-dimensional word embeddings using skip-gram (Mikolov et al., 2013b). We apply the algorithm in Section 4.2 using a batch size of 1,000 documents for the smaller vocabulary dataset and 500 for the larger vocabulary. We use a fully connected feed-forward inference network for the topic proportions θ_d. The network has tanh activations and 2 layers of 400 hidden units each. We set the mean and log-variance for θ_d as linear maps of the output. For the latent means $\eta_{1:T}$, each bag-of-word representation $\widetilde{\mathbf{w}}_t$ is first linearly mapped to a low-dimensional space of dimensionality 200. This conforms the input of an LSTM that has 2 layers of 200 hidden units each. The LSTM output is then concatenated with the previous latent mean η_{t-1}, and the result is linearly mapped to a K-dimensional space to get the mean and log-variance for η_t. We apply a small weight decay of $1.2 \cdot 10^{-6}$ on all network parameters. We run Algorithm 1 for 100 epochs. The stopping criterion is based on the held-out log-likelihood on the validation set. The learning rate is 0.01 for the smaller vocabulary and 0.001 for the larger vocabulary. We anneal this learning rate by dividing by 4 if there is no improvement for 10 consecutive epochs.

For D-LDA, Blei and Lafferty (2006) derive a bound of the ELBO to enable a coordinate-ascent inference algorithm that also uses Kalman filtering and smoothing as a subroutine. Besides loosening the variational bound on the log-evidence, this algorithm presents scalability issues. Thus, we leverage recent advances in variational inference to overcome these issues. We use stochastic optimization based on reparameterization gradients and we obtain batches of 1,000 documents at each iteration. We collapse the discrete latent topic indicators z_{dn} to enable the reparameterization gradients, and we use a fully factorized Gaussian approximation for the rest of the latent variables, except for $\eta_{1:T}$, for which we use a full-covariance Gaussian for each of its dimensions. We initialize D-LDA using LDA. In particular, we run 5 epochs of LDA followed by 120 epochs of D-LDA. For D-LDA, we use RMSProp (Tieleman and Hinton, 2012) to set the step size, setting the learning rate to 0.05 for the mean parameters and to 0.005 for the variance parameters.

Quantitative results. We compare the D-ETM and D-LDA according to two metrics: perplexity on a document completion task and topic quality. The perplexity is obtained by computing the probability of each word in the second half of a test document, conditioned on the first half (Rosen-Zvi et al., 2004; Wallach et al., 2009). To obtain the topic quality, we combine two metrics. The first metric is topic coherence; it provides a quantitative measure of the interpretability of a topic (Mimno et al., 2011). We obtain the coherence by taking the average pointwise mutual information of two words drawn randomly from the same document (Lau et al., 2014); this requires to approximate word probabilities with empirical counts. The second metric is topic diversity; it is the percentage of unique words in the top 25 words of all topics (Dieng et al., 2019). Diversity close to 0 indicates redundant topics. We obtain both topic coherence and topic diversity by averaging over time. Finally, topic quality is defined as the product between topic coherence and diversity (Dieng et al., 2019).

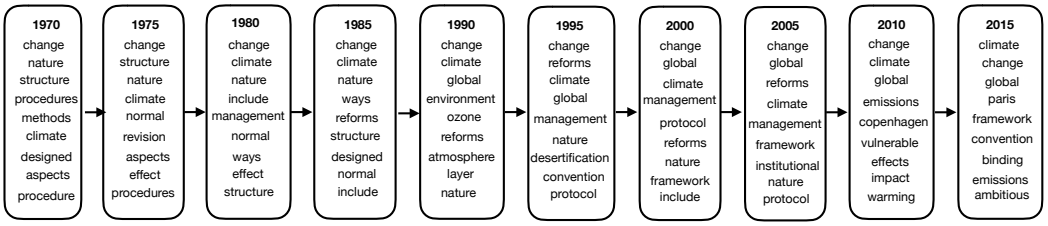

Figure 2. Temporal evolution of the top-10 words from a topic about climate change learned by the D-ETM. This topic is in agreement with historical events. In the 1990s the destruction of the ozone layer was of major concern. More recently the concern is about global warming. Events such as the Kyoto protocol and the Paris convention are also reflected in this topic's evolution.

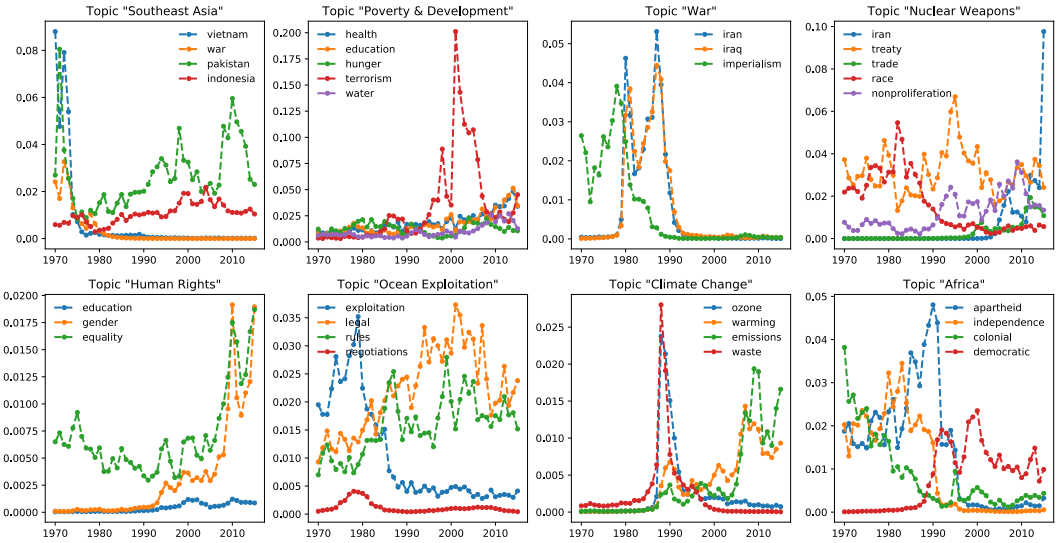

Figure 3. Evolution of word probability across time for eight different topics learned by the D-ETM. For each topic, we choose a set of words whose probability shift aligns with historical events (these are not the words with the highest probability in each topic). For example, one interesting finding is the increased relevance of the words "gender" and "equality" in a topic about human rights.

Table 1 shows that the D-ETM outperforms D-LDA according to both metrics—perplexity and topic quality. The topic coherence of D-LDA is slightly better, but the D-ETM finds more varied topics, as evidenced by a significantly higher topic diversity score. However, note that the topic coherence numbers may be rough approximations, as each individual time step might not contain enough documents to accurately approximate the pointwise mutual information.

Qualitative results. The D-ETM finds that the topics' evolution over time are in agreement with historical events. As an example, Figure 2 shows a topic on climate change. In the 1990s, protecting the ozone layer was the primary concern; more recently the topic has shifted towards global warming and reducing the greenhouse gas emissions. Some events on climate change, such as the Kyoto protocol (1997) or the Paris convention (2016), are also reflected in the topic's evolution.

We now examine the evolution of the probability of individual words. Figure 3 shows these probabilities for a variety of words and topics. For example, the probability of the word "Vietnam" in a topic on Southeast Asia decays after the end of the war in 1975. In a topic about nuclear weapons, the concern about the arms "race" between the USA and the Soviet Union eventually decays, and "Iran" becomes more relevant in recent years. Similarly, words like "equality" and "gender" become more important in recent years within a topic about human rights. Note that the names of the topics are subjective; we assigned the names inspired by the top words in each topic (the words in Figure 3 are not necessarily the most likely words within each topic). One example is the topic on climate change, whose top words are shown in Figure 2. Another example is the topic on human rights, which exhibits the words "human" and "rights" consistently at the top across all time steps.

6 Conclusion

We developed the D-ETM, a probabilistic model of documents that uses semantic representation of words to define the topics. Under the D-ETM, each topic is a time-varying vector in the embedding space of words. Using a probabilistic time series, the D-ETM uncovers topics that vary smoothly over time. We use the D-ETM to analyze a dataset of UN debates and find that the D-ETM has better held-out perplexity and provides higher topic quality than D-LDA.

Acknowledgments

This work is funded by ONR N00014-17-1-2131, NIH 1U01MH115727-01, DARPA SD2 FA8750-18-C-0130, ONR N00014-15-1-2209, NSF CCF-1740833, the Alfred P. Sloan Foundation, 2Sigma, Amazon, and NVIDIA. FJRR is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 706760. ABD is supported by a Google PhD Fellowship.

References

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

Bamler, R. and Mandt, S. (2017). Dynamic word embeddings. In *International Conference on Machine Learning*.

Batmanghelich, K., Saeedi, A., Narasimhan, K., and Gershman, S. (2016). Nonparametric spherical topic modeling with word embeddings. In *Association for Computational Linguistics*.

Baturo, A., Dasandi, N., and Mikhaylov, S. (2017). Understanding state preferences with text as data: Introducing the UN general debate corpus. *Research & Politics*, 4:1–9.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F., and Gauvain, J.-L. (2006). Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer.

Bhadury, A., Chen, J., Zhu, J., and Liu, S. (2016). Scaling up dynamic topic models. In *International World Wide Web Conference*.

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

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

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *International Conference on Machine Learning*.

Blei, D. M. and McAuliffe, J. D. (2007). Supervised topic models. In *Advances in Neural Information Processing Systems*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

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

Bunk, S. and Krestel, R. (2018). Welda: Enhancing topic models by incorporating local word context. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages 293–302. ACM.

Card, D., Tan, C., and Smith, N. A. (2017). A neural framework for generalized topic models. In *arXiv:1705.09296*.

Cong, Y., Chen, B., Liu, H., and Zhou, M. (2017). Deep latent Dirichlet allocation with topic-layer-adaptive stochastic gradient Riemannian MCMC. In *International Conference on Machine Learning*.

Das, R., Zaheer, M., and Dyer, C. (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)*.

Dieng, A. B., Ruiz, F. J. R., and Blei, D. M. (2019). Topic modeling in embedding spaces. *arXiv preprint arXiv:1907.04907*.

Gershman, S. J. and Goodman, N. D. (2014). Amortized inference in probabilistic reasoning. In *Annual Meeting of the Cognitive Science Society*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347.

Jähnichen, P., Wenzel, F., Kloft, M., and Mandt, S. (2018). Scalable generalized dynamic topic models. In *Artificial Intelligence and Statistics*.

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

Kingma, D. P. and Ba, J. L. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Lafferty, J. D. and Blei, D. M. (2005). Correlated topic models. In *Advances in Neural Information Processing Systems*.

Lau, J. H., Newman, D., and Baldwin, T. (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*.

Lefebure, L. (2018). Exploring the UN General Debates with dynamic topic models. Available online at https://towardsdatascience.com/exploring-the-un-general-debates-with-dynamic-topic-models-72dc0e307696.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Neural Information Processing Systems*, pages 2177–2185.

Miao, Y., Yu, L., and Blunsom, P. (2016). Neural variational inference for text processing. In *International Conference on Machine Learning*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *ICLR Workshop Proceedings. arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems*.

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). Polylingual topic models. In *Conference on Empirical Methods in Natural Language Processing*.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Conference on Empirical Methods in Natural Language Processing*.

Nguyen, D. Q., Billingsley, R., Du, L., and Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Conference on Empirical Methods on Natural Language Processing*.

Petterson, J., Buntine, W., Narayanamurthy, S. M., Caetano, T. S., and Smola, A. J. (2010). Word features for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 1921–1929.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Uncertainty in Artificial Intelligence*.

Rudolph, M. and Blei, D. M. (2018). Dynamic embeddings for language evolution. In *International World Wide Web Conference*.

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

Saul, L. K. and Jordan, M. I. (1996). Exploiting tractable substructures in intractable networks. In *Advances in Neural Information Processing Systems*.

Shi, B., Lam, W., Jameel, S., Schockaert, S., and Lai, K. P. (2017). Jointly learning word embeddings and latent topics. In *ACM SIGIR Conference on Research and Development in Information Retrieval*.

Srivastava, A. and Sutton, C. (2017). Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5-RMSPROP: Divide the gradient by a running average of its recent magnitude. Coursera: Neural Networks for Machine Learning, 4.

Titsias, M. K. and Lázaro-Gredilla, M. (2014). Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *International Conference on Machine Learning*.

Wang, C., Blei, D. M., and Heckerman, D. (2008). Continuous time dynamic topic models. In *Uncertainty in Artificial Intelligence*.

Wang, X. and McCallum, A. (2006). Topics over time: a non-Markov continuous-time model of topical trends. In *ACM SIGKDD*.

Xie, P., Yang, D., and Xing, E. (2015). Incorporating word correlation knowledge into topic modeling. In *Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies*.

Xu, H., Wang, W., Liu, W., and Carin, L. (2018). Distilled Wasserstein learning for word embedding and topic modeling. In *Advances in Neural Information Processing Systems*.

Xun, G., Gopalakrishnan, V., Ma, F., Li, Y., Gao, J., and Zhang, A. (2016). Topic discovery for short texts using word embeddings. In *International Conference on Data Mining*.

Xun, G., Li, Y., Zhao, W. X., Gao, J., and Zhang, A. (2017). A correlated topic model using word embeddings. In *IJCAI*, pages 4207–4213.

Zhang, H., Chen, B., Guo, D., and Zhou, M. (2018). WHAI: Weibull hybrid autoencoding inference for deep topic modeling. In *International Conference on Learning Representations*.

Zhao, H., Du, L., and Buntine, W. (2017a). A word embeddings informed focused topic model. In *Asian Conference on Machine Learning*.

Zhao, H., Du, L., Buntine, W., and Liu, G. (2017b). MetaLDA: A topic model that efficiently incorporates meta information. In *International Conference on Data Mining*.