The Dynamic Embedded Topic Model

Adji B. Dieng[1, *], Francisco J. R. Ruiz[2, 3, *], and
David M. Blei[1, 2]

[1]Department of Statistics, Columbia University
[2]Department of Computer Science, Columbia University
[3]Department of Engineering, University of Cambridge
[*]Equal Contributions

October 14, 2019

Abstract

Topic modeling analyzes documents to learn meaningful patterns of words. For documents collected in sequence, dynamic topic models capture how these patterns vary over time. 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 parameterized by the inner product between the word embedding and a per-time-step embedding representation of its assigned topic. The D-ETM learns smooth topic trajectories by defining a random walk prior over the embedding representations of the topics. We fit the D-ETM using structured amortized variational inference with a recurrent neural network. On three different corpora—a collection of United Nations debates, a set of ACL abstracts, and a dataset of Science Magazine articles—we found that the D-ETM outperforms D-LDA on a document completion task. We further found that the D-ETM learns more diverse and coherent topics than D-LDA while requiring significantly less time to fit.[1]

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 documents from many fields, including marketing, sociology, political science, and the digital humanities; see Boyd-Graber et al. (2017) for a review. 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).

[1]**Code:** The code for this paper can be found at https://github.com/adjidieng/DETM

In this paper, we focus on analyzing the temporal evolution of topics in large document collections. 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.[2] 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).

The embedded topic model (ETM) aims to solve these problems (Dieng et al., 2019). It uses continuous representations of words (Bengio et al., 2006; Mikolov et al., 2013b) to improve LDA in terms of predictive performance and topic quality. The ETM defines each topic as a vector on the word embedding space; it then uses the dot product between each word and the topic embedding to define the per-topic distribution over words. However, while the ETM better fits large document collections, it cannot analyze a corpus whose topics shift over time.

In this paper 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. As in the ETM, the probability of each word under the D-ETM is a categorical distribution whose natural parameter depends on the inner product between the word's embedding and a per-topic embedding representation of its assigned topic. In contrast to the ETM, the topic embeddings of the D-ETM vary over time.

Given a time-series corpus of documents, we are interested in the posterior distribution of the topic proportions and the per-time-point topic embeddings. As for most interesting probabilistic models, the posterior distribution is intractable to compute; 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 speed up the learning procedure and reduce the number of variational parameters. 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). Qualitatively, the D-ETM reveals the topics discussed in the political debates and their trajectories, which are aligned with historical events. For example Figure 2 in Section 5 shows a topic about climate change found by the D-ETM that transitions from being mainly about the ozone layer in the 1990s to global warming and emissions in 2015.

We also used the D-ETM to analyze a dataset of articles from Science Magazine (1990-1999) and a corpus of ACL abstracts (1973-2006). We quantitatively assess

[2] Blei and Lafferty (2006) called it a *dynamic topic model*, but we refer to it as D-LDA because it is motivated as a dynamic extension of LDA.

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 in general.

To validate that the gains in performance of the D-ETM is due to the model and not to the inference procedure used to fit it, we compare to a baseline that applies the same inference procedure as the D-ETM to D-LDA. We call this baseline D-LDA-REP. On all three corpora, we found the D-ETM and D-LDA both outperform D-LDA-REP and that the only advantage of D-LDA-REP over D-LDA is that it is significantly faster to fit.

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 inference algorithm used to fit it. Finally, Section 5 details the empirical study and Section 6 concludes the paper.

2 Related Work

The D-ETM builds on word embeddings, topic models, and dynamic topic models.

Word embeddings are low-dimensional continuous representations of words that capture their semantics (Rumelhart and Abrahamson, 1973; Bengio et al., 2003, 2006; Mikolov et al., 2013a,b; Pennington et al., 2014; Levy and Goldberg, 2014). Some recent work finds embedding representations 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. (In future research, the D-ETM developed here could be used in concert with these methods.)

There has been a surge of methods that combine word embeddings and probabilistic topic models. Some methods modify the prior distributions over topics in LDA (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." 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. Still 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). In contrast to all these methods, the D-ETM uses sequential priors and is a probabilistic model of discrete data that directly models the words.

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.

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 developed for a different purpose. 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 topics to vary over time to analyze time-series corpora (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 $\beta^{(t)}_{1:K}$ and η_t,

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

The variables $\widetilde{\beta}^{(t)}_k \in \mathbb{R}^V$ are the transformed topics; the topics $\beta^{(t)}_k$ are obtained after mapping $\widetilde{\beta}^{(t)}_k$ 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, allowing it to generalize better than D-LDA and improving its topics. We describe the model in Section 4.1 and then we develop an efficient structured variational 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,

$$p(w_{dn} = v \,|\, z_{dn} = k, \alpha_k^{(t_d)}) \propto \exp\{\rho_v^\top \alpha_k^{(t_d)}\}. \tag{2}$$

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 $\alpha_k^{(t)}$. The topic representations evolve under Gaussian noise with variance γ^2,

$$p(\alpha_k^{(t)} \,|\, \alpha_k^{(t-1)}) = \mathcal{N}(\alpha_k^{(t-1)}, \gamma^2 I). \tag{3}$$

Similarly to D-LDA, the D-ETM considers time-varying priors over the topic proportions θ_d. In addition to time-varying topics, this construction allows the model to capture how the general topic usage evolves over time. The prior over θ_d depends on a latent variable η_{t_d}, where recall that t_d is the time stamp of document d,

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

Figure 1 depicts the graphical model for the D-ETM. The generative process is as follows:

1. Draw initial topic embedding $\alpha_k^{(0)} \sim \mathcal{N}(0, I)$
2. Draw initial topic proportion mean $\eta_0 \sim \mathcal{N}(0, I)$
3. For time step $t = 1, \ldots, T$:
   (a) Draw topic embeddings $\alpha_k^{(t)} \sim \mathcal{N}(\alpha_k^{(t-1)}, \gamma^2 I)$ for $k = 1, \ldots, K$
   (b) Draw topic proportion means $\eta_t \sim \mathcal{N}(\eta_{t-1}, \delta^2 I)$
4. 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 $w_{dn} \sim \text{Cat}(\text{softmax}(\rho^\top \alpha_{z_{dn}}^{(t_d)}))$.

Steps 1 and 3a give the prior over the topic embeddings; it encourages smoothness on the resulting topics. Steps 2 and 3b is shared with D-LDA; it describes the evolution of the prior mean over the topic proportions. Steps 4a and 4b-i are standard

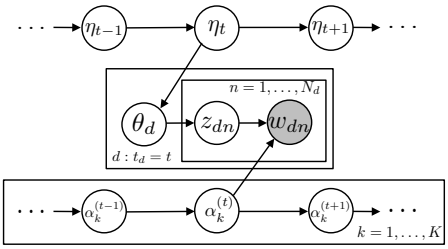

Figure 1: Graphical representation of D-ETM. The topic embeddings $\alpha_k^{(t)}$ and the latent means η_t evolve over time. For each document at time step t, the prior over the topic proportions θ_d depends on η_t. The variables z_{dn} denote the topic assignment; the variables w_{dn} denote the words.

for topic modeling; they represent documents as distributions over topics and draw a topic assignment for each word. Step 4b-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 aspect of the model is useful when the embedding of a new word is available, i.e., a word that does not appear in the corpus. Specifically, 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

We observe 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. 2 for convenience,[3]

$$p(w_{dn} \mid \alpha_k^{(t_d)}) = \sum_{k=1}^{K} p(w_{dn} \mid z_{dn} = k, \alpha_k^{(t_d)}). \tag{4}$$

The posterior is intractable. We approximate it with 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, and 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),

$$\mathcal{L}(\nu) = \mathbb{E}_q \left[\log p(\mathcal{D}, \theta, \eta, \alpha) - \log q_\nu(\theta, \eta, \alpha) \right]. \tag{5}$$

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

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

$$q(\theta, \eta, \alpha) = \prod_d q(\theta_d \mid \eta_{t_d}, \mathbf{w}_d) \times \prod_t q(\eta_t \mid \eta_{1:t-1}, \widetilde{\mathbf{w}}_t) \times \prod_k \prod_t q(\alpha_k^{(t)}). \quad (6)$$

(To avoid clutter, we suppress the notation for the variational parameters.)

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 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 an 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, unlike Blei and Lafferty (2006), we do not use structured variational inference for the topics. Instead, we use the mean-field family for the approximation over the topic embeddings, $q(\alpha_k^{(t)})$, for simplicity.

We optimize the ELBO with respect to the variational parameters. Because the expectations in Eq. 5 are intractable, we use black box variational inference, obtaining unbiased gradient estimators with Monte Carlo. 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

We use the D-ETM to analyze the transcriptions of the UN general debates from 1970 to 2015, a corpus of ACL abstracts from 1973 to 2006, and a set of articles from Science Magazine from 1990 to 1999. We found the D-ETM provides better predictive power and higher topic quality in general on these datasets when compared to D-LDA.

On the transcriptions of the UN general debates, we additionally carried out a qualitative analysis of the results. We found that the D-ETM reveals the temporal evolution of the topics discussed in the debates (such as climate change, war, poverty, or human rights).

Algorithm 1: Dynamic topic modeling with the D-ETM

input : Documents $\{\mathbf{w}_1, \dots, \mathbf{w}_D\}$ and their time stamps $\{t_1, \dots, t_D\}$
Initialize all variational parameters
for iteration $1, 2, 3, \dots$ **do**
  Sample the latent means and the topic embeddings, $\eta \sim q(\eta \mid \tilde{\mathbf{w}})$ and $\alpha \sim q(\alpha)$
  Compute the topics $\beta_k^{(t)} = \text{softmax}(\rho^\top \alpha_k^{(t)})$ for $k = 1, \dots, K$ and $t = 1, \dots, T$
  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. 5 and its gradient w.r.t. the variational parameters (backpropagation)
  Update the model and variational parameters using Adam
end

Table 1: Summary statistics of the different datasets under study.

Dataset	# Docs Train	# Docs Val	# Docs Test	# Timestamps	Vocabulary
UN	196,290	11,563	23,097	46	12,466
SCIENCE	13,894	819	1,634	10	25,987
ACL	8,936	527	1,051	31	35,108

We compared the D-ETM against two versions of D-LDA, labeled as D-LDA and D-LDA-REP, which differ only in the inference method (the details are below). The comparison of the D-ETM against D-LDA-REP reveals that the key to the D-ETM's performance is the model and not the inference procedure.

Datasets. We study the D-ETM on three datasets. The UN debates corpus[4] spans 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. We follow Lefebure (2018) and split the speeches into paragraphs, treating each paragraph as a separate document.

The second dataset is ten years of SCIENCE articles, 1990 to 1999. The articles are from JSTOR, an on-line archive of scholarly journals that scans bound volumes and runs optical character recognition algorithms on the scans. This data was used by Blei and Lafferty (2007).

The third dataset is a collection of articles from 1973 to 2006 from the ACL Anthology (Bird et al., 2008). This anthology is a repository of computational linguistics and natural language processing papers.

[4] Available at https://www.kaggle.com/unitednations/un-general-debates.

Table 2: Predictive performance as measured by held-out perplexity (lower is better) on a document completion task. The D-ETM outperforms both D-LDA and D-LDA-REP on all but one corpus. These results also show that the D-ETM gains its advantage through its modeling assumptions and not through its inference procedure.

Method	UN	SCIENCE	ACL
D-LDA (Blei and Lafferty, 2006)	2393.5	**3600.7**	4324.2
D-LDA-REP	2931.3	8377.4	5836.7
D-ETM	**1970.7**	4206.1	**4120.6**

Table 3: Qualitative performance on the UN dataset as measured by topic coherence (TC), topic diversity (TD), and topic quality (TQ). The higher these metrics the better. The D-ETM achieves better overall topic quality than D-LDA and D-LDA-REP.

Method	TC	TD	TQ
D-LDA (Blei and Lafferty, 2006)	**0.1317**	0.6065	0.0799
D-LDA-REP	0.1180	0.2691	0.0318
D-ETM	0.1206	**0.6703**	**0.0809**

For each dataset, we apply standard preprocessing techniques, such as tokenization and removal of numbers and punctuation marks. We also filter out stop words, i.e., words with document frequency above 70%, as well as standard stop words from a list. Additionally, we remove low-frequency words, i.e., words that appear in less than a certain number of documents (30 documents for UN debates, 100 documents for the SCIENCE corpus, and 10 documents for the ACL dataset). We use 85% randomly chosen documents for training, 10% for testing, and 5% for validation, and we remove one-word documents from the validation and test sets. Table 1 summarizes the characteristics of each dataset.

Methods. We compare the D-ETM against two variants of D-LDA. One variant is the original model and algorithm of Blei and Lafferty (2006). The other variant, which we call D-LDA-REP, is the D-LDA model of Blei and Lafferty (2006) fitted using mean-field variational inference with the reparameterization trick. The comparison against D-LDA-REP helps us delineate between performance due to the model and performance due to the inference algorithm.

Settings. We use 50 topics for all the experiments and follow Blei and Lafferty (2006) to set the variances of the different priors 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)[5]. We apply the algorithm in Section 4.2 using a batch size of 200 documents for all datasets except for ACL for which we used 100. We use a fully connected feed-forward inference network for the topic proportions θ_d. The network has ReLU activations and 2 layers of 800 hidden units each. We set the mean and log-variance for θ_d as linear maps of the output. We applied a small

[5]More advanced methods can be used to learn word embeddings. We used skip-gram for simplicity and found it leads to good performance.

Table 4: Qualitative performance on the SCIENCE dataset as measured by topic coherence (TC), topic diversity (TD), and topic quality (TQ). The higher these metrics the better. The D-ETM achieves better overall topic quality than D-LDA and D-LDA-REP.

Method	TC	TD	TQ
D-LDA (Blei and Lafferty, 2006)	**0.2392**	0.6502	0.1556
D-LDA-REP	0.0611	0.2290	0.0140
D-ETM	0.2298	**0.8215**	**0.1888**

Table 5: Qualitative performance on the ACL dataset as measured by topic coherence (TC), topic diversity (TD), and topic quality (TQ). The higher these metrics the better. The D-ETM achieves better overall topic quality than D-LDA and D-LDA-REP.

Method	TC	TD	TQ
D-LDA (Blei and Lafferty, 2006)	0.1429	0.5904	0.0844
D-LDA-REP	0.1011	0.2589	0.0262
D-ETM	**0.1630**	**0.8286**	**0.1351**

dropout rate of 0.1 to the output of this network before using it to compute the mean and the log-variance. 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 400. This conforms the input of an LSTM that has 4 layers of 400 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 1000 epochs on SCIENCE and ACL and for 400 epochs on the UN dataset. The stopping criterion is based on the held-out log-likelihood on the validation set. The learning rate is set to 0.001 for the UN and SCIENCE datasets and to 0.0008 on the ACL corpus. We fixed the learning rate throughout training. We clip the norm of the gradients of the ELBO to 2.0 to stabilize training.

We fit D-LDA using the published code of Blei and Lafferty (2006). (See https://github.com/blei-lab/dtm.) To fit D-LDA, Blei and Lafferty (2006) derived 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-marginal likelihood of the data, this algorithm presents scalability issues both in terms of the number of topics and in terms of the vocabulary. For example fitting D-LDA took almost two days on each dataset whereas we only required less than 6 hours for the D-ETM.

To fit D-LDA-REP we leverage recent advances in variational inference to overcome these issues. We use stochastic optimization based on reparameterization gradients and we draw 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

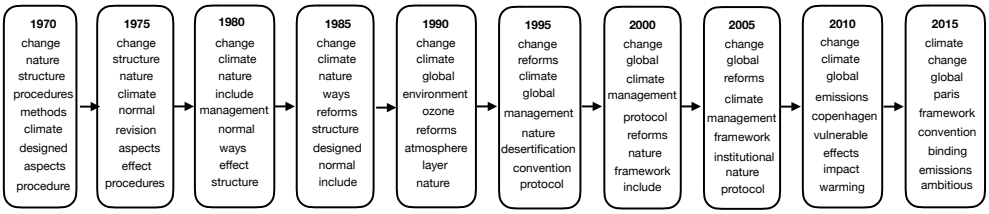

Figure 2: Temporal evolution of the top-10 words from a topic about climate change learned by the D-ETM. This topic is in agreement with historical events. In the 1990s the destruction of the ozone layer was of major concern. More recently the concern is about global warming. Events such as the Kyoto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

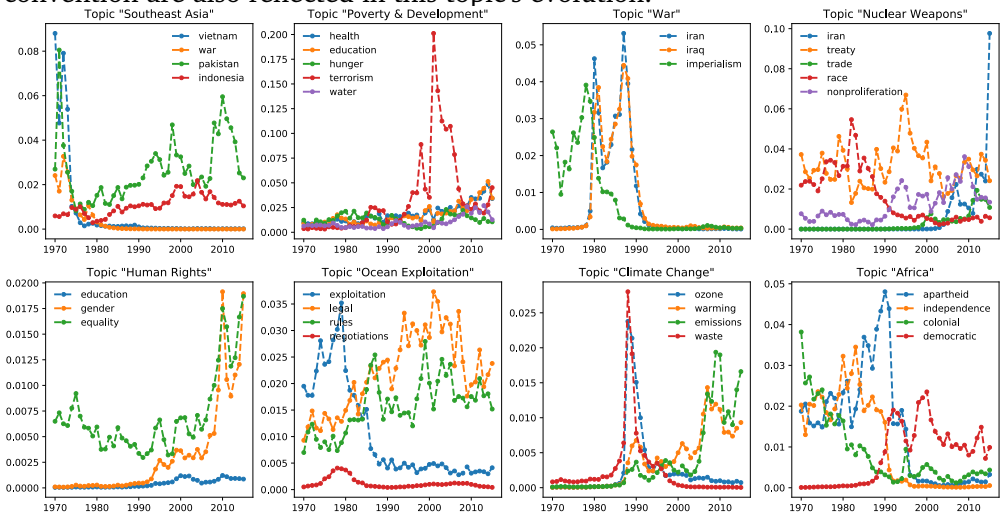

Figure 3: Evolution of word probability across time for eight different topics learned by the D-ETM. For each topic, we choose a set of words whose probability shift aligns with historical events (these are not the words with the highest probability in each topic). For example, one interesting finding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

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, D-LDA, and D-LDA-REP 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 topic 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).

Table 2, Table 3, Table 4, and Table 5 show that the D-ETM outperforms both D-LDA and D-LDA-REP according to both perplexity and topic quality on almost all datasets. In particular, the D-ETM finds more diverse and coherent topics. We posit this is due to its use of embeddings.

Qualitative results. The D-ETM finds that the topics' evolution over time are in agreement with historical events. As an example, Figure 2 shows the trajectory of 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 combines word embeddings and dynamic latent Dirichlet allocation (D-LDA). The D-ETM models each word with a categorical distribution parameterized by the dot product between the embedding of the word and an embedding representation of its assigned topic. Each topic embedding is a time-varying vector in the embedding space of words. Using a random walk prior over these topic embeddings, the D-ETM uncovers smooth topic trajectories. We applied the D-ETM to analyze three different corpora and found that the D-ETM outperforms D-LDA both in terms of predictive performance and topic quality while requiring significantly less time to fit.

Acknowledgements

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. Francisco J. R. Ruiz is supported

by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 706760. Adji B. Dieng 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*.

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D., and Tan, Y. F. (2008). The ACL anthology reference corpus: a reference dataset for bibliographic research in computational linguistics. In *International Conference on Language Resources and Evaluation*.

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

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*, 3:993–1022.

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., 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.

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*.