Neural Dynamic Focused Topic Model

Kostadin Cvejoski[1, 2], **Ramsés J. Sánchez**[1, 4], **César Ojeda**[3]
[1] Lamarr-Institute for Machine Learning and Artificial Intelligence
[2] Fraunhofer-Institute for Intelligent Analysis and Information Systems (IAIS)
[3] University of Potsdam
[4] BIT University of Bonn
kostadin.cvejoski@iais.fraunhofer.de, sanchez@bit.uni-bonn.de, ojedamarin@uni-potsdam.de

Abstract

Topic models and all their variants analyse text by learning meaningful representations through word co-occurrences. As pointed out by previous work, such models implicitly assume that the probability of a topic to be active and its proportion within each document are positively correlated. This correlation can be strongly detrimental in the case of documents created over time, simply because recent documents are likely better described by new and hence rare topics. In this work we leverage recent advances in neural variational inference and present an alternative neural approach to the dynamic Focused Topic Model. Indeed, we develop a neural model for topic evolution which exploits sequences of Bernoulli random variables in order to track the appearances of topics, thereby decoupling their activities from their proportions. We evaluate our model on three different datasets (the UN general debates, the collection of NeurIPS papers, and the ACL Anthology dataset) and show that it (i) outperforms state-of-the-art topic models in generalization tasks and (ii) performs comparably to them on prediction tasks, while employing roughly the same number of parameters, and converging about two times faster.

Introduction

Probabilistic topic models, the likes of Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), are generative models of word co-occurrence that analyse large document collections by learning latent representations (topics) encoding their themes. These models represent the documents of the collection as mixtures of latent topics, and group semantically related words into single topics by means of word-pair frequency information within the collection. Such a generic generative structure has been successfully applied to problems ranging from information retrieval, visualization and multilingual modelling to linguistic understanding in fiction and non-fiction, scientific publications and political texts (see e.g. Boyd-Graber et al. (2017) for a review), and keeps being extended to new domains (Rezaee and Ferraro 2020; Zhao et al. 2021).

Topic models implicitly assume that the documents within a given collection are exchangeable. Yet document collections such as magazines, academic journals, news articles and social media content not only feature trends and themes that change with time, but also employ their language differently as time evolves (Danescu-Niculescu-Mizil et al. 2013). The exchangeability assumption along the time component is hence inappropriate in these cases and topic models have been extended to account for changes in both topic (Blei and Lafferty 2006; Wang, Blei, and Heckerman 2012; Jähnichen et al. 2018) and word (Bamler and Mandt 2017; Rudolph and Blei 2018; Dieng, Ruiz, and Blei 2019) distributions, among documents collected over long periods of time.

It is easy to imagine, however, that if one analyses the collection's content as one moves forward in time, one would find that (some of) the topics describing those documents appear, disappear or reappear with time. This simple intuition entails that one should not only model the time- and document-dependent topic proportions, but also *the probabilities for the topics to be active*, and how such probabilities change with time. Previous work has already pointed out that existing topic models implicitly assume that the probability of a topic being active and its proportion within each document are positively correlated (Williamson et al. 2010; Perrone et al. 2017). This assumption is generally unwanted, simply because rare topics may account for a large part of the words in the few documents in which they are active. It is particularly detrimental (for both modelling and prediction) in a dynamic setting, because recent documents are likely better described by new and hence rare topics.

Indeed, whenever the topic distribution over documents is strongly skewed, topic models tend to learn the more general topics held by the big majority of documents in the collection, rather than the rare topics contained only by fewer documents (Jagarlamudi, Daumé III, and Udupa 2012; Tang et al. 2014; Zuo, Zhao, and Xu 2014). Document collections that reflect evolving content typically feature skew topic distribution over its documents, with the newly added documents being well described by new, rare topics. Dynamic topic models that feature the topic proportion-activity coupling are then expected to perform badly, simply because these will not be able to infer the new topics characteristic of recent documents. To properly model such recent documents one should therefore allow rarely seen topics to be active with high proportion and frequently seem topics to be active with low proportion.

In this work we seek to decouple the probability for a topic to be active from its proportion with the introduction of sequences of Bernoulli random variables, which select the active topics for a given document at a particular instant of time. Earlier models attained such a decoupling via nonparametric priors, such as the Indian Buffet Process prior over infinite binary matrices, in both static (Williamson et al. 2010) and dynamic (Perrone et al. 2017) settings. Our construction roughly follows a similar logic, but leverages the reparametrization trick to perform neural variational inference (Kingma and Welling 2013). The result is a scalable model that allows the instantaneous number of active topics per document to fluctuate, and explicitly decouples the topic proportion from its activity, thereby offering some novel layers of interpretability and transparency into the evolution of topics over time.

We introduce the Neural Dynamic Focused Topic Model (`NDF-TM`) which builds on top of Neural Variational Topic models (Miao, Yu, and Blunsom 2016) and uses Deep Kalman Filters (Krishnan, Shalit, and Sontag 2015) to model the independent dynamics of both topic proportion and topic activities. We train and test our model on three datasets, namely the UN general debates, the collection of NeurIPS papers and the ACL Anthology dataset. Our results show via different metrics that `NDF-TM` outperforms state-of-the-art topic models in generalization tasks, and performs comparably to them on prediction tasks. Very importantly, `NDF-TM` does this while employing roughly the same number of parameters and converging two times faster than the strongest baseline.

Related Work

The `NDF-TM` model merges concepts from dynamic topic models, dynamic embeddings and neural topic models.

Dynamic topic models. The seminal work of Blei and Lafferty (2006) introduced the Dynamic Topic Model (DTM), which uses a state space model on the natural parameters of the distribution representing the topics, thus allowing the latter to change with time. The DTM methodology was first extended by Caron, Davy, and Doucet (2007) to a nonparametric setting, via the correlation of Dirichlet process mixture models in time. Later Wang, Blei, and Heckerman (2012) replaced the discrete state space model of DTM with a Diffusion process, thereby extending the approach to a continuous time setting. Jähnichen et al. (2018) further extended DTM by introducing Gaussian process priors that allowed for a non-Markovian representation of the dynamics. Other recent work on dynamic topic models is that of Hida et al. (2018)

Dynamic embeddings. Rather than modelling the content evolution of document collections like DTM, other works focus on modelling how word semantics change with time (Bamler and Mandt 2017; Rudolph and Blei 2018). These works use continuous representation of words capturing their semantics (as e.g. those of Pennington, Socher, and Manning (2014)) and evolve such representation via diffusion processes. More recently, Dieng, Ruiz, and Blei (2019) represent topics as dynamic embeddings, and model words via categorical distributions whose parameters are given by the inner product between the static word embeddings and the dynamic topic embeddings. As such, this model corresponds to the dynamic extension of Dieng, Ruiz, and Blei (2020).

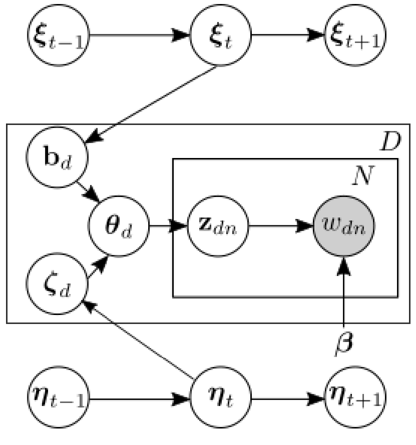

Figure 1: Graphical model representation of `NDF-TM`.

Neural topic models. Another line of research leverages neural networks to improve the performance of topic models, the so-called neural topic models (Miao, Yu, and Blunsom 2016; Srivastava and Sutton 2017; Zhang et al. 2018; Dieng, Ruiz, and Blei 2020, 2019) which deploy neural variational inference (Kingma and Welling 2013) for training.

Decoupling topic activity from its proportion. Williamson et al. (2010) noted the implicit and undesirable correlation between topic activity and proportion assumed by standard topic models and introduced the Focused Topic Model (FTM). FTM uses the Indian Buffet Process (IBP) to decouple across-data prevalence and within-data proportion in mixed membership models. Later Perrone et al. (2017) extended FTM to a dynamic setting by using the Poisson Random Fields model from population genetics to generate dependent IBPs, which allow them to model temporal correlations in data. Both of these models are trained using complex sampling schemes, which can make the fast and accurate inference of their model parameters difficult (Miao, Grefenstette, and Blunsom 2017).

In what follows we propose an alternative neural approach to the dynamic Focused Topic model of Perrone et al. (2017), trainable via backpropagation, which learns to decouple the dynamic topic activity from its dynamic topic proportion.

Neural Dynamic Focused Topic Model

Suppose we are given an ordered collection of corpora $\mathcal{D} = \{D_1, D_2, \ldots, D_T\}$, so that the tth corpus D_t is composed of N_t documents, all received within the tth time window. Let $\mathbf{W}_t$ denote the Bag-of-word (BoW) representation for the whole document set within D_t and let $\mathbf{w}_{t,d}$ denote the BoW representation of the d-th document in D_t.

Let us now suppose that the corpora collection is described by a set of K unknown topics. We then assume there are two sequences of continuous hidden variables $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T \in \mathbb{R}^{\dim(\eta)}$ and $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_T \in \mathbb{R}^{\dim(\xi)}$ which encode, respectively, how the topic proportions and the topic activities change among corpora as time evolves (i.e. as one moves from D_t to D_{t+1}). That is, $\boldsymbol{\eta}_t$ and $\boldsymbol{\xi}_t$ encode the *global dynamics* of semantic content. We also assume there are two *local* hidden variables, conditioned on the global ones, namely a continuous variable $\boldsymbol{\zeta}_{t,d} \in \mathbb{R}^K$ which encodes the content of the dth document in D_t, in terms of the available topics, and a binary variable $\mathbf{b}_{t,d} \in \{0,1\}^K$ which encodes which topics are active in the document in question. We combine these local variables to compute the topic proportions $\boldsymbol{\theta}_{t,d} \in [0,1]^K$ from which the dth document in D_t is generated.

Generation

Let us denote with ψ the set of parameters of our generative model. We are first of all interested in modelling the topic activity per document at each time step, directly from the data. One could, for example, use a K-dimensional mask (i.e. a K-dimensional vector, whose kth entry is either 1 or 0 depending on whether the kth topic is active or inactive) for each document d, at each time step t. To account for the variability of the data, one could also make this mask stochastic. We thus introduce K time- and document-dependent Bernoulli variables $\mathbf{b}_{t,d} \in \{0,1\}^K$ whose generation process is given by

$$\begin{aligned}
\boldsymbol{\xi}_t &\sim \mathcal{N}\left(\boldsymbol{\mu}^{\xi}_{\psi}(\boldsymbol{\xi}_{t-1}), \boldsymbol{\delta}\,\mathbf{I}\right), && (1)\\
\boldsymbol{\pi}_t &= \alpha_0 \,\text{Sigmoid}\left(\mathbf{W}_{\xi}\,\boldsymbol{\xi}_t + \mathbf{c}_{\xi}\right), && (2)\\
\mathbf{b}_{t,d} &\sim \text{Bernoulli}(\boldsymbol{\pi}_t), && (3)
\end{aligned}$$

where α_0 is a hyperparameter controlling the percentage of active topics, and $\mathbf{W}_{\xi} \in \mathbb{R}^{K \times \dim(\xi)}, \mathbf{c}_{\xi} \in \mathbb{R}^K \subset \psi$ are trainable parameters. Also note that, just as in Deep Kalman Filters (Krishnan, Shalit, and Sontag 2015), $\boldsymbol{\xi}_t$ is Markovian and evolves under a Gaussian noise with mean $\boldsymbol{\mu}^{\xi}_{\psi}$, defined via a neural network with parameters in ψ, and variance $\boldsymbol{\delta}$. The latter being a hyperparameter of the model. Finally, we choose $\boldsymbol{\xi}_1 \sim \mathcal{N}(0,1)$.

Analogously, we generate the topic proportions $\boldsymbol{\theta}_{t,d}$ as

$$\begin{aligned}
\boldsymbol{\eta}_t &\sim \mathcal{N}\left(\boldsymbol{\mu}^{\eta}_{\psi}(\boldsymbol{\eta}_{t-1}), \boldsymbol{\delta}\,\mathbf{I}\right), && (4)\\
\boldsymbol{\zeta}_{t,d} &\sim \mathcal{N}\left(\mathbf{W}_{\zeta}\,\boldsymbol{\eta}_t + \mathbf{c}_{\zeta},\, 1\right), && (5)\\
\boldsymbol{\theta}_{t,d} &= \frac{\mathbf{b}_{t,d} \odot \exp\left(\boldsymbol{\zeta}_{t,d}\right)}{\sum_k^K b^k_{t,d} \odot \exp\left(\zeta^k_{t,d}\right)}, && (6)
\end{aligned}$$

where $\mathbf{b}_{t,d}$ is defined in (3) and $\odot$ labels element-wise product, $\mathbf{W}_{\zeta} \in \mathbb{R}^{K \times \dim(\eta)}, \mathbf{c}_{\zeta} \in \mathbb{R}^K \subset \psi$ are trainable, and $\boldsymbol{\mu}^{\eta}_{\psi}$ is modelled via a neural network. Here $\boldsymbol{\eta}_t$ is also Markovian and we set $\boldsymbol{\eta}_1 \sim \mathcal{N}(0,1)$. Note that the topic proportion thus defined can be *sparse vectors*. That is, the model has the flexibility to completely mask some of the topics out of a given document, at a given time.

Once we have $\boldsymbol{\theta}_{t,d}$ we generate the corpora sequence by sampling

$$\begin{aligned}
z_{t,d,n} &\sim \text{Categorical}(\boldsymbol{\theta}_{t,d}), && (7)\\
w_{t,d,n} &\sim \text{Categorical}(\boldsymbol{\beta}_{z_{t,d,n}}), && (8)
\end{aligned}$$

where $z_{t,d,n}$ is the time-dependent topic assignment for $w_{t,d,n}$, which labels the nth word in document $d \in D_t$, and $\boldsymbol{\beta} \in \mathbb{R}^{K \times V}$ is a learnable topic distribution over words. We define the latter as

$$\boldsymbol{\beta} = \text{softmax}(\boldsymbol{\alpha} \otimes \boldsymbol{\rho}), \tag{9}$$

with $\boldsymbol{\alpha} \in \mathbb{R}^{K \times E}, \boldsymbol{\rho} \in \mathbb{R}^{V \times E}$ learnable topic and word embeddings, respectively, for some embedding dimension E, and $\otimes$ denoting tensor product.

`NDF-TM` is summarized in Figure 1.

Inference

The generative model above involves two independent global hidden variables $\boldsymbol{\xi}_t, \boldsymbol{\eta}_t$, and two local hidden variables $\boldsymbol{\zeta}_{t,d}$ and $\mathbf{b}_{t,d}$. Our task is to infer the posterior distributions of all these variables.[1] Denoting with $\boldsymbol{\Gamma}_{t,d}$ the set $\{\boldsymbol{\xi}_t, \boldsymbol{\eta}_t, \boldsymbol{\zeta}_{t,d}, \mathbf{b}_{t,d}\}$, we approximate the true posterior distribution of the model with a variational (and structured) posterior of the form

$$\begin{aligned}
&q_{\varphi}(\boldsymbol{\Gamma}_{t,d}|\mathbf{w}_{t,d}, \mathbf{W}_{1:T}) = \\
&\quad \prod_t^T q_{\varphi}(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{1:t-1}, \mathbf{W}_{1:T})\, q_{\varphi}(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{1:t-1}, \mathbf{W}_{1:T}) \\
&\quad\quad \times \prod_d^{N_t} q_{\varphi}(\boldsymbol{\zeta}_{t,d}|\mathbf{w}_{t,d}, \boldsymbol{\eta}_t)\, q_{\varphi}(\mathbf{b}_{t,d}|\mathbf{w}_{t,d}, \boldsymbol{\xi}_t), \qquad (10)
\end{aligned}$$

where $\mathbf{W}_{1:T} = (\mathbf{W}_1, \ldots, \mathbf{W}_T)$ is the ordered sequence of BoW representations for the corpus collection and φ labels the variational parameters.

Local variables. The posterior distribution over the local variables $\boldsymbol{\zeta}_{t,d}, \mathbf{b}_{t,d}$ are chosen as Gaussian and Bernoulli, respectively, each parametrized by neural networks taking as input their conditional variables.

Global variables. The posterior distribution over the dynamic global variables $\boldsymbol{\xi}_t, \boldsymbol{\eta}_t$ are also Gaussian, but now depend not only on the latent variables at time $t-1$, but also on the entire sequence of BoW representations $\mathbf{W}_{1:T}$. This follows directly from the graphical model in Figure 1, as noted by e.g. Krishnan, Shalit, and Sontag (2015). We shall use LSTM networks (Hochreiter and Schmidhuber 1997) to model these dependencies. Specifically let

$$q_{\varphi}(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t-1}, \mathbf{W}_{1:T}) = \mathcal{N}(\boldsymbol{\mu}^{\xi}_{\varphi}, \boldsymbol{\sigma}^{\xi}_{\varphi}), \tag{11}$$

where $\boldsymbol{\mu}^{\xi}_{\varphi}$, $\boldsymbol{\sigma}^{\xi}_{\varphi}$ are neural networks which take as input the pair $(\boldsymbol{\xi}_{t-1}, \mathbf{h}^{\xi}_t)$, with $\mathbf{h}^{\xi}_t$ a hidden representation encoding the sequence $\mathbf{W}_{1:T}$. Similarly

$$q_{\varphi}(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1}, \mathbf{W}_{1:T}) = \mathcal{N}(\boldsymbol{\mu}^{\eta}_{\varphi}, \boldsymbol{\sigma}^{\eta}_{\varphi}), \tag{12}$$

where $\boldsymbol{\mu}^{\eta}_{\varphi}, \boldsymbol{\sigma}^{\eta}_{\varphi}$, again neural networks, take as input the pair $(\boldsymbol{\eta}_{t-1}, \mathbf{h}^{\eta}_t)$, with $\mathbf{h}^{\eta}_t$ a second hidden representation also encoding $\mathbf{W}_{1:T}$. These hidden representations $\mathbf{h}^i_t$, with $i = \{\xi, \eta\}$, correspond to the hidden states of LSTM networks whose update equation read

$$\mathbf{h}^i_t = f^i_{\varphi}(\mathbf{W}_t, \mathbf{h}^i_{t-1}). \tag{13}$$

[1] Note in passing that we do not need to perform inference of the latent topics $z_{t,d,n}$, simply because these can be integrated out (aka marginalized).

Models	UN		NeurIPS		ACL	
	PPL-DC ↓	P-NLL ↓	PPL-DC ↓	P-NLL ↓	PPL-DC ↓	P-NLL ↓
DTM*	2393.5	-	-	-	-	-
DTM-REP	3 012 ± 14	8.331 ± 0.003	8 107±907	9.5 ± 0.4	8 503 ± 875	9.7 ± 0.5
D-ETM	1 748 ± 13	**7.615 ± 0.005**	7 746±699	8.983 ± 0.003	7 805±182	**8.84 ± 0.02**
NDF-LT-TM	1 578 ± 29	7.682± 0.080	6 549±21	8.923 ± 0.002	7 877±213	8.91 ± 0.03
NDF-TM	**1 527 ± 36**	7.640± 0.004	**6 529±26**	**8.901 ± 0.001**	**7 690±215**	8.88 ± 0.03

Table 1: Perplexity on document completion (PPL-DC) and predictive negative log likelihood (P-NLL). PPL-DC is calculated by conditioning the model on the first half of the document and evaluating the perplexity on the second half of the document. P-NLL is estimated using equation 16. The DTM* results are taken from (Dieng, Ruiz, and Blei 2019). All other results are obtained by training the models on 5 different random splits of the datasets.

Training Objective

To optimize the model parameters $\{\psi, \varphi\}$ we minimize the variational lower bound on the logarithm of the marginal likelihood $p_\psi(w_{t,d,n}|\boldsymbol{\beta})$. Following standard methods (Bishop 2006), the latter can readily be shown to be

$$\begin{aligned}\mathcal{L}[\boldsymbol{\beta}, \psi, \varphi] &= \sum_{t=1}^{T}\sum_{d=1}^{N_t}\sum_{n=1}^{N_d} \mathbb{E}_{\boldsymbol{\Gamma}}\Big\{ \log p_\psi(w_{t,d,n}|\boldsymbol{\beta}, \boldsymbol{\Gamma})\Big\} \\ &- \mathrm{KL}\left[q_\varphi(\boldsymbol{\eta}_1|\mathbf{W}_{1:T}); p(\boldsymbol{\eta}_1)\right] - \mathrm{KL}\left[q_\varphi(\boldsymbol{\xi}_1|\mathbf{W}_{1:T}); p(\boldsymbol{\xi}_1)\right] \\ &- \sum_{t=2}^{T} \mathrm{KL}\left[q_\varphi(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{1:t-1}, \mathbf{W}_{1:T}); p_\psi(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1})\right] \\ &- \sum_{t=2}^{T} \mathrm{KL}\left[q_\varphi(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{1:t-1}, \mathbf{W}_{1:T}); p_\psi(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t-1})\right] \\ &- \sum_{t=1}^{T}\sum_{d=1}^{N_t}\Bigg(\mathbb{E}_{\boldsymbol{\eta}_t}\Big\{\mathrm{KL}\left[q_\varphi(\boldsymbol{\zeta}_{t,d}|\mathbf{w}_{t,d}, \boldsymbol{\eta}_t); p_\psi(\boldsymbol{\zeta}_{t,d}|\boldsymbol{\eta}_t)\right]\Big\} \\ &+ \mathbb{E}_{\boldsymbol{\xi}_t}\Big\{\mathrm{KL}\left[q_\varphi(\mathbf{b}_{t,d}|\mathbf{w}_{t,d}, \boldsymbol{\xi}_t); p_\psi(\mathbf{b}_{t,d}|\boldsymbol{\xi}_t)\right]\Big\}\Bigg), \quad (14)\end{aligned}$$

where KL labels the Kullback-Leibler divergence and $\boldsymbol{\beta}$ is defined in Eq. 9.

Experiments

In this section we introduce our datasets and define our baselines. Details about pre-processing and experimental setup can be found in the supplementary material, provided within the repository of our code. Nevertheless, let us mention here that two important hyperparameters of the model are the maximum topic number K and the percentage of active topics α_0. Both these hyperpameters are chosen via cross-validation, with $K = 50$ and $\alpha_0 = 0.5$ given the best results[2]. Source code to reproduce our experiments is available online.[2]

Datasets

We evaluate our model on three datasets, namely the collection of UN speeches, NeurIPS papers and the ACL Anthology. The UN dataset[3] (Baturo, Dasandi, and Mikhaylov 2017) contains the transcription of the speeches given at the UN General Assembly during the period between the years 1970 and 2016. It consists of about 230950 documents. The NeurIPS dataset[4] contains the collection of papers published in between the years 1987 and 2016. It consists of about 6562 documents. Finally, the ACL Anthology (Bird et al. 2008) contains a collection of computational linguistic and natural language processing papers published between 1973 and 2006. It consists of about 10514 documents.

Baselines

Our main aim is to study the effect of the topic proportion-activity coupling in the performance of *dynamic topic models*[5] on data collections displaying evolving content. To do so we compare against three models:

(1) DTM-REP — the neural extension of DTM, fitted using neural variational inference (Dieng, Ruiz, and Blei 2019). This model uses a logistic-normal distribution, parametrized with feedforward neural networks, as posterior for the topic proportion distribution; as in Miao, Grefenstette, and Blunsom (2017). It also uses Kalman Filters to model the topic dynamics, but parametrizes the posterior distribution over the dynamic latent variables with LSTM networks, just as in Deep Kalman Filters (Krishnan, Shalit, and Sontag 2015) (and just as NDF-TM too, see e.g. Eq. 13). As such, DTM-REP works as the dynamic extension of Miao, Grefenstette, and Blunsom (2017)'s model. It follows that the DTM-REP model thus defined only differs from NDF-TM in the way we model the topic proportions. Comparing our model against DTM-REP should therefore explicitly show the effect of lifting the topic proportion-activity coupling in dynamic neural topic models.

(2) D-ETM — the Dynamic Embedded Topic Model (Dieng, Ruiz, and Blei 2019), which captures the evolution of topics in such a way that both the content of topics and their proportions evolve over time. This model adds complexity

[2] Specifically, K was chosen from the set 50, 100 and 200. We found 50 to be the best value for all models, i.e. including the baselines. Similarly α_0 was chosen from the set 0.1, 0.5, 1.0

[2] Source code: https://github.com/cvejoski/Neural-Dynamic-Focused-Topic-Model

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

[4] https://www.kaggle.com/benhamner/nips-papers

[5] This means we do not consider static topic models

	UN		NEURIPS		ACL	
Models	TC ↑	TD ↑	TC ↑	TD ↑	TC ↑	TD ↑
DTM*	0.1317	0.0799	-	-	-	-
DTM-REP	0.11 ± 0.30	0.59 ± 0.10	-0.62±0.07	0.15±0.01	-0.82 ± 0.08	0.55 ± 0.02
D-ETM	0.43 ± 0.20	0.61 ±0.01	-0.54±0.09	0.82 ±0.01	-0.71±0.16	0.63±0.05
NDF-LT-TM	0.43 ± 0.18	0.56 ± 0.03	-0.53±0.02	**0.90±0.01**	-0.74±0.11	0.73±0.01
NDF-TM	**0.46 ± 0.20**	**0.63 ± 0.01**	**-0.50±0.04**	0.85±0.02	**-0.64±0.12**	**0.74±0.01**

Table 2: Topic coherence (TC) and Topic diversity (TD) for all models. TC is calculated by taking the average pointwise mutual information between two words drawn randomly from the same topic. TD is the percentage of unique words in the top 25 words of all topics. The DTM* results are taken from (Dieng, Ruiz, and Blei 2019). All other results are obtained by training the models on 5 different random splits of the datasets.

to DTM-REP by modelling words via categorical distributions whose parameters are given by the inner product between the static word embeddings and the dynamic topic embeddings. In this way, D-ETM does not (necessarily) suffers from the topic proportion-activity coupling, for it can implicitly model their decoupling via its additional degrees of freedom.

(3) NDF-LT-TM — the Neural Dynamic Focused topic model *with linear transition*. We introduce this last baseline for the sake of ablation, viz. to investigate the effect of the neural networks $\boldsymbol{\mu}^{\xi}_{\psi}, \boldsymbol{\mu}^{\eta}_{\psi}$ in Eqs. 1 and 4. NDF-LT-TM is defined by replacing these neural networks with the identity function.

Results

In order to quantify the performance of our models, we first focus on two aspects, namely its prediction capabilities and its ability to generalize to unseen data. Later we also (qualitatively) discuss how the model actually performs the decoupling between topic activities and proportions.

(1) To test how well our models perform on a prediction task we compute the *predictive negative log likelihood* (P-NLL). Since to our knowledge the latter does not appear explicitly in the dynamic topic model literature, we briefly revisit how to estimate it in what follows.

In order to predict N steps into the future we rely on the generative process of our model, albeit conditioned on the past. Essentially, one must generate Monte Carlo samples from the posterior distribution and propagate the latent representations ($\boldsymbol{\xi}_t$ and $\boldsymbol{\eta}_t$ in our model) into the future with the help of the prior transition function (Eqs. 1 and 4, respectively)[6]. This procedure is depicted on the conditional predictive distribution of our model

$$p(\mathbf{W}_{T+1}|\mathbf{W}_{1:T}) = \int p_{\psi}(\mathbf{W}_{T+1}|\mathbf{\Gamma}_{T+1}) \times p_{\psi}(\mathbf{\Gamma}_{T+1}|\mathbf{\Gamma}_{T}) q_{\varphi}(\mathbf{\Gamma}_{1:T}|\mathbf{W}_{1:T}) d\mathbf{\Gamma}_{1:T}, \quad (15)$$

where we replaced the true (intractable) posterior with the approximate posterior $q_{\varphi}(\mathbf{\Gamma}_{1:T}|\mathbf{W}_{1:T})$, and where $\mathbf{\Gamma}_{t,d}$ labels the set $\{\boldsymbol{\xi}_t, \boldsymbol{\eta}_t, \boldsymbol{\zeta}_{t,d}, \mathbf{b}_{t,d}\}$ as before.

Models	0.5	0.6	0.7	0.8
WF-IBP	5.2	5.5	6.2	13.8
D-ETM	27.2	26.8	26.8	25.1
NDF-TM	**35.3**	**27.8**	**27.8**	**27.3**

Table 3: Percentage (wrt. the score of the static model) of the PPL-DC difference between dynamic and static models on the NEURIPS dataset, as the percentage of held-out words was increased from 50% to 80%. Higher is better.

We can now define the predictive log likelihood as

$$\text{P-NLL} = \mathbb{E}_{p_{\psi}(\mathbf{\Gamma}_{T+1}|\mathbf{\Gamma}_{T})}\mathbb{E}_{q_{\varphi}(\mathbf{\Gamma}_{1:T}|\mathbf{W}_{1:T})} \Big\{ \log p_{\psi}(\mathbf{W}_{T+1}|\mathbf{\Gamma}_{T+1}) \Big\}. \quad (16)$$

(2) To test generalization we use three metrics, namely *perplexity* (PPL) on document completion, *topic coherence* (TC) and *topic diversity* (TD). The document completion PPL is calculated on the second half of the documents in the test set, conditioned on their first half (Rosen-Zvi et al. 2012). The TC is calculated by taking the average pointwise mutual information between two words drawn randomly from the same topic (Lau, Newman, and Baldwin 2014) and measures the interpretability of the topic. In contrast, TD is the percentage of unique words in the top 25 words of all topics (Dieng, Ruiz, and Blei 2020). Note that one also often finds in the literature the *topic quality* metric (TQ), defined as the product of TC with TD.

Comparison With Baselines

The results on both P-NLL and PPL tasks are shown in Table 1. Both our models (NDF-TM and NDF-LT-TM) outperform all baselines on the completion PPL metric, on all the datasets. Similarly, our models outperform all baselines on both the TC and TD metrics, on all datasets, as shown in Table 2. These results (empirically) demonstrate that decoupling the topic activity from the topic proportion generically improves the performance of topic models on generalization tasks. In particular, we see that adding a non-linear transformation to the prior transition functions (Eqs. 1 and 4) overall improves the model performance (i.e. compare NDF-TM against NDF-LT-TM).

Regarding the prediction task we first notice that NDF-TM outperforms DTM-REP in all datasets. As explained in the

[6] Note that one is effectively performing a sequential Monte Carlo sample (Speekenbrink 2016), in which future steps are particles sampled from the posterior and propagated by the prior.

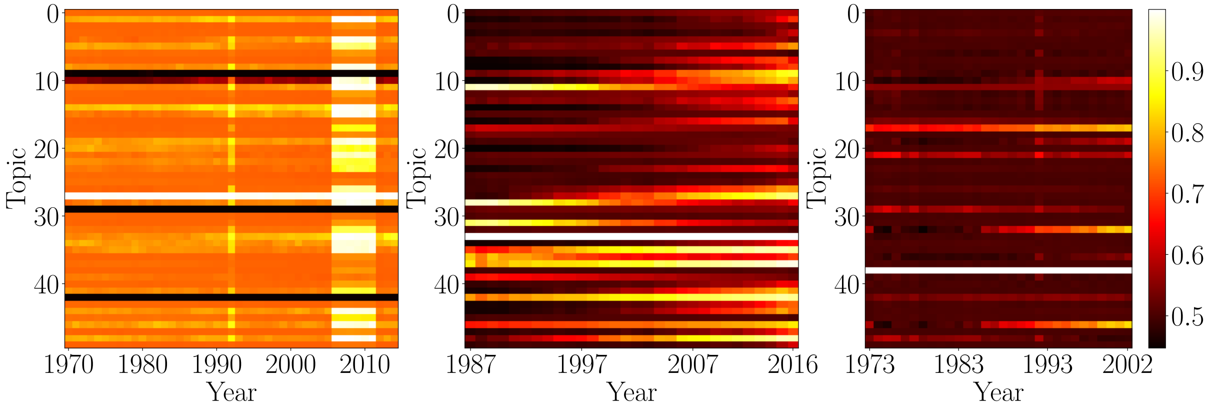

Figure 2: Average time-dependent topic activity $\mathbf{b}_{t,d}$ of all $K = 50$ topics in `NDF-TM` for UN (left), NIPS (middle) and ACL (right) datasets.

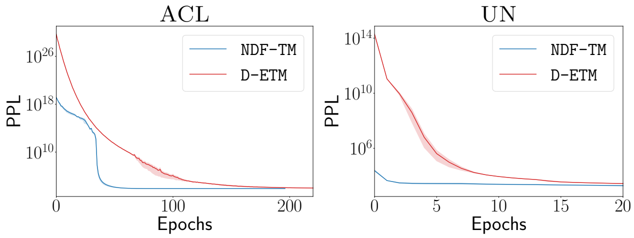

Figure 3: Learning curves for `D-ETM` and `NDF-TM` (100 topics each) on the ACL and UN evaluation datasets. The mean and the 2x std are obtained by training the models on 5 different random splits of the data.

Baselines subsection, `DTM-REP` and `NDF-TM` only differ in the topic proportion-activity coupling, from which one can infer that lifting the coupling explicitly helps when predicting the content of future documents. Yet `NDF-TM` only performs comparably to `D-ETM`, the strongest baseline, on this task. Note that `D-ETM` learns different embeddings for each topic *at each time step* (i.e. $K * T$ embeddings in total). One can argue that the flexibility to change the semantic content of topics as time evolves gives `D-ETM` the possibility to implicitly model rare yet relevant topics. In comparison, `NDF-TM` learns only K topic embeddings, and has only about $\alpha_0 K$ *active* embeddings (in average), at each time step. The number of parameter for both models is about the same however, because `NDF-TM` embeds the (fairly large) BoW vectors for the inference of its two global variables. Learning a single, global embedding for these BoW vectors would lower the number of needed parameters in `NDF-TM`, way below those needed in `D-ETM`, and we shall explore such an approach in the future. Nevertheless, in practice, and as shown in Figure 3, `NDF-TM` converges ∼2.8x faster than `D-ETM` in the ACL dataset (left figure). It also converges ∼2x faster than `D-ETM` in the UN dataset (right figure), and this is the worst case we have observed. Thus, ultimately, `NDF-TM` is more efficient than `D-ETM`.

We have also tried to compare against the non-parametric model of Perrone et al. (2017). In their work they evaluated the PPL-DC on four splits of a NEURIPS datasets.[7] The splits differ from each other on the percentage of held-out words used to define their test sets. Intuitively, the larger the percentage of held-out words, the more a dynamic topic model has to rely on its inferred temporal representations. The reported results seem however to be in a completely different scale from those we get (e.g. their simplest, static model yields PPL-DC values of the order of 1000, whereas our best models yield results twice as large). We therefore decided to compare the difference in performance between their dynamic `WF-IBP` model and their static baseline, against the difference in performance between our neural dynamic models and a static LDA model (`LDA-REP`), fitted with the reparametrization trick. Table 3 shows our results.

Qualitative Results

One of our main claims is that decoupling topic activity from topic proportion helps the model better describe sequentially collected data. We have seen above this is indeed the case from a quantitative point of view. Nevertheless, one could ask whether (or how) this decoupling is effectively taking place as time evolves. To study how the model encodes the temporal aspects of the data, we track the time evolution of both (i) the probability for topics to be active and (ii) the topic proportions. Figure 2 shows the first of these. Immediately we notice there is much more structure on the topic activities in both the NEURIPS and ACL datasets, as compared to the UN dataset. We can understand these findings by arguing (*a posteriori*) that NEURIPS and ACL feature more emergent and volatile topics (wrt. their activity) as compared to those characteristic of the UN dataset. Typically, (dynamic) topic models fitted on the UN dataset tend

[7] Note that this dataset is different from the NEURIPS dataset in our main experiments. We only used this new one to compare against Perrone et al. (2017). The dataset is available at https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015.

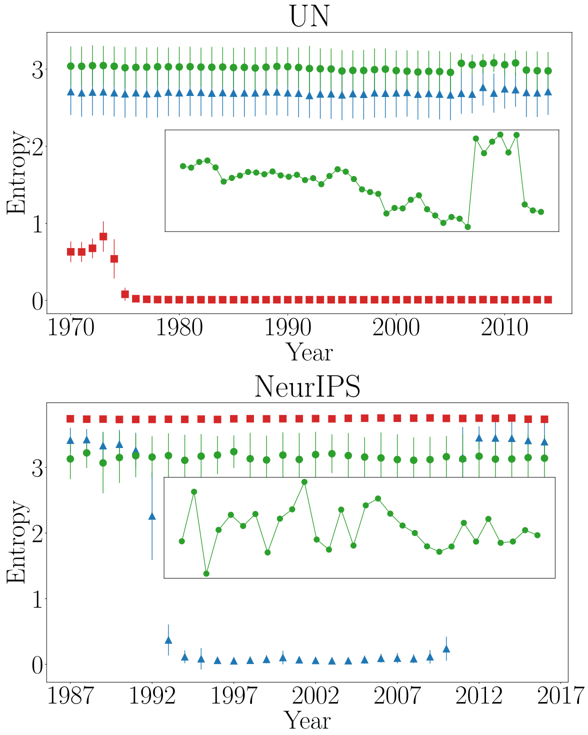

Figure 4: Entropy of topic distribution inferred by DTM−REP (squares), D−ETM (triangles) and NDF−TM (circles), averaged over documents as time evolves. Values shown with one standard deviation for both UN (above) and NEURIPS (below) datasets. Note that the maximum entropy value is $\log(K = 50) \approx 3.9$. The inset shows the details of the time-dependent topic-proportion entropy featured by NDF−TM. Note how the entropy decreases with time in the UN dataset (leaving aside the pick around the window 2005-2010) but fluctuates strongly for the (skewed) NEURIPS dataset.

to infer topics which circle about e.g. war, peace or climate. In contrast, topic models trained on, say, NEURIPS, generically infer more varied topics, ranging from e.g. Neural Networks and their training to Reinforcement Learning. See, for example, Table 6 in the supplementary material provided within the repository of our code, which shows six randomly sampled topics from each dataset as inferred by NDF−TM.

It is easy to imagine that the more generic topics in the UN dataset (like war, climate, etc) have reached some type of equilibrium and thus display overall a less skewed distribution over the document collection. If this were the case, topic models featuring the proportion-activity coupling would fit well the data by only inferring the more generic topics. Figure 4 shows the (Shannon) entropy of the topic distribution, averaged over documents as time evolves, as inferred by all models.[8] Note how the entropy inferred by DTM−REP (which features the proportion-activity coupling) for UN is close to zero, meaning that DTM−REP usually describes the documents with few topics, whereas for NEURIPS the entropy of the average topic distribution is close to its maximum value ($\log(K = 50) \approx 3.9$), meaning that it allocates almost equal probability for all K topics (that is, the model needs all topics to fit the data well), as expected for a skew topic distributions. In contrast, NDF−TM uses the additional Bernoulli variable sequences to redistribute the noise in the topic dynamics. Note also how the topic entropy of D−ETM is often similar to that of NDF−TM, meaning D−ETM does in fact implicitly lift the proportion-activity coupling.

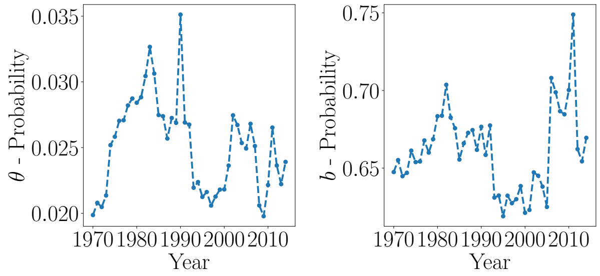

Figure 5: Evolution of topic proportion and activity probability for the topic *middle east* inferred from the UN dataset via NDF−TM.

Figure 5 shows our results for one topic inferred from the UN dataset, namely *middle east*. Note, for example, that the topic proportion for this topic peaks in the year 1990, which coincides with the Gulf War, to then drop right after. Such a drop is also reflected in the topic activity. Later, in 2011, the Syrian Civil War started. This event is captured by the topic activity which peaks at 2011, even though the topic proportion probability is decreasing. That is, even when the proportion of the *middle east* topic is low within the documents of that year, it must remain active to properly describe the data.

Conclusion

We have introduced the Neural Dynamic Focused Topic Model for sequentially collected data, which explicitly decouples the dynamic topic proportions from the topic activities through the addition of sequences of Bernoulli variables. We have shown that our approach consistently yields coherent and diverse topics, which correctly capture historical events. Future work includes using NDF−TM together with Variational Autoencoders for topic-guided text generation.

Acknowledgments

This research has been funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence, LAMARR22B.

César Ojeda is supported by Deutsche Forschungsgemeinschaft (DFG) - Project-ID 318763901 - SFB1294.

[8] The Shannon entropy of the topic distribution per document and time is defined here by $H_{t,d} = -\sum_i^K \theta_{t,d}^{(i)} \log \theta_{t,d}^{(i)}$, where $\theta_{t,d}^{(i)}$ is the ith component of $\boldsymbol{\theta}_{t,d}$.

References

Bamler, R.; and Mandt, S. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 380–389. JMLR. org.

Baturo, A.; Dasandi, N.; and Mikhaylov, S. J. 2017. Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Research & Politics*, 4(2): 2053168017712821.

Bird, S.; Dale, R.; Dorr, B. J.; Gibson, B.; Joseph, M. T.; Kan, M.-Y.; Lee, D.; Powley, B.; Radev, D. R.; and Tan, Y. F. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. springer.

Blei, D. M.; and Lafferty, J. D. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, 113–120.

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. L.; Hu, Y.; Mimno, D.; et al. 2017. *Applications of topic models*, volume 11. Now Publishers Incorporated.

Caron, F.; Davy, M.; and Doucet, A. 2007. Generalized Polya Urn for Time-Varying Dirichlet Process Mixtures. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'07, 33–40. Arlington, Virginia, USA: AUAI Press. ISBN 0974903930.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. In *Proceedings of the 22nd International Conference on World Wide Web*, 307–318. New York, NY, USA: Association for Computing Machinery. ISBN 9781450320351.

Dieng, A. B.; Ruiz, F. J.; and Blei, D. M. 2019. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*.

Dieng, A. B.; Ruiz, F. J.; and Blei, D. M. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8: 439–453.

Hida, R.; Takeishi, N.; Yairi, T.; and Hori, K. 2018. Dynamic and Static Topic Model for Analyzing Time-Series Document Collections. *CoRR*, abs/1805.02203.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Jagarlamudi, J.; Daumé III, H.; and Udupa, R. 2012. Incorporating Lexical Priors into Topic Models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 204–213. Avignon, France: Association for Computational Linguistics.

Jähnichen, P.; Wenzel, F.; Kloft, M.; and Mandt, S. 2018. Scalable generalized dynamic topic models. *arXiv preprint arXiv:1803.07868*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krishnan, R. G.; Shalit, U.; and Sontag, D. 2015. Deep Kalman Filters. arXiv:1511.05121.

Lau, J. H.; Newman, D.; and Baldwin, T. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 530–539.

Miao, Y.; Grefenstette, E.; and Blunsom, P. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2410–2419. JMLR. org.

Miao, Y.; Yu, L.; and Blunsom, P. 2016. Neural variational inference for text processing. In *International conference on machine learning*, 1727–1736.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Perrone, V.; Jenkins, P. A.; Spano, D.; and Teh, Y. W. 2017. Poisson random fields for dynamic feature models. *Journal of Machine Learning Research*, 18.

Rezaee, M.; and Ferraro, F. 2020. A Discrete Variational Recurrent Topic Model without the Reparametrization Trick. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 13831–13843. Curran Associates, Inc.

Rosen-Zvi, M.; Griffiths, T.; Steyvers, M.; and Smyth, P. 2012. The author-topic model for authors and documents. *arXiv preprint arXiv:1207.4169*.

Rudolph, M.; and Blei, D. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, 1003–1011.

Speekenbrink, M. 2016. A tutorial on particle filters. *Journal of Mathematical Psychology*, 73: 140–152.

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

Tang, J.; Meng, Z.; Nguyen, X.; Mei, Q.; and Zhang, M. 2014. Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis. In Xing, E. P.; and Jebara, T., eds., *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, 190–198. PMLR.

Wang, C.; Blei, D.; and Heckerman, D. 2012. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*.

Williamson, S.; Wang, C.; Heller, K. A.; and Blei, D. M. 2010. The IBP compound Dirichlet process and its application to focused topic modeling. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, 1151–1158.

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.; Phung, D.; Huynh, V.; Le, T.; and Buntine, W. 2021. Neural Topic Model via Optimal Transport. In *International Conference on Learning Representations*.

Zuo, Y.; Zhao, J.; and Xu, K. 2014. Word Network Topic Model: A Simple but General Solution for Short and Imbalanced Texts. arXiv:1412.5404.