Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
Survey Track

Topic Modelling Meets Deep Neural Networks: A Survey

He Zhao[1] , **Dinh Phung**[1,2] , **Viet Huynh**[1] , **Yuan Jin**[1] , **Lan Du**[1] , **Wray Buntine**[1]
[1]Department of Data Science and Artificial Intelligence, Monash University, Australia
[2]VinAI Research, Vietnam
{ethan.zhao, dinh.phung, viet.huynh, yuan.jin, lan.du, wray.buntine}@monash.edu

Abstract

Topic modelling has been a successful technique for text analysis for almost twenty years. When topic modelling met deep neural networks, there emerged a new and increasingly popular research area, *neural topic models*, with nearly a hundred models developed and a wide range of applications in neural language understanding such as text generation, summarisation and language models. There is a need to summarise research developments and discuss open problems and future directions. In this paper, we provide a focused yet comprehensive overview of neural topic models for interested researchers in the AI community, so as to facilitate them to navigate and innovate in this fast-growing research area. To the best of our knowledge, ours is the first review on this specific topic.

1 Introduction

A powerful technique for text analysis, topic modelling has enjoyed success in various applications in machine learning, natural language processing (NLP), and data mining for almost two decades. A topic model is applied to a collection of documents and aims to discover a set of latent topics, each of which describes an interpretable semantic concept. Bayesian probabilistic topic models (BPTMs) have been the most popular and successful series of models, with latent Dirichlet allocation (LDA) as representative. A BPTM usually specifies a probabilistic generative model that generates the data of a document with a structure of latent variables sampled from pre-specified distributions connected by Bayes' theorem. Topics are captured by these latent variables. Like other Bayesian models, the learning of a BPTM is done by a (Bayesian) inference process (e.g. variational inference (VI) or Monte Carlo Markov chain sampling).

Despite their success, conventional BPTMs started to show signs of fatigue in the era of big data and deep learning: **1)** Given a specific BPTM, its inference process usually needs to be customised accordingly and the inference complexity may grow significantly as the model complexity grows. Unfortunately, it is also hard to automate the design of the inference processes. **2)** The inference processes for conventional BPTMs can be hard to scale efficiently on large text collections or to leverage parallel computing facilities like GPUs. **3)** It is usually inconvenient to integrate BPTMs with other deep neural networks (DNNs) for joint training.

With the recent developments in DNNs and deep generative models, there has been an emerging research direction that aims to leverage DNNs to boost performance, efficiency, and usability of topic modelling, named *neural topic models* (NTMs). With appealing flexibility and scalability, NTMs have gained a huge research following, with more than a hundred models and variants developed to date. Moreover, NTMs have been used in important NLP tasks including text generation, document summarisation, and translation, areas to which conventional topic models are harder to apply. Therefore, it is important to properly summarise research developments, categorise existing approaches, identify remaining issues, and discuss open problems and future directions. To the best of our knowledge, a comprehensive review specifically focusing on NTMs has not been published. In this paper, we would like to fill this gap by providing an overview for interested researchers who want to develop new NTMs and/or to apply NTMs in their domains. The notable contributions of our paper can be summarised as follows: **1)** We propose a taxonomy of NTMs where we categorise existing models based on their backbone framework. **2)** We provide an informative discussion and overview of the background and evaluation methods for NTMs and conduct a focused yet comprehensive review, offering detailed comparisons of the variants of NTMs. **3)** We identify the limitations of existing methods and analyse possible future research directions for NTMs.

The rest of this paper is organised as follows. Section 2 introduces the background, definitions, and evaluations. Section 3 and 4 review NTMs with various backbone frameworks. The current challenges and future directions are discussed in Section 5.

2 Background, Definition, and Evaluation

2.1 Background and Definition

The most important idea of a topic model is modelling of three key entities: *document*, *word*, and *topic*.

Notations of Data. A topic model works on a corpus (i.e., a collection of documents), where a document, by its nature,

can be represented as a sequence of words, which can be denoted by a vector of natural numbers, $\boldsymbol{s} \in \mathbb{N}^L$, where L is the length of the document and $s_j \in \{1, \cdots, V\}$ is the index in the vocabulary (with the size of V) of the token for the j^{th} ($j \in \{1, \cdots, L\}$) word. A more common representation in topic modelling is the bag-of-words model, which represents a document by a vector of word counts, $\boldsymbol{b} \in \mathbb{Z}_{\geq 0}^V$, where b_v indicates the occurrences of the vocabulary token $v \in \{1, \cdots, V\}$ in the document. One can readily obtain $\boldsymbol{b}$ for a document from its word sequence vector $\boldsymbol{s}$.

Notations of Latent Variables. A central concept is a *topic*, which is usually interpreted as a cluster of words, describing a specific semantic meaning. A topic is or can be normalised into a distribution over the tokens in the vocabulary, named *word distribution*, $\boldsymbol{t} \in \Delta^V$, where Δ^V is a V dimensional simplex and t_v indicates the weight or relevance of token v under this topic. Usually, a document's semantic content is assumed to be captured or generated by one or more topics shared across the corpus. Therefore, a document is commonly associated with a distribution (or a vector that can be normalised into a distribution) over K ($K \geq 1$) topics, named *topic distribution*, $\boldsymbol{z} \in \Delta^K$, where z_k indicates the weight of the k^{th} topic for this document. We further use $\mathcal{D}$, $\mathcal{Z}$, and $\mathcal{T}$ to denote the corpus with all the document data, the collections of topic distributions of all the documents, and the collections of word distributions of all the topics, respectively.

Notations of Architectures and Learning. With these notations, the task for a topic model is to learn the latent variables of $\mathcal{Z}$ and parameters of $\mathcal{T}$ from the observed data $\mathcal{D}$. More formally, a topic model learns a projection parameterised by θ from a document's data to its topic distribution: $\boldsymbol{z} = \theta(\boldsymbol{b})$ and a set of global variables for the word distributions of the topics: $\mathcal{T}$. To learn these parameters, one can generate or reconstruct a document's BoW data from its topic distribution, which is modelled by another projection parameterised by ϕ: $\tilde{\boldsymbol{b}} = \phi(\boldsymbol{z}, \mathcal{T})$. Note that the majority of topic models belong to the category of probabilistic generative models, where $\boldsymbol{z}$ and $\boldsymbol{b}$ are latent and observed random variables assumed to be generated from certain distributions respectively. The projection from the latent variables to the observed ones is named the generative process, which we further denote as: $\tilde{\boldsymbol{b}} \sim p_\phi^b(\boldsymbol{z}, \mathcal{T})$ where $\boldsymbol{z}$ is sampled from the prior distribution $\boldsymbol{z} \sim p^z$. While the inverse projection is named the inference process, denoted as $\boldsymbol{z} \sim q_\theta^z(\boldsymbol{b})$, where q^z is the posterior distribution of $\boldsymbol{z}$. For NTMs, these probabilities are typically parameterised by deep neural networks.

2.2 Evaluation

It is still challenging to comprehensively evaluate and compare the performance of topic models including NTMs. Based on the nature and applications of topic models, the commonly-used metrics are as follows.

Predictive accuracy. It has been common to measure the log-likelihood of a model on held-out test documents, i.e., the predictive accuracy. A more popular metric based on log-likelihood is perplexity, which captures how surprised a model is of new (test) data and is inversely proportional to average log-likelihood per word. Although log-likelihood or perplexity gives a straight numerical comparison between models, there remain issues: **1)** As topic models are not for predicting unseen data but learning interpretable topics and representations of seen data, predictive accuracy does not reflect the main use of topic models. **2)** Predictive accuracy does not capture topic quality. Predictive accuracy and human judgement on topic quality are often not correlated [Chang *et al.*, 2009], and even sometimes slightly anti-correlated. **3)** The estimation of the predictive probability is usually intractable for Bayesian models and different papers may apply different sampling or approximation techniques [Wallach *et al.*, 2009; Buntine, 2009]. For NTMs, the computation of log-likelihood is even more inconsistent, making it harder to compare the results across different papers.

Topic Coherence. Experiments show topic coherence (TC) computed with the coherence between a topic's most representative words (e.g, top 10 words) is in line with human evaluation of topic interpretability [Lau *et al.*, 2014]. As various formulations have been proposed to compute TC, we refer readers to [Röder *et al.*, 2015] for more details. Most formulations require to compute the general coherence between two words, which are estimated based on word co-occurrence counts in a reference corpus. Regarding TC, we have the following remarks: **1)** The ranking of TC scores may vary under different formulations. Therefore, it is encouraged to report TC scores of different formulations or report the average score. **2)** The choice of the reference corpus can also affect the TC scores, due to the change of lexical usage, i.e, the shift of word distribution. For example, computing TC for a machine learning paper collection with a tweet dataset as reference may generate inaccurate results. Popular choices of the reference corpus are the target corpus itself or an external corpus such as a large dump of Wikipedia. **3)** To exclude less interpretable "background" topics, one can select the topics (e.g., top 50%) with the highest TC or the largest proportions and report the average score over those selected topics [Zhao *et al.*, 2018a] or to vary the proportion of the selected topics (e.g, from 10% to 100%) and plot TC score at each proportion [Zhao *et al.*, 2021].

Topic Diversity. Topic diversity (TD), as its name implies, measures how diverse the discovered topics are. It is preferable that discovered topics describe different semantic topical meanings. Specifically, [Dieng *et al.*, 2020] defines topic diversity to be the percentage of unique words in the top 25 words.

Downstream Application Performance. The topic distribution $\boldsymbol{z}$ of a document learned by a topic model can be viewed as the semantic representation of the document, which can be used in document classification, clustering, retrieval, visualisation, and elsewhere. For document classification, one can train a classification model with the topic distributions learned by a topic model as features and report the classification performance to compare different topic models. Document clustering can be conducted by two strategies: **1)** Similar to classification, one can perform a clustering model (e.g. K-means with different numbers of clusters) on

the topic distributions, such as in [Zhao *et al.*, 2021]; **2)** Alternatively, topics can actually be viewed as clusters of documents. Thus, one can use the most significant topic of a document (i.e., the topic with the largest weight in the topic distribution) as the cluster assignment, such as in [Nguyen *et al.*, 2015]. For document retrieval, we can use the distance of the topic distributions of two documents as their semantic distance and report retrieval accuracy as a metric of topic modelling [Larochelle and Lauly, 2012]. For qualitative analysis, a straight-forward way is to plot the most significant words of topics. Recently, [Doogan and Buntine, 2021] shows that it can be more insightful to show and analyse the typical documents for a topic.

3 Neural Topic Models with Amortised Variational Inference

The recent success of deep generative models such as variational autoencoders (VAEs) and amortised variational inference (AVI) has shed light on extending the generative process and amortising the inference process of BPTMs, which is the most popular framework for NTMs. We name this series of models VAE-NTMs. The basic framework of a VAE follows the description in Section 2.1, where $\boldsymbol{b}$ and $\boldsymbol{z}$ are the observed and latent variables respectively and the generative and inference processes are modelled by the DNN-based decoder and encoder respectively. Following [Kingma and Welling, 2014; Rezende *et al.*, 2014], one can learn a VAE model by maximising the Evidence Lower BOund (ELBO) of the marginal likelihood of the BoW data $\boldsymbol{b}$ in terms of θ, ϕ, and $\mathcal{T}$: $\mathbb{E}_{\boldsymbol{z}\sim q^z}\left[\log \mathrm{p}(\boldsymbol{b} \mid \boldsymbol{z})\right] - \mathbb{KL}\left[q^z \parallel p^z\right]$, where the RHS term is the Kullback-Leiber (KL) divergence. To compute/estimate gradients, tricks like reparameterisations are usually used to back-propagate gradients through the expectation in the LHS term and approximations are applied when the analytical form of the KL divergence is unavailable.

To adapt the VAE framework for topic modelling, there are two key questions to be answered: **1)** Different from other applications, the input data of topic modelling has its unique properties, i.e., $\boldsymbol{b}$ is a high-dimensional, sparse, count-valued vector and $\boldsymbol{s}$ is a variable-length sequential data. How to deal with such data is the first question for designing a VAE topic model. **2)** Interpretability of topics is extremely important in topic modelling. When it comes to a VAE model, how to explicitly or implicitly incorporate the word distributions of topics (i.e., $\mathcal{T}$) to interpret the latent representations or each dimension remains another question. [Miao *et al.*, 2016] proposes the first answers to the above questions, where the decoder is developed by specifying the data distribution p^b as: $p^b := \text{Multi}\left(\text{softmax}\left(\mathbf{T}^T\boldsymbol{z} + \boldsymbol{c}\right)\right)$. Here $\boldsymbol{z} \in \mathbb{R}^K$ models the topic distribution of a document, $\mathbf{T} \in \mathbb{R}^{K\times V}$ models the words distributions of the topics, and $\boldsymbol{c} \in \mathbb{R}^V$ is the bias. That is to say, $\phi := \{\boldsymbol{c}\}$[1] and $\mathcal{T} := \{\mathbf{T}\}$. For the encoder which takes $\boldsymbol{b}$ as input and outputs (the samples of) $\boldsymbol{z}$, the paper follows the original VAE: $p^z := \mathcal{N}(\mathbf{0}, \text{diag}_K(\mathbf{1}))$; $q^z := \mathcal{N}(\boldsymbol{\mu}, \text{diag}_K(\boldsymbol{\sigma}^2))$, where $\boldsymbol{\pi} = \theta_0(\boldsymbol{b})$, $\boldsymbol{\mu} = \theta_1(\boldsymbol{\pi})$, and $\log \boldsymbol{\sigma} = \theta_2(\boldsymbol{\pi})$. Here, $\theta := \{\theta_0, \theta_1, \theta_2\}$, all of which are multi-layer perceptrons (MLPs). To better address the above questions, various configurations of the prior distribution p^z, data distribution p^b, posterior distribution q^z, as well as different architectures of the decoder ϕ, encoder θ, word distributions of the topics $\mathcal{T}$, have been proposed for VAE-NTMs.

[1]With a slight abuse of notation, we use θ and ϕ to denote the projections or the parameters of the projections.

3.1 Variants of Distributions

Given the knowledge and experience of BPTMs, $\boldsymbol{z}$'s prior plays an important role in the quality of topics and document representations in topic models. Thus, various constructions of the prior distributions and their corresponding posterior distributions have been proposed for VAE-NTMs, aiming to be better alternatives to the normal distributions used in the original models.

Variants of Prior Distributions for $\boldsymbol{z}$. Note that the application of Dirichlet is one of the key successes of LDA for encouraging topic smoothness and sparsity. For VAE-NTMs, one can apply: $p^z := \text{Dir}(\alpha_0)$ and $q^z := \text{Dir}(\theta(\boldsymbol{b}))$. However, it is difficult to develop an effective reparameterisation function (RF) for Dirichlet, making it hard to compute the gradient of the expectation in ELBO. Therefore, various approximations have been proposed. For example, [Srivastava and Sutton, 2017] uses the Laplace approximation, where Dirichlet samples are approximated by these sampled from a logistic normal distribution, whose mean and co-variance are specifically configured. Recall that the Dirichlet distribution can be simulated by normalising gamma variables. Although the gamma distribution still does not have non-central differentiable RF, it is easier to approximate. Several works have been proposed in this line, such as using the Weibull distribution as the approximation of gamma in [Zhang *et al.*, 2018], approximating the cumulative distribution function of gamma with an auxiliary uniform variable in [Joo *et al.*, 2020], and leveraging the proposal function of a rejection sampler of the gamma distribution as the RF in [Burkhardt and Kramer, 2019]. Recently, [Tian *et al.*, 2020] proposes to tackle this challenge by using the so-called rounded RF, which approximates Dirichlet samples by those drawn from the rounded posterior distribution. Other than Dirichlet, [Miao *et al.*, 2017] introduces a Gaussian softmax (GSM) function in the encoder: $q^z := \text{softmax}\left(\mathcal{N}(\boldsymbol{\mu}, \text{diag}_K(\boldsymbol{\sigma}^2))\right)$ and [Silveira *et al.*, 2018] proposes to use a logistic-normal mixture distribution for the prior of $\boldsymbol{z}$. To further enhance the sparsity in $\boldsymbol{z}$, [Lin *et al.*, 2019] introduces to use the sparsemax function to replace the softmax in GSM.

Nonparametric Prior for $\boldsymbol{z}$. Bayesian Nonparametrics such as the Dirichlet processes, Indian Buffet Processes, and gamma processes have been successfully applied in Bayesian topic modelling, enabling to automatically infer the prior proportion and number of topics (i.e., K), e.g., in [Teh *et al.*, 2006; Williamson *et al.*, 2010; Buntine and Mishra, 2014; Zhou *et al.*, 2016; Zhao *et al.*, 2018b]. As a flexible construction of Dirichlet processes, the stick-breaking process (SBP) is able to generate probability vectors with infinite dimensions, which has been used to the prior of $\boldsymbol{z}$ in VAE-NTMs. Given $\boldsymbol{z} \sim \text{SBP}(\alpha_0)$, we have $z_1 = v_1$ and $z_k = v_k \prod_{j<k}(1 - v_j)$ for $k > 1$, where $v_k \sim \text{Beta}(1, \alpha_0)$. This

procedure can be viewed as iteratively breaking a length-one stick into multiple ones and the k^{th} iteration breaks the stick at the point of v_k. Although not for NTMs, [Nalisnick and Smyth, 2017] uses SBP to generate $\boldsymbol{z}$ for VAEs, where its VI is done by various approximations to the beta distribution of v_k with truncation. [Ning *et al.*, 2020] adapts this SBP construction for VAE-NTMs and also proposes to impose an SBP on the corpus level, which serves as the prior for the document-level SBP, forming into a hierarchical model. In [Miao *et al.*, 2017], the break points v_k are generated from a posterior modelled by a recurrent neural network (RNN) with normal noises as input, making the model able to automatically infer K in a truncation-free manner. Recently, [Wu *et al.*, 2020a] uses the (truncated) gamma negative binomial process to generate discrete vectors for $\boldsymbol{z}$ (i.e. each entry of $\boldsymbol{z}$ is equivalently generated by an independent Poisson distribution), which gives the model certain ability to be nonparametric.

Variants of Data Distribution p^b. In addition to manipulating distributions on $\boldsymbol{z}$, [Zhao *et al.*, 2020] proposes to replace the multinomial data distribution used in other NTMs with the negative-binomial distribution to capture overdispersion, making the model more robust: $\boldsymbol{b} \sim \text{NB}(\phi_0(\boldsymbol{z}), \phi_1(\boldsymbol{z}))$, where two separate decoders ϕ_0 and ϕ_1 are proposed to generate the two parameters of the negative-binomial distribution from $\boldsymbol{z}$.

Variants of Word Distributions $\mathcal{T}$. Conventionally, the collection of the word distributions of the topics $\mathcal{T}$ is a $K \times V$ matrix, i.e., $\mathbf{T} \in \mathbb{R}^{K \times V}$ with KV free parameters. In BPTMs, it has been popular to factorise the matrix into a product of topic and word embeddings, meaning that the relevance between a topic and a word is captured by their distance in the embedding space [Zhao *et al.*, 2017a]. This construction has been studied in NTMs, e.g., in [Jung and Choi, 2017; Dieng *et al.*, 2020; Ding *et al.*, 2018].

3.2 Correlated and Structured Topics

Topics discovered by conventional topic models like LDA are usually independent. An important research direction is to explicitly capture topic correlations (e.g. pairwise relations between topics) or structures (e.g. tree structures of topics), which has been studied in NTMs as well. Following the framework of VAE with Householder flow that enables to draw $\boldsymbol{z}$ from the normal posterior with a non-diagonal covariance matrix, [Liu *et al.*, 2019] develops a more efficient centralised transformation flow for NTMs, which is able to discover pairwise topic correlations by the covariance matrix. In terms of discovering tree-structured topics, [Isonuma *et al.*, 2020] introduces to generate a series of topics from the root to the leaf of a topic tree with a doubly-recurrent neural network [Alvarez-Melis and Jaakkola, 2017]. When applied in topic modelling, the gamma belief network (GBN) [Zhou *et al.*, 2016] can be viewed as a Bayesian model that also discovers three-structured topics, whose inference is done by Gibbs sampling. [Zhang *et al.*, 2018] introduces the NTM counterpart of GBN, which leverages AVI as the inference process and significantly improves the test time of GBN. [Esmaeili *et al.*, 2019] proposes an structured VAE-NTM that discovers topics with respect to different aspects, specialising in modelling user reviews.

3.3 NTMs with Meta-data

Conventionally, topic models learn from documents in an unsupervised way. However, documents are usually associated with rich sets of meta-data on both document and word levels, such as document labels, authorship, and pre-trained word embeddings, which can be used to improve topic quality or document representation quality [Zhao *et al.*, 2017b] for supervised tasks (e.g., accuracy of predicting document meta-data). [Card *et al.*, 2018] proposes a VAE-NTM that is able to incorporate various kinds of meta-data, where the BoW data $\boldsymbol{b}$ of a document and its labels (e.g., sentiment) are generated with a joint process conditioned on the document's covariates (e.g., publication year) in the decoder and the encoder generates $\boldsymbol{z}$ by conditioning on all types of data of the document: BoW, covariates, and labels. Instead of specifying the generative model as a directed network as in most of topic models, [Korshunova *et al.*, 2019] introduces the logistic LDA model whose generative process can be viewed as an undirected graph. In addition to the BoW data, a document's label is also an observed variable in the graph. Following a few assumptions of factorisation in the generative process, the paper manually specifies the complete conditional distributions in the graph with the interactions between the latent variables captured by neural networks. The inference is done by the mean-field VI and $\boldsymbol{z}$ in the model is further trained to be more discriminative for the classification of labels. Given a set of documents with labels, [Wang and Yang, 2020] uses a VAE-NTM to model a document's BoW data and an RNN classifier to predict a document's label based on its sequential data in a joint training process. The paper combines the two models by introducing an attention mechanism in the RNN which takes documents' topics into account. [Bai *et al.*, 2018] proposes to incorporate relational graphs (e.g. citation graph) of documents into NTMs, where the topic distributions of two document are fed into a neural network to predict whether they should be connected.

3.4 NTMs for Short Texts

Texts generated on the internet (e.g., tweets, news headlines and product reviews) can be short, meaning that each individual document contains insufficient word co-occurrence information. This results in degraded performance for both BPTMs and NTMs. To tackle this issue, one can limit a model's capacity and to enhance the contextual information of short texts. [Zeng *et al.*, 2018] proposes a combination of an NTM and a memory network for short text classification in a similar spirit to [Wang and Yang, 2020]. The main difference is the memory network instead of RNN is responsible for classification, which is informed by the topic distributions learned by the NTM. To enhance the contextual information of short documents, [Zhu *et al.*, 2018] proposes an NTM whose encoder is a graph neural network (GNN) taking the biterms graph of the words in sampled documents as inputs and outputting the topic distribution for the whole corpus. The model also learns a decoder that reconstructs the input biterms graph. Despite the novel idea, the model

might not be able to generate the topic distribution of an individual document. To limit a short document to focus on several salient topics, [Lin *et al.*, 2020] introduces to use the Archimedean copulas to regularise the discreteness of topic distributions for short texts. [Wu *et al.*, 2020b] introduces an NTM with vector quantisation over $\boldsymbol{z}$, i.e., a document's topic distribution can only be one vector in the learned dictionary in the vector quantisation process. In addition to maximising the likelihood of the input documents, the paper introduces to minimise the likelihood of the negatively-sampled "fake documents". Although not directly addressing the short text problem for topic modelling, [He *et al.*, 2018] introduces NTMs for modelling microblog conversations, by leveraging their unique meta data and structures.

3.5 Sequential NTMs

The flexibility of VAE-NTMs enables to leverage various neural network architectures for the encoder and decoder. With the help of sequential networks like RNNs, unlike other NTMs working with BoW data (i.e., $\boldsymbol{b}$), sequential NTMs (SNTMs) usually take sequences of words of documents (i.e., $\boldsymbol{s}$) and are able to capture the orders of words, sentences, and topics. [Nallapati *et al.*, 2017] proposes an SNTM working with $\boldsymbol{s}$, which samples a topic for each sentence of an input document according to $\boldsymbol{z}$ and then generates the word sequence of the sentence with an RNN conditioned on the sentence's topic. Note that $\boldsymbol{z}$ is attached to a document and shared across all its sentences. In [Zaheer *et al.*, 2017], given $\boldsymbol{s}$, a word's topic is conditioned on its previous word's and this order dependency is captured by a long short-term memory (LSTM) model. At the similar period of time, [Dieng *et al.*, 2017] independently proposes an SNTM whose generative process is similar to [Zaheer *et al.*, 2017], with an additional variable modelling stop words and several variants in the inference process. Recently, [Panwar *et al.*, 2020] proposes to use an LSTM with attentions as the encoder taking $\boldsymbol{s}$ as input, where the attention incorporates topical information with a context vector that is constructed by topic embeddings and document embeddings. [Rezaee and Ferraro, 2020] introduces an SNTM that is related to [Dieng *et al.*, 2017], where instead of marginalising out the discrete topic assignments, the paper proposes to generate them from an RNN model. This helps to avoid using reparameterisation tricks in the variational inference.

3.6 NTMs with Pre-trained Language Models

Recently, pre-trained transformer-based language models such as BERT are becoming ubiquitous in NLP. Pre-trained on large corpora, such models usually have a fine-grained ability to capture aspects of linguistic context, which can be partially represented by contextual word embeddings. These contextual word embeddings can provide richer context information than BoW or sequential data, which has been recently used to assist the training of topic models. Instead of using the BoW or sequential data of a document as the input of the encoder, [Bianchi *et al.*, 2020] proposes to use the document embedding vector generated by Sentence-BERT [Reimers and Gurevych, 2019] and to keep the remaining part of an NTM the same as [Srivastava and Sutton, 2017]. [Thompson and Mimno, 2020] shows that the clusters obtained by performing clustering algorithms (e.g., Kmeans) on the contextual word embeddings generated by various pre-trained models such as BERT and GPT-2 can be interpreted as topics, similar to those discovered by LDA. Having similar ideas with [Zeng *et al.*, 2018; Wang and Yang, 2020], [Chaudhary *et al.*, 2020] proposes to combine an NTM with a fine-tuned BERT model by concatenating the topic distribution and the learned BERT embedding of a document as the features for document classification. [Hoyle *et al.*, 2020] proposes an NTM learned by distilling knowledge from a pre-trained BERT model. Specifically, given a document, the BERT model generates the predicted probability for each word then the paper introduces to average those probabilities to generate a pseudo BoW vector for the document. An NTM following [Card *et al.*, 2018] is used to reconstruct both the actual and pseudo BoW data.

4 NTMs based on Other Frameworks

Besides VAE-NTMs, there are other frameworks for NTMs that also draw research attention.

NTMs based on Autoregressive Models. VAE-NTMs gained popularity after VAEs were invented. Before that, NTMs based on the autoregressive framework had been studied. Specifically, [Larochelle and Lauly, 2012] proposes an autoregressive NTM, named DocNADE, similar to the spirit of RNNs, where the predictive probability of a word in a document is conditioned on its hidden state, which is further conditioned on the previous words. A hidden unit can be interpreted as a topic and a document's hidden states capture its topic distribution. The learning is done by maximising the likelihood of the input documents. Recently, [Gupta *et al.*, 2019a] extends DocNADE by introducing a structure similar to the bi-directional RNN, which allows to model bi-directional dependencies between words. [Gupta *et al.*, 2019b] combines DocNADE with an LSTM for incorporating external knowledge. [Gupta *et al.*, 2020] extends DocNADE into the life long learning settings.

NTMs based on Generative Adversarial Nets. Besides VAEs, generative adversarial networks (GANs) are another popular series of deep generative models. Recently, there are a few attempts on adapting the GAN framework for topic modelling. [Wang *et al.*, 2019] proposes a GAN generator that takes a random sample of the Dirichlet distribution as a topic distribution $\tilde{\boldsymbol{z}}$ and generates the word distributions of a "fake" document conditioning on $\tilde{\boldsymbol{z}}$. A discriminator is introduced to distinguish between generated word distributions and real word distributions obtained by normalising the TF-IDF vectors of real documents. Although the proposed model is able to discover interpretable topics, it cannot learn topic distributions for documents. To address this issue, [Wang *et al.*, 2020] introduces an additional encoder that learns $\boldsymbol{z}$ for a given document. Moreover, $\boldsymbol{z}$ is concatenated with the word distribution of a document as a real datum and $\tilde{\boldsymbol{z}}$ is concatenated with the generated word distribution as a fake datum. The discriminator is designed to distinguish between the real and fake ones. [Hu *et al.*, 2020] further extends the above model with a CycleGAN framework.

NTMs based on Graph Neural Networks. Instead of viewing a document as a sequence or bag of words, one can consider the graph presentations of a corpus of documents. This perspective enables leveraging a variety of GNNs to discover latent topics. As discussed in Section 3.4, [Zhu *et al.*, 2018] views a collection of documents as a biterm word graph. While [Yang *et al.*, 2020; Zhou *et al.*, 2020] model a corpus by a bipartite graph with documents and words as two separate parties and connected by the occurrences of words in documents. For the former, it directly uses the word occurrences of documents as the weights of the connections between them and for the latter, it uses TF-IDF values instead.

Other NTMs. In addition to the above frameworks, other kinds of NTMs have also been developed. An NTM is developed in [Cao *et al.*, 2015] that takes n-gram embeddings (obtained from word embeddings) and a document index as input and then predicts whether an n-gram is in the document. [Chen and Zaki, 2017] proposes an autoencoder model for NTMs where the neurons in the hidden layer of the autoencoder compete with each other, focusing them to be specialised in recognising specific data patterns. [Peng *et al.*, 2018] proposes an NTM based on matrix factorisation. [Gui *et al.*, 2019] proposes a reinforcement learning framework for NTMs, where the encoder and decoder of an NTM are kept. In addition, an agent takes actions to select the topical-coherent words from a document and uses the selected words as the input document for the encoder. The reward to the agent is the topic coherence of the reconstructed document from the decoder. [Nan *et al.*, 2019] adapts the framework of Wasserstein auto-encoders (WAEs), which minimises the Wasserstein distance between reconstructed documents from the decoder and real documents, similarly to VAE-NTMs. Recently, topic models based optimal transport have been developed, such as in [Huynh *et al.*, 2020]. [Zhao *et al.*, 2021] introduces an NTM based on optimal transport, which minimises the optimal transport distance between the topic distribution learned by an encoder and the word distribution of a document.

5 Discussion

This paper is the first survey paper focusing on the specific area of neural topic models, which is the most popular research trend of topic modelling in the deep learning era. Due to the appealing flexibility, effectiveness, and efficiency, NTMs show a promising potential in a range of applications.

After providing an overview of existing approaches of NTMs, we in this section would like to discuss the following challenges and opportunities for NTMs.

Better evaluation. As stated in Section 2.2, evaluation of topic models is challenging. This is mainly because there has not been a unified system of evaluation metrics, and indeed some metrics are not always appropriate, making the comparisons across different NTMs harder due to the variety of frameworks, architectures and datasets. For example, VAE-NTMs calculate perplexity using the ELBO, attached to the models with variational inference, which cannot be compared with models without ELBO. Also for topic coherence and downstream performance, the evaluation processes, metrics, settings usually vary in different papers. A topic model should be evaluated with comprehensive metrics, including those on topic quality, predictive accuracy, document representation, and downstream applications. It could be tendentious to only use one kind of metric (e.g., topic coherence), which can reflect just one aspect of a model. Therefore, unified platforms and benchmarks for NTMs are needed.

Richer architectures and applications. Compared to BPTMs, NTMs offer better flexibility for representing topic distributions for documents and word distributions for topics. Particularly, projecting documents, topics, and words into a unified embedding space transforms the thinking of the relationships between the three. Given this flexibility, NTMs are expected to get integrated with the most recent neural architectures and play a unique role in richer applications.

More external knowledge. With the development of topic models including NTMs, people have not stopped seeking to leverage external knowledge to help the learning, from document meta-data to pre-trained word embeddings. Recently-proposed pre-trained language models (e.g., BERT) provide more advanced, finer-grained, and higher-level representations of semantic knowledge (e.g., contextual word embeddings over global embeddings), which can be leveraged in NTMs to boost performance. Although the marriage between NTMs and language models is still an emerging area, we expect to see more developments in this important direction.

References

[Alvarez-Melis and Jaakkola, 2017] David Alvarez-Melis and Tommi S Jaakkola. Tree-structured decoding with doubly-recurrent neural networks. In *ICLR*, 2017.

[Bai *et al.*, 2018] Haoli Bai, Zhuangbin Chen, Michael R Lyu, Irwin King, and Zenglin Xu. Neural relational topic models for scientific article analysis. In *CIKM*, 2018.

[Bianchi *et al.*, 2020] Federico Bianchi, Silvia Terragni, and Dirk Hovy. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv*, 2020.

[Buntine and Mishra, 2014] Wray L Buntine and Swapnil Mishra. Experiments with non-parametric topic models. In *SIGKDD*, pages 881–890, 2014.

[Buntine, 2009] Wray Buntine. Estimating likelihoods for topic models. In *ACML*, 2009.

[Burkhardt and Kramer, 2019] Sophie Burkhardt and Stefan Kramer. Decoupling sparsity and smoothness in the Dirichlet variational autoencoder topic model. *JMLR*, 2019.

[Cao *et al.*, 2015] Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. A novel neural topic model and its supervised extension. In *AAAI*, 2015.

[Card *et al.*, 2018] Dallas Card, Chenhao Tan, and Noah A Smith. Neural models for documents with metadata. In *ACL*, 2018.

[Chang *et al.*, 2009] Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M Blei. Reading tea leaves: How humans interpret topic models. In *NeurIPS*, 2009.

[Chaudhary *et al.*, 2020] Yatin Chaudhary, Pankaj Gupta, Khushbu Saxena, Vivek Kulkarni, Thomas Runkler, and Hinrich Schütze. TopicBERT for energy efficient document classification. In *EMNLP*, 2020.

[Chen and Zaki, 2017] Yu Chen and Mohammed J Zaki. KATE: K-competitive autoencoder for text. In *SIGKDD*, 2017.

[Dieng *et al.*, 2017] Adji B Dieng, Chong Wang, Jianfeng Gao, and John Paisley. TopicRNN: A recurrent neural network with long-range semantic dependency. In *ICLR*, 2017.

[Dieng *et al.*, 2020] Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *TACL*, 2020.

[Ding *et al.*, 2018] Ran Ding, Ramesh Nallapati, and Bing Xiang. Coherence-aware neural topic modeling. In *EMNLP*, 2018.

[Doogan and Buntine, 2021] Caitlin Doogan and Wray Buntine. Topic model or topic twaddle? re-evaluating semantic interpretability measures. In *NAACL*, 2021.

[Esmaeili *et al.*, 2019] Babak Esmaeili, Hongyi Huang, Byron Wallace, and Jan-Willem van de Meent. Structured neural topic models for reviews. In *AISTATS*, 2019.

[Gui *et al.*, 2019] Lin Gui, Jia Leng, Gabriele Pergola, Ruifeng Xu, and Yulan He. Neural topic model with reinforcement learning. In *EMNLP-IJCNLP*, 2019.

[Gupta *et al.*, 2019a] Pankaj Gupta, Yatin Chaudhary, Florian Buettner, and Hinrich Schütze. Document informed neural autoregressive topic models with distributional prior. In *AAAI*, 2019.

[Gupta *et al.*, 2019b] Pankaj Gupta, Yatin Chaudhary, Florian Buettner, and Hinrich Schütze. Texttovec: Deep contextualized neural autoregressive topic models of language with distributed compositional prior. In *ICLR*, 2019.

[Gupta *et al.*, 2020] Pankaj Gupta, Yatin Chaudhary, Thomas Runkler, and Hinrich Schuetze. Neural topic modeling with continual lifelong learning. In *ICML*, 2020.

[He *et al.*, 2018] Ruifang He, Xuefei Zhang, Di Jin, Longbiao Wang, Jianwu Dang, and Xiangang Li. Interaction-aware topic model for microblog conversations through network embedding and user attention. In *COLING*, 2018.

[Hoyle *et al.*, 2020] Alexander Miserlis Hoyle, Pranav Goel, and Philip Resnik. Improving neural topic models using knowledge distillation. In *EMNLP*, 2020.

[Hu *et al.*, 2020] Xuemeng Hu, Rui Wang, Deyu Zhou, and Yuxuan Xiong. Neural topic modeling with cycle-consistent adversarial training. In *EMNLP*, 2020.

[Huynh *et al.*, 2020] Viet Huynh, He Zhao, and Dinh Phung. OTLDA: A geometry-aware optimal transport approach for topic modeling. *NeurIPS*, 2020.

[Isonuma *et al.*, 2020] Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. Tree-structured neural topic model. In *ACL*, 2020.

[Joo *et al.*, 2020] Weonyoung Joo, Wonsung Lee, Sungrae Park, and Il-Chul Moon. Dirichlet variational autoencoder. *Pattern Recognition*, 2020.

[Jung and Choi, 2017] Namkyu Jung and Hyeong In Choi. Continuous semantic topic embedding model using variational autoencoder. *arXiv*, 2017.

[Kingma and Welling, 2014] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

[Korshunova *et al.*, 2019] Iryna Korshunova, Hanchen Xiong, Mateusz Fedoryszak, and Lucas Theis. Discriminative topic modeling with logistic LDA. In *NeurIPS*, 2019.

[Larochelle and Lauly, 2012] Hugo Larochelle and Stanislas Lauly. A neural autoregressive topic model. *NeurIPS*, 2012.

[Lau *et al.*, 2014] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *ACL*, 2014.

[Lin *et al.*, 2019] Tianyi Lin, Zhiyue Hu, and Xin Guo. Sparsemax and relaxed Wasserstein for topic sparsity. In *WSDM*, 2019.

[Lin *et al.*, 2020] Lihui Lin, Hongyu Jiang, and Yanghui Rao. Copula guided neural topic modelling for short texts. In *SIGIR*, 2020.

[Liu *et al.*, 2019] Luyang Liu, Heyan Huang, Yang Gao, Yongfeng Zhang, and Xiaochi Wei. Neural variational correlated topic modeling. In *WWW*, 2019.

[Miao *et al.*, 2016] Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *ICML*, 2016.

[Miao *et al.*, 2017] Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In *ICML*, 2017.

[Nalisnick and Smyth, 2017] Eric Nalisnick and Padhraic Smyth. Stick-breaking variational autoencoders. In *ICLR*, 2017.

[Nallapati *et al.*, 2017] Ramesh Nallapati, Igor Melnyk, Abhishek Kumar, and Bowen Zhou. Sengen: Sentence generating neural variational topic model. *arXiv*, 2017.

[Nan *et al.*, 2019] Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. Topic modeling with Wasserstein autoencoders. In *ACL*, 2019.

[Nguyen *et al.*, 2015] Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. Improving topic models with latent feature word representations. *TACL*, 2015.

[Ning *et al.*, 2020] Xuefei Ning, Yin Zheng, Zhuxi Jiang, Yu Wang, Huazhong Yang, Junzhou Huang, and Peilin Zhao. Nonparametric topic modeling with neural inference. *Neurocomputing*, 2020.

[Panwar *et al.*, 2020] Madhur Panwar, Shashank Shailabh, Milan Aggarwal, and Balaji Krishnamurthy. TAN-NTM: Topic attention networks for neural topic modeling. *arXiv*, 2020.

[Peng *et al.*, 2018] Min Peng, Qianqian Xie, Yanchun Zhang, Hua Wang, Xiuzhen Jenny Zhang, Jimin Huang, and Gang Tian. Neural sparse topical coding. In *ACL*, 2018.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP-IJCNLP*, 2019.

[Rezaee and Ferraro, 2020] Mehdi Rezaee and Francis Ferraro. A discrete variational recurrent topic model without the reparametrization trick. *NeurIPS*, 2020.

[Rezende *et al.*, 2014] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

[Röder *et al.*, 2015] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *WSDM*, 2015.

[Silveira *et al.*, 2018] Denys Silveira, Andr'e Carvalho, Marco Cristo, and Marie-Francine Moens. Topic modeling using variational auto-encoders with Gumbel-softmax and logistic-normal mixture distributions. In *IJCNN*, 2018.

[Srivastava and Sutton, 2017] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. In *ICLR*, 2017.

[Teh *et al.*, 2006] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *JASA*, 101(476):1566–1581, 2006.

[Thompson and Mimno, 2020] Laure Thompson and David Mimno. Topic modeling with contextualized word representation clusters. *arXiv*, 2020.

[Tian *et al.*, 2020] Runzhi Tian, Yongyi Mao, and Richong Zhang. Learning VAE-LDA models with rounded reparameterization trick. In *EMNLP*, 2020.

[Wallach *et al.*, 2009] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *ICML*, 2009.

[Wang and Yang, 2020] Xinyi Wang and Yi Yang. Neural topic model with attention for supervised learning. In *AISTATS*, 2020.

[Wang *et al.*, 2019] Rui Wang, Deyu Zhou, and Yulan He. ATM: Adversarial-neural topic model. *Information Processing & Management*, 2019.

[Wang *et al.*, 2020] Rui Wang, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu. Neural topic modeling with bidirectional adversarial training. In *ACL*, 2020.

[Williamson *et al.*, 2010] Sinead Williamson, Chong Wang, Katherine A Heller, and David M Blei. The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*, 2010.

[Wu *et al.*, 2020a] Jiemin Wu, Yanghui Rao, Zusheng Zhang, Haoran Xie, Qing Li, Fu Lee Wang, and Ziye Chen. Neural mixed counting models for dispersed topic discovery. In *ACL*, 2020.

[Wu *et al.*, 2020b] Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *EMNLP*, 2020.

[Yang *et al.*, 2020] Liang Yang, Fan Wu, Junhua Gu, Chuan Wang, Xiaochun Cao, Di Jin, and Yuanfang Guo. Graph attention topic modeling network. In *WWW*, 2020.

[Zaheer *et al.*, 2017] Manzil Zaheer, Amr Ahmed, and Alexander J Smola. Latent LSTM allocation: Joint clustering and non-linear dynamic modeling of sequence data. In *ICML*, 2017.

[Zeng *et al.*, 2018] Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R Lyu, and Irwin King. Topic memory networks for short text classification. In *EMNLP*, 2018.

[Zhang *et al.*, 2018] Hao Zhang, Bo Chen, Dandan Guo, and Mingyuan Zhou. Whai: Weibull hybrid autoencoding inference for deep topic modeling. In *ICLR*, 2018.

[Zhao *et al.*, 2017a] He Zhao, Lan Du, and Wray Buntine. A word embeddings informed focused topic model. In *ACML*, 2017.

[Zhao *et al.*, 2017b] He Zhao, Lan Du, Wray Buntine, and Gang Liu. MetaLDA: A topic model that efficiently incorporates meta information. In *ICDM*, 2017.

[Zhao *et al.*, 2018a] He Zhao, Lan Du, Wray Buntine, and Mingyuan Zhou. Dirichlet belief networks for topic structure learning. In *NeurIPS*, 2018.

[Zhao *et al.*, 2018b] He Zhao, Lan Du, Wray Buntine, and Mingyuan Zhou. Inter and intra topic structure learning with word embeddings. In *ICML*, 2018.

[Zhao *et al.*, 2020] He Zhao, Piyush Rai, Lan Du, Wray Buntine, Dinh Phung, and Mingyuan Zhou. Variational autoencoders for sparse and overdispersed discrete data. In *AISTATS*, 2020.

[Zhao *et al.*, 2021] He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. Neural topic model via optimal transport. In *ICLR*, 2021.

[Zhou *et al.*, 2016] Mingyuan Zhou, Yulai Cong, and Bo Chen. Augmentable gamma belief networks. *JMLR*, 2016.

[Zhou *et al.*, 2020] Deyu Zhou, Xuemeng Hu, and Rui Wang. Neural topic modeling by incorporating document relationship graph. In *EMNLP*, 2020.

[Zhu *et al.*, 2018] Qile Zhu, Zheng Feng, and Xiaolin Li. Graphbtm: Graph enhanced autoencoded variational inference for biterm topic model. In *EMNLP*, 2018.