# Supervised Topic Compositional Neural Language Model for Clinical Narrative Understanding

Xiao Qin[1,3], Cao Xiao[2], Tengfei Ma[1], Tabassum kakar[3], Susmitha Wunnava[3], Xiangnan Kong[3], Elke Rundensteiner[3] and Fei Wang[4]

IBM Research[1], IQVIA[2], Worcester Polytechnic Institute[3], Weill Cornell Medicine[4]

{xiao.qin,tengfei.ma1}@ibm.com[1], cao.xiao@iqvia.com [2],
{xqin,tkakar,swunnava,xkong,rundenst}@wpi.edu[3], few2001@med.cornell.edu[4]

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9378347

***Abstract*—Clinical narratives that describe complex medical events are often accompanied by meta-information such as a patient's demographics, diagnoses and medications. This structured information implicitly relates to the logical and semantic structure of the entire narrative, and thus affects vocabulary choices for the narrative composition. To leverage this meta-information, we propose a supervised topic compositional neural language model, called MeTRNN, that integrates the strength of supervised topic modeling in capturing global semantics with the capacity of contextual recurrent neural networks (RNN) in modeling local word dependencies. MeTRNN generates interpretable topics from global meta-information and uses them to facilitate contextual RNNs in modeling local dependencies of text. For efficient training of MeTRNN, we develop an autoencoding variational Bayes inference method. We evaluate MeTRNN on the word prediction tasks using public text datasets. MeTRNN consistently outperforms all baselines across all datasets in perplexity ranging from 5% to 40%. Our case studies on real world electronic health records (EHR) data show that MeTRNN can learn and benefit from meaningful topics.**

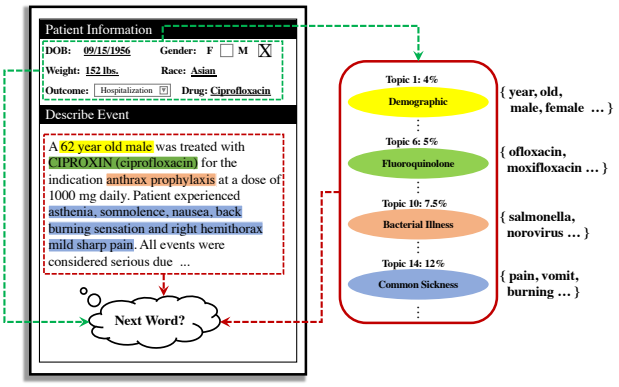



Fig. 1. The generative process of a clinical narrative. Green dashed box highlights document meta-information, red box the latent topics of the narrative and their associated vocabulary, while red dashed box the preceding text.

## I. Introduction

The era of big data has seen an explosion in the amount of digital information presented in the electronic health records (EHR) [1] providing the platform on which advanced analytics can be built to facilitate better care practices. An EHR is a patient-centered record consisting of heterogeneous data elements, including patient demographic information, diagnoses, laboratory test results, medication prescriptions, medical images and free-form clinical narratives [2]. In particular, the clinical narratives provide a diagram that concatenates complex medical events via natural language which encode critical insight very often not presented or missed from the structured fields, e.g. description of *Challenge-Dechallenge-Rechallenge* (CDR) [3] phenomenon that verifies adverse drug reactions. The problem of text mining clinical narratives through natural language processing (NLP) has attracted increasing attention in recent years [4].

Language models (LMs) whose goal is to learn the joint probability function of sequences of words in a language are one of the key enablers to many NLP applications including machine translation, named entity recognition and text summarization. The capability of capturing long term relationships among text is crucial to the performance of LMs [5]. Recurrent neural network (RNN) based language models in particular have demonstrated promising results in modeling complex and long dependencies. Recently, RNN based methods have been widely used in processing medical text [6]. In theory [7], RNNs such as Long Short-Term Memory (LSTM) [8] and Gated Recurrent Unit (GRU) [9] can "remember" arbitrarily long span of history if provided with enough capacity. However, they do not perform well on very long sequences in practice as the gradient computation for RNNs becomes increasingly ill-behaved as the expected dependency becomes longer [10]. One way of tackling this problem is to feed succinct information that encodes the semantic structure of the document such as latent topics as context to guide the modeling process [11], as illustrated in Figure 2(a). In this vein, existing works [12], [13], [14] focus

This work was done when Xiao Qin was a Ph.D. student at Worcester Polytechnic Insititute. Xiao Qin and Tabassum Kakar were supported by ORISE fellowships with the United States Food and Drug Administration. This work was partly supported by the National Science Foundation under grant CNS-1852498.

on the global context obtained from the text itself, overlooking the opportunity to exploit existing document meta-information which may provide explicit insight into the global context.

**Motivating Example.** Let's consider the generative process of a clinical narrative describing a patient's adverse drug events as illustrated in Figure 1. Before drafting the narrative, the structured template form with the "central ingredients" of the narrative such as the patient's demographics, suspected drugs, severity, etc are filled first. With this descriptive information and the observed events such as "experiencing nausea after taking Ciproxin" in mind, the overall story is then composed by considering the relevant topics and their corresponding vocabulary. Finally, the narrative summarizing all information is drafted with appropriate words in order. This motivating example highlights the following insights: (1) Latent topic information such "Bacterial Illness" topic and its proportion in the text as the global context to guide and regulate the language modeling process; (2) Document meta-information can be leveraged to learn more accurate and relevant topic information with respect to the key medical information.

**Limitations of State-of-the-Art.** Contextual RNNs (cRNNs) [11] obtain topic information from latent Dirichlet allocation (LDA) [15] and feed it into an additional *feature layer* connected to the recurrent unit to guide the modeling process. To ensure that the learned topics are in favor of those that indeed improve the language modeling performance, TopicRNN [12] and TCNLM [14] further extends cRNN by combing topic model and cRNN into a unified model that trains the two components simultaneously. However, these models only focus on the semantic structure inferred from the text itself. Hence, they miss the opportunity of obtaining a more complete context that also incorporates document meta-information. On this front, supervised topic models (sTMs) [16], [17], [18], [19], [20], [21], [22], illustrated in Figure 2(b), use observable document meta-information to supervise the learning of better topic representations. However, these models are bag-of-word models that do not account for word ordering, which is essential to our problem.

**Challenges.** To integrate the strength of sTMs into cRNNs for better clinical narrative modeling performance, the following research challenges need to be tackled: (1) *Flexible supervised topic model component.* Existing latent Dirichlet allocation (LDA) variations [16], [17], [18], [19], [20], [22] that incorporate document meta-information focus on specially constructed models. Even small changes to these ad-hoc solutions require deriving new inference methods which can be onerous for practitioners to freely experiment with different modeling assumptions. Moreover, existing solutions cannot accommodate combinations of modalities of data beyond their original intention. The lack of capability to manage arbitrary meta-data limits their effectiveness on complex inputs such as EHR which contains meta-information coded in various formats. (2) *End-to-end Framework.* sTMs learn the topics from the bag-of-words representation of the text and their corresponding meta-information. Although the learned topics representing the underlying semantic structure of a document can encode long-range dependencies for cRNNs, such topics do not reflect on information indicated by the ordering of the words (e.g "eat to live" vs. "live to eat") missing the opportunity to capture the true semantics of the text. In order to better facilitate cRNNs on sequence modeling task, establishing direct connection between sTMs to the goal of language model becomes critically important.

**Contribution.** To tackle the above challenges, we propose a neural language model called MeTRNN (Figure 2(c)) which enhances RNN-based language models' capability of establishing long-range dependencies by leveraging arbitrary document meta-information through their *implicit* influence via supervised latent topics and through *explicit* influence via a feature layer that directly connects to the RNN cells. It is worthwhile to highlight the following contributions of the proposed approach.

1) MeTRNN defines and explicitly models the text generative process based on the observation of the composition of the clinical narrative in an EHR.
2) MeTRNN captures the latent topics in text by leveraging the associated meta-information, which serves as the global context of the text that leads to better language modeling performance. To cope with various structured information in the EHRs, we propose a flexible supervised topic model component that can take on arbitrary meta-information.
3) We design a joint model that connects sTMs to cRNNs with an end-to-end autoencoding variational Bayes inference method using the conditional variational autoencoder framework [23]. It can be easily adjusted or extended.
4) We demonstrate the effectiveness of MeTRNN in the word prediction using publicly available text datasets as well as real world Electronic Health Records (EHRs). MeTRNN achieves improvement in perplexity from 5% to 40% against baselines. We also conduct a case study that demonstrates MeTRNN's ability to learn useful global context for better language modeling performance and more relevant topics to the structured meta-information.

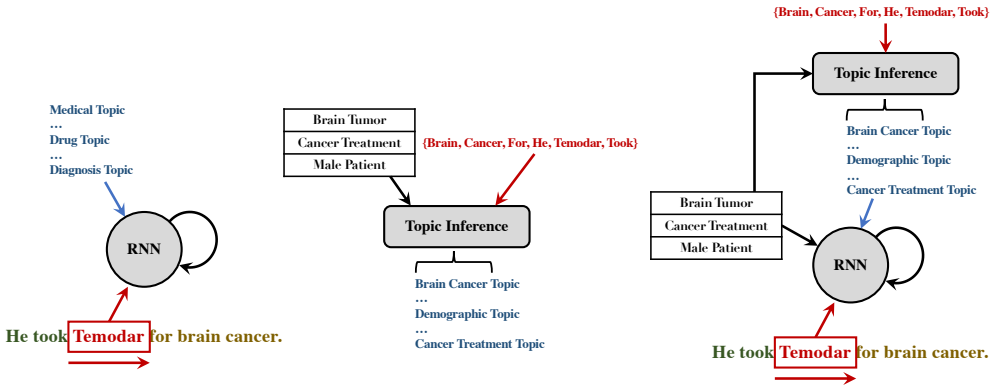
## II. Preliminary

### A. RNN-based Language Models

Traditional $n$-gram and feed-forward neural network-based language models make the Markov assumption about the dependencies between consecutive words where the chain rule limits conditioning to a fixed size context window. RNN-based language models overcome the Markov assumption by defining the conditional probability of each word $w_t$ given all the previous words $w_{1:t-1}$ through a hidden state $h_t$:

$$\begin{aligned} p(w_t|w_{1:t-1}) &\triangleq p(w_t|h_t), \\ h_t &= f(h_{t-1}, w_{t-1}). \end{aligned} \tag{1}$$

The function $f(\cdot)$ can be a standard RNN cell or a more complex cell such as GRU or LSTM. While in principle RNN is good at remembering the long-term dependencies, in practice, training a large-scale neural network on long histories can be difficult. Contextual RNN (cRNN) [11] tackles this




**(a) Contextual RNN** **(b) Supervised Topic Model** **(c) MeTRNN**

Fig. 2. Intuitions of different text modeling approaches.

problem by adding a *feature layer* that regulates the model by introducing the side information as additional context. The feature layer is connected to the hidden and output layers of the RNN [11]. Side information refers to information in or reasoned from the text such as document topic information obtained from latent Dirichlet allocation (LDA) [15]:

$$p(w_t|w_{1:t-1}) \triangleq p(w_t|h_t, x), \\ h_t = f(h_{t-1}, w_{t-1}, x), \tag{2}$$

where $x$ denotes the side information. While this study focuses on the RNN-based approach, we leave the exploration of other sequence models such as Transformer [24] as a future work.

### B. Latent Dirichlet Allocation

Probabilistic topic models are a family of models that aim to find groups of words that tend to co-occur within a document. These groups of words are called topics. Each topic $\beta_k$ represents a probability distribution that puts most of its mass on this topic related vocabulary. A document can then be represented as a mixture over these topics $\beta = (\beta_1 \cdots \beta_K)$. $\beta$ is said to encode the global semantics. Topic models are *bag-of-words* models where the word order is ignored.

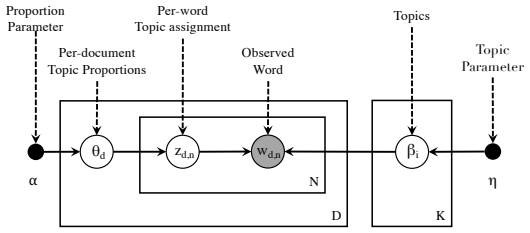



Fig. 3. Plate notation for LDA with Dirichlet-distributed topic-word distributions. D denotes the number of documents in a corpus, N is the number of words in a document and K is the specified number of topics.

For the most popular topic model, latent Dirichlet allocation (LDA) [15], its generative process of a document $w_{1:T}$ is: The marginal likelihood of a document $w_{1:T}$ is:

$$p(w_{1:T}|\alpha, \beta) = \int_\theta \Big( \prod_{t=1}^{T} \sum_{z_t=1}^{K} p(w_t|z_t, \beta) p(z_t|\theta) \Big) d\theta. \tag{3}$$

**for** *each document* $w_{1:T}$ **do**
  Draw topic distribution $\theta \sim$ Dirichlet($\alpha$)
  **for** *each word* $w_t$ **do**
    Draw topic assignment $z_t \sim$ Multinomial(1,$\theta$)
    Draw word $w_t \sim$ Multinomial(1,$\beta_{z_t}$)
  **end**
**end**

Posterior inference over the hidden variables $\theta$ and $z$ is intractable due to the coupling between $\theta$ and $\beta$ under the multinomial assumption. A popular approximation for efficient inference is *mean field variational inference* [25] which sidesteps this issue by introducing free variational parameters $\gamma$ over $\theta$ and $\phi$ over $z$ and dropping the edge between them. This results in an approximate variational posterior $q(\theta, z|\gamma, \phi)$ = $q_\gamma(\theta) \prod_t q_\phi(z_t)$, which is optimized to best approximate the true posterior $p(\theta, z|w_{1:T}, \alpha, \beta)$. The optimization problem is to minimize the *evidence lower bound* (ELBO):

$$\mathcal{L}(\gamma, \phi|\alpha, \beta) = -D_{KL}[q(\theta, z|\gamma, \phi)||p(\theta, z|\alpha)] + \\ \mathbb{E}_{q(\theta,z|\gamma,\phi)}[\log p(w_{1:T}|z, \theta, \alpha, \beta)]. \tag{4}$$

The first term in Equation 4 tries to match the variational posterior over latent variables to the prior on the latent variables, while the second term ensures that the variational posterior favors values of the latent variable that are good at explaining the data. Recently, several methods are proposed to "black box" the inference by using the variational autoencoder framework [26]. The variational parameters are computed by using a neural network called an inference network that takes the observed data as input. The second term in Equation 4 is referred to as a *reconstruction term* in the autoencoder network. The expectation w.r.t. $q$ is computed by using a Monte Carlo estimator, called *reparameterization trick*.

Supervised topic models (sTMs) [27] are a group of topic models for incorporating side information. They can be categorized into two classes, namely, *downstream supervised topic* (DsTM) and *upstream supervised topic model* (UsTM). In a DsTM such as [28], [29], [30], [31], [32], meta-information,

a.k.a. the response, is predicted based on the latent representation of the document, whereas in a UsTM such as [16], [33], [17], [21], [34] the response variable is being conditioned on to generate the latent representation.

## III. The Proposed Approach

Next, we describe our proposed supervised topic compositional neural language model (MeTRNN). The realization of MeTRNN is a deep learning framework that integrates a sTM like component into a cRNN for improving the language modeling capacity. First, we introduce the general principle of how we utilize the meta-information in our model in Section III-A. Second, we formally define the MeTRNN model including a sTM like component, a cRNN component and how they interact with each other in Section III-B. Third, we propose an inference method for MeTRNN in Section III-C. Finally, we discuss our strategy for training MeTRNN III-D.

### A. Document Meta-Information

Document meta-information, as motivated in the clinical narrative scenario, provides the central ingredients of the narrative text as well as a clue in semantic structure of the entire narrative. Based on this observation, we design our model such that meta-information has both *explicit* and *implicit* influence on language modeling. For *explicit* influence, we add a *feature layer* similar to [11] that takes meta-information directly connected to the recurrent unit in RNN. For *implicit* influence, we introduce a sTM like component where the meta-information is used as a response to produce relevant topic information. In this study, we adopt the idea of UsTM approach where meta-information is being conditioned on to generate the topic information of the narrative. The widely used UsTM approach is considered closer to the generative process [21] in the clinical narrative scenario where all meta-information is predefined and is used for defining the topics. MeTRNN works with arbitrary meta-information including medical images, lab test measurements, medication information and etc as long as there exists appropriate embeddings for these information. MeTRNN is able to take in embeddings generated from the pre-trained models as they are or connect to these models as the embedding layers allowing fine tuning while learning the final language model. The exact computation of *explicit* and *implicit* influence is formalized next.

### B. MeTRNN Model

We define MeTRNN as a generative probabilistic model of an EHR corpus. The idea is that the semantic structure of a document is represented as a random mixture of latent topics conditioned on some document meta-information. Each topic is characterized by a distribution over words. The distribution of a word in the text narrative is then estimated given all the preceding words, latent topics and the document meta-information. For each document $d = (x_d, w_{1:T})$ where $x_d$ is a vector that encodes the meta-information of $d$, e.g. representation of the structured information in an EHR, and

**for** *each document* $d = (x_d, w_{1:T})$ **do**
  I. Draw a topic proportion vector $\theta \sim p(\theta|x_d)$
  **for** *each word* $w_t$ **do**
    II. Compute the hidden state $h_t = f(w_{t-1}, h_{t-1})$
    III. Draw word $w_t \sim p(w_t|h_t, \theta, x_d)$ where $p(w_t{=}i|h_t, \theta, x_d) \propto \exp(v_i^\top h_t + b_i^\top \theta + c_i^\top x_d)$
  **end**
**end**

$w_{1:T}$ is the associated narrative text, the generation process of $w_{1:T}$ is defined as follows:
$\theta$ is drawn from a Dirichlet distribution over $\theta$ conditioned on the document meta-information $x_d$. $\theta$ is the topic proportions influenced by the document meta-information which encodes the semantic structure of the document $d$. $f$ computes the hidden state of the RNN (Equation 1) based on the previous word and hidden state. The current hidden state $h_t$ encodes the local dynamics of the composed word sequence up to time $t-1$. Finally, the next word $w_t$ is decided based on the hidden state $h_t$, topic proportions $\theta$ and document meta-information $x_d$ through an additive procedure. In [11], $x_d$ and $\theta$ are referred as additional side information to affect the word choices in the language model. Following [12], instead of passing them into the hidden state of the RNN, they are used as bias to have their global semantic contributions to the word choices clearly separated from those of local dynamics. The contextual contribution is measured by the summation of the dot products between $\theta$, $x_d$ and respective latent word vectors $b_i \in W_{\theta w}$ and $c_i \in W_{mw}$ for the $i$th vocabulary word. And $v_i \in W_{hw}$.

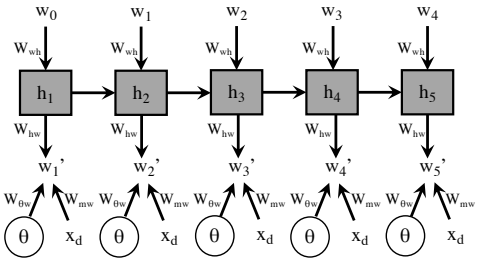



Fig. 4. The unrolled MeTRNN architecture: $w_0, \cdots, w_4$ are words in the document, $h_t$ is the state of the RNN at time step $t$, $\theta$ is the latent representation of the EHR and $x_d$ is the meta-information.

The unrolled graphical representation of MeTRNN is depicted in Figure 4. The log marginal likelihood of the word sequence $w_{1:T}$ composing a document $d$ is:

$$\log p(w_{1:T}|x_d) = \log \int p(\theta|x_d) \prod_{t=1}^{T} p(w_t|h_t, \theta, x_d) \mathrm{d}\theta \quad (5)$$

### C. The Model Inference

Since directly optimizing Equation 5 is intractable, we use variational inference for approximating this marginal. Let $q(\theta)$ be the variational distribution on the marginalized $\theta$. The

variational lower bound of the model is written as follows (derivation can be found in Section A):

$$\log p(w_{1:T}|x_d) \geq -D_{KL}(q(\theta|\tilde{w}_{1:T}, x_d)||p(\theta|x_d)) + \mathbb{E}_{q(\theta|w_{1:T},x_d)}[\log p(w_{1:T}|\theta, x_d)]. \tag{6}$$

ELBO is written as:

$$\mathcal{L}(x_d, w_{1:T}) \triangleq -D_{KL}(q(\theta|\tilde{w}_{1:T}, x_d)||p(\theta|x_d)) + \mathbb{E}_{q(\theta|w_{1:T},x_d)}\Big[\sum_{t=1}^{T} \log p(w_t|h_t, \theta, x_d)\Big]. \tag{7}$$

Following the proposed conditional variational autoencoder (CVAE) [23], we choose the form of $q(\theta)$ to be a "black box" inference network using a feed-forward neural network. Specifically, the MeTRNN inference network consists of a recognition network $q(\theta|\tilde{w}_{1:T}, x_d)$ where $\tilde{w}_{1:T} \in d$ is a bag-of-words representation of $w_{1:T}$, a prior network $p(\theta|x_d)$ and a generation network $p(w_{1:T}|\theta, x_d)$ that reconstructs the word sequence. In our formulation, the prior of the latent variable $\theta$ is modulated by the meta-information. This can be relaxed to make the latent variables statistically independent of $x_d$ [35], i.e., $p(\theta|x_d) = p(\theta)$. We show the graphical representation of MeTRNN inference network in Figure 5.

$q(\theta)$ is reparameterized with a deterministic, differentiable function $g(\cdot, \cdot, \cdot)$, whose arguments are meta-information $x_d$, words $\tilde{w}_{1:T}$ and the noise variable $\epsilon$. This is known as *reparameterization trick* [26], allowing for error backpropagation through the latent variables, essential in variational autoencoder training. In MeTRNN, the latent variable $\theta$ follows a Dirichlet distribution as suggested by the classical topic models [15] due to its flexibility. However, the Dirichlet distribution does not belong to the *location-scale* family which makes *reparameterization trick* difficult to use. We solve this by constructing a Laplace approximation to the Dirichlet prior [36]. We approximate the prior distribution with $\hat{p}(\theta|\mu_1, \Sigma_1) = \mathcal{LN}(\theta|\mu_1, \Sigma_1)$ where $\mathcal{LN}$ is a logistic normal distribution,

$$\begin{aligned} \mu_{1k} &= \log \alpha_k - \frac{1}{K}\sum_i \log \alpha_i, \\ \Sigma_{1kk} &= \frac{1}{\alpha_k}\Big(1 - \frac{2}{K}\Big) + \frac{1}{K^2}\sum_i \frac{1}{\alpha_i}, \end{aligned} \tag{8}$$

with $\alpha = (\alpha_1, \cdots, \alpha_K)$ being the parameter of the Dirichlet prior and $K$ the dimension of the hidden space, a.k.a. specified number of topics. Finally, $\theta = g(x_d, w_{1:T}, \epsilon), \epsilon \sim \mathcal{N}(0, \mathrm{I})$.

According to the defined prior network, the input of the recognition network $\tilde{w}_{1:T}$ and the meta-information vector $x_d$ is first projected into a $K$-dimensional latent space. Specifically, we have:

$$\begin{aligned} q(\theta|\tilde{w}_{1:T}, x_d) &= \mathcal{LN}(\theta|\mu(\tilde{w}_{1:T}, x_d), diag(\sigma^2(\tilde{w}_{1:T}, x_d))), \\ \mu(\tilde{w}_{1:T}, x_d) &= W_{w\mu}\tilde{g}(\tilde{w}_{1:T}) + W_{m\mu}\tilde{g}(x_d) + b_\mu, \\ \log \sigma(\tilde{w}_{1:T}, x_d) &= W_{w\sigma}\tilde{g}(\tilde{w}_{1:T}) + W_{m\sigma}\tilde{g}(x_d) + b_\sigma, \end{aligned} \tag{9}$$

where $\tilde{g}(\cdot)$ denotes the feed-forward neural network. The weight matrices $W_{w\mu}, W_{m\mu}, W_{w\sigma}, W_{m\sigma}$ and biases $b_\mu$, $b_\sigma$ are shared across documents. Each document has its own parameter setting $\mu(\tilde{w}_{1:T}, x_d)$ and $\sigma(\tilde{w}_{1:T}, x_d)$ resulting in a unique distribution $q(\theta|\tilde{w}_{1:T}, x_d)$ for each document. The output of the inference network is a topic proportion vector $\theta$ that represents the global semantics of the document.

The generation network is in the form of a recurrent neural network. It learns the local dynamics of the word sequence for each topic proportion vector $\theta$. Here we show the specification with a vanilla RNN cell and it can be easily extended to other structures such as a GRU or LSTM cell since $\theta$ and $x_d$ are only utilized as bias in the output layer:

$$\begin{aligned} h_t &= \sigma_h(W_{wh}w_{t-1} + W_{hh}h_{t-1} + b_h), \\ w_t &= \sigma_w(W_{hw}h_t + W_{\theta w}\theta + W_{mw}x_d + b_w), \end{aligned} \tag{10}$$

where $\sigma(\cdot)$ denotes the activation functions. The weight matrices $W_{wh}, W_{hh}, W_{hw}, W_{\theta w}, W_{mw}$ and biases $b_h, b_w$ are shared across words. The hidden state of the recurrent unit, the topic proportion vector $\theta$ and the document meta-information $x_d$ affect the output through an additive procedure.

During training, the parameters of the inference network and the model are jointly learned and updated via truncated backpropagation throughout time using the AdaGrad algorithm [37]. The dimension of the parameters in MeTRNN are reported in Section B.

### *D. Training MeTRNN*

Each training instance for MeTRNN consists of (1) the meta-information, (2) the words in bag-of-words representation and (3) the word sequence. Following [12], we truncate the document into shorter subsequences for RNN training. In word prediction task, MeTRNN is given the preceding word sequence $w_{1:t-1}$ and the meta-information $x_d$ from which MeTRNN has an estimation of $q(\theta|\tilde{w}_{1:t-1}, x_d)$. To predict the next word $w_t$, MeTRNN computes the probability distribution of $w_t$ incrementally. After the predicted word $w_t$ being sampled from the predictive distribution, MeTRNN update $q(\theta)$ by including $w_t$. MeTRNN is then go on to predict the next word $w_{t+1}$.

Similar to [38], we find that using RNN as a decoder under the conditional variational autoencoder framework fails to produce meaningful information in $\theta$ due to the *vanishing latent variable problem*. Following [38], we apply a small weight on the $D_{KL}$ term and gradually increase it during training. The idea of having a constrained $D_{KL}$ cost in VAE to obtain better latent representations is studied in [39]. Specifically, we have:

$$\mathcal{L}(x_d, w_{1:T}) \triangleq -\beta D_{KL}(q(\theta|\tilde{w}_{1:T}, x_d)||p(\theta|x_d)) + \mathbb{E}_{q(\theta|w_{1:T},x_d)}\Big[\sum_{t=1}^{T} \log p(w_t|h_t, \theta, x_d)\Big], \tag{11}$$

where $\beta$ is a hyper-parameter that balances the latent channel capacity and independence constraints.

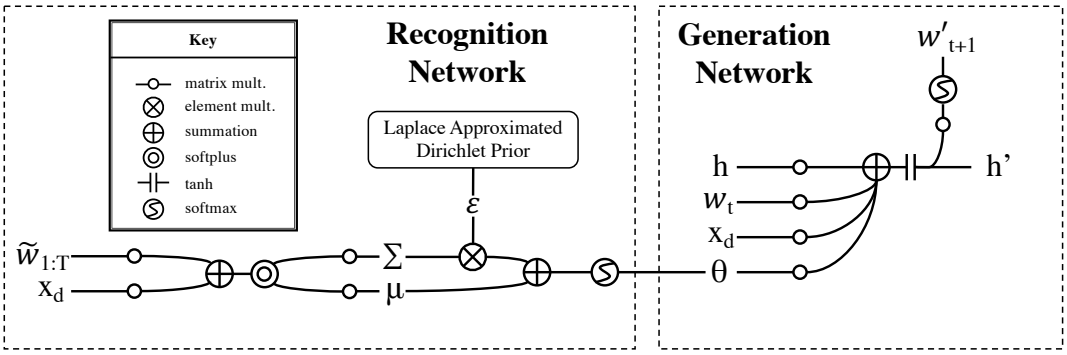



Fig. 5. An example of MeTRNN inference network with a vanilla recurrent neural network cell. The input of the recognition network are $\tilde{w}_{1:T}$ (or $\tilde{w}_{1:t}$) the bag-of-words representation of the text and $x_d$ the vector representation of the meta-information. The input of the generation network at time $t$ includes the hidden state $h$ from the previous time stamp, current word $w_t$, topic vector $\theta$ and meta-information $x_d$.

## IV. Experimental Evaluation

### A. Experimental Setup & Methodology

We evaluate our proposed MeTRNN model with publicly available text datasets as well as EHRs by comparing its performance on the word prediction tasks against other baselines. We also conduct a case study on EHRs that shows the effectiveness of MeTRNN for learning meaningful and useful topics. All methods are implemented in PyTorch [40] and trained on an Ubuntu server with Intel Xeon E-5 2680v2 @2.8GHz CPUs and Nvidia Tesla K40m GPUs. Note that our model training can be easily implemented in a multi-node and multi-GPU environment by using data parallel approach supported by torch.distributed API.

TABLE I
Size in number of words and document meta-information type. M=million, K=thousand.

| Dataset | Train | Valid | Test | Vocabulary |
|---|---|---|---|---|
| 20NG | 2M | 248K | 266K | 10K |
| R52 | 465K | 90K | 77K | 10K |
| MADE | 306K | 53K | 53K | 11K |

*1) Datasets:* For reproducibility, we use two well known labeled datasets, namely *20 Newsgroups* (20NG) [41] and *Reuters-21578* (R52) [42] for the word prediction task. The category information of each document is used as the document meta-information. The category distributions over these two corpus are also publicly available [1]. We also use a labeled EHR dataset *MADE* for an adverse drug event detection competition [2]. *MADE* consists of total of 1089 de-identified EHR narratives from 21 cancer patients. Each EHR comes with medical information such as medication name, adverse events, indications and other signs and symptoms. In this experiment, we select 102 unique indications to describe the document meta-information. Each EHR contains at least one indication from our selection. Basic statistics of the datasets are summarized in Table I. We partition each document into sliding windows with length of 50. 20NG and R52 datasets are preprocessed with *stopword removal* and *stemming* (preprocessed datasets are publicly available [43]). MADE corpus is preprocessed with *stopword removal*.

*2) Baselines:* For the word prediction tasks, we compare our MeTRNN with GRU and LSTM cells denoted as MeTGRU and MeTLSTM respectively against:

- **RNNs.** LSTM and GRU, commonly used in language modeling, are proved to be superior than vanilla RNN for long documents. Therefore, we include these two as baselines.
- **Contextual RNNs.** We implemented the contextual RNN (cRNN) from [11] and extended it using LSTM and GRU cells denoted as cLSTM and cGRU respectively. We consider three features for cRNNs: (1) topic information obtained separately from ProdLDA [36] (with an existing Pytorch implementation[3]); (2) document meta-information; (3) combination of (1) and (2). Topic information is inferred from the text.
- **TopicRNNs.** TopicRNN [12] falls into the same category as TCNLM [14] as they share a similar generative process and the inference network. We implemented TopicRNNs as a representative of such kind. We experimented TopicRNNs with LSTM and GRU cells as they have been shown to achieve better performance than the ones with vanilla RNN cell [12]. Since stopwords are excluded from our datasets, the mechanism that explicitly models stopwords is ignored. Topic information is inferred from the text.

*3) Metric:* For the word prediction, we measure the word *perplexity* (PPL) typical metric for language model evaluation:

$$\text{PPL} = \exp\Big(-\frac{1}{T}\sum_{t=1}^{T}\log(p(w_t|w_{1:t-1}))\Big), \qquad (12)$$

where $T$ is the length of the test document. Lower PPL indicates better prediction performance.

*4) Word Prediction on 20NG and R52:* We evaluate MeTRNN against other baselines on the word prediction task by varying the complexity of the models in the number of neurons used in each layer. We use 1 RNN layer for all

[1] http://ana.cachopo.org/datasets-for-single-label-text-categorization

[2] http://bio-nlp.org/index.php/announcements/39-nlp-challenges

[3] https://github.com/hyqneuron/pytorch-avitm

TABLE II

TEST PERPLEXITIES OF DIFFERENT MODELS BY VARYING THE NUMBER OF NEURONS. THE LOWER THE PERPLEXITY THE BETTER THE PERFORMANCE. $(\cdots)$ AFTER EACH PERPLEXITY INDICATES THE RANKING OF THE METHOD W.R.T. THE SPECIFIC SETTING. "T" DENOTES THE TOPIC FEATURE OBTAINED FROM PRODLDA TRAINED SEPARATELY USING AVITM AND "F" DENOTES THE DOCUMENT META-INFORMATION. † OR ⋆ INDICATES THAT THE BASELINE IS IMPLEMENTED BY OTHERS OR OURSELVES.

| | **20 NG** | | | **R52** | | | **MADE** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Methods** | **n=128** | **n=256** | **n=512** | **n=64** | **n=128** | **n=256** | **n=16** | **n=32** | **n=64** |
| GRU† | 360.76(12) | 352.79(12) | 345.68(12) | 163.04(15) | 151.70(15) | 149.20(15) | 174.17(12) | 122.57(12) | 99.42(12) |
| LSTM† | 352.15(11) | 337.15(11) | 333.95(11) | 154.26(14) | 145.62(14) | 143.29(14) | 170.81(11) | 115.97(11) | 98.28(11) |
| cRNN(T)⋆ | 370.26(14) | 365.47(15) | 353.64(15) | 146.36(13) | 143.13(13) | 142.56(13) | 177.40(14) | 130.99(13) | 109.18(13) |
| cRNN(F)⋆ | 363.59(13) | 362.43(13) | 352.12(13) | 134.57(11) | 134.20(11) | 132.35(11) | 186.87(15) | 137.06(15) | 110.83(14) |
| cRNN(T+F)⋆ | 371.81(15) | 364.28(14) | 353.37(14) | 137.22(12) | 134.30(12) | 133.98(12) | 175.90(13) | 132.46(14) | 113.23(15) |
| cGRU(T)⋆ | 316.93(6) | 299.99(10) | 280.53(5) | 118.79(6) | 110.20(8) | 104.74(5) | 151.66(8) | 108.44(4) | 90.87(8) |
| cGRU(F)⋆ | 314.69(3) | 297.79(8) | 279.21(4) | 115.38(3) | 109.70(6) | 106.96(8) | 159.96(10) | 114.58(10) | 93.29(10) |
| cGRU(T+F)⋆ | 320.49(7) | 298.12(9) | 281.78(7) | 118.34(5) | 111.30(9) | 105.76(6) | 147.36(6) | 112.25(9) | 92.89(9) |
| cLSTM(T)⋆ | 322.13(9) | 289.54(5) | 284.58(10) | 119.46(7) | 109.96(7) | 108.42(9) | 144.89(4) | 108.72(5) | 88.30(3) |
| cLSTM(F)⋆ | 315.36(5) | 293.77(6) | 281.74(6) | 117.58(4) | 108.56(4) | 106.50(7) | 158.88(9) | 111.69(8) | 89.52(4) |
| cLSTM(T+F)⋆ | 321.63(8) | 289.14(4) | 282.89(9) | 127.09(8) | 116.85(10) | 113.63(10) | 145.79(5) | 108.93(6) | 90.86(7) |
| TopicGRU⋆ | 315.28(4) | 296.31(7) | 278.13(3) | 117.32(9) | 108.72(5) | 103.79(3) | 148.66(7) | 111.32(7) | 90.45(6) |
| TopicLSTM⋆ | 323.31(10) | 286.30(3) | 282.38(8) | 121.29(10) | 107.72(3) | 104.02(4) | 144.13(3) | 108.05(3) | 90.20(5) |
| MeTGRU⋆ | **309.30**(1) | 283.90(2) | 273.60(2) | 108.29(2) | **96.34**(1) | **90.34**(1) | **139.10**(1) | 101.93(2) | 82.48(2) |
| MeTLSTM⋆ | 309.98(2) | **281.59**(1) | **272.29**(1) | **107.25**(1) | 98.34(2) | 95.13(2) | 141.05(2) | **99.84**(1) | **80.73**(1) |

methods and do not apply dropout for comparison purpose. For TopicRNNs and MeTRNNs, we use a multilayer perception with 2 hidden layers for the inference network. For comparability, we specify the number of topics for TopicRNNs and MeTRNNs to be equal to the number of categories in 20NG and R52 respectively. The validation set is used for early stopping. Hyperparameters including *learning rate*, *batch size*, $\alpha$ (parameter of Dirichlet prior) and $\beta$ (scaling parameter for $D_{KL}$) are properly tuned for each method with different complexities. The specific hyperparameter settings are reported in Section C.

As shown in Table II, MeTRNN consistently outperforms all other baselines. In general, the models with the capability of incorporating extra context information perform better than the ones that do not account for such information. Specifically, GRU and LSTM cannot achieve lower PPL than others with the same type of recurrent units. In the experiments with R52, cRNNs conditioned on various combinations of features achieve lower PPL than GRU and LSTM. When testing cRNNs, cGRUs and cLSTMs, we find that the document meta-information can better help the model as compared to the topic features obtained from ProdLDA. The reason is that the category label in these two datasets can be seen as a better representation of the semantic structure of the document. It uniquely identifies the theme of the document and the underlying vocabulary used for the content.

As opposed to using the topic information obtained separately, TopicGRU and TopicLSTM learn the latent topics simultaneously during language modeling. Although they outperform their comparable methods cGRU(T) and cGRU(T) in a few experiments, the performances are not consistent across different settings. The closest methods to MeTRNN in context information leveraged in the model are cRNN(T+F), cGRU(T+F) and cLSTM(T+F). Interestingly, these methods which take both features by simple feature concatenation do not outperform the ones that consider only one feature. Worse yet, in some cases, their PPLs are higher than all of those which take a single feature. The reason is that the topic proportions $\theta$ obtained separately from ProdLDA and the meta-information associated with the document may not entirely "agree" with each other. In an extreme case, a topic representing some common words used in corpus may not be helpful for language modeling. Worse yet, it may diminish the contribution of the meta-information which encodes the central ingredients of the narrative. One naive solution is to obtain topic proportions $\theta$ from a supervised topic model so that the learned topics information balances the information from the text itself as well as the meta-information. MeTRNN extends this idea by combining a supervised topic model component with the language model to make sure that the learned semantic structure is helpful for the word prediction.

### *B. Case Study: EHR Narrative Modeling and Generation*

Next we will take a deep dive into the experiments on a real EHR dataset to demonstrate how MeTRNN can learn meaningful and useful topics. Besides the structured information provided with an EHR (See Introduction), the narrative text provides a full story about the medical events of a patient. Modeling such narrative text is a fundamental task for many applications in healthcare systems [44]. We conduct a case study using MADE – a labeled EHR dataset that reports adverse drug reactions. An adverse drug reaction corresponds to an unwanted and often dangerous effect caused by the administration of a drug. MADE's labels include drug name, indication, adverse reactions, etc. In this study, we use the indication as the meta-information of the narrative. In medicine, an indication is a valid reason to use a certain medication. An indication can correspond to a certain type of medication

TABLE III

TOP 10 WORDS OF 5 TOPICS (RANDOMLY SELECTED OUT OF 20) LEARNED BY 3 METHODS. THE ORIGINAL WORDS ARE ALL IN LOWERCASE. LETTERS ARE MANUALLY CAPITALIZED FOR BETTER INTERPRETATION. † OR ⋆ INDICATES THAT THE BASELINE IS IMPLEMENTED BY OTHERS OR OURSELVES.

| Methods | Topic | Vocabulary |
|---|---|---|
| ProdLDA† | 1 | AbdPelvis, Island, Oxymizer, Aids, Acidophilus, Hotline, Things, Greens, CCU, Hypoxemia ... |
| | 2 | Laparotomy, Excercize, Striae, Reduce, Cecectomy, Noninflamed, Dipstick Counseled, Transaminitis, DOs... |
| | 3 | Nephrectomy, Amplitudes, Hysterectomy, Stinging, Amplitude, Unimproved, Crease, Prepped, Flexed, Pasty ... |
| | 4 | Nonsteroidals, Onethird, Ascertain, Upward, NP, Advancing, Excess, Leaflet, Twothirds, Outflow ... |
| | 5 | Deal, Clustered, Proves, Demonstration, Desire, Thinned, Extent, Familysocial, Lobulated, Exclude ... |
| TopicLSTM⋆ | 1 | Autoimmune, Splenectomy, Marginal, Folic, Reticulocyte, Elbow, Furosemide Calcitonin, Celexa, Losartan ... |
| | 2 | Comments, Modified, PO, Medicalsurgical, Laboratorystudies, Communication, SOB, Agree, Temp, Reclast ... |
| | 3 | Pediatric, Amitriptyline, Burkitt, Wound, Med, PO, Broviac, Community Headache, Mom ... |
| | 4 | Plasmacytoid, Impacted, Badly, Ideal, Priority, Reviews, Fremitus, Expiratory, Accessory, Tactile ... |
| | 5 | Testosterone, Lymphoplasmacytoid, Androderm, Bendamustine, Hypogonadism, Acknowledgement, Diltiazem, Kyphoplasties, Alprazolam, Salmonella ... |
| MeTLSTM⋆ | 1 | eGFR, Antiresorptive, Well, Leery, Equation, MDRDs, SQ, Velcade, Performing, Injuries ... |
| | 2 | NP, Amitriptyline, Reports, Pediatric, Burkitt, Palpated, CKD, Kidney, Supervising, Comments ... |
| | 3 | Sinuses, Infectious, Transplant, ABVD, Autologous, Acyclovir, Natural, Nasal, Hodgkins, Patient ... |
| | 4 | Underwent, Laminectomy, Brachial, Radiation, Intrathecal, Vertebral, Compression, Shoulder, Spondylolisthesis, Insurance ... |
| | 5 | Quite, Actually, Breaths, Panic, Attacks, Anxiety, Well, Velcade, Increase, Twice ... |

TABLE IV

TOP 10 WORDS OF 5 (OUT OF 102) INDICATION TYPES LEARNED BY METRNN (OBTAINED FROM WEIGHT MATRIX $W_{mw}$). THE ORIGINAL WORDS ARE ALL IN LOWERCASE. LETTERS ARE MANUALLY CAPITALIZED FOR BETTER INTERPRETATION.

| Indications | Vocabulary |
|---|---|
| Hodgkin's Lymphoma | Hodgkins, ABVD, Chest, Omeprazole, Chemotherapy, MD, FI, MR, Told, Port ... |
| Peripheral Neuropathy | Transplant, Peripheral, P, Levels, Neurontin, Marrow, Therapy, Done, Copay, MR ... |
| Mantle Cell Lymphoma | Cycles, Velcade, Mantel, Location, Therapy, Allogeneic, MD, Positive, Status, Cycle ... |
| Cellulitis | Cellulitis, Currently, Doxycycline, Redness, Foot, Lymph, Ankle, Anxiety, Rule, Doxazosin ... |
| Hypercalcemia | Continues, Hypercalcemic, Pamidronate, Radiation, Due, Hospitalization, Weekly, Taking, Schedule, Potassium ... |

which may trigger specific reactions commonly associated with these drugs. The indication can reveal the semantics of the narrative. We include 102 unique indications in this dataset to encode a narrative's meta-information vector.

For the word prediction task, words are not stemmed in order to generate interpretable topics. Comparing to the meta-information used for 20NG and R52, indication can capture partial or different semantics from the topic information learned from the narrative as confirmed by the results shown in Table II. The cRNNs conditioned on topic feature achieve lower PPL than those conditioned on the indication feature. However, it remains true that cRNNs is not further improved by simply concatenating those features. MeTRNN outperforms all other baselines while incorporating both self generated feature and indication information into consideration.

Next, we show the vocabulary for different indication types obtained from the weight matrix $W_{mw}$ learned by MeTLSTM. We randomly select 5 indications from MADE in Table IV. We observe that the vocabulary is closely related to the corresponding indication type. For example, the learned vocabulary for Hodgkin's Lymphoma includes "Hodgkins", "ABVD" (ABVD is a chemotherapy regimen used in the first-line treatment of Hodgkin lymphoma), etc. Later we show that the topics learned by MeTRNN are indeed influenced by the indication feature.

Table III shows the vocabulary of selected topics generated by ProdLDA, TopicLSTM and MeTLSTM. Topics learned by ProdLDA and TopicLSTM are similar as they exhibit similar diversity in types of words across topics. Within each topic, we observe more common words, e.g., "deal", "upward" and "med", from ProfLDA and TopicLSTM than from MeTLSTM which is not ideal for capturing unique topics. The topics learned by MeTLSTM emphasize more on different diseases and symptoms as they are influenced by the indication feature. More importantly, such influence mechanized by our proposed MeTRNN improves the modeling performance confirmed by the previous word prediction results.

## V. RELATED WORK

### A. Context Dependent Neural Language Models

[11] augments contextual information into a conventional RNNLM [5] by adding an extra layer connected to the recurrent unit. The contextual information in this work is obtained by using LDA from a block of proceeding text. TopicRNN [12] extends this idea by integrating a topic model like unit to model the contextual information and the word sequence simultaneously. The topic information is inferred from the document in the bag-of-words representation and is then fed to the recurrent unit to regulate the language modeling in every time step. It uses a variational autoencoder for model inference. [13] introduces an attention-based convolutional neural network to extract semantic topics. [14] incorporates global context of the document obtained from a topic model like unit through a Mixture-of-Experts model design. However, these model do not account for document meta-information for either topic inference or language modeling.

### B. Supervised Topic Models

Author-Topic model [16] assumes words are generated by an author uniformly selected from an observed author list and then a topic selected from a distribution over topics that is specific to that author. [17] models expertise by multiple topical mixtures associated with each individual author. Supervised LDA (sLDA) [18] models document with single label by learning a generalized linear model with an appropriate link function and exponential family dispersion function. Labelled LDA (LLDA) [19] assumes a multi-label document such that each label has a corresponding topic and a document is generated by a mixture of the topics. As an extension to LLDA, Partially Labelled LDA (PLLDA) [20] assigns multiple topics to a label. The Dirichlet Multinomial Regression (DMR) [21] incorporates document meta-information on the prior of the topic distributions with the logistic-normal transformation. [22] introduces a Poisson factorization model with hierarchical document labels. However, these models are *bag-of-words* models that do not consider word ordering.

## VI. Conclusion

In this paper, we propose MeTRNN which is a supervised topic compositional neural language model for modeling clinical narratives supported by meta-information. The main idea is to leverage the meta-information which hints the semantics of the entire document to regulate the RNN-based language model. We integrate a supervised topic model-like component to allow the meta-information to make implicit impact on language modeling via hidden topics. We also propose a black box deep Bayesian inference network for MeTRNN which is easily extendable to new models. Through our extensive experiments with several datasets, we show the effectiveness of MeTRNN on language modeling as well as the ability of generating useful and meaningful topics.

## References


[1] G. S. Birkhead, M. Klompas, and N. R. Shah, "Uses of electronic health records for public health surveillance to advance public health," *Annual Review of Public Health*, vol. 36, pp. 345–359, 2015.

[2] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1419–1428, 2018.

[3] W. Spitzer, "Importance of valid measurements of benefit and risk." *Medical toxicology*, vol. 1, pp. 74–78, 1986.

[4] Y. Wang, A. Tafti, S. Sohn, and R. Zhang, "Applications of natural language processing in clinical research and practice," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 2019, pp. 22–25.

[5] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Annual Conference of the International Speech Communication Association*, 2010, pp. 1045–1048.

[6] F. Liu, A. Jagannatha, and H. Yu, "Towards drug safety surveillance and pharmacovigilance: Current progress in detecting medication and adverse drug events from electronic health records," 2019.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Advances in Neural Information Processing Systems Workshop on Deep Learning*, 2014.

[10] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, 2013, pp. 1310–1318.

[11] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model," in *Spoken Language Technology Workshop*, 2012, pp. 234–239.

[12] A. B. Dieng, C. Wang, J. Gao, and J. W. Paisley, "TopicRNN: A recurrent neural network with long-range semantic dependency," *International Conference on Learning Representations*, vol. abs/1611.01702, 2017.

[13] J. H. Lau, T. Baldwin, and T. Cohn, "Topically driven neural language model," in *Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 355–365.

[14] W. Wang, Z. Gan, W. Wang, D. Shen, J. Huang, W. Ping, S. Satheesh, and L. Carin, "Topic compositional neural language model," in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 356–365.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003. [Online]. Available: http://www.jmlr.org/papers/v3/blei03a.html

[16] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2004, pp. 487–494.

[17] D. M. Mimno and A. McCallum, "Expertise modeling for matching papers with reviewers," in *International Conference on Knowledge Discovery and Data Mining*, 2007, pp. 500–509.

[18] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Advances in neural information processing systems*, 2008, pp. 121–128.

[19] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," in *Conference on Empirical Methods in Natural Language Processing*, 2009, pp. 248–256.

[20] D. Ramage, C. D. Manning, and S. Dumais, "Partially labeled topic models for interpretable text mining," in *International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 457–465.

[21] D. M. Mimno and A. McCallum, "Topic models conditioned on arbitrary features with dirichlet-multinomial regression," in *Conference on Uncertainty in Artificial Intelligence*, 2008, pp. 411–418.

[22] C. Hu, P. Rai, and L. Carin, "Non-negative matrix factorization for discrete data with hierarchical side-information," in *International Conference on Artificial Intelligence and Statistics*, 2016, pp. 1124–1132.

[23] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, 2015, pp. 3483–3491.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.

[25] D. M. Blei, M. I. Jordan *et al.*, "Variational inference for dirichlet process mixtures," *Bayesian analysis*, vol. 1, no. 1, pp. 121–143, 2006.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, vol. abs/1312.6114, 2013.

[27] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[28] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed-membership models of scientific publications," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5220–5227, 2004.

[29] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 424–433.

[30] X. Wang, N. Mohanty, and A. McCallum, "Group and topic discovery from relations and their attributes," in *Advances in Neural Information Processing Systems*, 2006, pp. 1449–1456.

[31] D. Newman, C. Chemudugunta, and P. Smyth, "Statistical entity-topic models," in *Proceedings of the 12th ACM SIGKDD international*

*conference on Knowledge discovery and data mining*. ACM, 2006, pp. 680–686.
[32] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 127–134.
[33] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "Topic and role discovery in social networks," in *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30 - August 5, 2005*, 2005, pp. 786–791.
[34] L. Dietz, S. Bickel, and T. Scheffer, "Unsupervised prediction of citation influences," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 233–240.
[35] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in Neural Information Processing Systems*, 2014, pp. 3581–3589.
[36] A. Srivastava and C. Sutton, "Autoencoding variational inference for topic models," in *International Conference on Learning Representations*, 2017.
[37] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.
[38] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Józefowicz, and S. Bengio, "Generating sentences from a continuous space," in *Conference on Computational Natural Language Learning*, 2016, pp. 10–21.
[39] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "$\beta$-vae: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2016.
[40] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.
[41] K. Lang, "Newsweeder: Learning to filter netnews," in *International Conference on Machine Learning*, 1995, pp. 331–339.
[42] D. D. Lewis, "Reuters 21578 dataset," 1997.
[43] A. Cardoso-Cachopo, "Improving Methods for Single-label Text Categorization," PdD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa, 2007.
[44] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?" *Journal of Biomedical Informatics*, vol. 42, no. 5, pp. 760–772, 2009.

## APPENDIX

### A. Derivation

The derivation for variational lower bound of the conditional log-likelihood (Equation 6) is given bellow:

$$\begin{aligned}
\log p(w_{1:T}|x_d) &= D_{KL}(q(\theta|x_d, w_{1:T})||p(\theta|x_d, w_{1:T})) \\
&\quad + \mathbb{E}_{q(\theta|x_d,w_{1:T})}[-\log q(\theta|x_d, w_{1:T}) \\
&\quad + \log p(w_{1:T}, \theta|x_d)] \\
&\geq \mathbb{E}_{q(\theta|x_d,w_{1:T})}[-\log q(\theta|x_d, w_{1:T}) \\
&\quad + \log p(w_{1:T}, \theta|x_d)] \\
&= \mathbb{E}_{q(\theta|x_d,w_{1:T})}[-\log q(\theta|x_d, w_{1:T}) \\
&\quad + \log p(\theta|x_d)] \\
&\quad + \mathbb{E}_{q(\theta|x_d,w_{1:T})}[\log p(w_{1:T}|x_d, \theta)] \\
&= -D_{KL}(q(\theta|x_d, w_{1:T})||p(\theta|x_d)) \\
&\quad + \mathbb{E}_{q(\theta|x_d,w_{1:T})}[\log p(w_{1:T}|x_d, \theta)] \\
&= -D_{KL}(q(\theta|x_d, \tilde{w}_{1:T})||p(\theta|x_d)) \\
&\quad + \mathbb{E}_{q(\theta|x_d,w_{1:T})}[\log p(w_{1:T}|x_d, \theta)]
\end{aligned} \tag{13}$$

### B. Dimension of the parameter of MeTRNN

The dimension of the weight matrices in MeTRNN are summarized in Table V.

TABLE V
DIMENSION OF THE PARAMETERS OF METRNN. $H$ IS THE SIZE OF THE HIDDEN STATE, $V$ IS THE SIZE OF THE VOCABULARY, $K$ IS THE NUMBER OF TOPICS AND $M$ IS THE SIZE OF THE META-INFORMATION VECTOR.

| | | | |
|---|---|---|---|
| Matrix | $W_{hh}$ | $W_{wh}$ | $W_{hw}$ |
| Dimension | $H \times H$ | $V \times H$ | $H \times V$ |
| Matrix | $W_{\theta w}$ | $W_{mw}$ | $W_{w\mu}$ |
| Dimension | $K \times V$ | $M \times V$ | $V \times K$ |
| Matrix | $W_{w\sigma}$ | $W_{m\mu}$ | $W_{m\sigma}$ |
| Dimension | $V \times K$ | $M \times K$ | $M \times K$ |

### C. Hyper-parameter

The hyper-parameter configurations of all the baselines and variations on three different datasets previously not reported in Section IV are listed in Table VI for reproducibility.

TABLE VI
HYPER-PARAMETER SETTINGS OF DIFFERENT METHODS. "LR" STANDS FOR LEARNING RATE. "BZ" STANDS FOR BATCH SIZE. $\alpha$ IS THE PARAMETER FOR DIRICHLET PRIOR. $\beta$ IS A SCALING PARAMETER IN METRNN EXPLAINED IN SECTION III-D.

| | **20 NG** | | | |
|---|---|---|---|---|
| **Methods** | **lr** | **bz** | $\alpha$ | $\beta$ |
| GRU | 0.05 | 32 | N/A | N/A |
| LSTM | 0.05 | 32 | N/A | N/A |
| cRNN | 0.05 | 32 | N/A | N/A |
| cGRU | 0.05 | 32 | N/A | N/A |
| cLSTM | 0.05 | 32 | N/A | N/A |
| TopicGRU | 0.05 | 32 | N/A | N/A |
| TopicLSTM | 0.05 | 32 | N/A | N/A |
| ProdLDA | 0.1 | 32 | 1 | N/A |
| MeTGRU | 0.05 | 32 | 1 | 0.01 |
| MeTLSTM | 0.05 | 32 | 1 | 0.01 |
| | **R52** | | | |
| GRU | 0.01 | 32 | N/A | N/A |
| LSTM | 0.01 | 32 | N/A | N/A |
| cRNN | 0.01 | 32 | N/A | N/A |
| cGRU | 0.01 | 32 | N/A | N/A |
| cLSTM | 0.01 | 32 | N/A | N/A |
| TopicGRU | 0.01 | 32 | N/A | N/A |
| TopicLSTM | 0.01 | 32 | N/A | N/A |
| ProdLDA | 0.1 | 32 | 1 | N/A |
| MeTGRU | 0.01 | 32 | 1 | 0.01 |
| MeTLSTM | 0.01 | 32 | 1 | 0.01 |
| | **MADE** | | | |
| GRU | 0.15 | 16 | N/A | N/A |
| LSTM | 0.15 | 16 | N/A | N/A |
| cRNN | 0.15 | 16 | N/A | N/A |
| cGRU | 0.15 | 16 | N/A | N/A |
| cLSTM | 0.15 | 16 | N/A | N/A |
| TopicGRU | 0.15 | 16 | N/A | N/A |
| TopicLSTM | 0.15 | 16 | N/A | N/A |
| ProdLDA | 0.1 | 32 | 1 | N/A |
| MeTGRU | 0.15 | 16 | 1 | 0.1 |
| MeTLSTM | 0.15 | 16 | 1 | 0.1 |