Neural Topic Modeling with Continual Lifelong Learning

Pankaj Gupta [1] **Yatin Chaudhary** [1 2] **Thomas Runkler** [1] **Hinrich Schütze** [2]

Abstract

Lifelong learning has recently attracted attention in building machine learning systems that continually accumulate and transfer knowledge to help future learning. Unsupervised topic modeling has been popularly used to discover topics from document collections. However, the application of topic modeling is challenging due to data sparsity, e.g., in a small collection of (short) documents and thus, generate incoherent topics and sub-optimal document representations. To address the problem, we propose a lifelong learning framework for neural topic modeling that can continuously process streams of document collections, accumulate topics and guide future topic modeling tasks by knowledge transfer from several sources to better deal with the sparse data. In the lifelong process, we particularly investigate jointly: (1) sharing generative homologies (latent topics) over lifetime to transfer prior knowledge, and (2) minimizing catastrophic forgetting to retain the past learning via novel selective data augmentation, co-training and topic regularization approaches. Given a stream of document collections, we apply the proposed Lifelong Neural Topic Modeling (LNTM) framework in modeling three sparse document collections as future tasks and demonstrate improved performance quantified by perplexity, topic coherence and information retrieval task. Code: `https://github.com/pgcool/Lifelong-Neural-Topic-Modeling`

1. Introduction

Unsupervised topic models, such as LDA (Blei et al., 2003), RSM (Salakhutdinov & Hinton, 2009), DocNADE (Lauly et al., 2017), NVDM (Srivastava & Sutton, 2017), etc. have been popularly used to discover topics from large document collections. However in sparse data settings, the application of topic modeling is challenging due to limited context in a small document collection or short documents (e.g., tweets, headlines, etc.) and the topic models produce incoherent topics. To deal with this problem, there have been several attempts (Petterson et al., 2010; Das et al., 2015; Nguyen et al., 2015; Gupta et al., 2019) that introduce prior knowledge such as pre-trained word embeddings (Pennington et al., 2014) to guide meaningful learning.

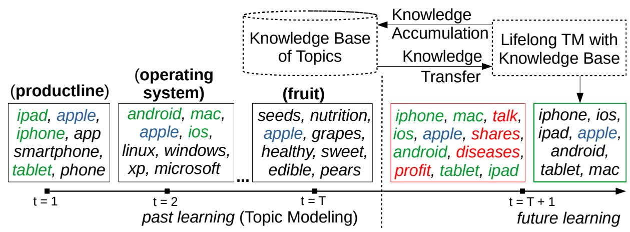


Figure 1. Motivation for Lifelong Topic Modeling

Lifelong Machine Learning (LML) (Thrun & Mitchell, 1995; Mitchell et al., 2015; Hassabis et al., 2017; Parisi et al., 2019) has recently attracted attention in building adaptive computational systems that can continually acquire, retain and transfer knowledge over life time when exposed to modeling continuous streams of information. In contrast, the traditional machine learning is based on isolated learning i.e., a one-shot task learning (OTL) using a single dataset and thus, lacks ability to continually learn from incrementally available heterogeneous data. The application of LML framework has shown potential for supervised natural language processing (NLP) tasks (Chen & Liu, 2016) such as in sentiment analysis (Chen et al., 2015), relation extraction (Wang et al., 2019), text classification (de Masson d'Autume et al., 2019), etc. Existing works in topic modeling are either based on the OTL approach or transfer learning (Chen & Liu, 2014) using stationary batches of training data and prior knowledge without accounting for streams of document collections. The unsupervised document (neural) topic modeling still remains unexplored regarding lifelong learning.

In this work, we explore unsupervised document (neural) topic modeling within a continual lifelong learning paradigm to enable knowledge-augmented topic learning over lifetime. We show that *Lifelong Neural Topic Modeling* (LNTM) is capable of mining and retaining prior knowledge

[1]Corporate Technology, Siemens AG Munich, Germany [2]CIS, University of Munich (LMU) Munich, Germany. Correspondence to: Pankaj Gupta <pankaj.gupta@drimco.net>.

Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

(topics) from streams of large document collections, and particularly guiding topic modeling on sparse datasets using accumulated knowledge of several domains over lifespan. For example in Figure 1, we have a stream of coherent topics associated with *apple* extracted from a stream of large document collections over time $t \in [1, T]$ (i.e., past learning). Observe that the word *apple* is topically contextualized by several domains, i.e., `productline`, `operating system` and `fruit` at tasks $t = 1$, $t = 2$ and $t = T$, respectively. For the future task $T + 1$ on a small document collection, the topic (red box) produced without LNTM is incoherent, containing some irrelevant words (marked in red) from various topics. Given a sufficient overlap (marked in green) in the past and future topic words, we aim to help topic modeling for the future task $T + 1$ such that the topic (red box) becomes semantically coherent (green box), leading to generate an improved document representation.

Therefore, the goal of LNTM is to (1) detect topic overlap in prior topics $t \in [1, T]$ of the knowledge base (KB) and topics of future task $T+1$, (2) positively transfer prior topic information in modeling future task, (3) retain or minimize forgetting of prior topic knowledge, and (4) continually accumulate topics in KB over life time. In this work, we particularly focus on addressing the challenge: *how to simultaneously mine relevant knowledge from prior topics, transfer mined topical knowledge and also retain prior topic information under domain shifts over lifespan?*

Contributions: We present a novel lifelong neural topic modeling framework that learns topics for a future task with proposed approaches of: (1) *Topic Regularization* that enables topical knowledge transfer from several domains and prevents catastrophic forgetting in the past topics, (2) *Word-embedding Guided Topic Learning* that introduces prior multi-domain knowledge encoded in word-embeddings, and (3) *Selective-data Augmentation Learning* that identifies relevant documents from historical collections, learns topics simultaneously with a future task and controls forgetting due to selective data replay. We apply the proposed framework in modeling three sparse (future task) and four large (past tasks) document collections in sequence. Intensive experimental results show improved topic modeling on future task while retaining past learning, quantified by information retrieval, topic coherence and generalization capabilities.

2. Methodology: Lifelong Topic Modeling

In following section, we describe our contributions in building Lifelong Neural Topic Modeling framework including: topic extraction, knowledge mining, retention, transfer and accumulation. *See Table 1 for the description of notations.*

Consider a stream of document collections **S** = $\{\Omega^1, \Omega^2, ..., \Omega^T, \Omega^{T+1}\}$ over lifetime $t \in [1, ..., T, T+1]$, where Ω^{T+1} is used to perform future learning. During the lifelong learning, we sequentially iterate over $\mathbf{S}$ and essentially analyze a document collection $\Omega^t \in \mathbf{S}$ using a novel topic modeling framework that can leverage and retain prior knowledge extracted from each of the lifelong steps $\{1, ..., t-1\}$.

Table 1. Description of the notations used in this work

Notation	Description
LNTM	Lifelong Neural Topic Modeling
EmbTF	Word Embedding based transfer
TR	Topic Regularization
SAL	Selective-data Augmentation Learning
`TopicPool`	Pool of accumulated topics
`WordPool`	Pool of accumulated word embeddings
Ω^t	A document collection at time/task t
$(T+1)$	Future task
$\{1, ..., T\}$	Past tasks
$\mathbf{Z}^t \in \mathbb{R}^{H \times K}$	Topic Embedding matrix for task t
$\mathbf{E}^t \in \mathbb{R}^{E \times K}$	Word Embedding matrix for task t
$\mathbf{\Theta}$	LNTM parameters
$\mathbf{\Phi}$	LNTM hyper-parameters
λ^t_{EmbTF}	Degree of relevance of $\mathbf{E}^t \in$ `WordPool` for $(T+1)$
λ^t_{TR}	Degree of topic imitation/forgetting of $\mathbf{Z}^t$ by $\mathbf{Z}^{T+1}$
λ^t_{SAL}	Degree of domain-overlap in Ω^t and Ω^{T+1}
$\mathbf{A}^t \in \mathbb{R}^{H \times H}$	Topic-alignment in $\mathbf{Z}^t$ and Z^{T+1}
K, D	Vocabulary size, document size
E, H	Word embedding dimension, #topics
$\mathbf{b} \in \mathbb{R}^K$	Visible (input) bias vector
$\mathbf{c} \in \mathbb{R}^H$	Hidden (input) bias vector
$\mathbf{v}$	An input document (visible units)
$\mathcal{L}^t$	Loss (negative log-likelihood) for task t
$\mathbf{W} \in \mathbb{R}^{H \times K}$	Encoding matrix of DocNADE for task $(T+1)$
$\mathbf{U} \in \mathbb{R}^{K \times H}$	Decoding matrix of DocNADE for task $(T+1)$

2.1. Topic Learning via Neural Topic Model

Within the OTL framework, an unsupervised neural-network based topic model named as Document Neural Autoregressive Distribution Estimation (DocNADE) (Larochelle & Lauly, 2012; Lauly et al., 2017) has shown to outperform existing topic models based on LDA (Blei et al., 2003; Srivastava & Sutton, 2017) or neural networks such as Replicated Softmax (RSM) (Salakhutdinov & Hinton, 2009), Autoencoders (Lauly et al., 2017), NVDM (Miao et al., 2016) etc. Additionally, Gupta et al. (2019) have recently demonstrated competitiveness of DocNADE in transfer learning settings. Thus, we adopt DocNADE as the backbone in discovering topics and building lifelong topic learning framework.

DocNADE Formulation: For a document (observation vector) $\mathbf{v} \in \Omega$ of size D such that $\mathbf{v} = (v_1, ...v_D)$, each word index v_i takes a value in vocabulary $\{1, ..., K\}$ of size K. Inspired by NADE (Larochelle & Murray, 2011) and RSM (Salakhutdinov & Hinton, 2009) generative modeling architectures, DocNADE computes the joint probability distribution $p(\mathbf{v}; \mathbf{\Theta}) = \prod_{i=1}^{D} p(v_i|\mathbf{v}_{<i}; \mathbf{\Theta})$ of words in the

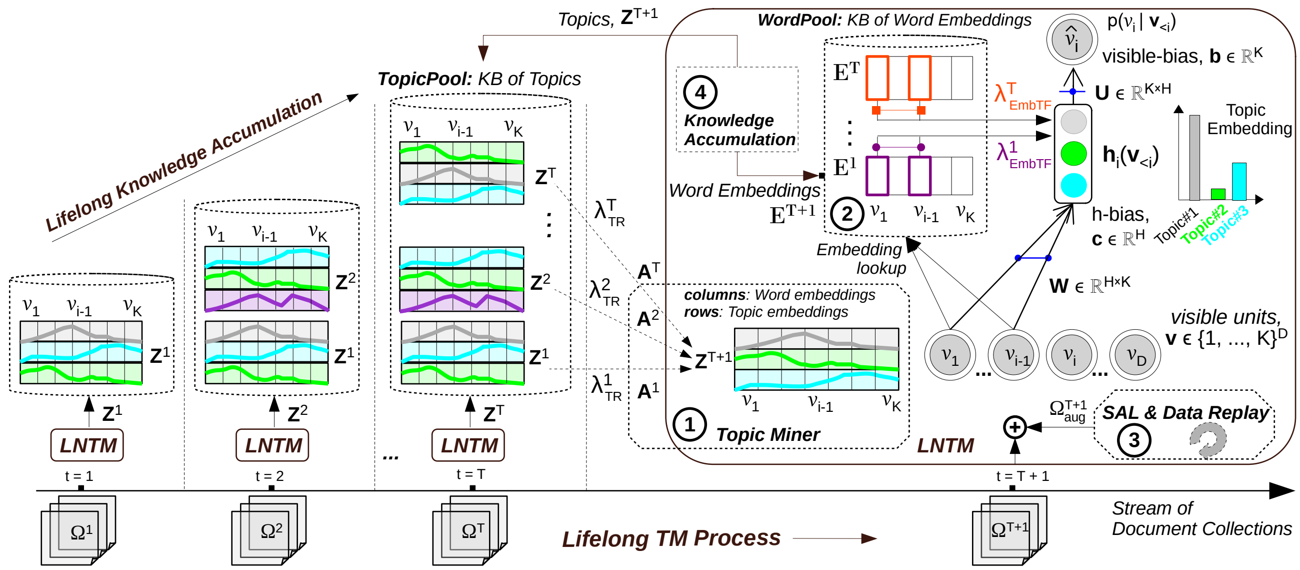


Figure 2. An illustration of the proposed Lifelong Neural Topic Modeling (LNTM) framework over s stream of document collections

document $\mathbf{v}$ by factorizing it as a product of conditional distributions $p(v_i|\mathbf{v}_{<i};\mathbf{\Theta})$, where each conditional is efficiently modeled via a feed-forward neural network using proceeding word $\mathbf{v}_{<i}$ in the sequence.

Following reconstruction principle, the DocNADE computes a hidden vector $\mathbf{h}_i(\mathbf{v}_{<i})$ at each autoregressive step:

$$\mathbf{h}_i(\mathbf{v}_{<i}) = g(\mathbf{c}+\sum_{q<i}\mathbf{W}_{:,v_q}) \text{ and } g = \{sigmoid, tanh\}$$

$$p(v_i = w|\mathbf{v}_{<i};\mathbf{\Theta}) = \frac{\exp(b_w + \mathbf{U}_{w,:}\mathbf{h}_i(\mathbf{v}_{<i}))}{\sum_{w'}\exp(b_{w'} + \mathbf{U}_{w',:}\mathbf{h}_i(\mathbf{v}_{<i}))}$$

for each $i \in \{1, ...D\}$, where $\mathbf{v}_{<i} \in \{v_1, ..., v_{i-1}\}$ is a subvector consisting of all v_q such that $q < i$. $\mathbf{\Theta}$ is a collection of parameters including weight matrices $\{\mathbf{W} \in \mathbb{R}^{H\times K}$, $\mathbf{U} \in \mathbb{R}^{K\times H}\}$ and biases $\{\mathbf{c} \in \mathbb{R}^H$, $\mathbf{b} \in \mathbb{R}^K\}$. H and K are the number of hidden units (topics) and vocabulary size.

Figure 2 (rightmost; without components ①, ②, ③ and ④) illustrates the DocNADE architecture, computing the probability $\hat{v}_i = p(\mathbf{v}_i|\mathbf{v}_{<i};\mathbf{\Theta})$ of the i^{th} word v_i conditioned on position dependent hidden layer $\mathbf{h}_i(\mathbf{v}_{<i})$. The parameter $\mathbf{W}$ is shared in the feed-forward networks and $\mathbf{h}_i$ encodes topic proportion for the document $\mathbf{v}$.

Algorithm 1 (lines #1-4) and TOPIC-LEARNING utility (algorithm 2) describe the computation of objective function: negative log-likelihood $\mathcal{L}(\mathbf{v};\mathbf{\Theta})$ that is minimized using stochastic gradient descent. In terms of model complexity, computing $\mathbf{h}_i(\mathbf{v}_{<i})$ is efficient (linear complexity) due to NADE (Larochelle & Murray, 2011) architecture that leverages the pre-activation $\mathbf{a}_{i-1}$ of $(i-1)$th step in computing $\mathbf{a}_i$. The complexity of computing all hidden layers $\mathbf{h}_i(\mathbf{v}_{<i})$ is in $O(DH)$ and all $p(\mathbf{v}_i|\mathbf{v}_{<i};\mathbf{\Theta})$ in $O(KDH)$ for D words in the document $\mathbf{v}$. Thus, the total *complexity* of computing the joint distribution $p(\mathbf{v})$ is in $O(DH + KDH)$.

Importantly, the topic-word matrix $\mathbf{W} \in \mathbb{R}^{H\times K}$ has a property that the row-vector $\mathbf{W}_{j,:}$ encodes jth topic (distribution over vocabulary words), i.e., topic-embedding whereas the column-vector $\mathbf{W}_{:,v_i}$ corresponds to embedding of the word v_i, i.e., word-embedding. We leverage this property to introduce prior knowledge via topic and word embeddings during lifelong learning. Additionally, we accumulate all topic and word embeddings in `TopicPool` and `WordPool`, respectively learned over lifetime.

2.2. Lifelong Learning in Neural Topic Modeling

Given the prior knowledge (`TopicPool` and `WordPool`), a stream of document collections $\mathbf{S}$ and a new (future) topic learning task on document collection Ω^{T+1}, the proposed LNTM framework operates in two phases:

Phase 1: Joint Topic Mining, Transfer and Retention: The task of topic modeling with lifelong learning capabilities is prone to three main challenges: (a) mining prior knowledge relevant for the future task $T + 1$, (b) learning with prior knowledge, and (c) minimizing catastrophic forgetting, i.e., retaining of prior knowledge. Here, the prior knowledge refers to topic and word embeddings extracted from the historical tasks $\{1, ...T\}$. In modeling a future task $T + 1$, we address the above challenges by jointly mining, transferring and retaining prior knowledge. Algorithms 1 and 2 demonstrate the following three **approaches** within lifelong neural topic modeling, `LNTM`= {`TR`, `EmbTF`, `SAL`}:

① **Topic Regularization with** `TopicPool` (`TR`): To address the learning without forgetting, several works (Jung et al., 2016; Kirkpatrick et al., 2017; Zenke et al., 2017; Li & Hoiem, 2018) have investigated regularization approaches in building LML systems that constrain the updates of neural weights in modeling the future task (i.e., $T+1$) such that catastrophic forgetting with all the previously learned tasks is minimized. These existing works majorly focus on building LML systems dealing with computer vision tasks mostly in supervised fashion. Lifelong topic and document representation learning in unsupervised fashion has received considerable less attention. Thus inspired by the regularization strategies, we regularize topics of the past and future tasks in a way that not only minimizes forgetting of prior topics but also maximizes topical knowledge transfer for the future task (i.e., unsupervised topic modeling).

Given a pool of prior topics, i.e., `TopicPool` built by accumulating topics from each of the past tasks, we perform topic mining for the future task $T+1$ using DocNADE. In doing so, the topic learning $\mathbf{Z}^{T+1}$ on document collection Ω^{T+1} is guided by all the past topics $[\mathbf{Z}^1, ...\mathbf{Z}^T] \in$ `TopicPool`, building a *Topic Miner* that consists of:

(a) *Topic Extractor*: A topic is essentially a distribution over vocabulary that explains thematic structures in the document collection. In modeling a stream of document collections, the vocabulary size may not be same in tasks over lifetime and thus topic analogy (e.g., shifts, overlap, etc.) requires common vocabulary words in the participating topics. Illustrated in Figure 2, each latent topic vector in $\mathbf{Z}^{T+1}$ (marked by ①) of the future task $T+1$ encodes a distribution over words appearing in the past tasks, e.g., $\mathbf{Z}^T \in$ `TopicPool`. As discussed in section 2.1, the topics $\mathbf{Z}^{T+1}$ can be obtained from the row-vectors of $\mathbf{W} \in \mathbf{\Theta}^{T+1}$ by masking all its column-vectors v_i not in the past.

(b) *Topic Regularizer*: Given `TopicPool`, we model the future task by introducing an additional topic-regularization term Δ_{TR} in its objective function $\mathcal{L}(\Omega^{T+1}; \mathbf{\Theta}^{T+1})$:

$$\Delta_{TR} = \sum_{t=1}^{T} \lambda_{TR}^t \big(\underbrace{||\mathbf{Z}^t - \mathbf{A}^t \mathbf{Z}^{T+1}||_2^2}_{topic-imitation} + \underbrace{||\mathbf{U}^t - \mathbf{P}^t \mathbf{U}||_2^2}_{decoder-proximity} \big)$$

$$\mathcal{L}(\Omega^{T+1}; \mathbf{\Theta}^{T+1}) = \textstyle\sum_{\mathbf{v} \in \Omega^{T+1}} \mathcal{L}(\mathbf{v}; \mathbf{\Theta}^{T+1}) + \Delta_{TR}$$

such that the first term (*topic-imitation*) allows controlled knowledge transfer by inheriting relevant topic(s) in $\mathbf{Z}^{T+1}$ from `TopicPool`, accounting for domain-shifts via a topic-alignment matrix $\mathbf{A}^t \in \mathbb{R}^{H \times H}$ for every prior task. Moreover, the two terms together preserve the prior learning with encoder and decoder proximity, respectively due to a quadratic penalty on the selective difference between the parameters for the past and future topic modeling tasks, such that the parameters $\mathbf{\Theta}^{T+1}$ also retain representation capabilities for the document collections in the past, e.g., $\mathcal{L}(\Omega^t; \mathbf{\Theta}^t) \sim \mathcal{L}(\Omega^t; \mathbf{\Theta}^{T+1})$. Here, λ_{TR}^t is per-task regularization strength that controls the degree of topic-imitation and forgetting of prior learning t by the future task $T+1$. $(\mathbf{Z}^t, \mathbf{U}^t) \in \mathbf{\Theta}^t$ are parameters at the end of the past task t.

Algorithm 1 Lifelong Neural Topic Modeling using DocNADE

input Sequence of document collections $\{\Omega^1, ...\Omega^T, ...\Omega^{T+1}\}$
input Past learning: $\{\mathbf{\Theta}^1, ..., \mathbf{\Theta}^T\}$
input `TopicPool`: $\{\mathbf{Z}^1, ..., \mathbf{Z}^T\}$
input `WordPool`: $\{\mathbf{E}^1, ..., \mathbf{E}^T\}$
parameters $\mathbf{\Theta}^{T+1} = \{\mathbf{b}, \mathbf{c}, \mathbf{W}, \mathbf{U}, \mathbf{A}^1, ..., \mathbf{A}^T, \mathbf{P}^1, ..., \mathbf{P}^T\}$
hyper-paramaters $\mathbf{\Phi}^{T+1} = \{H, \lambda_{LNTM}^1, ..., \lambda_{LNTM}^T\}$
1: **Neural Topic Modeling**:
2: `LNTM` = {}
3: Train a topic model and get PPL on test set Ω_{test}^{T+1}:
4: PPL^{T+1}, $\mathbf{\Theta}^{T+1} \leftarrow$ topic-learning(Ω^{T+1}, $\mathbf{\Theta}^{T+1}$)

5: **Lifelong Neural Topic Modeling (LNTM) framework**:
6: `LNTM` = {`EmbTF, TR, SAL`}
7: For a document $\mathbf{v} \in \Omega^{T+1}$:
8: Compute loss (negative log-likelihood):
9: $\mathcal{L}(\mathbf{v}|\mathbf{\Theta}^{T+1}) \leftarrow$ compute-NLL($\mathbf{v}$, $\mathbf{\Theta}^{T+1}$, `LNTM`)
10: **if** `TR` in `LNTM` **then**
11: Jointly minimize-forgetting and learn with `TopicPool`:
12: $\Delta_{TR} \leftarrow \sum_{t=1}^{T} \lambda_{TR}^t \, (||\mathbf{Z}^t - \mathbf{A}^t \mathbf{Z}^{T+1}||_2^2 + ||\mathbf{U}^t - \mathbf{P}^t \mathbf{U}||_2^2)$
13: $\mathcal{L}(\mathbf{v}; \mathbf{\Theta}^{T+1}) \leftarrow \mathcal{L}(\mathbf{v}; \mathbf{\Theta}^{T+1}) + \Delta_{TR}$
14: **end if**
15: **if** `SAL` in `LNTM` **then**
16: Detect domain-overlap and select relevant historical documents from $[\Omega^1, ..., \Omega^T]$ for augmentation at task (T+1):
17: $\Omega_{aug}^{T+1} \leftarrow$ distill-documents($\mathbf{\Theta}^{T+1}$, PPL^{T+1}, $[\Omega^1, ..., \Omega^T]$)
18: Perform augmented learning (co-training) with Ω_{aug}^{T+1}:
19: $\Delta_{SAL} \leftarrow \sum_{(\mathbf{v}^t, t) \in \Omega_{aug}^{T+1}} \lambda_{SAL}^t \, \mathcal{L}(\mathbf{v}^t; \mathbf{\Theta}^{T+1})$
20: $\mathcal{L}(\mathbf{v}; \mathbf{\Theta}^{T+1}) \leftarrow \mathcal{L}(\mathbf{v}; \mathbf{\Theta}^{T+1}) + \Delta_{SAL}$
21: **end if**
22: Minimize $\mathcal{L}(\mathbf{v}; \mathbf{\Theta}^{T+1})$ using stochastic gradient-descent
23: **Knowledge Accumulation**:
24: `TopicPool` $\leftarrow$ accumulate-topics($\mathbf{\Theta}^{T+1}$)
25: `WordPool` $\leftarrow$ accumulate-word-embeddings($\mathbf{\Theta}^{T+1}$)

Figure 2 (*Topic Miner* component ①) and Algorithm 1 (lines #10-14) demonstrate the `TR` approach in LNTM framework. The topic regularization Δ_{TR} approach enables jointly mining, transferring and retaining prior topics when learning future topics continually over lifetime.

② **Transfer Learning with** `WordPool` (`EmbTF`): Beyond topical knowledge, we also leverage pre-trained word embeddings (complementary to topics) accumulated in `WordPool` during lifelong learning. Essentially, we pool word embedding representation for every word v_i learned while topic modeling over a stream of document collections from several domains. Thus, we have in total T number of embeddings (encoding different semantics) for a word v_i in `WordPool`, if the word appears in all the past collections. Following Gupta et al. (2019), we introduce prior knowledge in form of pre-trained word embeddings $[\mathbf{E}^1, ..., \mathbf{E}^T]$

Algorithm 2 Lifelong Learning Utilities

1: **function** topic-learning (Ω, $\mathbf{\Theta}$)
2: Build a DocNADE neural topic model: Initialize $\mathbf{\Theta}$
3: **for** $\mathbf{v} \in \Omega_{train}$ **do**
4: Forward-pass:
5: Compute loss, $\mathcal{L}(\mathbf{v}; \mathbf{\Theta}) \leftarrow$ compute-NLL($\mathbf{v}$, $\mathbf{\Theta}$)
6: Backward-pass:
7: Minimize $\mathcal{L}(\mathbf{v}; \mathbf{\Theta})$ using stochastic gradient-descent
8: **end for**
9: Compute perplexity PPL of test set Ω_{test}:
10: PPL $\leftarrow \exp(\frac{1}{|\Omega_{test}|} \sum_{\mathbf{v} \in \Omega_{test}} \frac{\mathcal{L}(\mathbf{v};\mathbf{\Theta})}{|\mathbf{v}|})$
11: return PPL, $\mathbf{\Theta}$
12: **end function**

13: **function** compute-NLL ($\mathbf{v}$, $\mathbf{\Theta}$, `LNTM` = {})
14: Initialize $\mathbf{a} \leftarrow \mathbf{c}$ and $p(\mathbf{v}) \leftarrow 1$
15: **for** word $i \in [1, ..., N]$ **do**
16: $\mathbf{h}_i(\mathbf{v}_{<i}) \leftarrow g(\mathbf{a})$, where g = {sigmoid, tanh}
17: $p(v_i = w|\mathbf{v}_{<i}) \leftarrow \frac{\exp(b_w + \mathbf{U}_{w,:}\mathbf{h}_i(\mathbf{v}_{<i}))}{\sum_{w'} \exp(b_{w'} + \mathbf{U}_{w',:}\mathbf{h}_i(\mathbf{v}_{<i}))}$
18: $p(\mathbf{v}) \leftarrow p(\mathbf{v}) p(v_i|\mathbf{v}_{<i})$
19: Compute pre-activation at i^{th} step: $\mathbf{a} \leftarrow \mathbf{a} + \mathbf{W}_{:,v_i}$
20: **if** `EmbTF` in `LNTM` **then**
21: Get word-embedding vectors for v_i from `WordPool`:
22: $\mathbf{a} \leftarrow \mathbf{a} + \sum_{t=1}^{T} \lambda^t_{EmbTF}\ \mathbf{W}^t_{:,v_i}$
23: **end if**
24: **end for**
25: return $-\log p(\mathbf{v}; \mathbf{\Theta})$
26: **end function**

27: **function** distill-documents ($\mathbf{\Theta}^{T+1}$, PPL^{T+1}, $[\Omega^1, ..., \Omega^T]$)
28: Initialize a set of selected documents: $\Omega^{T+1}_{aug} \leftarrow \{\}$
29: **for** task $t \in [1, ..., T]$ and document $\mathbf{v}^t \in \Omega^t$ **do**
30: $\mathcal{L}(\mathbf{v}^t; \mathbf{\Theta}^{T+1}) \leftarrow$ compute-NLL($\mathbf{v}^t$, $\mathbf{\Theta}^{T+1}$, `LNTM` = {})
31: PPL($\mathbf{v}^t; \mathbf{\Theta}^{T+1}$) $\leftarrow \exp(\frac{\mathcal{L}(\mathbf{v}^t; \mathbf{\Theta}^{T+1})}{|\mathbf{v}^t|})$
32: Select document $\mathbf{v}^t$ for augmentation in task $T+1$:
33: **if** PPL($\mathbf{v}^t; \mathbf{\Theta}^{T+1}$) $\leq$ PPL^{T+1} **then**
34: Document selected: $\Omega^{T+1}_{aug} \leftarrow \Omega^{T+1}_{aug} \cup (\mathbf{v}^t, t)$
35: **end if**
36: **end for**
37: return Ω^{T+1}_{aug}
38: **end function**

in each hidden layer of DocNADE when analyzing Ω^{T+1}:

$$\mathbf{h}(\mathbf{v}_{<i}) = g(\mathbf{c} + \sum_{q<i} \mathbf{W}_{:,v_q} + \sum_{q<i} \sum_{t=1}^{T} \lambda^t_{EmbTF}\ \mathbf{E}^t_{:,v_q})$$

Observe that the topic learning for task $T+1$ is guided by an embedding vector $\mathbf{E}_{:,v_q}$ for the word v_q from each of the T domains (sources), where λ^t_{EmbTF} is per-task transfer strength that controls the amount of prior (relevant) knowledge transferred to $T+1$ based on domain overlap with the past task t. Discussed in section 2.1, the word embedding representation $\mathbf{E}^t \in$ `WordPool` is obtained from the column-vectors of parameter $\mathbf{W}$ at the end of the task t.

Figure 2 (component ②), Algorithm 1 (lines #7-9) and Algorithm 2 (lines #20-23) illustrate the mechanism of topic modeling (DocNADE) with pre-trained word embeddings $\mathbf{E}$ from several sources (i.e., multi-source transfer learning) when learning topics $\mathbf{Z}^{T+1}$ for the future task $T+1$.

③ **Selective-Data Augmentation Learning** (`SAL`): Beyond the weight-based approaches in LML, the data-based approaches (Robins, 1995) augment the training data of a future task with the data collected from the past tasks, allowing for (a) multi-task learning (MTL) (Collobert & Weston, 2008; Ruder, 2017) to share representations among tasks and (b) minimizing catastrophic forgetting by data replay (augmentation). However, the data augmentation (DA) approaches are inefficient when the data collection grows large and often penalize positive transfer in MTL due to domain shifts in the stream of data over lifetime.

Our approach of SAL works in the following two steps:

Step 1 *Document Distillation* (Algorithm 1: line #17 and Algorithm 2: lines #27-38): Given document collections $[\Omega^1,...,\Omega^T]$ of the past tasks, we ignore documents found not relevant in modeling a future task due to domain shifts. To do so, we first build a topic model with parameters $\mathbf{\Theta}^{\mathbf{T+1}}$ over Ω^{T+1} and compute an average perplexity (PPL^{T+1}) score on its test set Ω^{T+1}_{test}. Then, we prepare an augmented set $\Omega^{T+1}_{aug} \subset [\Omega^1, ...\Omega^T]$ such that each document $\mathbf{v}^t \in \Omega^{T+1}_{aug}$ of a past task t satisfies: PPL($\mathbf{v}^t; \mathbf{\Theta}^{T+1}$) $\leq$ PPL^{t+1}. In essence, this unsupervised document distillation scheme detects domain-overlap in the past and future tasks based on representation ability of $\mathbf{\Theta}^{T+1}$ for documents of the past.

Step 2 *Selective Co-training* (Algorithm 1: lines #18-20): We re-train topic modeling over Ω^{T+1} simultaneously using Ω^{T+1}_{aug}, leveraging topical homologies in (selective) documents of the past and future tasks, as:

$$\Delta_{SAL} = \sum_{(\mathbf{v}^t, t) \in \Omega^{T+1}_{aug}} \lambda^t_{SAL}\ \mathcal{L}(\mathbf{v}^t; \mathbf{\Theta}^{T+1})$$

$$\mathcal{L}(\Omega^{T+1}; \mathbf{\Theta}^{T+1}) = \sum_{\mathbf{v} \in \Omega^{T+1}} \mathcal{L}(\mathbf{v}; \mathbf{\Theta}^{T+1}) +\ \Delta_{SAL}$$

Here, λ^t_{SAL} is per-task contribution that modulates influence of shared representations while co-training with selected documents of the past task t. The SAL approach jointly helps in transferring prior knowledge from several domains, minimizing catastrophic forgetting and reduce training time due to selective data replay over lifetime.

Overall loss in LNTM framework: Combining the different approaches within the proposed lifelong learning paradigm, the overall loss in modeling documents Ω^{T+1} being the future (new) task $T+1$ is given by:

$$\mathcal{L}(\Omega^{T+1}; \mathbf{\Theta}^{T+1}) = \sum_{\mathbf{v} \in \Omega^{T+1}} \mathcal{L}(\mathbf{v}; \mathbf{\Theta}^{T+1}) + \Delta_{TR} + \Delta_{SAL}$$

Computation complexity of LNTM: In DocNADE (section 2.1) without LNTM, the complexity of computing the joint

Lifelong Neural Topic Modeling over a stream of document collections →

	Scores on historical data incurring Catastrophic Forgetting								Scores with Lifelong Knowledge Transfer					
	AGnews		TMN		R21578		20NS		20NSshort (Sparse Data)					
	PPL	P@0.02	PPL	P@0.02	PPL	P@0.02	PPL	P@0.02	PPL	P@5	P@10	P@0.02	COH	r-time (second)
LNTM + EmbTF + TR + SAL	471	0.786	696	0.647	380	0.724	541	0.248	**641**	**0.418**	**0.375**	**0.324**	**0.735**	50.1
LNTM + EmbTF + TR	469	0.786	698	0.649	382	0.724	541	0.247	672	0.376	0.341	0.297	0.728	2.80
LNTM + TR	466	0.786	690	0.649	382	0.724	538	0.249	655	0.380	0.345	0.306	0.719	2.76
LNTM + EmbTF	468	0.786	696	0.649	383	0.724	540	0.247	647	0.386	0.348	0.310	0.670	2.37
NTM without Lifelong Learning	454	0.785	584	0.651	311	0.726	470	0.268	646	0.375	0.335	0.290	0.667	1.98

Figure 3. PPL, P@R (precision@Recall), COH and r-time of LNTM system on future task, i.e., `20NSshort` over the stream **S1**

Lifelong Neural Topic Modeling over a stream of document collections →

	Scores on historical data incurring Catastrophic Forgetting								Scores with Lifelong Knowledge Transfer					
	AGnews		TMN		R21578		20NS		TMNtitle (Sparse Data)					
	PPL	P@0.02	PPL	P@0.02	PPL	P@0.02	PPL	P@0.02	PPL	P@5	P@10	P@0.02	COH	r-time (second)
LNTM + EmbTF + TR + SAL	493	0.786	698	0.650	390	0.720	535	0.249	707	**0.689**	**0.662**	**0.562**	0.745	852
LNTM + EmbTF + TR	484	0.786	701	0.650	390	0.720	534	0.249	723	0.679	0.656	0.556	**0.750**	29.38
LNTM + TR	483	0.784	700	0.650	389	0.725	534	0.249	736	0.676	0.652	0.554	0.743	29.75
LNTM + EmbTF	485	0.784	700	0.650	390	0.722	534	0.249	**666**	0.675	0.649	0.548	0.726	23.21
NTM without Lifelong Learning	454	0.785	584	0.651	311	0.726	470	0.268	706	0.665	0.634	0.521	0.709	22.23

Figure 4. PPL, P@R (precision@Recall), COH and r-time of LNTM system on future task, i.e., `20TMNtitle` over the stream **S2**

distribution $p(\mathbf{v})$ is in $O(DH + KDH)$. The complexity of computing Δ_{TR} and Δ_{SAL} are in $O(KH + KH)$ and $O(DH + KDH)$, respectively. The overall complexity of `LNTM` = {`EmbTF`, `TR`, `SAL`} is in $O(DH + KDH + KH + KH + DH + KDH) \sim O(DH + KDH + KH)$.

Phase 2: Lifelong Knowledge Accumulation: For each topic modeling task t, the phase 1 generates knowledge in form of topic and word embeddings that is respectively accumulated in `TopicPool` $\leftarrow$ row-vectors($\mathbf{W} \in \mathbf{\Theta}^t$) and `WordPool` $\leftarrow$ column-vectors($\mathbf{W} \in \mathbf{\Theta}^t$). Additionally, each decoding parameter $\mathbf{U} \in \mathbf{\Theta}^t$ is retained to be used in minimizing catastrophic forgetting (i.e., Δ_{TR}).

3. Experiments and Analysis

Streams of Document Collections: To demonstrate the applicability of our proposed LNTM framework, we prepare a stream of document collections consisting of four long-text (high-resource) corpora in sequence: `AGnews`, `TMN`, `R21578` and `20NS` (20NewsGroups), and three short-text (low-resource) corpora Ω^{T+1} as **future tasks** $T+1$: `20NSshort`, `TMNtitle` and `R21578title`. Thus, we perform lifelong topic learning over following three streams:
S1: `AGnews` $\rightarrow$ `TMN` $\rightarrow$ `R21578` $\rightarrow$ `20NS` $\rightarrow$ `20NSshort`
S2: `AGnews` $\rightarrow$ `TMN` $\rightarrow$ `R21578` $\rightarrow$ `20NS` $\rightarrow$ `TMNtitle`
S3: `AGnews` $\rightarrow$ `TMN` $\rightarrow$ `R21578` $\rightarrow$ `20NS` $\rightarrow$ `R21578title`
such that we demonstrate improved topic modeling for the three sparse document collections (Ω^{T+1}) at $T+1$. The order of Ωs is based on their decreasing sizes. See the *supplementary* for data description and domain overlap.

Baselines: Discussed in section 2.1, we adopt DocNADE (NTM: a neural topic modeling tool) and compare it with the proposed framework `LNTM` = {`EmbTF`, `TR`, `SAL`}. Moreover, we show topic learning in *zero-shot*, *few-shot* and *data augmentation* settings in the following section.

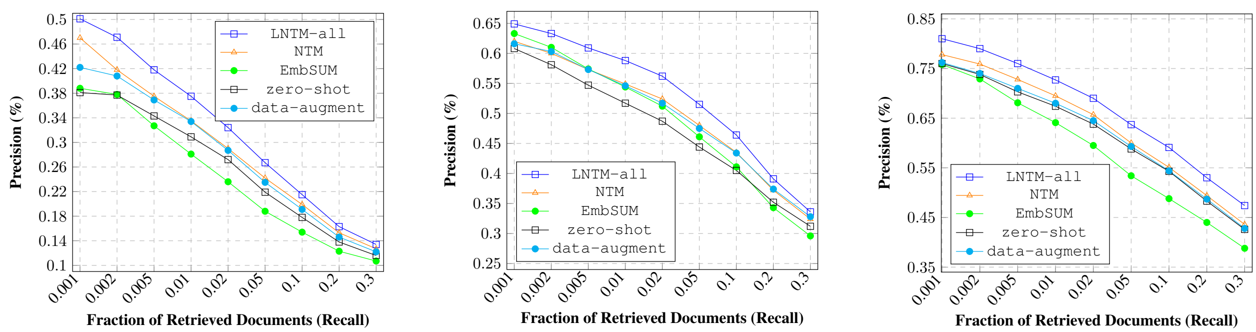
Reproducibility: PPL (Algorithm 2: line #10) is used for model selection and adjusting parameters $\mathbf{\Theta}^t$ and hyperparameters $\mathbf{\Phi}^t$. See the *supplementary* for the hyperparameters settings. Figures 3, 4 and 5 show average runtime (r-time) for each training epoch of different LNTM approaches, run on an NVIDIA Tesla K80 Processor (RAM: 12 GB) to a maximum of 100 epochs.

To evaluate the capabilities of LNTM framework, we employ three measures: precision@recall (P@R) in information retrieval (IR) task for *document representation*, topic coherence (COH) for *topic quality* and perplexity (PPL) for *generative performance* of topic modeling over lifetime.

	← Scores on historical data incurring Catastrophic Forgetting →								Scores with Lifelong Knowledge Transfer					r-time (second)
	AGnews		TMN		R21578		20NS		R21578title (Sparse Data)					
	PPL	P@0.02	PPL	P@0.02	PPL	P@0.02	PPL	P@0.02	PPL	P@5	P@10	P@0.02	COH	
LNTM + EmbTF + TR + SAL	533	0.789	699	0.648	438	0.721	531	0.251	194	**0.828**	**0.810**	**0.690**	0.747	519
LNTM + EmbTF + TR	550	0.788	703	0.650	444	0.721	532	0.251	203	0.812	0.786	0.676	**0.752**	12.63
LNTM + TR	571	0.787	704	0.649	451	0.722	532	0.251	208	0.810	0.770	0.668	0.742	12.18
LNTM + EmbTF	555	0.784	702	0.650	446	0.722	532	0.251	**183**	0.814	0.790	0.678	0.709	11.42
NTM without Lifelong Learning	454	0.785	584	0.651	311	0.726	470	0.268	192	0.799	0.778	0.657	0.713	10.49

Lifelong Neural Topic Modeling over a stream of document collections →

Figure 5. PPL, P@R (precision@Recall), COH and r-time of LNTM system on future task, i.e., `R21578title` over the stream **S3**



(a) **IR:** 20NSshort (b) **IR:** TMNtitle (c) **IR:** R21578title

Figure 6. Precision-recall curve on three future task datasets at different recall fractions. `LNTM-all` → `LNTM` + `EmbTF` + `TR` + `SAL`

3.1. Document Representation via Retrieval (IR)

To evaluate the quality of document representation learned within LNTM, we perform an unsupervised document retrieval task for each collection over lifetime. In doing so, we compute average P@R on the test set for a task t, where each test document is treated as a test query to retrieve a fraction/top R of the closest documents in the training set. We compute cosine similarity between document vectors (i.e., the last hidden $\mathbf{h}_D$ of DocNADE) and average the number of retrieved documents with the same label as the query. Figures 3, 4 and 5 show P@5, P@10 and P@0.02 on the all test collections of the streams **S1**, **S2** and **S3**, respectively accounting for knowledge transfer and forgetting.

Precision@Recall on future tasks: Figures 3, 4 and 5 report P@5, P@10 and P@0.02 scores (green boxes) on three future tasks: `20NSshort`, `TMNtitle` and `R21578title`, respectively leveraging prior knowledge over lifetime. Compared to NTM without lifelong learning (blue boxes), all the proposed approaches: `EmbTF`, `TR` and `SAL` (green boxes) within LNTM outperform it for all the future tasks, e.g., P@0.02: (.324 vs .290), (.562 vs .521) and (.690 vs .657) on `20NSshort`, `TMNtitle` and `R21578title`, respectively due to `LNTM+EmbTF+TR+SAL`. Observe that the `SAL` leads to higher gains when combined with the other LNTM approaches, suggesting a positive knowledge transfer from both the past learning and document collections.

Precision@Recall on past tasks incurring forgetting (orange boxes): To demonstrate the ability of LNTM framework in minimizing catastrophic forgetting, we also report P@R scores on the past tasks[1] using parameters of a future task $\mathbf{\Theta}^{T+1}$. Figures 3, 4 and 5 report P@0.02 for each of the past tasks over lifetime using **S1**, **S2** and **S3** streams, suggesting that the proposed approaches in LNTM help in preventing catastrophic forgetting. For each stream, compare scores in the orange and blue boxes column-wise correspondingly for each task. E.g., P@0.02 for `TMN` in **S1**, **S2** and **S3** incurring forgetting are (.647 vs .651), (.650 vs .651) and (.648 vs .651), respectively advocating for representation capabilities of the future tasks for the past learning within LNTM.

Zero-shot and Data-augmentation Investigations: Additionally, we analyze representation capabilities of LNTM in zero-shot and data-augmentation settings, where we compute P@R on all future tasks $T+1$ respectively using pa-

[1]Due to partially overlapping vocabulary in Ωs over a stream, we overwrite column-vectors of $\mathbf{W}{\in}\mathbf{\Theta}^t$ by column-vectors of $\mathbf{W}{\in}\mathbf{\Theta}^{T+1}$ for all words v_i appearing in both tasks t and $T+1$

rameters: (a) $\mathbf{\Theta}^T$ learned from the past task T and no Ω^{T+1} used, and (b) $\mathbf{\Theta}^{T+1}$ learned on a future task by combining all document collections $[\Omega^1, ..., \Omega^{T+1}]$ in a stream. Figures 6a, 6b and 6c show precision-recall plots for `20NSshort`, `TMNtitle` and `R21578title` datasets, respectively. Observe that the proposed approach `LNTM-all` (i.e., `LNTM` + `EmbTF` + `TR`+ `SAL`) outperforms NTM (i.e., DocNADE without lifelong learning), zero-shot, data-augment and `EmbSUM` baselines at all the retrieval fractions. Here, `EmbSUM` represents a document by summing the embedding vectors of its words using Glove embeddings (Pennington et al., 2014).

3.2. Topic Quality via Coherence (COH)

Beyond document representation, topic models essentially enable interpretability by generating topics (sets of key terms) that explain thematic structures hidden in document collections. Topics are often incoherent when captured in data sparsity (low-resource) settings, leading to restrict the interpretability. Thus, we compute topic coherence (COH) scores proposed by Röder et al. (2015) to estimate the quality (meaningfulness of words) of topics captured within LNTM framework. Following Gupta et al. (2019), we compute COH (Figures 3, 4 and 5) on the top-10 words in each topic. The higher scores imply topic coherency.

COH scores on future tasks: Within LNTM, we show a gain of 10.2% (0.735 vs 0.667), 5.8%(0.750 vs 0.709) and 5.5%(0.752 vs 0.713) respectively on the three sparse datasets, suggesting quality topics discovered.

Figures 7, 8 and 9 show topic coherence (COH) scores on document collections in streams S1,S2 and S3, respectively. We also show scores incurring forgetting on past tasks in each of the three streams. Our proposed lifelong topic modeling framework reports gains in topic coherence scores for each of the target (future) tasks and also minimizes catastrophic forgetting on the past tasks.

3.3. Generalization via Perplexity (PPL)

To evaluate generative performance of topic models, we estimate the log-probabilities for unseen test documents Ω^{T+1}_{test} of the future tasks, and compute the average held-out perplexity per word (Algorithm 2: line #10). Note that lower the PPL (negative log-likelihood), better the topic model. Figures 3, 4 and 5 show PPL scores on all (test) document collections in the streams **S1**, **S2** and **S3**, respectively.

PPL on future tasks: Figure 3 shows PPL scores on the future task using `20NSshort` without (blue boxes) and with (green boxes) lifelong settings. Compared to NTM, the configuration `LNTM+EmbTF+TR+SAL` reports an improved score of (641 vs 646). Similarly, Figures 4 and 5 depict that the generalization capability is boosted, i.e., (666 vs 706) and (183 vs 192) on `TMNtitle` and `R21578title`, respectively due to word-embedding based multi-domain multi-source knowledge transfer (`LNTM+EmbTF`) over lifetime.

PPL on past tasks incurring forgetting(orange boxes): We also report PPL on the all past document collections of the streams **S1**, **S2** and **S3** using parameters $\mathbf{\Theta}^{T+1}$ of a future task. Comparing the proposed approaches of LNTM, we observe that they retain PPL over lifetime for each document collection in each of the streams; however at the cost of forgetting due to sensitivity of the log-likelihood computation towards neural network parameters. Note that $\mathbf{\Theta}^{T+1}$ retains representation ability for all $t < T + 1$ quantified by IR.

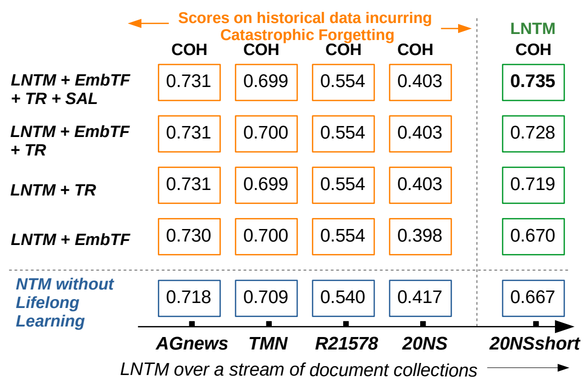


Figure 7. COH on future (`20NSshort`) and past tasks for **S1**

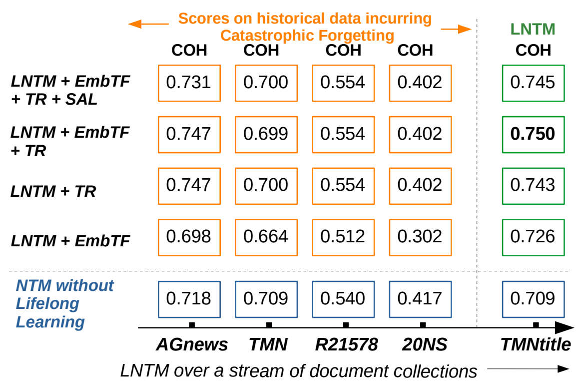


Figure 8. COH on future (`TMNtitle`) and past tasks for **S2**

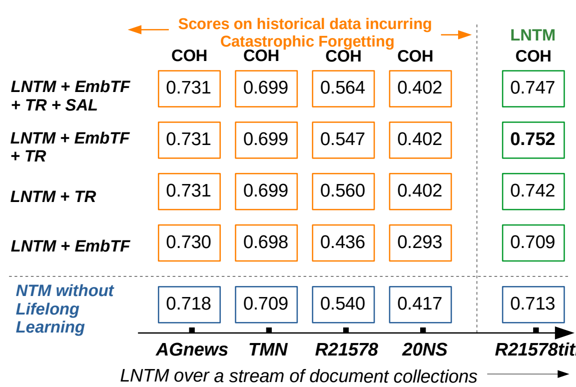


Figure 9. COH on future (`R21578title`) and past tasks for **S3**

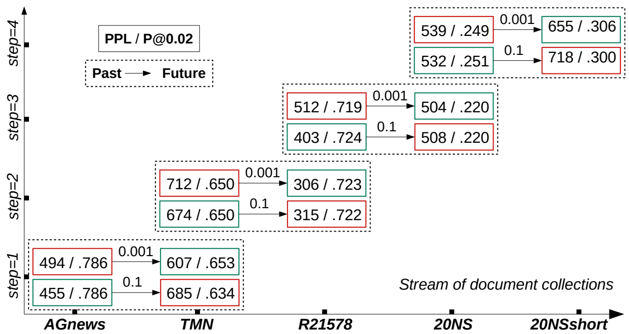


Figure 10. Ablation for λ_{TR}: Maximum knowledge transfer vs minimum catastrophic forgetting over lifetime using stream **S1**

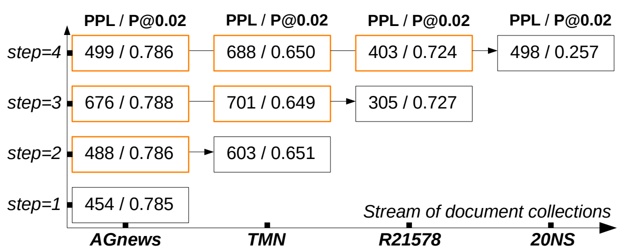


Figure 11. Illustration of LNTM (`LNTM-all`) for each task over stream **S1**, where orange colored boxes indicate scores incurring forgetting while modeling a future task (gray boxes) at each step

3.4. Analysis: Quantitative and Qualitative

Knowledge Transfer vs Forgetting: While learning within lifelong framework, there is a trade-off in knowledge transfer for future and forgetting of past. In Figure 10, we provide an ablation over $\lambda_{TR} \in \{0.001, 0.1\}$ for `LNTM` = $\{$TR$\}$ approach to show how λ_{TR} regulates the trade-off. Observe that the lower values of λ_{TR} leads to maximizing knowledge transfer (green boxes) for a future task (within a gray box), however at the cost of forgetting (red) of past learning and vice-versa when λ_{TR} increases. Here at each step (y-axis), we perform the ablation in pairs of document collections over the stream **S1** and show PPL and P@R. The study suggests to set λ_{TR} such that the trade-off is balanced.

Lifelong Topic Learning over the stream S1: Similar to Figures (3, 4 and 5), we additionally provide an illustration of lifelong topic learning over **S1**, where each task in sequence is treated as a future task accounting for the trade-off and forgetting over lifetime. Figure 11 provides illustration of LNTM (`LNTM-all`) for each dataset (as future task in gray box) in the streams of document collection used. Here, we show scores: PPL and P@0.02 of generalization and IR task, where the y-axis indicates each step of the lifelong learning process of topic modeling over a stream of document collections. Observe that the orange color box indicates scores incurring forgetting while modeling a target (in gray box). Once the step 4 is executed, we use three sparse targets (as step 5) to show applicability of lifelong topic modeling to address data-sparsity issues.

Table 2. Analysis: Qualitative topics of `TMNtitle`

Model	Topic-words (Top 5)
`NTM`	T1: nuclear, break, jobs, afghanistan, ipad T2: gulf, bruins, japanese, michigan, radiation
`LNTM + TR`	T1: arts, android, iphone, tablet, ipad T2: rail, medicare, wildfire, radioactive, recession
`LNTM-all`	T1: linkedin, android, tablet, ipad, iphone T2: tornadoes, fukushima, radioactive, radiation, medicare

Qualitative Topics: Table 2 shows topics (top-5 words) captured on `TMNtitle` (sparse) document collection of the stream **S2**, extracted using row-vectors of $\mathbf{W} \in \mathbf{\Theta}^{T+1}$. Observe that NTM generates incoherent topics (terms marked in red); however the two topics (T1 and T2) becomes coherent within LNTM framework, representing thematic structures about *product-line* and *disaster*, respectively. It suggest that the quality of topics is improved due to a positive transfer of knowledge via `EmbTF`, `TR` and `SAL` approaches.

4. Conclusion

We have presented a novel lifelong neural topic modeling framework that models a stream of document collections and exploits prior knowledge from several domains over lifetime in form of pre-trained topics, word embeddings and generative homologies in historical collections. Experimental results show that our proposed approaches of joint topic regularization, selective-data augmented learning and word-embedding guided topic learning within the lifelong framework help modeling three sparse datasets, quantified by information retrieval, topic coherence and generalization.

Acknowledgment

This research was supported by Bundeswirtschaftsministerium (bmwi.de), grant 01MD19003E (PLASS-Platform for Analytical Supply Chain Management Services (plass.io)) at Siemens AG- CT Machine Intelligence, Munich Germany.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022, 2003.

Chen, Z. and Liu, B. Topic modeling using topics from many domains, lifelong learning and big data. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pp. 703–711, 2014.

Chen, Z. and Liu, B. Lifelong machine learning for natural language processing. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, Austin, Texas, November 2016. Association for Computational Linguistics.

Chen, Z., Ma, N., and Liu, B. Lifelong learning for sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pp. 750–756, 2015.

Collobert, R. and Weston, J. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, pp. 160–167, 2008.

Das, R., Zaheer, M., and Dyer, C. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 795–804. Association for Computational Linguistics, 2015. doi: 10.3115/v1/P15-1077.

de Masson d'Autume, C., Ruder, S., Kong, L., and Yogatama, D. Episodic memory in lifelong language learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 13122–13131, 2019.

Gupta, P., Chaudhary, Y., Buettner, F., and Schütze, H. Document informed neural autoregressive topic models with distributional prior. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 6505–6512, 2019.

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

Jung, H., Ju, J., Jung, M., and Kim, J. Less-forgetting learning in deep neural networks. *CoRR*, abs/1607.00122, 2016.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Larochelle, H. and Lauly, S. A neural autoregressive topic model. In Bartlett, P. L., Pereira, F. C. N., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems*, pp. 2717–2725, 2012.

Larochelle, H. and Murray, I. The neural autoregressive distribution estimator. In Gordon, G. J., Dunson, D. B., and Dudík, M. (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 15 of *JMLR Proceedings*, pp. 29–37. JMLR.org, 2011.

Lauly, S., Zheng, Y., Allauzen, A., and Larochelle, H. Document neural autoregressive distribution estimation. *Journal of Machine Learning Research*, 18:113:1–113:24, 2017.

Li, Z. and Hoiem, D. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947, 2018.

Miao, Y., Yu, L., and Blunsom, P. Neural variational inference for text processing. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1727–1736. JMLR.org, 2016.

Mitchell, T. M., Cohen, W. W., Jr., E. R. H., Talukdar, P. P., Betteridge, J., Carlson, A., Mishra, B. D., Gardner, M., Kisiel, B., Krishnamurthy, J., Lao, N., Mazaitis, K., Mohamed, T., Nakashole, N., Platanios, E. A., Ritter, A., Samadi, M., Settles, B., Wang, R. C., Wijaya, D., Gupta, A., Chen, X., Saparov, A., Greaves, M., and Welling, J. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pp. 2302–2310, 2015.

Nguyen, D. Q., Billingsley, R., Du, L., and Johnson, M. Improving topic models with latent feature word representations. *TACL*, 3:299–313, 2015.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Pennington, J., Socher, R., and Manning, C. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. Association for Computational Linguistics, 2014.

Petterson, J., Smola, A. J., Caetano, T. S., Buntine, W. L., and Narayanamurthy, S. M. Word features for latent dirichlet allocation. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems*, pp. 1921–1929. Curran Associates, Inc., 2010.

Robins, A. V. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connect. Sci.*, 7(2):123–146, 1995.

Röder, M., Both, A., and Hinneburg, A. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, Shanghai, China, February 2-6, 2015*, pp. 399–408. ACM, 2015.

Ruder, S. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098, 2017.

Salakhutdinov, R. and Hinton, G. E. Replicated softmax: an undirected topic model. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems*, pp. 1607–1614. Curran Associates, Inc., 2009.

Srivastava, A. and Sutton, C. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR*, 2017.

Thrun, S. and Mitchell, T. M. Lifelong robot learning. *Robotics and autonomous systems*, 15(1-2):25–46, 1995.

Wang, H., Xiong, W., Yu, M., Guo, X., Chang, S., and Wang, W. Y. Sentence embedding alignment for lifelong relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 796–806, 2019.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 3987–3995, 2017.