Deep Temporal-Recurrent-Replicated-Softmax for Topical Trends over Time

Pankaj Gupta[1,2], Subburam Rajaram[1], Hinrich Schütze[2], Bernt Andrassy[1]

[1]Corporate Technology, Machine-Intelligence (MIC-DE), Siemens AG Munich, Germany
[2]CIS, University of Munich (LMU) Munich, Germany
{pankaj.gupta, subburam.rajaram, bernt.andrassy}@siemens.com
pankaj.gupta@campus.lmu.de | inquiries@cislmu.org

Abstract

Dynamic topic modeling facilitates the identification of topical trends over time in temporal collections of unstructured documents. We introduce a novel unsupervised neural dynamic topic model named as Recurrent Neural Network-Replicated Softmax Model (RNN-RSM), where the discovered topics at each time influence the topic discovery in the subsequent time steps. We account for the temporal ordering of documents by explicitly modeling a joint distribution of latent topical dependencies over time, using distributional estimators with temporal recurrent connections. Applying RNN-RSM to 19 years of articles on NLP research, we demonstrate that compared to state-of-the art topic models, RNN-RSM shows better generalization, topic interpretation, evolution and trends. We also introduce a metric (named as SPAN) to quantify the capability of dynamic topic model to capture word evolution in topics over time.

1 Introduction

Topic Detection and Tracking (Allan et al., 1998) is an important area of natural language processing to find topically related ideas that evolve over time in a sequence of text collections and exhibit temporal relationships. The temporal aspects of these collections can present valuable insight into the topical structure of the collections and can be quantified by modeling the dynamics of the underlying topics discovered over time.

Problem Statement: We aim to generate temporal topical trends or automatic overview timelines of topics for a time sequence collection of documents. This involves the following three tasks in dynamic topic analysis: **(1)** *Topic Structure Detection* (TSD): Identifying main topics in the document collection. **(2)** *Topic Evolution Detection* (TED): Detecting the emergence of a new topic
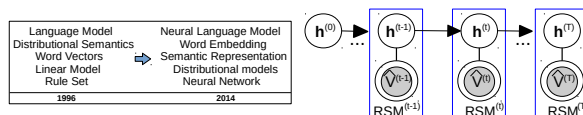

Figure 1: (Left): Word Usage over time for Topic (*Word Representation*) in scholarly articles. (Right): RSM-based dynamic topic model with explicit temporal topic dependence

and recognizing how it grows or decays over time (Allan, 2002). **(3)** *Temporal Topic Characterization* (TTC): Identifying the characteristics for each of the main topics in order to track the words' usage (*keyword trends*) for a topic over time i.e. *topical trend analysis for word evolution* (Fig 1, Left).

Probabilistic static topic models, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and its variants (Wang and McCallum, 2006; Hall et al., 2008; Gollapalli and Li, 2015) have been investigated to examine the emergence of topics from historical documents. Another variant known as Replicated Softmax (RSM) (Hinton and Salakhutdinov, 2009) has demonstrated better generalization in log-probability and retrieval, compared to LDA. Prior works (Iwata et al., 2010; Pruteanu-Malinici et al., 2010; Saha and Sindhwani, 2012; Schein et al., 2016) have investigated Bayesian modeling of topics in time-stamped documents. Particularly, Blei and Lafferty (2006) developed a LDA based dynamic topic model (DTM) to capture the evolution of topics in a time sequence collection of documents; however they do not capture explicitly the topic popularity and usage of specific terms over time. We propose a family of probabilistic time series models with distributional estimators to explicitly model the dynamics of the underlying topics, introducing temporal latent topic dependencies (Fig 1, Right).

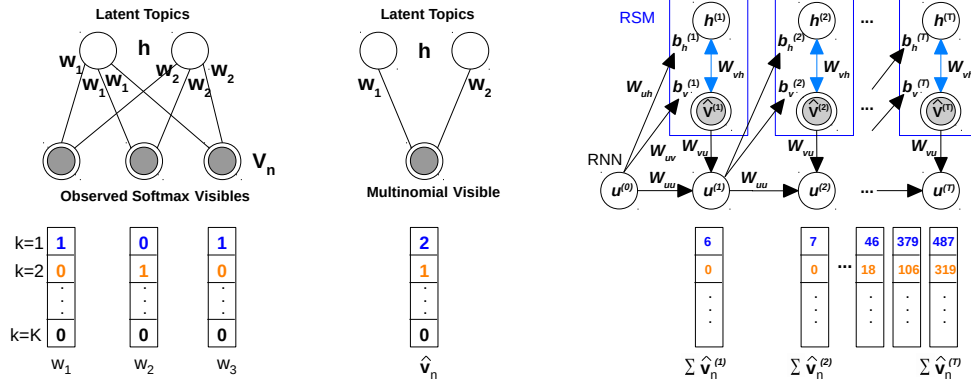To model temporal dependencies in high dimen-

Figure 2: (Left): RSM for a document $\mathbf{V}_n$ of $D_n=3$ words (w). The bottom layer represents the softmax visible units, that share the same set of weights connected to binary hidden units $\mathbf{h}$. (Middle): Interpretation of RSM in which D_n softmax units with identical weights are replaced by a single multinomial unit, sampled D_n times. (Right): Graphical structure of 2-layered **RNN-RSM**, unfolded in time. Single and double headed arrows represent deterministic and stochastic-symmetric connections, respectively. $\widehat{\mathbf{V}}^{(t)}$ and $\mathbf{h}^{(t)}$ are binary visible and hidden layers of RSM for a document collection at time, t. $\mathbf{u}$ is RNN hidden layer. k: dictionary index for a word w

sional sequences, such as polyphonic music, the temporal stack of RBMs (Smolensky, 1986; Hinton, 2002) has been investigated to model complex distributions. The Temporal RBM (Taylor et al., 2007; Sutskever and Hinton, 2007), Recurrent Temporal RBM (RTRBM) (Sutskever et al., 2009) and RNN-RBM (Boulanger-Lewandowski et al., 2012) show success in modeling the temporal dependencies in such symbolic sequences. In addition, RNNs (Gupta et al., 2015a; Vu et al., 2016a,b; Gupta et al., 2016) have been recognized for sentence modeling in natural language tasks. We aspire to build neural dynamic topic model called RNN-RSM to model document collections over time and learn temporal topic correlations.

We consider RSM for TSD and introduce the explicit latent topical dependencies for TED and TTC tasks. Fig 1 illustrates our *motivation*, where temporal ordering in document collection $\widehat{\mathbf{V}}^{(t)}$ at each time step t, is modeled by conditioning the latent topic $\mathbf{h}^{(t)}$ on the sequence history of latent topics $\mathbf{h}^{(0)}, ..., \mathbf{h}^{(t-1)}$, accumulated with temporal lag. Each RSM discovers latent topics, where the introduction of a bias term in each RSM via the time-feedback latent topic dependencies enables to explicitly model topic evolution and specific topic term usage over time. The temporal connections and RSM biases allow to convey topical information and model relation among the words, in order to deeply analyze the dynamics of the underlying topics. We demonstrate the applicability of proposed **RNN-RSM** by analyzing 19 years of scientific articles from NLP research.

The *contributions* in this work are:
(1) Introduce an unsupervised neural dynamic topic model based on recurrent neural network and RSMs, named as RNN-RSM to explicitly model discovered latent topics (evolution) and word relations (topic characterization) over time.
(2) Demonstrate better generalization (log-probability and time stamp prediction), topic interpretation (coherence), evolution and characterization, compared to the state-of-the-art.
(3) It is the first work in dynamic topic modeling using undirected stochastic graphical models and deterministic recurrent neural network to model collections of different-sized documents over time, within the generative and neural network framework. The code and data are available at `https://github.com/pgcool/RNN-RSM`.

2   The RNN-RSM model

RSM (Fig 2, Left) models are a family of different-sized Restricted Boltzmann Machines (RBMs) (Gehler et al., 2006; Xing et al., 2005; Gupta et al., 2015b,c) that models *word counts* by sharing the same parameters with multinomial distribution over the observable i.e. it can be interpreted as a single multinomial unit (Fig 2, Middle) sampled as many times as the document size. This facilitates in dealing with the documents of different lengths.

The proposed RNN-RSM model (Fig 2, Right) is a sequence of conditional RSMs[1] such that at any time step t, the RSM's bias parameters $\mathbf{b}_\mathbf{v}^{(t)}$

---

[1]Notations: $\widehat{\mathbf{U}}=\{\mathbf{U}_n\}_{n=1}^N$; $\mathbf{U}$:2D-Matrix; $\mathbf{l}$:vector; **U**/**l**:**U**pper/**l**ower-case; Scalars in unbold

and $\mathbf{b_h}^{(t)}$ depend on the output of a deterministic RNN with hidden layer $\mathbf{u}^{(t-1)}$ in the previous time step, $t-1$. Similar to RNN-RBM (Boulanger-Lewandowski et al., 2012), we constrain RNN hidden units ($\mathbf{u}^{(t)}$) to convey temporal information, while RSM hidden units ($\mathbf{h}^{(t)}$) to model conditional distributions. Therefore, parameters ($\mathbf{b_v}^{(t)}$, $\mathbf{b_h}^{(t)}$) are time-dependent on the sequence history at time t (via a series of conditional RSMs) denoted by $\Theta^{(t)} \equiv \{\widehat{\mathbf{V}}^{(\tau)}, \mathbf{u}^{(\tau)} | \tau < t\}$, that captures temporal dependencies. The RNN-RSM is defined by its joint probability distribution:

$$P(\widehat{\mathfrak{V}}, \mathbf{H}) = P(\{\widehat{\mathbf{V}}^{(t)}, \mathbf{h}^{(t)}\}_{t=1}^{T}) = \prod_{t=1}^{T} P(\widehat{\mathbf{V}}^{(t)}, \mathbf{h}^{(t)} | \Theta^{(t)})$$

where $\widehat{\mathfrak{V}} = [\widehat{\mathbf{V}}^{(1)}, ... \widehat{\mathbf{V}}^{(T)}]$ and $\mathbf{H} = [\mathbf{h}^{(1)}, ... \mathbf{h}^{(T)}]$. Each $\mathbf{h}^{(t)} \in \{0, 1\}^{F}$ be a binary stochastic hidden topic vector with size F and $\widehat{\mathbf{V}}^{(t)} = \{\mathbf{V}_n^{(t)}\}_{n=1}^{N^{(t)}}$ be a collection of N documents at time step t. Let $\mathbf{V}_n^{(t)}$ be a $K \times D_n^{(t)}$ observed binary matrix of the nth document in the collection where, $D_n^{(t)}$ is the document size and K is the dictionary size over all the time steps. The conditional distribution (for each unit in hidden or visible) in each RSM at time step, is given by softmax and logistic functions:

$$P(v_{n,i}^{k,(t)} = 1 | \mathbf{h}_n^{(t)}) = \frac{\exp(b_{v,i}^{k,(t)} + \sum_{j=1}^{F} h_{n,j}^{(t)} W_{ij}^k)}{\sum_{q=1}^{K} \exp(b_{v,i}^{q,(t)} + \sum_{j=1}^{F} h_{n,j}^{(t)} W_{ij}^q)}$$

$$P(h_{n,j}^{(t)} = 1 | \mathbf{V}_n^{(t)}) = \sigma(b_{h,j}^{(t)} + \sum_{i=1}^{D_n^{(t)}} \sum_{k=1}^{K} v_{n,i}^{k,(t)} W_{ij}^k)$$

where $P(v_{n,i}^{k,(t)} = 1 | \mathbf{h}_n^{(t)})$ and $P(h_{n,j}^{(t)} = 1 | \mathbf{V}_n^{(t)})$ are conditional distributions for i^{th} visible $v_{n,i}$ and j^{th} hidden unit $h_{n,j}$ for the nth document at t. W_{ij}^k is a symmetric interaction term between i that takes on value k and j. $v_n^{k,(t)}$ is sampled $D_n^{(t)}$ times with identical weights connected to binary hidden units, resulting in multinomial visibles, therefore the name *Replicated Softmax*. The conditionals across layers are factorized as: $P(\mathbf{V}_n^{(t)} | \mathbf{h}_n^{(t)}) = \prod_{i=1}^{D_n^{(t)}} P(\mathbf{v}_{n,i}^{(t)} | \mathbf{h}_n^{(t)})$; $P(\mathbf{h}_n^{(t)} | \mathbf{V}_n^{(t)}) = \prod_j P(h_{n,j}^{(t)} | \mathbf{V}_n^{(t)})$.

Since biases of RSM depend on the output of RNN at previous time steps, that allows to propagate the estimated gradient at each RSM backward through time (BPTT). The *RSM biases* and RNN hidden state $\mathbf{u}^{(t)}$ at each time step t are given by-

$$\mathbf{b_v}^{(t)} = \mathbf{b_v} + \mathbf{W_{uv}} \mathbf{u}^{(t-1)}$$
$$\mathbf{b_h}^{(t)} = \mathbf{b_h} + \mathbf{W_{uh}} \mathbf{u}^{(t-1)} \qquad (1)$$

$$\mathbf{u}^{(t)} = \tanh(\mathbf{b_u} + \mathbf{W_{uu}} \mathbf{u}^{(t-1)} + \mathbf{W_{vu}} \sum_{n=1}^{N^{(t)}} \hat{\mathbf{v}}_n^{(t)}) \qquad (2)$$

Algorithm 1 *Training RNN-RSM with BPTT*

Input: *Observed visibles*, $\widehat{\mathfrak{V}} = \{\widehat{\mathbf{V}}^{(0)}, \widehat{\mathbf{V}}^{(1)}, ..., \widehat{\mathbf{V}}^{(t)}, ..., \widehat{\mathbf{V}}^{(T)}\}$
RNN-RSM Parameters: $\theta = \{\mathbf{W_{uh}}, \mathbf{W_{vh}}, \mathbf{W_{uv}}, \mathbf{W_{vu}}, \mathbf{W_{uu}}, \mathbf{b_v}, \mathbf{b_u}, \mathbf{b_h}, \mathbf{b_v}^{(t)}, \mathbf{b_h}^{(t)}, \mathbf{u}^{(0)}\}$
1: Propagate $\mathbf{u}^{(t)}$ in RNN portion of the graph using eq 2.
2: Compute $\mathbf{b_v}^{(t)}$ and $\mathbf{b_h}^{(t)}$ using eq 1.
3: Generate negatives $\mathbf{V}^{(t)*}$ using k-step Gibbs sampling.
4: Estimate the gradient of the cost C w.r.t. parameters of RSM $\mathbf{W_{vh}}$, $\mathbf{b_v}^{(t)}$ and $\mathbf{b_h}^{(t)}$ using eq 5.
5: Compute gradients (eq 6) w.r.t. RNN connections ($\mathbf{W_{uh}}$, $\mathbf{W_{uv}}$, $\mathbf{W_{uu}}$, $\mathbf{W_{vu}}$, $\mathbf{u}^0$) and biases ($\mathbf{b_v}$, $\mathbf{b_h}$, $\mathbf{b_u}$).
6: **Goto** *step 1* until stopping_criteria (early stopping or maximum iterations reached)

---

where $\mathbf{W_{uv}}$, $\mathbf{W_{uh}}$ and $\mathbf{W_{vu}}$ are weights connecting RNN and RSM portions (Figure 2). $\mathbf{b_u}$ is the bias of $\mathbf{u}$ and $\mathbf{W_{uu}}$ is the weight between RNN hidden units. $\hat{v}_n^{(t)}$ is a vector of $\hat{v}_n^k$ (denotes the count for the kth word in n^{th} document). $\sum_{n=1}^{N^{(t)}} \hat{v}_n^{(t)}$ refers to the sum of observed vectors across documents at time step t where each document is represented as-

$$\hat{\mathbf{v}}_n^{(t)} = [\{\hat{v}_n^{k,(t)}\}_{k=1}^{K}] \text{ and } \hat{v}_n^{k,(t)} = \sum_{i=1}^{D_n^{(t)}} v_{n,i}^{k,(t)} \quad (3)$$

where $v_{n,i}^{k,(t)} = 1$ if visible unit i takes on k^{th} value.

In each RSM, a separate RBM is created for each document in the collection at time step t with $D_n^{(t)}$ softmax units, where $D_n^{(t)}$ is the count of words in the nth document. Consider a document of $D_n^{(t)}$ words, the *energy* of the state $\{\mathbf{V}_n^{(t)}, \mathbf{h}_n^{(t)}\}$ at time step, t is given by-

$$E(\mathbf{V}_n^{(t)}, \mathbf{h}_n^{(t)}) = - \sum_{j=1}^{F} \sum_{k=1}^{K} h_{n,j}^{(t)} W_j^k \hat{v}_n^{k,(t)}$$
$$- \sum_{k=1}^{K} \hat{v}_n^{k,(t)} b_v^k - D_n^{(t)} \sum_{j=1}^{F} b_{h,j} h_{n,j}^{(t)}$$

Observe that the bias terms on hidden units are scaled up by document length to allow hidden units to stabilize when dealing with different-sized documents. The corresponding energy-probability relation in the energy-based model is-

$$P(\mathbf{V}_n^{(t)}) = \frac{1}{Z_n^{(t)}} \sum_{\mathbf{h}_n^{(t)}} \exp(-E(\mathbf{V}_n^{(t)}, \mathbf{h}_n^{(t)})) \quad (4)$$

where $Z_n^{(t)} = \sum_{\mathbf{V}_n^{(t)}} \sum_{\mathbf{h}_n^{(t)}} \exp(-E(\mathbf{V}_n^{(t)}, \mathbf{h}_n^{(t)}))$ is the normalization constant. The lower bound on the log likelihood of the data takes the form:

$$\ln P(\mathbf{V}_n^{(t)}) \geq \sum_{\mathbf{h}^{(t)}} Q(\mathbf{h}_n^{(t)} | \mathbf{V}_n^{(t)}) \ln P(\mathbf{V}_n^{(t)}, \mathbf{h}_n^{(t)}) + H(Q)$$
$$= \ln P(\mathbf{V}_n^{(t)}) - KL[Q(\mathbf{h}_n^{(t)} | \mathbf{V}_n^{(t)}) || P(\mathbf{h}_n^{(t)} | \mathbf{V}_n^{(t)})]$$

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	Total
ACL	58	73	250	83	79	70	177	112	134	134	307	204	214	243	270	349	227	398	331	3713
EMNLP	15	24	15	36	29	21	42	29	58	28	75	132	115	164	125	149	140	206	228	1756
ACL+EMNLP	73	97	265	119	108	91	219	141	192	162	382	336	329	407	395	498	367	604	559	**5469**

Table 1: Number of papers from ACL and EMNLP conferences over the years

where $H(\cdot)$ is the entropy and Q is the approximating posterior. Similar to Deep Belief Networks (Hinton et al., 2006), adding an extra layer improves lower bound on the log probability of data, we introduce the extra layer via RSM biases that propagates the prior via RNN connections. The dependence analogy follows-

$$E(\mathbf{V}_n^{(t)}, \mathbf{h}_n^{(t)}) \propto \frac{1}{\mathbf{b_v}^{(t)}} \text{ and } E(\mathbf{V}_n^{(t)}, \mathbf{h}_n^{(t)}) \propto \frac{1}{\mathbf{b_h}^{(t)}}$$

$$\ln P(\mathbf{V}_n^{(t)}) \propto \frac{1}{E(\mathbf{V}_n^{(t)}, \mathbf{h}_n^{(t)})}; \ln P(\widehat{\mathbf{V}}_n^{(t)}) \propto \ln P(\{\widehat{\mathbf{V}}_n^\tau\}_{\tau < t})$$

Observe that the prior is seen as the deterministic hidden representation of latent topics and injected into each hidden state of RSMs, that enables the likelihood of the data to model complex temporal densities i.e. heteroscedasticity in document collections ($\widehat{\mathfrak{V}}$) and temporal topics ($\mathbf{H}$).

Gradient Approximations: The *cost* in RNN-RSM is: $C = \sum_{t=1}^{T} C_t \equiv \sum_{t=1}^{T} - \ln P(\widehat{\mathbf{V}}^{(t)})$

Due to intractable Z, the gradient of cost at time step t w.r.t. (with respect to) RSM parameters are approximated by k-step Contrastive Divergence (CD) (Hinton, 2002). The gradient of the negative log-likelihood of a document collection $\{\mathbf{V}_n^{(t)}\}_{n=1}^{N^{(t)}}$ w.r.t. RSM parameter $\mathbf{W_{vh}}$,

$$\frac{1}{N^{(t)}} \sum_{n=1}^{N^{(t)}} \frac{\partial(-\ln P(\mathbf{V}_n^{(t)}))}{\partial \mathbf{W_{vh}}}$$

$$= \frac{1}{N^{(t)}} \sum_{n=1}^{N^{(t)}} \frac{\partial \mathfrak{F}(\mathbf{V}_n^{(t)})}{\partial \mathbf{W_{vh}}} - \frac{\partial(-\ln Z_n^{(t)})}{\partial \mathbf{W_{vh}}}$$

$$= \underbrace{\mathsf{E}_{P_{data}}[\frac{\partial \mathfrak{F}(\mathbf{V}_n^{(t)})}{\partial \mathbf{W_{vh}}}]}_{\text{data-dependent expectation}} - \underbrace{\mathsf{E}_{P_{model}}[\frac{\partial \mathfrak{F}(\mathbf{V}_n^{(t)})}{\partial \mathbf{W_{vh}}}]}_{\text{model's expectation}}$$

$$\simeq \frac{1}{N^{(t)}} \sum_{n=1}^{N^{(t)}} \frac{\partial \mathfrak{F}(\mathbf{V}_n^{(t)})}{\partial \mathbf{W_{vh}}} - \frac{\partial \mathfrak{F}(\mathbf{V}_n^{(t)*})}{\partial \mathbf{W_{vh}}}$$

The second term is estimated by negative samples $\mathbf{V}_n^{(t)*}$ obtained from k-step Gibbs chain starting at $\mathbf{V}_n^{(t)}$ samples. $P_{data}(\widehat{\mathbf{V}}^{(t)}, \mathbf{h}^{(t)}) = P(\mathbf{h}^{(t)}|\widehat{\mathbf{V}}^{(t)})P_{data}(\widehat{\mathbf{V}}^{(t)})$ and $P_{data}(\widehat{\mathbf{V}}^{(t)}) = \frac{1}{N^{(t)}} \sum_n^{N^{(t)}} \delta(\widehat{\mathbf{V}}^{(t)} - \mathbf{V}_n^{(t)})$ is the empirical distribution on the observable. $P_{model}(\mathbf{V}_n^{(t)*}, \mathbf{h}_n^{(t)})$ is

defined in eq. 4. The free energy $\mathfrak{F}(\mathbf{V}_n^{(t)})$ is related to normalized probability of $\mathbf{V}_n^{(t)}$ as $P(\mathbf{V}_n^{(t)}) \equiv \exp^{-\mathfrak{F}(\mathbf{V}_n^{(t)})}/Z_n^{(t)}$ and as follows-

$$\mathfrak{F}(\mathbf{V}_n^{(t)}) = -\sum_{k=1}^{K} \hat{v}_n^{k,(t)} b_v^k - \sum_{j=1}^{F} \log(1+$$

$$\exp(D_n^{(t)} b_{h,j} + \sum_{k=1}^{K} \hat{v}_n^{k,(t)} W_j^k))$$

Gradient approximations w.r.t. RSM parameters,

$$\frac{\partial C_t}{\partial \mathbf{b_v}^{(t)}} \simeq \sum_{n=1}^{N^{(t)}} \hat{\mathbf{v}}_n^{(t)*} - \hat{\mathbf{v}}_n^{(t)}$$

$$\frac{\partial C_t}{\partial \mathbf{b_h}^{(t)}} \simeq \sum_{n=1}^{N^{(t)}} \sigma(\mathbf{W_{vh}}\hat{\mathbf{v}}_n^{(t)*} - D_n^{(t)}\mathbf{b_h}^{(t)})$$

$$- \sigma(\mathbf{W_{vh}}\hat{\mathbf{v}}_n^{(t)} - D_n^{(t)}\mathbf{b_h}^{(t)})$$

$$\frac{\partial C_t}{\partial \mathbf{W_{vh}}} \simeq \sum_{t=1}^{T} \sum_{n=1}^{N^{(t)}} \sigma(\mathbf{W_{vh}}\hat{\mathbf{v}}_n^{(t)*} - D_n^{(t)}\mathbf{b_h}^{(t)})$$

$$\hat{\mathbf{v}}_n^{(t)*T} - \sigma(\mathbf{W_{vh}}\hat{\mathbf{v}}_n^{(t)} - D_n^{(t)}\mathbf{b_h}^{(t)})\hat{\mathbf{v}}_n^{(t)T} \tag{5}$$

The estimated gradients w.r.t. RSM biases are back-propagated via hidden-to-bias parameters (eq 1) to compute gradients w.r.t. RNN connections ($\mathbf{W_{uh}}$, $\mathbf{W_{uv}}$, $\mathbf{W_{vu}}$ and $\mathbf{W_{uu}}$) and biases ($\mathbf{b_h}$, $\mathbf{b_v}$ and $\mathbf{b_u}$).

$$\frac{\partial C}{\partial \mathbf{W_{uh}}} = \sum_{t=1}^{T} \frac{\partial C_t}{\partial \mathbf{b_h}^{(t)}} \mathbf{u}^{(t-1)T}$$

$$\frac{\partial C}{\partial \mathbf{W_{uv}}} = \sum_{t=1}^{T} \frac{\partial C_t}{\partial \mathbf{b_v}^{(t)}} \mathbf{u}^{(t-1)T}$$

$$\frac{\partial C}{\partial \mathbf{W_{vu}}} = \sum_{t=1}^{T} \frac{\partial C_t}{\partial \mathbf{u}^{(t)}} \mathbf{u}^{(t)}(1 - \mathbf{u}^{(t)}) \sum_{n=1}^{N^{(t)}} \hat{\mathbf{v}}_n^{(t)T}$$

$$\frac{\partial C}{\partial \mathbf{b_h}} = \sum_{t=1}^{T} \frac{\partial C_t}{\partial \mathbf{b_h}^{(t)}} \text{ and } \frac{\partial C}{\partial \mathbf{b_v}} = \sum_{t=1}^{T} \frac{\partial C_t}{\partial \mathbf{b_v}^{(t)}}$$

$$\frac{\partial C}{\partial \mathbf{b_u}} = \sum_{t=1}^{T} \frac{\partial C_t}{\partial \mathbf{u}^{(t)}} \mathbf{u}^{(t)}(1 - \mathbf{u}^{(t)})$$

$$\frac{\partial C}{\partial \mathbf{W_{uu}}} = \sum_{t=1}^{T} \frac{\partial C_t}{\partial \mathbf{u}^{(t)}} \mathbf{u}^{(t)}(1 - \mathbf{u}^{(t)})\mathbf{u}^{(t-1)T}$$

$$\tag{6}$$

Parameter	Value(s)	Optimal
epochs	1000	1000
CD iterations	15	15
learning rate	0.1, 0.03, 0.001	0.001
hidden size	20, 30, 50	30

Table 2: Hyperparameters for RNN-RSM model

model	metric				
	SumPPL	*Err*	*mean-COH*	*median-COH*	*TTD*
DTM	10.9	8.10	0.1514	0.1379	0.084
RNN-RSM	**3.8**	**7.58**	**0.1620**	**0.1552**	0.268

Table 3: State-of-the-art Comparison: Generalization (*PPL* and *Err*), Topic Interpretation (*COH*) and Evolution (*TTD*) in DTM and RNN-RSM models

For the single-layer RNN-RSM, the BPTT recurrence relation for $0 \leq t < T$ is given by-

$$\frac{\partial C_t}{\partial \mathbf{u}^{(t)}} = \mathbf{W_{uu}} \frac{\partial C_{t+1}}{\partial \mathbf{u}^{(t+1)}} \mathbf{u}^{(t+1)} (1 - \mathbf{u}^{(t+1)})$$
$$+ \mathbf{W_{uh}} \frac{\partial C_{t+1}}{\partial \mathbf{b_h}^{(t+1)}} + \mathbf{W_{uv}} \frac{\partial C_{t+1}}{\partial \mathbf{b_v}^{(t+1)}}$$

where $\mathbf{u}^{(0)}$ being a parameter and $\frac{\partial C_T}{\partial \mathbf{u}^{(T)}} = 0$.

See *Training RNN-RSM with BPTT* in Algo 1.

3 Evaluation

3.1 Dataset and Experimental Setup

We use the processed dataset (Gollapalli and Li, 2015), consisting of EMNLP and ACL conference papers from the year 1996 through 2014 (Table 1). We combine papers for each year from the two venues to prepare the document collections over time. We use ExpandRank (Wan and Xiao, 2008) to extract top 100 keyphrases for each paper, including unigrams and bigrams. We split the bigrams to unigrams to create a dictionary of all unigrams and bigrams. The dictionary size (K) and word count are 3390 and 5.19 M, respectively.

We evaluate RNN-RSM against static (RSM, LDA) and dynamic (DTM) topics models for topic and keyword evolution in NLP research over time. Individual 19 different RSM and LDA models are trained for each year, while DTM[2] and RNN-RSM are trained over the years with 19 time steps, where paper collections for a year is input at each time step. RNN-RSM is initialized with RSM ($\mathbf{W_{vh}}$, $\mathbf{b_v}$, $\mathbf{b_h}$) trained for the year 2014.

We use perplexity to choose the number of topics (=30). See Table 2 for hyperparameters.

3.2 Generalization in Dynamic Topic Models

Perplexity: We compute the perplexity on unobserved documents ($\widehat{\mathbf{V}}^{(t)}$) at each time step as

$$\text{PPL}(\widehat{\mathbf{V}}^{(t)}, t) = \exp\left(-\frac{1}{N^{(t)}} \frac{\sum_{n=1}^{N^{(t)}} \log P(\mathbf{V}_n^{(t)})}{\sum_{n=1}^{N^{(t)}} D_n^{(t)}}\right)$$

where t is the time step. $N^{(t)}$ is the number of documents in a collection ($\widehat{\mathbf{V}}^{(t)}$) at time t. Better models have lower perplexity values, suggesting less uncertainties about the documents. For held-out documents, we take 10 documents from each time step i.e. total 190 documents and compute perplexity for 30 topics. Fig 3d shows the comparison of perplexity values for unobserved documents from DTM and RNN-RSM at each time step. The *SumPPL* (Table 3) is the sum of PPL values for the held-out sets of each time step.

Document Time Stamp Prediction: To further assess the dynamic topics models, we split the document collections at each time step into 80-20% train-test, resulting in 1067 held-out documents. We predict the time stamp (dating) of a document by finding the most likely (with the lowest perplexity) location over the time line. See the *mean absolute error* (*Err*) in year for the held-out in Table 3. Note, we do not use the time stamp as observables during training.

3.3 TSD, TED: Topic Evolution over Time

Topic Detection: To extract topics from each RSM, we compute posterior $P(\widehat{\mathbf{V}}^{(t)} | h_j = 1)$ by activating a hidden unit and deactivating the rest in a hidden layer. We extract the top 20 terms for every 30 topic set from 1996-2014, resulting in $|Q|_{max} = 19 \times 30 \times 20$ possible topic terms.

Topic Popularity: To determine topic *popularity*, we selected three popular topics (*Sentiment Analysis*, *Word Vector* and *Dependency Parsing*) in NLP research and create a set[3] of key-terms (including unigrams and bigrams) for each topic. We compute cosine similarity of the key-terms defined for each selected topic and topics discovered by the topic models over the years. We consider the discovered topic that is the most similar to the key-terms in the target topic and plot the similarity values in Figure 3a, 3b and 3b. Observe that RNN-RSM shows better topic evolution for the three emerging topics. LDA and RSM show

---

[2]https://radimrehurek.com/gensim/models/dtmmodel.html

[3]topic-terms to be released with code

(a) Topic: Sentiment Analysis   (b) Topic: Word Vector   (c) Topic: Dependency Parsing   (d) Perplexity on Unobserved

(e) COH (mean) Over Time   (f) COH (median) Over Time   (g) RNN-RSM Adj Topic Sim   (h) DTM Adj Topic Sim
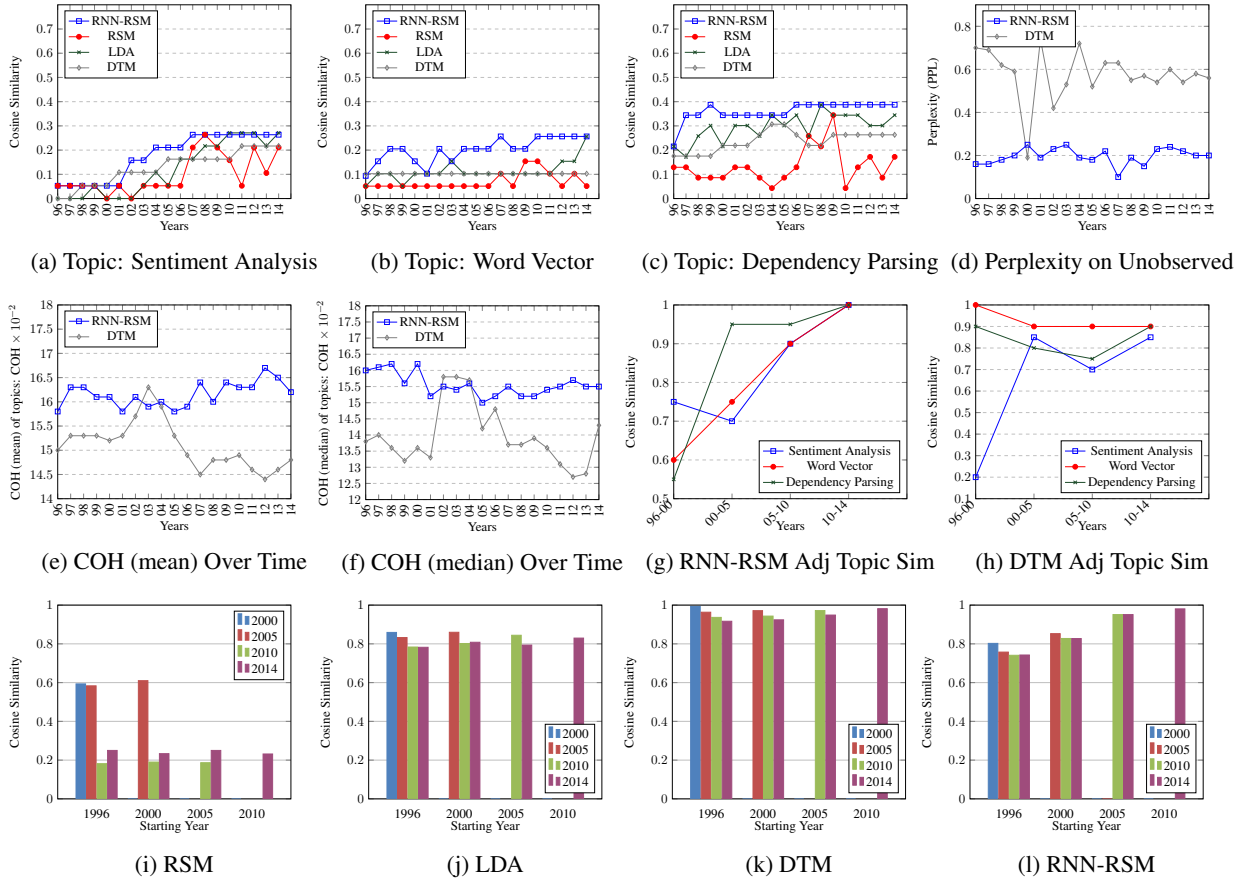
(i) RSM   (j) LDA   (k) DTM   (l) RNN-RSM

Figure 3: (a, b, c): Topic popularity by LDA, RSM, DTM and RNN-RSM over time (d): Perplexity on the unobserved document collections over time (e, f): Mean and Median Topic Coherence (g, h): Topic Evolution (i,j,k,l): Topic focus change over time. Adj- Adjacent; Sim- Similarity

topical locality in Figure 3c attributed to no correlation in topic dynamics over time, while in Figure 3b, DTM does not capture evolution of topic *Word Vector*.

Topic Drift (Focus Change): To compute the topic *focus* change over the years, we first split the time period 1996-2014 into five parts:{1996, 2000, 2005, 2010, 2014}. The cosine similarity scores are computed between the topic sets discovered in a particular year and the years preceding it in the above set, for example the similarity scores between the topic-terms in (1996, 2000), (1996, 2005), (1996, 2010) and (1996, 2014), respectively. Figure 3i, 3j, 3k and 3l demonstrate that RNN-RSM shows higher *convergence* in topic focus over the years, compared to LDA and RSM. In RNN-RSM, the topic similarity is gradually increased over time, however not in DTM. The higher similarities in the topic sets indicate that new/existing topics and words do not appear/disappear over time.

We compute topic-term drift (TTD) to show

the changing topics from initial to final year, as

$$TTD = 1.0 - cosineSimilarity(\mathbf{Q}^{(t)}, \mathbf{Q}^{(t')})$$

where $\mathbf{Q}$ is the set of all topic-terms for time step t. Table 3 shows that TTD (where t=1996 and t'=2014) are 0.268 and 0.084 for RNN-RSM and DTM, respectively. It suggests that the higher number of new topic-terms evolved in RNN-RSM, compared to DTM. Qualitatively, the Table 4 shows the topics observed with the highest and lowest cosine drifts in DTM and RNN-RSM.

In Figure 3g and 3h, we also illustrate the temporal *evolution* (drift) in the selected topics by computing cosine similarity on their adjacent topic vectors over time. The topic vectors are selected similarly as in computing topic popularity. We observe better TED in RNN-RSM than DTM for the three emerging topics in NLP research. For instance, for the selected topic *Word Vector*, the red line in DTM (Fig 3h) shows no drift (for x-axis 00-05, 05-10 and 10-14), suggesting the topic-terms in the adjacent years are similar and does not evolve.

Drift	Model (year)	Topic Terms
0.20	**DTM** (1996)	document, retrieval, query, documents, information, search, information retrieval, queries, terms, words, system, results, performance, method, approach
	DTM (2014)	document, query, search, documents, queries, information, retrieval, method, results, information retrieval, research, terms, other, approach, knowledge
0.53	**DTM** (1996)	semantic, lexical, structure, syntactic, argument, frame, example, lexicon, information, approach, source, function, figure, verbs, semantic representation
	DTM (2014)	semantic, argument, frame, sentence, syntactic, semantic parsing, structure, semantic role, example, role labeling, language, learning, logical form, system, lexicon
0.20	**RNN-RSM** (1996)	reordering, statistical machine, translation model, translations, arabic, word align, translation probability, word alignment, translation system, source word, ibm model, source sentence, english translation, target language, word segmentation
	RNN-RSM (2014)	reordering, statistical machine, translation model, translations, arabic, word align, translation probability, word alignment, translation system, source word, reordering model, bleu score, smt system, english translation, target language
0.53	**RNN-RSM** (1996)	input, inference, semantic representation, distributional models, logical forms, space model, clustering algorithm, space models, similar word, frequent word, meaning representation, lexical acquisition, new algorithm, same context, multiple words
	RNN-RSM (2014)	input, inference, word vector, word vectors, vector representation, semantic representation, distributional models, semantic space, space model, semantic parser, vector representations, neural language, logical forms, cosine similarity, clustering algorithm

Table 4: Topics (top 15 words) with the highest and lowest drifts (cosine) observed in DTM and RNN-RSM

3.4 Topic Interpretability

Beyond perplexities, we also compute topic coherence (Chang et al., 2009; Newman et al., 2009; Das et al., 2015) to determine the meaningful topics captured. We use the coherence measure proposed by Aletras and Stevenson (2013) that retrieves co-occurrence counts for the set of topic words using Wikipedia as a reference corpus to identify context features (window=5) for each topic word. Relatedness between topic words and context features is measured using normalized pointwise mutual information (NPMI), resulting in a single vector for every topic word. The coherence (*COH*) score is computed as the arithmetic mean of the cosine similarities between all word pairs. Higher scores imply more coherent topics. We use Palmetto[4] library to estimate coherence.

Quantitative: We compute mean and median coherence scores for each time step using the corresponding topics, as shown in Fig 3e and 3f. Table 3 shows *mean-COH* and *median-COH* scores, computed by mean and median of scores from Fig 3e and 3f, respectively. Observe that RNN-RSM captures topics with higher coherence.

Qualitative: Table 5 shows topics (top-10 words) with the highest and lowest coherence scores.

3.5 TTC: Trending Keywords over time

We demonstrate the capability of RNN-RSM to capture word evolution (usage) in topics over time. We define: *keyword-trend* and SPAN. The *keyword-trend* is the appearance/disappearance of the keyword in topic-terms detected over time, while SPAN is the length of the longest sequence of the keyword appearance in its keyword trend.

DTM (2001)	RNN-RSM (2001)	DTM (2012)	RNN-RSM (1997)
semantic	words	discourse	parse
frame	models	relation	cluster
argument	grammar	relations	clustering
syntactic	trees	structure	results
structure	dependency parsing	sentence	query
lexical	parsers	class	pos tag
example	dependency trees	lexical	queries
information	parsing	argument	retrieval
annotation	parse trees	corpus	coreference
lexicon	dependency parse	other	logical form
COH: 0.268	<u>0.284</u>	0.064	<u>0.071</u>

Table 5: Topics with the highest and lowest coherence

Let $\widehat{\mathbf{Q}}_{model} = \{\mathbf{Q}_{model}^{(t)}\}_{t=1}^{T}$ be a set of sets[5] of topic-terms discovered by the $model$ (LDA, RSM, DTM and RNN-RSM) over different time steps. Let $\mathbf{Q}^{(t)} \in \widehat{\mathbf{Q}}_{model}$ be the topic-terms at time step t. The keyword-trend for a keyword k is a time-ordered sequence of 0s and 1s, as

$$\text{trend}_k(\widehat{\mathbf{Q}}) = [\text{find}(k, \mathbf{Q}^{(t)})]_{t=1}^{T}$$

$$\text{where;}\quad \text{find}(k, \mathbf{Q}^{(t)}) = \begin{cases} 1 & \text{if } k \in \mathbf{Q}^{(t)} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

And the SPAN (S_k) for the kth keyword is-

$$S_k(\widehat{\mathbf{Q}}) = \text{length}\big(\text{longestOnesSeq}(\text{trend}_k(\widehat{\mathbf{Q}}))\big)$$

We compute keyword-trend and SPAN for each term from the set of some popular terms. We define average-SPAN for all the topic-terms appearing in the topics discovered over the years,

$$\text{avg-SPAN}(\widehat{\mathbf{Q}}) = \frac{1}{||\widehat{\mathbf{Q}}||} \sum_{\{k|\mathbf{Q}^{(t)} \in \widehat{\mathbf{Q}} \wedge k \in \mathbf{Q}^{(t)}\}} \frac{S_k(\widehat{\mathbf{Q}})}{\hat{v}^k}$$

$$= \frac{1}{||\widehat{\mathbf{Q}}||} \sum_{\{k|\mathbf{Q}^{(t)} \in \widehat{\mathbf{Q}} \wedge k \in \mathbf{Q}^{(t)}\}} S_k^{dict}(\widehat{\mathbf{Q}})$$

---

[4]

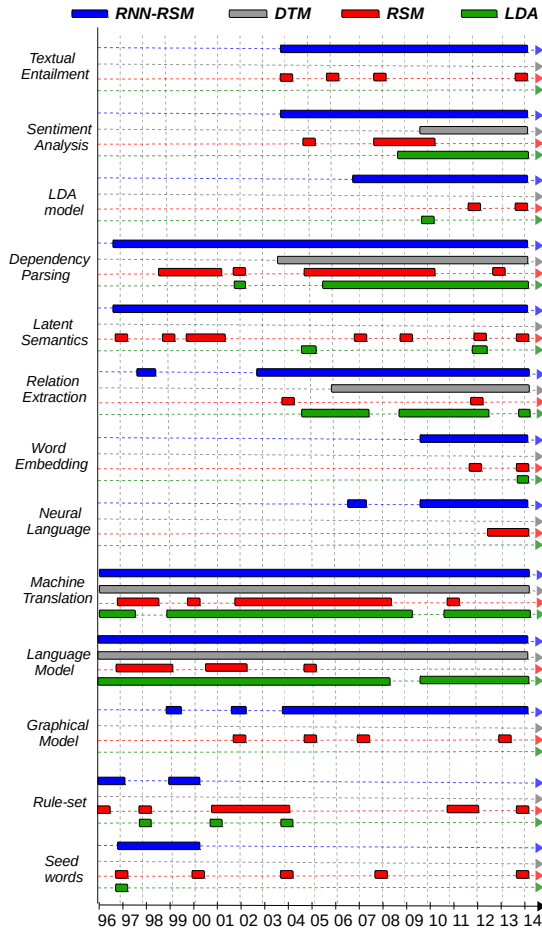[5] a set by **bold** and set of sets by $\widehat{\textbf{bold}}$

Figure 5: Key-term frequency in the input over years

Figure 4: Keyword-trend by RNN-RSM, DTM, RSM, LDA. Bar: Keyword presence in topics for the year

Term	$\hat{v}^k$	LDA		RSM		DTM		RNN-RSM	
		S_k	S_k^{dict}	S_k	S_k^{dict}	S_k	S_k^{dict}	S_k	S_k^{dict}
Textual entailment	918	0	.000	1	.001	0	.000	**11**	.011
Sentiment analysis	1543	6	.004	3	.002	5	.0032	**11**	0.007
Lda model	392	1	.003	1	.002	0	.000	**8**	.020
Dependency parsing	3409	9	.003	5	.001	11	.0032	**18**	.005
Latent semantic	974	1	.001	2	.002	0	.000	**18**	.018
Relation extraction	1734	4	.002	1	.001	9	.0052	**12**	.007
Word embedding	534	1	.002	1	.002	0	.000	**5**	.009
Neural language	121	0	.000	3	.025	0	.000	**5**	.041
Machine translation	11741	11	.001	7	.001	**19**	.0016	19	.002
Language model	11768	13	.001	3	.000	**19**	.0016	19	.002
Graphical model	680	0	.000	1	.001	0	.000	**11**	.016
Rule set	589	1	.0017	**4**	.0068	0	.000	2	.0034
Seed words	396	1	.0025	1	.0025	0	.000	**4**	.0101
avg-SPAN($\widehat{\mathbf{Q}}$)		.002		.007		.003		**.018**	
$\|\widehat{\mathbf{Q}}_{model}\|$		926		2274		335		731	

Table 6: SPAN (S_k) for selected terms, avg-SPAN and set $\|\widehat{\mathbf{Q}}\|$ by LDA, RSM, DTM and RNN-RSM

where $\|\widehat{\mathbf{Q}}\| = |\{k|\mathbf{Q}^{(t)} \in \widehat{\mathbf{Q}} \wedge k \in \mathbf{Q}^{(t)}\}|$ is the count of unique topic-terms and $\hat{v}^k = \sum_{t=1}^{T} \sum_{j=1}^{D_t} v_{j,t}^k$ denotes the count of kth keyword.

In Figure 4, the keyword-trends indicate emergence (appearance/disappearance) of the selected popular terms in topics discovered in ACL and EMNLP papers over time. Observe that RNN-RSM captures longer SPANs for popular keywords and better word usage in NLP research. For example: *Word Embedding* is one of the top keywords, appeared locally (Figure 5) in the recent years. RNN-RSM detects it in the topics from 2010 to 2014, however DTM does not. Similarly, for *Neural Language*. However, *Machine Translation* and *Language Model* are globally appeared in the input document collections over time and captured in the topics by RNN-RSM and DTM. We also show keywords (*Rule-set* and *Seed Words*) that disappeared in topics over time.

Higher SPAN suggests that the model is capable in capturing trending keywords. Table 6 shows corresponding comparison of SPANs for the 13
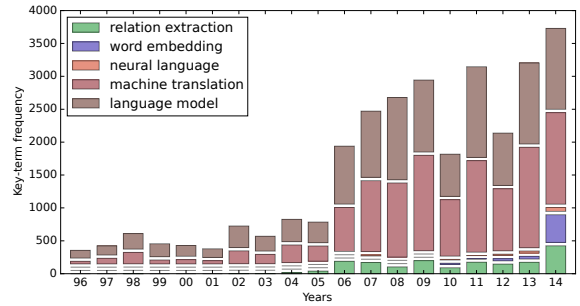
selected keywords. The SPAN S_k for each key-word is computed from Figure 4. Observe that $\|\widehat{\mathbf{Q}}\|_{DTM} < \|\widehat{\mathbf{Q}}\|_{RNN-RSM}$ suggests new topics and words emerged over time in RNN-RSM, while higher SPAN values in RNN-RSM suggest better trends. Figure 6 shows how the word usage, captured by DTM and RNN-RSM for the topic *Word Vector*, changes over 19 years in NLP research. RNN-RSM captures popular terms *Word Embedding* and *Word Representation* emerged in it.

4 Discussion: RNN-RSM vs DTM

Architecture: RNN-RSM treats document's stream as high dimensional sequences over time and models the complex conditional probability distribution i.e. *heteroscedasticity* in document collections and topics over time by a temporal stack of RSMs (undirected graphical model), conditioned on time-feedback connections using RNN (Rumelhart et al., 1985). It has two hidden layers: **h** (stochastic binary) to capture topical information, while **u** (deterministic) to convey temporal information via BPTT that models the topic dependence at a time step t on *all* the previous steps $\tau < t$. In contrast, DTM is built upon
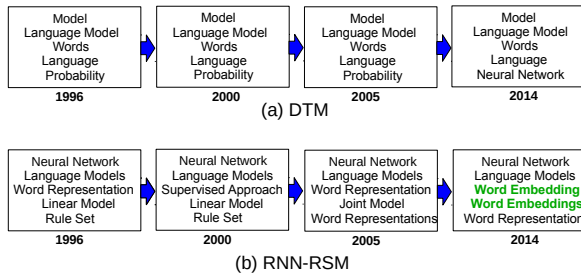
(a) DTM

1996	2000	2005	2014
Model Language Model Words Language Probability	Model Language Model Words Language Probability	Model Language Model Words Language Probability	Model Language Model Words Language Neural Network

(b) RNN-RSM

1996	2000	2005	2014
Neural Network Language Models Word Representation Linear Model Rule Set	Neural Network Language Models Supervised Approach Linear Model Rule Set	Neural Network Language Models Word Representation Joint Model Word Representations	Neural Network Language Models **Word Embedding** **Word Embeddings** Word Representation

Figure 6: Word usage for emerging topic *Word Vector* over time, captured by DTM and RNN-RSM

LDA (directed model), where Dirichlet distribution on words is not amenable to sequential modeling, therefore its natural parameters (topic and topic proportion distributions) for each topic are chained, instead of latent topics that results in intractable inference in topic detection and chaining.

Topic Dynamics: The introduction of explicit connection in latent topics in RNN-RSM allow new topics and words for the underlying topics to appear or disappear over time by the dynamics of topic correlations. As discussed, the distinction of $\mathbf{h}$ and $\mathbf{u}$ permits the latent topic $\mathbf{h}^{(t)}$ to capture new topics, that may not be captured by $\mathbf{h}^{(t-1)}$.

DTM assumes a fixed number of global topics and models their distribution over time. However, there is no such assumption in RNN-RSM. We fixed the topic count in RNN-RSM at each time step, since $\mathbf{W_{vh}}$ is fixed over time and RSM biases turn off/on terms in each topic. However, this is fundamentally different for DTM. E.g. a unique label be assigned to each of the 30 topics at any time steps t and t'. DTM follows the sets of topic labels: $\{TopicLabels^{(t)}\}_{k=1}^{30} = \{TopicLabels^{(t')}\}_{k=1}^{30}$, due to eq (1) in Blei and Lafferty (2006) (discussed in section 5) that limits DTM to capture new (or local) topics or words appeared over time. It corresponds to the keyword-trends (section 3.5).

Optimization: The RNN-RSM is based on Gibbs sampling and BPTT for inference while DTM employs complex variational methods, since applying Gibbs sampling is difficult due to the nonconjugacy of the Gaussian and multinomial distributions. Thus, easier learning in RNN-RSM.

For all models, approximations are solely used to compute the likelihood, either using variational approaches or contrastive divergence; perplexity was then computed based on the approximated likelihood. More specifically, we use variational approximations to compute the likelihood

for DTM (Blei and Lafferty, 2006). For RSM and RNN-RSM, the respective likelihoods are approximated using the standard Contrastive Divergence (CD). While there are substantial differences between variational approaches and CD, and thus in the manner the likelihood for different models is estimated - both approximations work well for the respective family of models in terms of approximating the true likelihood. Consequently, perplexities computed based on these approximated likelihoods are indeed comparable.

5   Conclusion and Future Work

We have proposed a neural temporal topic model which we name as RNN-RSM, based on probabilistic undirected graphical topic model RSM with time-feedback connections via deterministic RNN, to capture temporal relationships in historical documents. The model is the first of its kind that learns topic dynamics in collections of different-sized documents over time, within the generative and neural network framework. The experimental results have demonstrated that RNN-RSM shows better generalization (perplexity and time stamp prediction), topic interpretation (coherence) and evolution (popularity and drift) in scientific articles over time. We also introduced SPAN to illustrate topic characterization.

In future work, we forsee to investigate learning dynamics in variable number of topics over time. It would also be an interesting direction to investigate the effect of the skewness in the distribution of papers over all years. Further, we see a potential application of the proposed model in learning the time-aware i.e. dynamic word embeddings (Aitchison, 2001; Basile et al., 2014; Bamler and Mandt, 2017; Rudolph and Blei, 2018; Yao et al., 2018) in order to capture language evolution over time, instead of document topics.

References

Jean Aitchison. 2001. *Language change: progress or decay?*. Cambridge University Press.

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS)*. Potsdam, Germany, pages 13–22.

James Allan. 2002. Introduction to topic detection and tracking. In *Topic detection and tracking*, Springer, pages 1–16.

James Allan, Jaime G Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. Virginia, US, pages 194–218.

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning*. Sydney, Australia, pages 380–389.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. Analysing word meaning over time by exploiting temporal random indexing. In *Proceedings of the 1st Italian Conference on Computational Linguistics (CLiC-it)*. Pisa University Press, Pisa, Italy.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*. Association for Computing Machinery, Pittsburgh, Pennsylvania USA, pages 113–120.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Proceedings of Machine Learning Research* 3(Jan):993–1022.

Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. 2012. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *Proceedings of the 29th International Conference on Machine Learning*. Edinburgh, Scotland UK.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Vancouver, Canada, pages 288–296.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. 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*. Association for Computational Linguistics, Beijing, China, volume 1, pages 795–804.

Peter V. Gehler, Alex D. Holub, and Max Welling. 2006. The rate adapting poisson model for information retrieval and object recognition. In *Proceedings of the 23rd International Conference on Machine Learning*. Association for Computing Machinery, Pittsburgh, Pennsylvania USA, pages 337–344.

Sujatha Das Gollapalli and Xiaoli Li. 2015. Emnlp versus acl: Analyzing nlp research over time. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 2002–2006.

Pankaj Gupta, Thomas Runkler, Heike Adel, Bernt Andrassy, Hans-Georg Zimmermann, and Hinrich Schütze. 2015a. Deep learning methods for the extraction of relations in natural language text. Technical report, Technical University of Munich, Germany.

Pankaj Gupta, Thomas Runkler, and Bernt Andrassy. 2015b. Keyword learning for classifying requirements in tender documents. Technical report, Technical University of Munich, Germany.

Pankaj Gupta, Hinrich Schütze, and Bernt Andrassy. 2016. Table filling multi-task recurrent neural network for joint entity and relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan, pages 2537–2547.

Pankaj Gupta, Udhayaraj Sivalingam, Sebastian Pölsterl, and Nassir Navab. 2015c. Identifying patients with diabetes using discriminative restricted boltzmann machines. Technical report, Technical University of Munich, Germany.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, pages 363–371.

Geoffrey Hinton and Ruslan Salakhutdinov. 2009. Replicated softmax: an undirected topic model. In *Advances in Neural Information Processing Systems 22*. Curran Associates, Inc., Vancouver, Canada, pages 1607–1614.

Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation* 14(8):1771–1800.

Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural Computation* 18(7):1527–1554.

Tomoharu Iwata, Takeshi Yamada, Yasushi Sakurai, and Naonori Ueda. 2010. Online multiscale dynamic topic models. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, Washington DC, USA, pages 663–672.

David Newman, Sarvnaz Karimi, and Lawrence Cavedon. 2009. External evaluation of topic models. In *Proceedings of the 14th Australasian Document Computing Symposium*. Citeseer, Sydney, Australia.

Iulian Pruteanu-Malinici, Lu Ren, John Paisley, Eric Wang, and Lawrence Carin. 2010. Hierarchical bayesian modeling of topics in time-stamped documents. *IEEE transactions on pattern analysis and machine intelligence* 32(6):996–1011.

Maja Rudolph and David Blei. 2018. Dynamic bernoulli embeddings for language evolution. In *Proceedings of the 27th International Conference on World Wide Web Companion*. Lyon, France.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1985. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Ankan Saha and Vikas Sindhwani. 2012. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, Seattle, Washington USA, pages 693–702.

Aaron Schein, Hanna Wallach, and Mingyuan Zhou. 2016. Poisson-gamma dynamical systems. In *Advances in Neural Information Processing Systems 29*, Curran Associates, Inc., Barcelona, Spain, pages 5005–5013.

Paul Smolensky. 1986. Information processing in dynamical systems: Foundations of harmony theory. Technical report, Colorado University at Boulder Department of Computer Science.

Ilya Sutskever and Geoffrey Hinton. 2007. Learning multilevel distributed representations for high-dimensional sequences. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*. San Juan, Puerto Rico, pages 548–555.

Ilya Sutskever, Geoffrey E. Hinton, and Graham W. Taylor. 2009. The recurrent temporal restricted boltzmann machine. In *Advances in Neural Information Processing Systems 22*. Curran Associates, Inc., Vancouver, Canada, pages 1601–1608.

Graham W. Taylor, Geoffrey E. Hinton, and Sam T. Roweis. 2007. Modeling human motion using binary latent variables. In *Advances in Neural Information Processing Systems 20*. Curran Associates, Inc., Vancouver, Canada, pages 1345–1352.

Ngoc Thang Vu, Heike Adel, Pankaj Gupta, and Hinrich Schütze. 2016a. Combining recurrent and convolutional neural networks for relation classification. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California USA, pages 534–539.

Ngoc Thang Vu, Pankaj Gupta, Heike Adel, and Hinrich Schütze. 2016b. Bi-directional recurrent neural network with ranking loss for spoken language understanding. In *Proceedings of the Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Shanghai, China, pages 6060–6064.

Xiaojun Wan and Jianguo Xiao. 2008. Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence*. Chicago, Illinois USA, volume 8, pages 855–860.

Xuerui Wang and Andrew McCallum. 2006. 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*. Association for Computing Machinery, Philadelphia, Pennsylvania USA, pages 424–433.

Eric P. Xing, Rong Yan, and Alexander G. Hauptmann. 2005. Mining associated text and images with dual-wing harmoniums. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Edinburgh, Scotland UK.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM)*. Association for Computing Machinery, Los Angeles, California USA, pages 673–681.