Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Recurrent Attentional Topic Model

Shuangyin Li *
Department of Computer Science and Engineering
Hong Kong University of Science and Technology, China
shuangyinli@cse.ust.hk

Yu Zhang
Department of Computer Science and Engineering
Hong Kong University of
Science and Technology, China
zhangyu@cse.ust.hk

Rong Pan
School of Data and Computer Science
Sun Yat-sen University, China
panr@sysu.edu.cn

Mingzhi Mao
School of Data and Computer Science
Sun Yat-sen University, China
mcsmmz@mail.sysu.edu.cn

Yang Yang
iPIN
Shenzhen, China
yangyang@ipin.com

Abstract

In a document, the topic distribution of a sentence depends on both the topics of preceding sentences and its own content, and it is usually affected by the topics of the preceding sentences with different weights. It is natural that a document can be treated as a sequence of sentences. Most existing works for Bayesian document modeling do not take these points into consideration. To fill this gap, we propose a Recurrent Attentional Topic Model (RATM) for document embedding. The RATM not only takes advantage of the sequential orders among sentence but also use the attention mechanism to model the relations among successive sentences. In RATM, we propose a Recurrent Attentional Bayesian Process (RABP) to handle the sequences. Based on the RABP, RATM fully utilizes the sequential information of the sentences in a document. Experiments on two copora show that our model outperforms state-of-the-art methods on document modeling and classification.

1 Introduction

Probabilistic topic models provide a suite of algorithms to obtain good representations when facing a collection of documents. The representation obtained by a topic model often corresponds to latent topics in an interpretable space, which is an advantage over other models. Topic models have improved document classification and information retrieval (Wei and Croft 2006) on unstructured text, and many extended models have been applied to many structured text data and non-text data in computer vision (Fei-Fei and Perona 2005) and collaborative filtering (Marlin 2003). Topic models usually assume that words are interchangeable, which is helpful for efficient inference on large corpora (Blei 2012). Actually, documents are sequences of words, sentences, and paragraphs in a hierarchical manner and some works have modeled a document as a sequence of words, including the *n*-gram language modeling (Brown et al. 1992) and recurrent neural networks for language modeling (Sutskever, Martens, and Hinton 2011; Frinken et al. 2012). Moreover, some works consider the syntactic structure of sentences over words to model the document (Boyd-Graber and Blei 2009).

Although topic models have been widely used for document modeling, the topic coherence between sentences, which does exist in natural language, is ignored in existing works. To see this, let us consider the following four sentences describing “Machine Learning” from the Wikipedia: (1) Machine learning is closely related to and often overlaps with computational statistics, a discipline which also focuses in prediction-making through the use of computers. (2) It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. (3) Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms is infeasible. (4) Example applications include spam filtering, optical character recognition, search engines and computer vision. Obviously, sentence (4) is about the applications of machine learning, whose topics are highly coherent with those of the three preceding sentences and so the topics in a sentence could recurrently affect the following sentences. Moreover, the topics of sentence (4) are more relevant to those of sentences (2) and (3) since they all discuss the applications of machine learning. Thus, besides the topic relevance among sentences, it is intuitive that a sentence is related to the preceding sentences with different weights, which are called attention signals just like the attentional mechanism in deep neural networks (Mnih et al. 2014; Bahdanau, Cho, and Bengio 2014; Gregor et al. 2015).

To the best of our knowledge, there is no work to consider sentence coherence and attention signals in Bayesian modeling. To fill this gap, we develop a Recurrent Attentional Topic Model (RATM). Based on a proposed Recurrent Attentional Bayesian Process (RABP), the RATM can model sequences of sentences by considering the dependency between sentences as well as attention signals.

Specifically, the contributions of this work are follows. Firstly, We propose a novel RABP to handle sequential data and to allow a local recurrent information transmission

[*] Part of this work was performed at iPIN.

through a sequence. Secondly, We establish a previously unexplored connection between recurrent Bayesian methods and dynamic attention signals in the principled RATM model, where the attention signals are adaptively learned for sequences of sentences, and we develop an efficient variational inference algorithm. Lastly, a new topic model with RABP is proposed, and the experiments show that RABP can recover meaningful topics in the sequences of sentences. Based on it, RATM has better performance in terms of perplexity and classification accuracy.

The rest of the paper is organized as follows. In Section 2, we discuss related works. In Section 3, we propose the RABP mathematically. In Section 4, we present the RATM and its inference method. In Section 5, we present experimental results on two copora for document modeling and classification. Also, we show some case studies of attention signals among the sentences.

2 Related Works

Many probabilistic topic models have been proposed, including (Hofmann 1999; Blei, Ng, and Jordan 2003; Blei and Lafferty 2005; Blei and McAuliffe 2007; Boyd-Graber and Blei 2009; Hoffman, Blei, and Bach 2010; Li et al. 2015). These models and their extensions have been applied to many tasks such as information retrieval (Wei and Croft 2006; Li, Li, and Pan 2013), document classification (Cai et al. 2008; Li et al. 2016), and so on. Some models, such as the Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) and Correlated Topic Model (CTM) (Blei and Lafferty 2005), are used to model unstructured documents with assumptions that words in a document arise from a mixture of latent topics and that each topic is a distribution over the vocabulary. Many topic models such as (Griffiths et al. 2004; Gruber, Weiss, and Rosen-Zvi 2007) take the order of words and the syntactic of sentences into consideration. In (Griffiths et al. 2004), authors focus on short-range syntactic dependencies and long-range semantic dependencies between words. The HTMM proposed in (Gruber, Weiss, and Rosen-Zvi 2007) models the topics of words in a document as a Markov chain. The syntactic topic model proposed in (Boyd-Graber and Blei 2009) generates words via both thematically and syntactically constraints among them. Almost all the existing topic models consider the sequentiality of documents on the word level only.

Recurrent Neural Networks (RNN) (Sutskever, Martens, and Hinton 2011) provide an efficient way to handle the sequentiality of documents on both the word (Mikolov et al. 2010; 2013) and the sentence levels, and they have been applied to various tasks, including machine translation (Bahdanau, Cho, and Bengio 2014), summarization (Rush, Chopra, and Weston 2015), dialog system (Serban et al. 2016), document modeling (Lin et al. 2015) and so on.

Attention signals are widely applied to language modeling for many text mining tasks (Bahdanau, Cho, and Bengio 2014; Rush, Chopra, and Weston 2015; Ling et al. 2015; Serban et al. 2016) and speech tasks (Chorowski et al. 2015). However, all the proposed models with the attention mechanism are under the neural network framework and few Bayesian models focus on language modeling with attention signals.

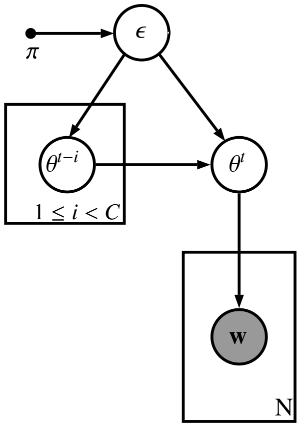


Figure 1: The recurrent attentional Bayesian process with the bag-of-words assumption. The shaded circles denote observed words and others are the hidden variables. ϵ denotes the attention signal, π is a Dirichlet parameter, and C is the length of time windows.

3 Recurrent Attentional Bayesian Process

A Recurrent Attentional Bayesian Process (RABP), denoted by RABP(G_0, π), is parameterized by a base measure G_0 and a concentration parameter π. The generative process for the RABP is defined as follows:

1. Draw θ^1 from G_0.
2. For $t > 1$
   (a) Draw $\epsilon = (\epsilon_1, \ldots, \epsilon_C)^T$ from Dir(π), where Dir(π) denotes a Dirichlet distribution with parameter π;
   (b) With probability ϵ_i, draw θ^t from $\delta_{\theta^{t-i}}$ for $i = 1, \ldots, C-1$, where δ_a denotes a discrete distribution whose probability mass function is equal to 1 at the point a;
   (c) With probability ϵ_C, draw θ^t from G_0.

In this generative process, G_0 is the base distribution and C is the length of the time window. Here ϵ_i for $i = 1, \ldots, C$ is defined as the attention signal and it captures the importance of a preceding neighbour θ^{t-i} to θ^t. ϵ satisfies $\sum_{i=1}^{C} \epsilon_i = 1$ as it follows a Dirichlet distribution with parameter π. The graphical model of RABP is shown in the right of Figure 1.

Based on the generative process, θ^t can be represented as

$$\theta^t | \theta^{t-C+1:t-1}, G_0, \pi \sim \sum_{i=1}^{C-1} \epsilon_i \cdot \theta^{t-i} + \epsilon_C \cdot G_0, \qquad (1)$$

where K is the length of each θ^i and $\theta^{j-C+1:j-1}$ is a $(C-1) \times K$ matrix containing $C-1$ preceding parameters. The attention signal ϵ_i reflects the importance of a preceding parameter in the sequence to current one and C indicates that θ^t can have dependency with the $C-1$ preceding ones.

The RABP is partly related to the Recurrent Chinese Restaurant Process (RCRP) (Ahmed and Xing 2008) and

Dirichlet-Hawkes Process (DHP) (Du et al. 2015). The RCRP, an extension of the Chinese restaurant process, defines a distribution over Dirichlet distribution. The main difference between the RABP and RCRP is that the RABP considers several preceding time points with dynamic attentional weights, while in the RCRP the dependency of the parameter over the preceding ones is invariant to both positions and the content information. The DHP focuses on modeling the intensity of discrete events using a Hawkes process but the RABP can model recurrent sequences in a discrete space with attention signals. The RABP considers the local reflection for the current state and can be used as a prior to model documents as sentence-level sequences, which will be shown in the next section.

4 Recurrent Attentional Topic Model

As a collection of M documents, a corpus is defined as $D = \{\mathbf{d}^1, \ldots, \mathbf{d}^M\}$, where $\mathbf{d}^i, i \in \{1, \ldots, M\}$ denotes the i-th document. A document $\mathbf{d}^i$ is a sequence of S_i sentences denoted by $\mathbf{d}^i = (\mathbf{s}_1^i, \ldots, \mathbf{s}_{S_i}^i)$, where $\mathbf{s}_j^i, j \in \{1, \ldots, S_i\}$ denotes the j-th sentence in $\mathbf{d}^i$. Let $\mathbf{s}_j^i = (w_{j,1}^i, \ldots, w_{j,N_j^i}^i)$ denotes the vector of N_j^i words associated with sentence $\mathbf{s}_j^i$.

It is clear that the topic distribution of one sentence is related to those of previous sentences, which is called cohesion or coherence in linguistics. This observation matches the motivation of the RABP and then based on the proposed RABP, we can model a document as a sequence of sentences, leading to the proposed RATM.

By considering a document as a sequence of sentences, the RATM attempts to capture the joint influences of previous sentences to current one. Moreover, the topics of a sentence are also affected by those of the host document that the sentence belongs to. Hence, the topic distribution of a sentence is generated from those of both its preceding sentences and the host document.

Let θ^{s_j} denote the topic distribution of the j-th sentence $\mathbf{s}_j$ in document $\mathbf{d}$, which follows the RABP with parameters G_0 and π. G_0 is defined as a K-dimensional Dirichlet distribution for θ and hence θ denotes the topic distribution of a sentence over K latent topics. Based on these notations, the generation process of RATM is defined as

1. For each topic $k \in \{1, \ldots, K\}$, draw $\beta_k \sim \text{Dir}(\eta)$, where η is a V-dimensional prior vector of β;
2. For each document $\mathbf{d}^i, i \in \{1, \ldots, M\}$:
   (a) Draw $\vartheta^d \sim \text{Dir}\ (\alpha)$;
   (b) For sentence $\mathbf{s}_j, j \in \{1, \ldots, S_i\}$ in the document $\mathbf{d}$:
      i. Draw $\theta^{s_j} \sim \text{RABP}(\delta_{\vartheta^d}, \pi)$;
      ii. For each word $w_n, n \in \{1, \ldots, N_j\}$ in sentence $\mathbf{s}_j$:
         A. Draw z_n ~Mult(θ^{s_j});
         B. Draw w_n ~Mult(β_{z_n}).

In this generative process, ϑ^d, a K-dimensional vector following a Dirichlet distribution, describes the topic distribution of a document and it is used as G_0 in the RABP for the topics of sentences. A topic is a distribution over a fixed vocabulary which is denoted by β_k. Attention signals are used in the RABP without explicitly introducing in the generative process and note that attention signals are dynamic in different sentences. z_n is the topic assignment for each word n and it describes the topic distribution of a word. $\theta^{s_j} \sim \text{RABP}(\delta_{\vartheta^d}, \pi)$ indicates that the topic distribution of current sentence $\mathbf{s}_j$ is generated by a RABP, which means that sentence $\mathbf{s}_j$ depends on the $C - 1$ preceding sentences via a vector of adaptive attention signals, ϵ^j, as described in RABP. Figure 2 shows the graphical model of the RATM.

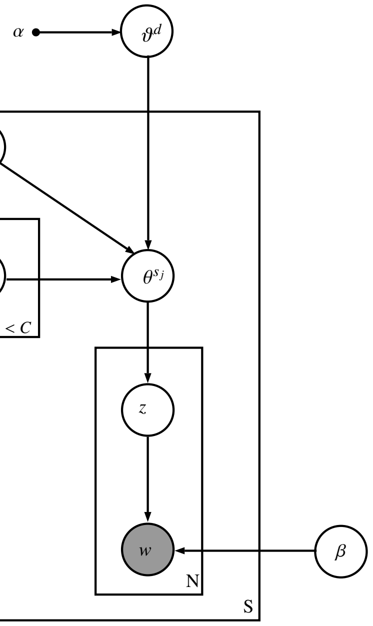


Figure 2: The graphical model of the RATM. ϑ^d is the topic distribution of a document. $\theta^{s_{j-i}}$ denotes the topic distribution of a preceding sentence where $1 \leq i < C$ and C is the length of time windows used in the RABP.

Inference

The main problem in the inference for the RATM is to estimate the posterior distribution of latent variables conditioned on observed data. Typical topic models can be learned by Gibbs sampling methods due to the conjugate property between the topic assignment and the prior over the topic distribution. However, the RATM does not enjoy such conjugate property due to the prior for the topic distribution of a sentence (see Eq. (1)), making the posterior distribution in the RATM intractable to compute. Thus, we resort to the variational inference.

In the variational inference, the posterior distribution is approximated by a group of variational distributions with free variational parameters and the group of variational distributions are enforced to be close to the true posterior. For each sentence $\mathbf{s}_j$ with N_j words in document d, we use the following fully factorized variational distribution:

$$q^s(\epsilon^j, \mathbf{z}|\xi, \gamma) = q(\epsilon^j|\xi) \prod_{n=1}^{N_j} q(z_n|\gamma_n),$$

where ξ is a variational parameter of a Dirichlet distribution for sentence $\mathbf{s}_j$ and $\{\gamma_n\}$ is a variational parameter of a multinomial distribution. Thus, the Jensen's lower bound on the log probability of sentence $\mathbf{s}_j$ can be computed as

$$\mathcal{L}^{s_j}(\beta, \pi; \xi, \gamma) = \mathbb{E}_q[\log p(\epsilon^j|\pi)] + \sum_{n=1}^{N_j} \mathbb{E}_q[\log p(z_n|\epsilon^j, \theta^{j-C+1:j-1})]$$

$$+ \sum_{n=1}^{N_j} \mathbb{E}_q[\log p(w_n|z_n, \beta)] - \mathbb{E}_q[\log q(\epsilon^j)] - \mathbb{E}_q[\log q(\mathbf{z})],$$

where the G_0 is ignored for the ease of presentation. Note that, even though it is difficult to compute $\mathbb{E}_q[\log p(z_n|\epsilon, \theta^{j-C+1:j-1})]$, we can obtain its lower-bound by following the method described in (Li et al. 2013). Then we need to maximize the lower-bound $\mathcal{L}^{s_j}(\beta, \pi; \xi, \gamma)$ to find the estimations of the variational parameters and model parameters, which are detailed in the following sections.

Variational Update for Attention Signals Based on $\mathcal{L}^{s_j}(\beta, \pi; \xi, \gamma)$, for the variational parameters ξ corresponding to the attention signals of sentence s_j, the objective is to maximize the following equation:

$$\mathcal{L}^{s_j}_{[\xi]} = \sum_{l=1}^{C-1}(\pi_l - 1)(\Psi(\xi_l) - \Psi(\sum_{l'=1}^{C-1} \xi_{l'})) - \log\Gamma(\sum_{l=1}^{C-1} \xi_l) + \sum_{l=1}^{C-1} \log\Gamma(\xi_l)$$

$$- \sum_{l=1}^{C-1}(\xi_l - 1)(\Psi(\xi_l) - \Psi(\sum_{l'=1}^{C-1} \xi_{l'})) + \sum_{n=1}^{N_j}\sum_{k=1}^{K} \gamma_{nk} \sum_{l=1}^{C-1} \log\theta_l^{j-C+1:j-1} \frac{\xi_l}{\sum_{l'=1}^{C-1} \xi_{l'}},$$

where $\Psi(\cdot)$ is the digamma function, which is the first derivative of the logarithm of the Gamma function. We use the gradient descent method to estimate ξ.

Variational Update for Word Assignment For each word w_n in sentence s_j, a topic index z_n is assigned to w_n and γ_{nk} is the variational parameter corresponding to the probability that the topic k is assigned to the word w_n. The variational update for γ_{nk} can easily be obtained as

$$\gamma_{nk} \propto \beta_{k,v^{w_n}} \exp \sum_{l=1}^{C-1} \log\theta_l^{j-C+1:j-1} \frac{\xi_l}{\sum_{l'} \xi_{l'}}, \qquad (2)$$

where v^{w_n} denotes the index of word w_n in the dictionary.

The traditional topic models based on the bag-of-words assumption would stumble when the document is too short, which has been discussed in (Tang et al. 2014). The proposed model is capable of handling short documents because it fully utilizes the topic information from the preceding sentences (see the summand in Eq. (2)) and adaptive attention signals to generate the topic distribution for current sentence.

Parameter Estimation The model parameters include π and β. Based on $\mathcal{L}^{s_j}(\beta, \pi; \xi, \gamma)$, the objective function for π to be maximized over the whole corpus is formulated as

$$\mathcal{L}_{[\pi]} = \sum_{i=1}^{M}\sum_{j=1}^{S_i}(\log\Gamma(\sum_{l=1}^{C-1} \pi_l) - \sum_{l'=1}^{C-1} \log\Gamma(\pi_{l'}) + \sum_{l=1}^{C-1}(\pi_l - 1)(\Psi(\xi_l) - \Psi(\sum_{l'=1}^{C-1} \xi_{l'}))).$$

We can invoke the linear-time Newton-Raphson algorithm described in the LDA to estimate π.

For β, we set the derivative of the variational lower-bound with respect to β_{kv} to 0, leading to the following solution:

$$\beta_{kv} = \sum_{i=1}^{M}\sum_{j=1}^{S_i}\sum_{n=1}^{N_j} \gamma_{nk} \cdot w_n^v,$$

where v is the index of w_n in the dictionary.

Document Embedding

As a kind of topic models, the RATM is to extract the topic distribution of each document for document embedding. In the above variational inference framework, we can define $G_0 = \delta_{\vartheta^d}$ and update the topic distribution for a whole document, ϑ^d, which is the embedding of one document. Note that we treat ϵ_C as the attention signal for ϑ^d as described in the RABP. Based on the above variational inference framework, we introduce a new variational variable for the document $\mathbf{d}^i$, $\rho^{\mathbf{d}^i}$, which follows a Dirichlet distribution. Thus, we can use the Jensen's inequality to lower-bound the log-probability of a document $\mathbf{d}^i$ as:

$$\mathcal{L}^{\mathbf{d}^i}(\beta, \pi, \vartheta^d; \xi, \gamma, \rho^{\mathbf{d}^i}) = \sum_{j}^{S_i} \mathcal{L}^{s_j}(\beta, \pi; \xi, \gamma) + \mathbb{E}_q[\log p(\vartheta^d|\alpha)] - \mathbb{E}_q[\log q(\rho^{\mathbf{d}^i})],$$

where α is initialized by LDA and then fixed as LDA did. We use an alternating optimization to solve the above objective function. When $\rho^{\mathbf{d}^i}$ and ϑ^d are fixed, the variational and model parameters for different sentences in the document are independent and we can follow the variational approach described in the previous sections to update $\{\beta, \pi, \xi, \gamma\}$. When $\{\beta, \pi, \xi, \gamma\}$ are fixed, we maximize $\mathcal{L}^{\mathbf{d}^i}(\beta, \pi, \vartheta^d; \xi, \gamma, \rho^{\mathbf{d}^i})$ with respect to ϑ^d and $\rho^{\mathbf{d}^i}$. By setting the derivative with respect to $\rho^{\mathbf{d}^i}$ to 0, we can obtain an analytical solution as

$$\rho_k^{\mathbf{d}^i} = \alpha_k + \sum_{j=1}^{S_i}\sum_{n=1}^{N_j} \gamma_{nk} \cdot \frac{\xi_C}{\sum_l^C \xi_l}. \qquad (3)$$

With $\rho_k^{\mathbf{d}^i}$, we can obtain ϑ^d as normalized $\{\rho_k^{\mathbf{d}^i}\}$, i.e., $[\vartheta^d]_k = \rho_k^{\mathbf{d}^i} / \sum_{k'} \rho_{k'}^{\mathbf{d}^i}$, according to (Blei, Ng, and Jordan 2003).

Discussion When we let $C = 1$, Eq (1) will be $\theta^t|\theta^{t-C+1:t-1}, G_0, \pi \sim \epsilon_C \cdot G_0$ where $\epsilon_C = 1$, which means that the topic distribution of current θ^t follows the base distribution G_0. Thus, with $G_0 = \delta_{\vartheta^d}$ and $C = 1$ in RATM, the variational parameter $\rho_k^{\mathbf{d}^i}$ will be

$$\rho_k^{\mathbf{d}^i} = \alpha_k + \sum_{j=1}^{S_i}\sum_{n=1}^{N_j} \gamma_{nk},$$

where γ_{nk} are the probability assignments for all the words in each sentence, and it is same as the equation of variational topic distribution in LDA. Thus, it is interesting to note that RATM degenerates into LDA when attentional signals are ignored.

5 Experiments

The proposed model is evaluated on two corpora. The first corpus is a subset of the Wikipedia. We extract abstracts of each page in the Wikipedia and remove abstracts with less than 5 sentences to form the corpus, which contains 241,290 documents. We remove stop words and obtain a vocabulary of 21,968 words. Each document belongs to one of 68 categories such as education, book, arts, and so on, and the average number of sentences in this corpus is about 7. The second corpus we used is news articles from New York Times (NYTimes) from January 1st, 2016 to May 8th, 2016. After removing news which contain less than 5 sentences, we obtain 27,523 articles, each of which belong to one of 42 categories such as world, movies, sports, magazine and so on. After removing stop words, we obtain a dictionary with 12,047 words and the average number of sentences in the NYTimes corpus is about 40.

Results

The baseline methods include the LDA, CTM, Hierarchical Dirichlet processes (HDP) (Teh et al. 2012), and Replicated Softmax Model (RSM) (Hinton and Salakhutdinov 2009). For the proposed RATM model, we trained two variants under different settings. A RATM model called the RATM-N is trained without using G_0 for the generation of each sentence and hence it does not update the ϑ^d for each document as well as the responding variational parameters ρ^{d^i}. Another RATM model called the RATM-D just uses the inference described in Section 4. In the RATM-N and RATM-D, the first $C-1$ sentences in a document do not have $C-1$ preceding sentences, which bring difficulties to the use of the attention signals in the RABP. In this case, we just use the topic distribution of the host document, ϑ^d, as the topics for the unused attention signals. For the Wikipedia corpus, C is set to 4, and it is 6 in the NYTimes corpus.

To compare the performance of different models, we use the held-out perplexity as a measure, which is defined for the RATM as

$$perplexity(D_{test}) = \exp\left(-\frac{\sum_{i=1}^{M}\sum_{j=1}^{S_i} \log p(\mathbf{s}_j^i)}{\sum_{i=1}^{M}\sum_{j=1}^{S_i} N_j^i}\right),$$

where the test set has M documents and $\sum_{j=1}^{S_i} N_j^i$ is the total number of words in document d^i. The lower the perplexity is, the better the performance is.

In each corpus, 80% documents are used for training and the rest is for testing. That is, for the Wikipedia corpus, there are 20,000 documents for training and 4,000 documents for testing. For the NYTimes corpus, 22,000 documents are used for training and 5,523 documents for testing.

To see the effect of the number of latent topics, Table 1 shows the held-out perplexities of different methods on the Wikipedia and NYTimes corpora when the number of latent topics takes three values, i.e., 50, 100 and 200. The results show that the RATM-N and RATM-D have much better performance than baseline models. The performance of the RATM-D is better than that of the RATM-N, which demonstrates that the incorporation of the topic modeling of the whole document can bring benefits for performance improvement. Moreover, when the number of topics increases, the performance of the RATM-N and RATM-D tends to become better due to the increasing capacity of the models. Besides, we compare different C on RATM-D based on the perplexity. We set $C = 2,3,4,5,6$ with $T = 50$ on Wikipedia. Note that, when $C = 1$, the RATM-D is equivalent to the LDA. As shown in Table 1, we find that RATM-D reaches the best result when $C = 4$, and then becomes worse due to the overfitting when C is increasing.

Analysis on Attention Signals

In the section, we show the effect of the attention signals used in the RATM model.

Models	Wikipedia			NYTimes		
	T=50	T=100	T=200	T=50	T=100	T=200
LDA	585.08	493.73	402.98	794.81	768.24	748.45
CTM	**524.11**	435.67	440.89	730.60	645.41	736.38
HDP	728.03	728.03	728.03	1582.67	1582.67	1582.67
RSM	752.36	750.08	767.08	1259.10	1251.71	1266.5
RATM-N	553.87	402.47	328.6	576.06	493.44	500.07
RATM-D	532.72	**392.05**	**314.47**	**529.08**	**442.65**	**440.68**

# on Wikipedia / T=100	C=2	C=3	C=4	C=5	C=6
RATM-D	424.56	396.68	392.05	398.51	430.15

Table 1: The top table shows the perplexity of different models on the two corpora with $T = 50, 100, 200$. The table below shows the perplexity of different C of RATM-D on Wikipedia with T =100.

The attention signals indicate the importance of the preceding sentences to current one in a document. We train the RATM-D model on the Wikipedia corpus and set C to be 4 and 5. We do not use the topic distribution of the host document as G_0 since we just want to show the local relations among sentences and hence we manually set all the probabilities to sample from G_0 in step 2(c) of the RABP to be 0. Thus, we can show the values of attention signals with 3 and 4 preceding sentences for each current sentence.

Table 3 shows the values of attention signals for some documents in the Wikipedia corpus and the numbers in red are the values of attention signals of the preceding sentences for the last sentence which is in italic. From the results, we can see that, in some cases, the values of attention signals increase for the closer sentences, for example, in the ”Machine learning” case. While, in other cases, the values of attention signals could be related to the similarities of topics between current sentence and the preceding ones.

To examine the robustness of the RATM-D model based on the attention signals, we randomly selected two sentences and inserted them into the document “Artificial intelligence” of the Wikipedia corpus as follows:
(1) The central problems or goals of AI research include reasoning knowledge planning learning natural language processing communication perception and the ability to move and manipulate objects. (2) Hype and glory is memoir from William Goldman which details his experiences as judge at the Cannes Film festival and miss America pageant. (3) The book includes an interview with Clint Eastwood and profile on Robert Redford. (4) There are large number of tools used in AI including versions of search and mathematical optimization logic methods based on probability and economics and many others. (5) The AI field is interdisciplinary

#	Values of attention signals		
Sentence (4)	(1) 0.73	(2) 0.15	(3) 0.12
Sentence (5)	(2) 0.03	(3) 0.01	(4) 0.96
Sentence (6)	(3) 0.09	(4) 0.48	(5) 0.43

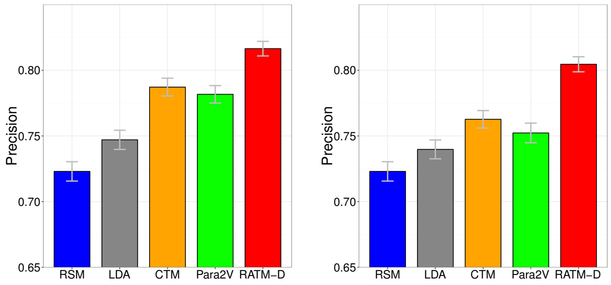
Table 2: The values of attention signals for each sentence in a row. The numbers in red are the values of the attention signals of noisy sentences.

Artificial intelligence	(0.591437) The central problems (or goals) of AI research include reasoning, knowledge, planning, learning, natural language processing (communication), perception and the ability to move and manipulate objects. (0.219417) General intelligence is still among the field's long-term goals. (0.189146) Currently popular approaches include statistical methods, computational intelligence and traditional symbolic AI. *There are a large number of tools used in AI, including versions of search and mathematical optimization, logic, methods based on probability and economics, and many others.*
Machine learning	(0.045228) Machine learning is closely related to and often overlaps with computational statistics; a discipline which also focuses in prediction-making through the use of computers. (0.280551) It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. (0.674221) Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms is infeasible. *Example applications include spam filtering, optical character recognition (OCR), search engines and computer vision.*
Human rights	(0.203636) They require empathy and the rule of law and impose an obligation on persons to respect the human rights of others. (0.216774) They should not be taken away except as a result of due process based on specific circumstances; for example, human rights may include freedom from unlawful imprisonment, torture, and execution. (0.050603) The doctrine of human rights has been highly influential within international law, global and regional institutions. (0.528987) Actions by states and non-governmental organizations form a basis of public policy worldwide. *The idea of human rights suggests that if the public discourse of peacetime global society can be said to have a common moral language, it is that of human rights.*
Mathematics	(0.313444) Rigorous arguments first appeared in Greek mathematics, most notably in Euclid's Elements. (0.336590) Since the pioneering work of Giuseppe Peano, David Hilbert, and others on axiomatic systems in the late 19th century, it has become customary to view mathematical research as establishing truth by rigorous deduction from appropriately chosen axioms and definitions. (0.102) Mathematics developed at a relatively slow pace until the Renaissance, when mathematical innovations interacting with new scientific discoveries led to a rapid increase in the rate of mathematical discovery that has continued to the present day.(0.247966) Galileo Galilei said, ”The universe cannot be read until we have learned the language and become familiar with the characters in which it is written.*It is written in mathematical language, and the letters are triangles, circles and other geometrical figures, without which means it is humanly impossible to comprehend a single word.*

Table 3: The values of attention signals for sentences in some documents of the Wikipedia corpus. The numbers in red are the values of attention signals of the preceding sentences for the last sentence which is in italic.

in which number of sciences and professions converge including computer science mathematics psychology linguistics philosophy and neuroscience as well as other specialized fields such as artificial psychology. (6) The field was founded on the claim that central property of human intelligence the sapience of Homo sapiens can be so precisely described that machine can be made to simulate it.

Sentences (2) and (3) are noises since they are unrelated to others. We record the values of the attention signals for the follow-up sentences (4), (5) and (6) in each row of Table 2. We can see that the attention signals of noisy sentences are much smaller than those of the normal sentences in the three cases and so our attention signals are robust to noisy sentences.



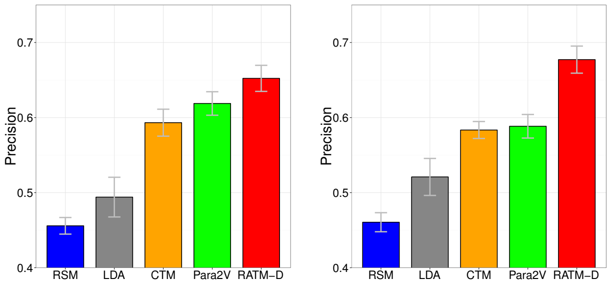
(a) $T = 100$

(b) $T = 200$

Figure 3: Classification results on the Wikipedia corpus for different models with 5-fold cross-validation.

Experiments on Document Classification

In this section, we evaluate the performance of different models on the Wikipedia and NYTimes corpora for the document classification task. We utilize the document features generated by RATM-D and baseline methods in two dimensions, 100 and 200, respectively. We use β generated by the LDA model to initialize the topic distributions over words in the proposed RATM-D. The baseline models we used here include the LDA, CTM, RSM and Para2V (Le and Mikolov 2014). Here we do not include the RATM-N for comparison since it cannot obtain embeddings for documents. The SVM with the LIBSVM implementation (Chang and Lin 2011) and the Gaussian kernel is used as the classifier. We use the held-out precision as the performance measure. From the results shown in Figures 3 and 4, we see that the performance of the RATM-D model is significantly better than that of baseline methods. One reason could be that compared with the LDA, CTM, RSM and Para2V, the proposed RATM-D uses not only the word counts in a document but also the sequential information between sentences, leading to more effective embeddings for documents.



(a) $T = 100$

(b) $T = 200$

Figure 4: Classification results on the NYTimes corpus for different models with 5-fold cross-validation.

6 Conclusion

In this work, we propose the RATM to handle sequences in a discrete space and apply it to document modeling by viewing a document as a sequence of sentences. We evaluate the approach on topic modeling based on two measures: the held-out perplexity and classification accuracy. Moreover, we analyze the attention signals learned from our model for sentences in two different corpora. A future direction is to devise parallel algorithms for the RATM to further improve its efficiency.

Acknowledgments

We thank the support of iPIN and Shenzhen project KC2016KCLT0009A. We also thank the great support of Prof. Qiang Yang.

References

Ahmed, A., and Xing, E. P. 2008. Dynamic non-parametric mixture models and the recurrent Chinese restaurant process: with applications to evolutionary clustering. In *ICDM*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Blei, D. M., and Lafferty, J. D. 2005. Correlated topic models. In *NIPS*.

Blei, D. M., and McAuliffe, J. D. 2007. Supervised topic models. In *NIPS*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*.

Blei, D. M. 2012. Probabilistic topic models. *Communications of the ACM*.

Boyd-Graber, J. L., and Blei, D. M. 2009. Syntactic topic models. In *NIPS*.

Brown, P. F.; Desouza, P. V.; Mercer, R. L.; Pietra, V. J. D.; and Lai, J. C. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics*.

Cai, D.; Mei, Q.; Han, J.; and Zhai, C. 2008. Modeling hidden topics on document manifold. In *Proceedings of International Conference on Information and Knowledge Management*.

Chang, C., and Lin, C. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*.

Chorowski, J. K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; and Bengio, Y. 2015. Attention-based models for speech recognition. In *NIPS*.

Du, N.; Farajtabar, M.; Ahmed, A.; Smola, A. J.; and Song, L. 2015. Dirichlet-Hawkes processes with applications to clustering continuous-time document streams. In *SIGKDD*. ACM.

Fei-Fei, L., and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Frinken, V.; Fischer, A.; Manmatha, R.; and Bunke, H. 2012. A novel word spotting method based on recurrent neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(2).

Gregor, K.; Danihelka, I.; Graves, A.; and Wierstra, D. 2015. DRAW: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*.

Griffiths, T. L.; Steyvers, M.; Blei, D. M.; and Tenenbaum, J. B. 2004. Integrating topics and syntax. In *NIPS*.

Gruber, A.; Weiss, Y.; and Rosen-Zvi, M. 2007. Hidden topic Markov models. In *Proceedings of International Conference on Artificial Intelligence and Statistics*.

Hinton, G. E., and Salakhutdinov, R. R. 2009. Replicated softmax: an undirected topic model. In *NIPS*.

Hoffman, M. D.; Blei, D. M.; and Bach, F. R. 2010. Online learning for latent Dirichlet allocation. In *NIPS*.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *SIGIR*.

Le, Q. V., and Mikolov, T. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Li, S.; Huang, G.; Tan, R.; and Pan, R. 2013. Tag-weighted Dirichlet allocation. In *Proceedings of 13th IEEE International Conference on Data Mining*.

Li, S.; Li, J.; Huang, G.; Tan, R.; and Pan, R. 2015. Tag-weighted topic model for large-scale semi-structured documents. *arXiv preprint arXiv:1507.08396*.

Li, S.; Pan, R.; Zhang, Y.; and Yang, Q. 2016. Correlated tag learning in topic model. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI)*.

Li, S.; Li, J.; and Pan, R. 2013. Tag-weighted topic model for mining semi-structured documents. In *IJCAI*.

Lin, R.; Liu, S.; Yang, M.; Li, M.; Zhou, M.; and Li, S. 2015. Hierarchical recurrent neural network for document modeling. In *EMNLP*.

Ling, W.; Chu-Cheng, L.; Tsvetkov, Y.; and Amir, S. 2015. Not all contexts are created equal: Better word representations with variable attention. In *EMNLP*.

Marlin, B. M. 2003. Modeling user rating profiles for collaborative filtering. In *NIPS*.

Mikolov, T.; Karafiát, M.; Burget, L.; Cernockỳ, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Mnih, V.; Heess, N.; Graves, A.; et al. 2014. Recurrent models of visual attention. In *NIPS*.

Rush, A. M.; Chopra, S.; and Weston, J. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*.

Sutskever, I.; Martens, J.; and Hinton, G. E. 2011. Generating text with recurrent neural networks. In *ICML*.

Tang, J.; Meng, Z.; Nguyen, X.; Mei, Q.; and Zhang, M. 2014. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *ICML*.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2012. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*.

Wei, X., and Croft, W. B. 2006. LDA-based document models for ad-hoc retrieval. In *SIGIR*.