World Wide Web (2019) 22:2545–2560
https://doi.org/10.1007/s11280-018-0639-1

Sentence level topic models for associated topics extraction

Haixin Jiang[1] · Rui Zhou[2] · Limeng Zhang[1] · Hua Wang[3] · Yanchun Zhang[3,4]

Received: 30 November 2017 / Revised: 4 August 2018 / Accepted: 19 September 2018 /
Published online: 18 October 2018

Abstract

In LDA model, independence assumptions in the Dirichlet distribution of the topic proportions lead to the inability to model the connections between topics. Some researchers have attempted to break them and thus obtained more powerful topic models. Following this strategy, by using an association matrix to measure the association between latent topics, we develop an associated topic model (ATM), in which consecutive sentences are considered important and the topic assignments for words are jointly determined by the association matrix and the sentence level topic distributions, instead of the document-specific topic distributions only. This approach gives a more realistic modeling of latent topic connections where the presence of a topic may be connected with the presence of another. We derive a collapsed Gibbs sampling algorithm for inference and parameter estimation for the ATM. The experimental results demonstrate that the ATM gives a more practical interpretation and is capable of learning more associated topics.

Keywords Topic models · Hierarchical models · Association measures · Gibbs sampling · Text analysis

1 Introduction

Topic modeling is an efficient tool to explore large collections of text by automatically extracting the underlying topics contained in the documents. Researchers in this field have proposed a suite of algorithms that uncover the hidden thematic structure in document collections and these algorithms help us develop new ways to search, browse and summarize large archives of texts. Nowadays topic models have achieved significant progress in the statistical analysis of documents collection and other discrete data.

In topic modeling field the basic approach is to model each word in a document as a sample from latent aspects or topics (e.g., [7, 11, 17, 21, 32]). They employ parameters on different levels, global parameters that are associated with the probability of words given

This article belongs to the Topical Collection: *Special Issue on Web and Big Data*
Guest Editors: Junjie Yao, Bin Cui, Christian S. Jensen, and Zhe Zhao

✉ Haixin Jiang
jianghaixin13b@mails.ucas.ac.cn

Extended author information available on the last page of the article.

topics, and a set of parameters for each of the documents that stand for the probability of topics in a document in the corpora. These models assume that each document is a multinomial distribution over topics and each topic is a multinomial distribution over words and uses the Dirichlet distribution to model the variability among topic proportions. This has gradually become a common framework in topic modeling field.

In these models, the order of words is usually neglected and text corpora is represented by a co-occurrence matrix of words and documents. That is, each word is generated from a single topic, and different words in a document may be generated from different topics and therefore each document is represented as mixing proportions for topics and thereby reduced to a multinomial distribution on a fixed set of topics. The distribution is considered as a short description of the document.

However, the strong independence assumption imposed by the Dirichlet that the random variables are independent and identically distributed in LDA is not realistic when analyzing real document collections and it fails to directly model relations between the occurrence of topics. Specifically, under a Dirichlet, the components of the proportions vector are nearly independent, which leads to the strong assumption that the presence of one topic is not correlated with the presence of another. However, it is natural to expect that the occurrences of the underlying latent topics will be highly correlated.

Relaxing these assumptions of exchangeability and dependence is expected to yield better models which can improve the ability to capture local dependencies between words. In recent years Markov models have been tried to solve this problem in which consecutive words are modeled by Markovian relations [1, 4, 19, 30, 32–35]. These models follow the framework by the LDA model and differently assume that each word generation depends on a latent topic assignment as well as on the previous words or sentences in the text and hence are able to capture relations between consecutive words or sentences. Griffiths et al. [18] employs a latent variable standing for syntactic classes whether words are generated from topics that are randomly drawn from the topic mixture of the document or from the syntactic classes that are drawn from the previous syntactic class. The model treats the latent variables of the syntactic classes as a sequence with local dependencies while latent assignments of topics are similar to the LDA model. Gruber et al. [19] models the topics of words in a document as a Markov chain where all words in a sentence have the same topic and consecutive sentences intend to have the same topic. Andrews and Vigliocco [1] assumes that the topics of words in a document form a Markov chain, and that consecutive words are more likely to have the same topics. Wang et al. [33] detects semantic relations by projecting the new relation's training instances onto a lower dimension topic space constructed from existing relation detectors through a three step process. All these models focus on what the word generations depend on and try to get better models.

Following the same lines we propose the associated topic model (ATM) to model unstructured data that contains stream of sentences and words. In this paper we assume sentences are bags of words and focus on the orders of sentences in each document because the transition of topic distributions between consecutive sentences can provide useful information about the possible meaning of words.

Unlike the LDA model and mixture of unigram models, which allow random topic transitions among words in a document, we assume consecutive sentences are more likely to have the similar topics and allow topics transitions through the sentences. Documents are no more than a random permutation of words. The input to the algorithm is the entire document, sentence by sentence, rather than a document-word co-occurrence matrix. This obviously increases the storage requirement for each document, but it allows us to capture unknown relations among topics. Hence the generated topics have more realistic meanings. It could be very useful for word sense disambiguation in many applications such as machine translation. Furthermore, the topic assignment will tend to be coherent and it in turn affects topics transitions in consecutive sentences.

The paper is organized as follows. We introduce basic notation and terminology in Section 3 and related work in Section 2. The ATM model is presented in Section 4 and inference and parameter estimation for ATM are discussed in Section 5. Empirical results in text modeling are presented in Section 6. Finally, Section 7 presents our discussions.

2 Related work

Conventional algorithms in topics models mostly assume each word in the document was generated by a hidden topic and explicitly model the word distribution of each topic as well as the prior distribution over topics in the document, which has formed a common framework. After that, various approaches, including purely unsupervised topics models and external knowledge based topic models, have been proposed to exploit the relations among words or topics to get meaningful topics.

In unsupervised topics models, latent variables are drawn in different ways from a fixed distribution. Blei and Lafferty [6] and [23] made the first attempt on breaking the independence assumption by substituting the logistic normal distribution for the Dirichlet prior distribution, which allows for a general pattern of variability between the components. CTM models the topic proportions with an alternative, more flexible distribution that allows for covariance structure among the components. Li and Mccallum [24], employing a directed acyclic graph (DAG) to represent the topic complex structure, introduced the pachinko allocation model (PAM) to captures arbitrary correlations. Afterwards [19] used Markov chain to model the topics of words in the document. They assume that all words in the same sentence have the same topic and successive sentences are more likely to have the same topics. Chong et al. [13] developed Markov topic models (MTMs) by applying Gaussian (Markov) random fields to model the correlations of different corpora. The MTMs could learn topics simultaneously from multiple corpora and capture the internal topic structure within each corpus and the relationships between topics across the corpora. Blei et al. [8] used nested Chinese restaurant process (nCRP) as a prior distribution in a Bayesian nonparametric model hLDA and.

In the knowledge based topic models, external knowledge are employed to obtain different models, although they are hard to acquire in the real word applications. Andrzejewski et al. [2] employs Dirichlet Forest prior over the topic-word multinomials to encode the Must-Links and Cannot-Links between words. Petterson et al. [29] used word information and a prior over the topic-word multinomials such that similar words share similar topic distributions. Newman et al. [26] proposed a quadratic regularizer and a convolved Dirichlet regularizer over topic-word multinomials to incorporate the correlation between words. Andrzejewski et al. [3] attempted to incorporate domain knowledge regarding documents, topics and side information into LDA. Chen et al. [12] models each topic as a probability distribution over domain knowledge. Jagarlamudi et al. [22] set a set of seed words in the beginning that users believe could represent certain topics.

Besides, some researchers tried new sentence based models to find better topics and applied them in new fields. Andrews and Vigliocco [1] proposed a hidden Markov topics model(HMTM) incorporating sequential and syntactic structures to model distributional structure. Hennig et al. [20] represented each sentence as a distribution over topics to identify sentential content with the same meaning. Tian et al. [31] took the one sentence one topic assumption where word generations in a sentence depend on both the topic of the sentence and the whole history of its preceding words in the sentence. Zhang et al. [36] proposed the sentence level topic model (SLTM), which contains a hidden layer, called the topic layer, between corpus and words to classify review sentences into different classes corresponding to different product features. Balikas et al. [5] incorporated the structure of the textual input in the generative and inference processes to encode much information hidden in coherent text spans such as sentences.These models take into account the inherent sequential nature of linguistic data. They focus on sentence generation and new application. Meanwhile, to our knowledge there are seldom work focusing on modeling topic relation.

Different from these models, we emphasis on the sentence orders to get their topic coherence and relatively ignore the word orders in each sentence. Our model does not introduce any external knowledge. We take into account sentences' influence to adjacent sentences in a document, we emphasis on the sentence orders to get their topic coherence and relatively ignore the word orders in each sentence. We make a small attempt to define the association relationship among topics that accords with topic transition and co-occurrence in or between sentence.

3 Notation and terminology

We use the language of text collections throughout the paper, referring to entities such as "words", "sentences", "documents" and "corpora" and we define the terms as follows. A word is the basic unit of discrete data, defined to be an item from a vocabulary denoted by w. A sentence is a sequence of N words denoted by $\mathbf{w}$= $(w_1, ...w_N)$, where w_n is the nth word in the sequence. A document is a sequence of L sentences denoted by $\mathbf{s}$= $(\mathbf{w}_1, ...\mathbf{w}_L)$, where $\mathbf{w}_l$ is the lth sentence in the document. A corpus is a collection of M documents denoted by $\mathbf{D}$= $\{\mathbf{s}_1, ..., \mathbf{s}_M\}$, where $\mathbf{s}_m$ is the mth document in the corpus. See Table 1 for detailed description.

4 Associated topic model (ATM) for text documents

The associated topic model(ATM) is a hierarchical generative model. In LDA as well as ATM each document is represented as a random mixture over latent topics and each topic is characterized by a random mixture over words. ATM differs from LDA in that ATM thinks of a document as a sequence of sentences, not a collection of words. ATM accounts for that consecutive sentences are more likely to have the same topics and meanwhile the order of words in the same topic is neglected. Accordingly, ATM neglects the orders of words but emphasizes the orders of sentences in a document.

We take the form of association measures, a square and symmetric matrix, to measure the association between topics [9]. Association measures are symmetric in the sense: the value of the association, referred to association coefficient, between the kth topic and the sth one is the same as the value of the association between the sth and the kth.

We use *previous sentence*, *current sentence* and *next sentence* to illustrate how a topic distribution for the current sentence is obtained. Rather than sampling topic identities for words in the current sentence from a probability distribution π_m for the mth document, these identities are generated by a specified topic distribution $\psi_{current}$ for the current sentence, which is determined commonly by $\boldsymbol{\phi}$ and the topic distributions $\psi_{previous}$, ψ_{next} for the previous and next sentences, except the first and the last sentences. Figure 1 and (1) show

Table 1 Definition of variables in the model

Variable	Meanings
K	number of topics
V	number of words in the vocabulary
M	number of documents in the corpus
N_l	number of words in the lth sentence $\mathbf{w}_l$
L_m	number of sentences in the mth document $\mathbf{s}_m$
T_m	number of words in the mth document $\mathbf{s}_m$, $T_m = \sum_{l=1}^{L} N_l$
w_{mln}	index of the nth word in the lth sentence of the mth document
z_{mln}	index of topic for the nth word in the lth sentence of the mth document
W	the chain of word index in a corpus in the sampling process
Z	the chain of topic index assignments in a corpus in the sampling process
α_k	prior weight of the kth topic in a document, α_k >0
$\boldsymbol{\alpha}$	a K-dimension vector of all α_k values
γ_k	prior weight of topic k associated with another topic, γ_k >0
$\boldsymbol{\gamma}$	a K-dimension vector, collection of all γ_k values
η_v	prior weight of the vth word in a topic, η_v >0
$\boldsymbol{\eta}$	a V-dimension vector, collection of all η_v values
π_{mk}	probability of the kth topic in the mth document for any word, $0 \leq \pi_{mk} \leq 1$
$\boldsymbol{\pi}_m$	a K-dimension probability distribution of topics in the mth document with $\sum_{k=1}^{K} \pi_{mk} = 1$
β_{kv}	probability of word w occurring in topic k, $0 \leq \beta_{kv} \leq 1$
$\boldsymbol{\beta}_k$	V-dimension probability distribution of words in the k topic with $\sum_{v=1}^{V} \beta_{kv} = 1$
ϕ_{ks}	probability of the kth topic associated with the sth topic, $0 \leq \phi_{ks} \leq 1$
$\boldsymbol{\phi}$	a $K \times K$ association matrix
$\psi_{current}$	K-dimension probability distribution of topics in the current sentence
$\psi_{previous}$	K-dimension probability distribution of topics in the previous sentence
ψ_{next}	K-dimension probability distribution of topics in the next sentence

how the topic distributions for sentences (except for the first sentence and the last sentence) are determined. The topic distribution for the first sentence in a document is determined by its next sentence and the topic distribution for the last sentence by its previous sentence. In (1) we think of the K-dimension probability distribution of topics ψ as a $1 \times K$ vector so that the matrix multiplication makes sense.

$$\psi_{current} = 0.5 * (\psi_{previous} + \psi_{next}) * \boldsymbol{\phi}. \tag{1}$$

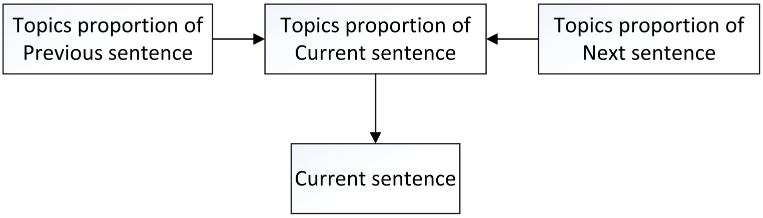
Topics proportion of Previous sentence
Topics proportion of Current sentence
Topics proportion of Next sentence
Current sentence

Figure 1 To describe how the topic distribution of previous sentence, the topic distribution of next sentence and the association matrix commonly determine the topic distribution of the current sentence

Now we turn to the description of generative process in this model. The key point is still topic allocations for all words. To generate a new word in the current sentence, one starts by firstly computing a multinomial distribution over topics corresponding to the current sentence $\psi_{current}$. After that, sample a hidden topic z for the word w from $\psi_{current}$ and then sample a word from the multinomial distribution over words with parameters β_k. Formally the generative process can be described as follows.

1. For each topic,
   - 1.1 Draw $\beta_k \sim \text{Dir}(\boldsymbol{\eta})$.
   - 1.2 Draw $\boldsymbol{\phi}_k \sim \text{Dir}(\boldsymbol{\gamma})$.
2. For each document,
   - 2.1 Draw $\pi_m \sim \text{Dir}(\boldsymbol{\alpha})$.
   - 2.2 For each sentence except the first and the last sentences
     - a Compute the topic distribution for this sentence $\psi_{current}$ by (1).
     - b For each word in this sentence assign a topic $z \sim Multinomial(\psi_{current})$.
     - c Choose a word $w \sim Multinomial(\boldsymbol{\beta}_z)$.

The associated topic model is represented as a probabilistic graphical model in Figure 2. In this model there are four levels of parameters related to a corpus, documents, sentences and words, separately. The corpus-level parameters $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are global variables and assumed to be sampled once in the process of generating a corpus. Specifically, βs are drawn from a common Dirichlet prior parameterized by $\boldsymbol{\eta}$ and $\boldsymbol{\phi}$ are drawn from a common Dirichlet prior parameterized by $\boldsymbol{\gamma}$. The document-level variables π_ms are local variables, sampled once per document. Finally, the word-level variables z_{dn} and w_{dn} are also local variables, sampled once for each word in each document. In order to make it more explicit, Figure 2 makes a graphical description of parameters and their connections.

The ATM makes several simplifying assumptions regarding parameters. The number of topics and thus the dimension of the Dirichlet distribution, often denoted by $Dir(\boldsymbol{\alpha})$, is

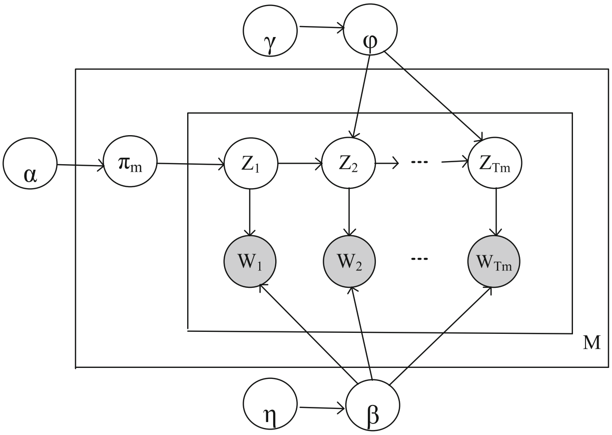

Figure 2 Graphical model representation of ATM. In our model the document-level variables π_ms are used in document representation only and will not participate in the process of generating topic assignments

fixed beforehand. The properties that the Dirichlet distribution is the conjugate prior of the multinomial distribution help facilitate efficient inference and estimation algorithms. Furthermore, when there is no prior knowledge favoring one component over another, the symmetric case might be useful, where all of the elements making up the parameter vector have the same value. The density function of symmetric Dirichlet distribution has the form

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha K)}{\Gamma^K(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha-1},$$

where the parameter $\boldsymbol{\alpha} = (\alpha, ..., \alpha)$ is a k-vector with its component α >0 and $\Gamma(.)$ is the Gamma function.

In this model, π_ms of all documents are drawn from a common Dirichlet prior distribution parameterized by $\boldsymbol{\alpha}$ and will be used to initialize topic assignments for all words of a document. Note that in our model the document-level variables π_ms are used to document representation only and will not participate in the process of generating topic assignments, so they do not appear on the graphical representation of ATM (Figure 2). After convergence of the Gibbs sampling, we will update them and each document is represented as mixing proportions for topics and thereby reduced to a multinomial distribution on a fixed set of topics. This distribution is the short description of the document and can be used in document modeling.

5 Inference and parameter estimation

The key inferential problem for this model is that of computing the posterior distribution of the hidden variables given a document

$$P(Z, \boldsymbol{\phi}, \boldsymbol{\beta}|W, \boldsymbol{\eta}, \boldsymbol{\gamma}) = \frac{P(Z, \boldsymbol{\phi}, \boldsymbol{\beta}, W|\boldsymbol{\eta}, \boldsymbol{\gamma})}{P(W|\boldsymbol{\eta}, \boldsymbol{\gamma})}. \tag{2}$$

and the likely function

$$P(W|\boldsymbol{\eta}, \boldsymbol{\gamma}) = \int P(\boldsymbol{\beta}|\boldsymbol{\eta})P(\boldsymbol{\phi}|\boldsymbol{\gamma}) \prod_{m,l,n} \sum_{z_{m,l,n}} P(z_{m,l,n}|\theta_{m,l})P(w_{m,l,n}|\beta_{z_{m,l,n}})d\beta d\boldsymbol{\phi}. \tag{3}$$

Unfortunately, the quality $P(W|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi})$ can not be computed tractably due to the coupling among $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ and we turn to approximate inference algorithms for the problem,such as variational approximation [6, 7] and Markov chain Monte Carlo (MCMC) [16, 18, 19].

In this section we employ MCMC to estimate the parameters because MCMC is an effective procedure for obtaining samples from complicated probability distributions, which allows a Markov chain to converge to the target distribution and then samples can be drawn from the Markov chain [15]. Each state of the chain is an assignment of values to the latent variables being sampled, and all variables will be sampled sequentially from their distribution, conditioned on the current values of all other variable and the data.

Gibbs Sampling is a member of a family of MCMC algorithms and this method is based on sampling from conditional distributions of the variables of the posterior. We choose Gibbs sampling aiming to construct a Markov chain that has the target posterior distribution as its stationary distribution. In other words, after a number of iterations of stepping through the chain, sampling from the distribution should converge to be close to sampling from the desired posterior.

For ATM, we need to estimate the latent document-topic portions π_m, the topic-word distributions $\boldsymbol{\beta}$ the association matrix $\boldsymbol{\phi}$ and the topic index assignments for each word Z. While conditional distributions can be derived from a chain of these latent variables, it should be noted that π_m, $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ can be calculated using just the chain of topic index assignments Z (i.e. Z is a sufficient statistic for both these distributions). That is, we use Gibbs sampling to sample only the assignments Z of words to topics. Therefore, a simpler algorithm, called a collapsed Gibbs sampler, can be used to compute the probability of a topic z being assigned to a word w, given all other topic assignments to all other words. What we will use in the sampling step is just the conditional posterior distribution $P(z_{m,l,n} = k|Z_{-l}, W)$.

By using Bayes' rule, for words in documents, the conditional posterior distribution for $z_{m,l,n}$ is given by

$$\begin{aligned}&P(z_{m,l,n} = k|Z_{-l}, W)\\ &\propto P(w_{m,l,n}|z_{m,l,n} = k, Z_{-l}, W_{-(m,l,n)})P(z_{m,l,n} = k|Z_{-l}, W_{-(m,l,n)}),\end{aligned} \tag{4}$$

where Z_{-l} is the assignments of words expect that in the lth sentence of the mth document. This is an application of Bayes' rule, where the first term on the right hand is a likelihood and the second a prior.

For the first term, we have

$$\begin{aligned}&P(w_{m,l,n}|z_{m,l,n} = k, Z_{-l}, W_{-(m,l,n)})\\ &= \int \beta^{(k)}_{w_{m,l,n}} P(\beta^{(k)}|Z_{-(m,l,n)}, W_{-(m,l,n)})d\beta^{(k)}\end{aligned} \tag{5}$$

Using the property of expectation of Dirichlet distribution, we have

$$P(w_{m,l,n}|z_{m,l,n} = k, Z_{-l}, W_{-(m,l,n)}) = \frac{t^{(w_{m,l,n})}_{-(m,l,n),k} + \gamma}{t^{(.)}_{-(m,l,n),k} + W\gamma}. \tag{6}$$

Here, $t^{(w_{m,l,n})}_{-(m,l,n),k}$ denotes the number of instances of word $w_{m,l,n}$ assigned to topic k and $t^{(.)}_{-(m,l,n),k}$ denotes the total number of words assigned to topic k, not including the current one.

Similarly, for the second term, we have

$$\begin{aligned}&P(z_{m,l,n} = k|Z_{-l}, W)\\ &\propto \sum_l \frac{1}{2}(z^l_{m,l-1} + z^l_{m,l-1}) \frac{n^{(l,k)}_{-(m,l)} + \alpha}{n^{(l)}_{-(m,l)} + K\alpha}.\end{aligned} \tag{7}$$

Here $\propto$ in this case means that the denominator is not a function of $z_{m,l,n}$ and thus is the same for all values of $z_{m,l,n}$; it forms part of the normalization constant for the distribution over $z_{m,l,n}$. Combining (4), (6) and (7), we get

$$\begin{aligned}&P(z_{m,l,n} = k|Z_{-l}, W) \propto\\ &\frac{t^{(w_{m,l,n})}_{-(m,l,n),k} + \gamma}{t^{(.)}_{-(m,l,n),k} + W\gamma} \sum_l \frac{1}{2}(z^l_{m,l-1} + z^l_{m,l-1}) \frac{t^{(l,k)}_{-(m,l)} + \alpha}{t^{(l)}_{-(m,l)} + K\alpha}.\end{aligned} \tag{8}$$

The Gibbs sampling algorithms generate an instance from (8) in turn, conditional on the current values of the other variables. It can be shown [14] that the sequence of samples constitutes a Markov chain, and the stationary distribution of that Markov chain is just the joint distribution. Additionally, the marginal distribution of any subset of variables can be approximated by simply considering the samples for that subset of variables, ignoring the rest.Therefore, on convergence of the Gibbs sampling, we will get samples from the joint posterior distribution $P(Z, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\beta}|W, \boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\gamma})$ and hence other variables of interest can be obtained from this posterior distribution.

6 Experiments

In this section, we empirically evaluate the effectiveness of our model, comparing it with three baseline methods on two datasets, which are wide benchmark in text analysis. The purpose of the experiments is to show the validity of the ATM and to demonstrate its better performance. We present the experimental results from three perspectives. First, we use perplexity curve to validate the convergence of the ATM on the NIPS dataset. Second, we demonstrate ATM's capability of obtaining more realistic topics on the NIPS dataset. Thirdly, we use document clustering on 20Newshome dataset and the Reuters-21578 dataset to measure the quality of the learned topical representations from different models.

6.1 Data sets

The experiments are conducted on the NIPS dataset , 20 Newsgroups dataset and the Reuters-21578 dataset. The NIPS dataset consists of 1740 documents, of which the train set consists of 1557 documents and the test set consists of the remaining 183. The vocabulary contains 12113 words. From the raw data we extracted the words that appear in the vocabulary and divided the text to sentences preserving their order. And we also discarded stop words from the input. The 20Newsgroups dataset is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. It has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering. The Reuters-21578 dataset contains 21578 documents which are grouped into 135 clusters. We use here the ModeApte version. Those documents with multiple category labels are discarded. It leaves us with 8293 documents in 65 categories. For ModeApte split, there are 5946 training documents and 2347 testing documents. Compared with TDT2 corpus, the Reuters corpus is more difficult for clustering.

6.2 Baselines

We compare our model with three baseline models: latent Dirichlet allocation(LDA) [7], the correlated topic model(CTM) [6] and the Markov topic models (MTMs) [13]. LDA is the most widely used topic model. CTM put first focus on relationship. As in MTMs, we employ Markov chain to describe topic transformation in a document. Besides, these models, like our model, did not introduce any external knowledge.

6.3 Experimental settings

Before explaining our experiments, we describe here the parameter specifications used to conduct our experiments. We run each Gibbs sampler for 500 iterations with 200 burn-in with the varying values of T, the number of topics. The value of 500 iterations is chosen to guarantee the convergence of posterior distribution so that topics are nearly drawn from the true distribution. We report the average perplexity of 5 randomly initialized runs on the NIPS dataset with $K = 50$ and for each run, 25% documents are held out for testing. The empirical prior parameters are set for all or part of models. We let $\eta = 0.02$, $\alpha = 50/K$ and $\gamma = 1$.

6.4 Evaluating topics

We compare our model with the baselines both qualitatively and quantitatively.

6.4.1 Perplexity

We follow the standard way in document modeling to evaluate the word predicative perplexity of our model and related models on the NIPS dataset. The perplexity of a testing collection of M documents is formally defined as:

$$perplexity(D_{test}) = \exp\{-\frac{\sum_{d=1}^{M} \log p(W_d)}{\sum_{d=1}^{M} N_d}\}.$$

Mathematics, the perplexity of a word distribution is the inverse of the geometric per-word average of the probability of the observations and it means that how the models predict the remaining words after observing a portion of the document. Therefore it reflects the

Table 2 Comparation of topics learned by ATM (top) and LDA (bottom) from NIPS dataset. Each column gives a topic represented by the top ten words from its distribution

#1	#2	#3	#4	#5	#6	#7	#8
network	case	probability	distance	input	parameters	data	figure
learning	hidden	set	linear	system	weight	output	networks
model	functions	patterns	vector	information	processing	problem	control
neural	algorithm	orientation	space	noise	neuron	performance	number
input	form	single	set	error	pattern	based	shown
data	analog	average	field	neurons	equation	features	values
figure	phase	cells	nonlinear	algorithm	classifier	task	order
function	classification	level	connections	spike	threshold	rules	visual
networks	time	object	images	point	result	paper	approximation
inputs	algorithms	local	regression	time	gaussian	gradient	estimation
units	strategy	noise	state	memory	data	feature	neurons
hidden	information	information	reinforcement	capacity	regression	features	neuron
unit	expected	code	action	associative	estimate	representation	connections
layer	approach	coding	policy	stored	method	level	neural
input	variables	channel	optimal	number	variance	structure	fig
weights	positive	input	states	storage	methods	representations	phase
output	cases	signal	actions	memories	clustering	figure	network
net	maker	spectral	control	high	estimation	similarity	activity
training	good	codes	function	recall	noise	information	activation
network	game	rates	time	fault	cluster	part	delay

Table 3 Association matrix for the eight topics shown in Table 2

0.11	0.08	0.06	0.06	0.1	0.07	0.08	0.03
0.08	0.12	0.05	0.07	0.08	0.09	0.08	0.03
0.06	0.05	0.09	0.04	0.05	0.07	0.08	0.07
0.06	0.07	0.04	0.1	0.04	0.06	0.07	0.05
0.1	0.08	0.05	0.04	0.13	0.08	0.08	0.04
0.07	0.09	0.07	0.06	0.08	0.09	0.08	0.05
0.08	0.08	0.08	0.07	0.08	0.08	0.07	0.07
0.03	0.03	0.07	0.05	0.04	0.05	0.07	0.06

difficulty of predicting a new unseen document and lower perplexity means more predictive capability.

We first demonstrate the performance of ATM on the NIPS dataset. Table 2 shows some associated topics learned by ATM, between which the associated coefficients are shown in Table 3 and makes a comparison with those by LDA. For example, *network, learning, model, neural, input, data, figure, function, networks, inputs* represent a topic related to *learning* and this is consistent with intuition. From this table we can see that more realistic semantic topics can be learned when sequential information is taken into account.

Perplexity as a function of the number of topics is depicted in Figure 3. The perplexities decrease rapidly at the beginning and then decrease gently to a stable value. This means that our algorithm can converge quickly to the local optimal value and the assigning process reaches a stable stage. Meanwhile, perplexities decrease rapidly with the number of topics varying from 10 to 50 and then decrease gently and the curves demonstrate ATM's better performance than LDA at a computational cost that is acceptable to us. These models are not only computationally efficient, but also seem to capture correlations between words via the topics.

6.4.2 Coherence measures

Topic models give no guarantees on well interpretable output, which extract topics from word counts in documents without requiring any semantic annotations. Therefore, coherence measures [10, 25, 27] were proposed and have been approved to distinguish between

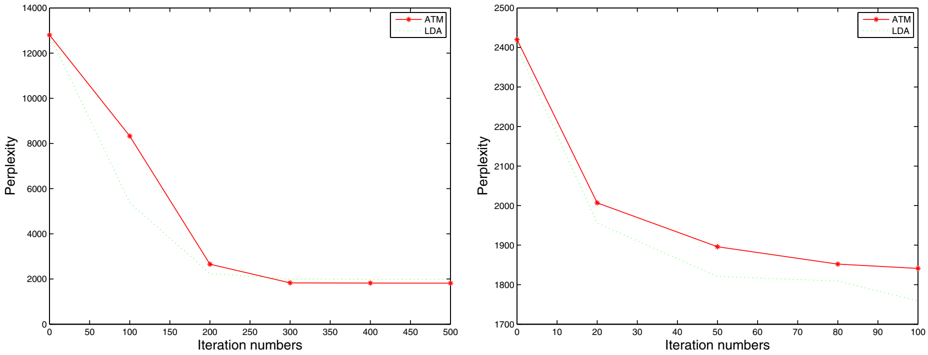

Figure 3 Perplexity vs iteration number (left) and the number of topics (right) on the NIPS dataset with $K = 50$

good and bad topics based on top words with respect to interpretability. Coherence measures compute a sum of scores over pairs of words from top words of a given topic:

$$coherence = \sum_{i<j} score(w_i, w_j).$$

The state-of-the-art measures in terms of topic coherence are the intrinsic measure UMass and the extrinsic measure UCI, both based on the same high-level idea.

The UCI-coherence measures the coherence of a topic based on pointwise mutual information(PMI) using large scale text data sets from external sources. Given the T most probable words of a topic k, $(w_1, \ldots, w_T)$, PMI-Score measures the pairwise association between them.

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j) + \frac{1}{M}}{P(w_i)P(w_j)}$$

where $P(w_i, w_j)$, $P(w_i)$ and $P(w_j)$ are the probabilities of co-occurring words pair (w_i, w_j) and w_i is estimated empirically from the external data sets, respectively. The smoothing count $\frac{1}{M}$ is added to avoid calculating the logarithm of zero. The UCI-coherence is calculated by:

$$UCI - coherence(topic) = \sum_{i=2}^{T} \sum_{j=1}^{i-1} PMI(w_i, w_j). \tag{9}$$

The coherence based on pointwise mutual information (PMI) gave large correlations with human ratings. The measure is extrinsic as it uses empirical probabilities from an external corpus such as Wikipedia. In our experiments, we extract topics and then compute UCI-coherence on the same dataset. Since these external data sets are model-independent, UCI-Score is fair for all the topic models.

Table 4 and Figure 4 show the average coherence measures of topic models with different fixed topic numbers on the NIPS dataset. It clearly shows the difference between the quality of the topics extracted by different models. For UCI-coherence our model always performs better than LDA, CTM and closely to MTMs. Besides, the three models get better results as the number of topics get larger.

6.4.3 Document clustering

To measure the quality of the learned topical representations from different models, we use k-means document clustering problem to see how accurate and discriminative the features obtained by different models are. k-means clustering aims to partition N observations into k

Table 4 Average topic coherence results on the NIPS dataset. The higher coherence score corresponds to a better topic quality

	100	200	300
ATM	**.4234**	**.4762**	**.4581**
LDA	.3852	.3975	.4234
CTM	.3729	.4134	.4177
MTMs	.4018	.4292	.4392

Values in bold in each column show that the algorithms corresponding to these numbers get the best performance

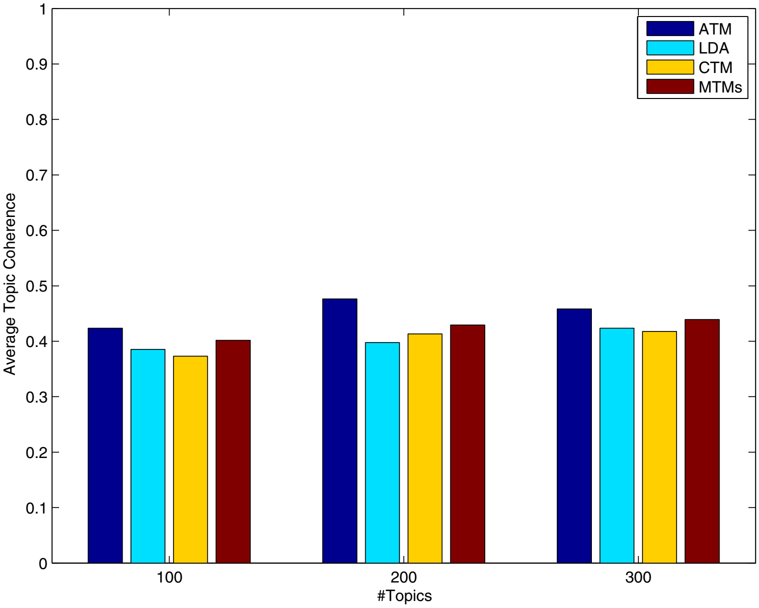

Figure 4 Average topic coherence results on the NIPS data set for ATM, LDA, CTM and MTMs

clusters in which each observation belongs to the cluster with the nearest mean. Document clustering is a well researched area that has several traditional methods available. In the text clustering problem, we wish to classify a document into two or more mutually exclusive classes. A challenging aspect of the document classification problem is the choice of features. Choosing features is essential in the document clustering problem. Treating all words as features yields a rich but very large feature set, which often causes great complexity. One way to reduce this feature set is to use topic models for dimension reduction and this is our focus in this section.

In these experiments, we split the data set into training and test subsets, and employed the k-means algorithm on the low-dimensional representations provided by LDA , ATM, CTM and MTMs respectively. It is of interest to see how much discriminatory information we leave in reducing the document description to topic-based features. The clustering result is evaluated by comparing the obtained label of each sample with that provided by the data set.

The accuracy(AC) and the normalized mutual information metric(NMI) are used to measure the clustering performance. Given a document d, let c_d and r_d be the obtained cluster label and the label by the corpus, respectively. The AC is defined as

$$AC = \frac{\sum_{d=1}^{D} \delta(c_d, r_d)}{D}$$

where D is the total number of documents and and $\delta(x, y)$ is the delta function that equals 1 if $x = y$ and equals 0 otherwise. Let X denote the set of clusters obtained from the ground truth and Y obtained from our models. $H(X)$, $p(x)$ and $p(x, y)$ denote the entropy of X, the probabilities that a document arbitrarily selected from the corpus belongs to the clusters x, and the joint probability that the arbitrarily selected document belongs to the clusters x

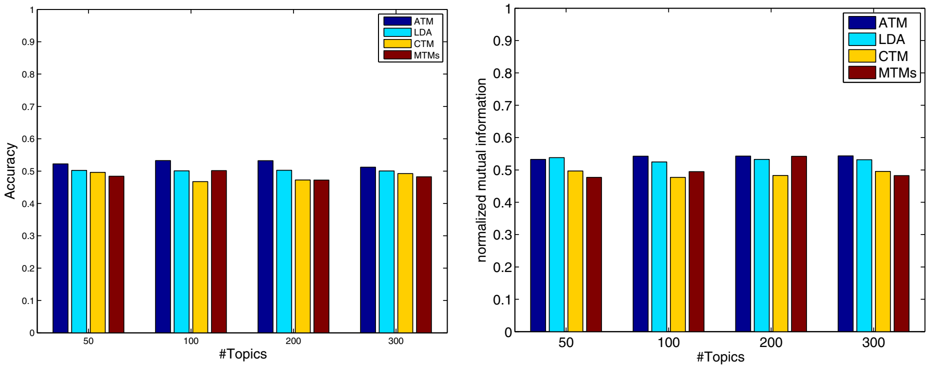

Figure 5 Accuracy (left) and normalized mutual information (right) results on the 20NewsHome data set for ATM, LDA, CTM and MTMs

as well as y at the same time, respectively. The mutual information metric $MI(X, Y)$ and the normalized mutual information metric $NMI(X, Y)$ are defined as follows:

$$MI(X, Y) = \sum_{x \in X, y \in Y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$$

$$NMI(X, Y) = \frac{MI(X, Y)}{\max\{H(X), H(Y)\}}.$$

The value of NMI ranges from 0 to 1 and a larger value means stronger independence.

Figure 5 and Table 5 show document clustering performance on the 20Newshome data set. The evaluations were conducted with the topics(features) numbers varying in {50, 100, 200, 300}. The performance confirms ATM's validity and effectiveness. The clustering results demonstrate that the topic-based representation provided by ATM can be thought as a filtering algorithm for feature selection in text analysis. Similar results appear on the Reuters-21578 dataset.

7 Discussion and future work

We have developed a hierarchical probabilistic model of documents that replaces the Dirichlet distribution of per-document topic proportions with an association matrix to generate topic assignments, which allows the model to capture associations between topics. We

Table 5 Clustering performance on 20NewsHome dataset. Each entry is the clustering accuracy(left) and NMI(right) of the column method on the corresponding row topic numbers

	ATM	LDA	CTM	MTMs
K=50	(.5224, .5327)	(.5324, .5425)	(.5322, .5429)	(.5120, .5435)
K=100	(.5021, .5380)	(.5010, .5247)	(.5026, .5324)	(.5004, .5316)
K=200	(.4961, .4967)	(.4676, .4770)	(.4727, .4828)	(.4924, .4953)
K=300	(.4840, .4771)	(.5014, .4947)	(.4723, .5421)	(.4824, .4826)

defined an associational relation between topics in terms of a joint distribution. The associated topic model gives better predictive performance and provides a rich way of exploring text collections.

It should be pointed out that in our model the number value of topics have to be set manually in advance. However, some topic models use nonparametric Bayesian methods based on the Dirichlet process to solve this problem that can accommodate new topics as more documents are observed.Seeking a suite of tools to tackle the model selection issue in ATM is an important area of future research. Another possible direction for future work is investigating whether there exist deep relations between topics, for instance, causal relations [28].

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (Grant No. 61672161, 61272480), and in part by the grants of Australian Research Council(DP170104747).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Andrews, M., Vigliocco, G.: The hidden Markov topic model: a probabilistic model of semantic representation. Top. Cogn. Sci. **2**(2), 101–113 (2010)
2. Andrzejewski, D., Zhu, X., Craven, M.: Incorporating domain knowledge into topic modeling via dirichlet forest priors. Intern. Conf. Machine Learn. **382**(26), 25–32 (2009)
3. Andrzejewski, D., Zhu, X., Craven, M., Recht, B.: A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic. In: International Joint Conference on Artificial Intelligence, pp. 1171–1177 (2011)
4. Bagheri, A.: Latent dirichlet Markov allocation. Thinklab University of Salford, Jong, F.D. (2013)
5. Balikas, G., Amini, M.R., Clausel, M.: On a topic model for sentences. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 921–924 (2016)
6. Blei, D., Lafferty, J.: Correlated topic models. Adv. Neural Inf. Proces. Syst. **18**, 147 (2006)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
8. Blei, D.M., Griffiths, T., L, Jordan, M.I.: The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. ACM (2010)
9. Borcard, D., Gillet, F., Legendre, P.: Association Measures and Matrices. Springer, New York (2011)
10. Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Eighth ACM International Conference on Web Search and Data Mining, pp. 399–408 (2015)
11. Buntine, W., Jakulin, A.: Applying discrete pca in data analysis. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, AUAI Press, pp. 59–66 (2004)
12. Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R.: Leveraging multi-domain prior knowledge in topic models. In: International Joint Conference on Artificial Intelligence, pp. 2071–2077 (2013)
13. Chong, W., Bo, T., Meek, C., Blei, D.M.: Markov topic models. In: International Conference on Artificial Intelligence and Statistics, pp. 583–590 (2009)
14. Gelman, A.: Bayesian data analysis. Biometrics **52**(3), 1160 (2000)
15. Gilks, W., Richardson, S., Spiegelhalter, D.: Markov chain monte carlo in practice, ser. Interdisciplinary statistics series (1996)
16. Griffiths, T.: Gibbs sampling in the generative model of latent dirichlet allocation. Standford University (2002)
17. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. **101**(suppl 1), 5228–5235 (2004)
18. Griffiths, T.L., Steyvers, M., Blei, D.M., Tenenbaum, J.B.: Integrating topics and syntax. In: Advances in Neural Information Processing Systems, pp. 537–544 (2004)
19. Gruber, A., Weiss, Y., Rosen-Zvi, M.: Hidden topic Markov models. In: Proceedings of Artificial Intelligence & Statistics, vol. 2007, pp. 163–170 (2007)
20. Hennig, L., Strecker, T., Narr, S., De Luca, E.W., Albayrak, S.: Identifying sentence-level semantic content units with topic models. In: Database and Expert Systems Applications, pp. 59–63 (2010)

21. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp. 50–57 (1999)
22. Jagarlamudi, J., Hal Daume, I., Udupa, R.: Incorporating lexical priors into topic models. In: Conference of the European Chapter of the Association for Computational Linguistics, pp. 204–213 (2012)
23. Lafferty, J.D.: A correlated topic model of science. Ann. Appl. Stat. **1**(1), 17–35 (2007)
24. Li, W., Mccallum, A.: Pachinko allocation: dag-structured mixture models of topic correlations. In: International Conference on Machine Learning, pp. 577–584 (2006)
25. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA, pp. 100-108 (2010)
26. Newman, D., Bonilla, E.V., Buntine, W.: Improving topic coherence with regularized topic models. In: International Conference on Neural Information Processing Systems, pp. 496–504 (2011)
27. O'Callaghan, D., Greene, D., Carthy, J.: An analysis of the coherence of descriptors in topic modeling. Expert Syst. Appl. **42**(13), 5645–5657 (2015)
28. Passos, A., Wallach, H.M., Mccallum, A.: Correlations and anticorrelations in lda inference. University of Massachusetts - Amherst 37(5):548–555 (2011)
29. Petterson, J., Buntine, W.L., Narayanamurthy, S.M., Caetano, T.S., Smola, A.J.: Word features for latent dirichlet allocation. In: Neural Information Processing Systems, vol. 2010, pp. 1921–1929 (2010)
30. Suh, S., Choi, S.: Two-dimensional correlated topic models. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2559–2563 (2016)
31. Tian, F., Gao, B., He, D., Liu, T.: Sentence level recurrent topic model: letting topics speak for themselves. arXiv: learning (2016)
32. Wallach, H.M.: Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd International Conference on Machine Learning, ACM, pp. 977–984 (2006)
33. Wang, C., Fan, J., Kalyanpur, A., Gondek, D.: Relation extraction with relation topics. In: Conference on Empirical Methods in Natural Language Processing, pp. 1426–1436 (2011)
34. Wang, X., McCallum, A.: A Note on Topical N-Grams. Tech. rep., DTIC Document (2005)
35. Xie, P., Yang, D., Xing, E.P.: Incorporating word correlation knowledge into topic modeling. In: North american chapter of the association for computational linguistics, pp. 725–734 (2015)
36. Zhang, Y., Xu, H.: Sltm: A sentence level topic model for analysis of online reviews. pp. 449–453 (2016)

Affiliations

Haixin Jiang[1] · Rui Zhou[2] · Limeng Zhang[1] · Hua Wang[3] · Yanchun Zhang[3,4]

Rui Zhou
rzhou@swin.edu.au

Limeng Zhang
zhanglimeng13@mails.ucas.ac.cn

Hua Wang
hua.wang@vu.edu.au

Yanchun Zhang
Yanchun.Zhang@vu.edu.au

[1] University of Chinese Academy of Sciences, Beijing, China

[2] Swinburne University of Technology, Hawthorn, Australia

[3] Victoria University, Footscray, Australia

[4] Fudan University, Shanghai, China