Sentence Level Recurrent Topic Model: Letting Topics Speak for Themselves

Abstract

We propose Sentence Level Recurrent Topic Model (SLRTM), a new topic model that assumes the generation of each word within a sentence to depend on both the topic of the sentence and the whole history of its preceding words in the sentence. Different from conventional topic models that largely ignore the sequential order of words or their topic coherence, SLRTM gives full characterization to them by using a Recurrent Neural Networks (RNN) based framework. Experimental results have shown that SLRTM outperforms several strong baselines on various tasks. Furthermore, SLRTM can automatically generate sentences given a topic (i.e., topics to sentences), which is a key technology for real world applications such as personalized short text conversation.

Introduction

Statistic topic models such as Latent Dirichlet Allocation (LDA) and its variants BIBREF0 , BIBREF1 , BIBREF2 , BIBREF3 , BIBREF4 have been proven to be effective in modeling textual documents. In these models, a word token in a document is assumed to be generated by a hidden mixture model, where the hidden variables are the topic indexes for each word and the topic assignments for words are related to document level topic weights. Due to the effectiveness and efficiency in modeling the document generation process, topic models are widely adopted in quite a lot of real world tasks such as sentiment classification BIBREF5 , social network analysis BIBREF6 , BIBREF5 , and recommendation systems BIBREF7 .

Most topic models take the bag-of-words assumption, in which every document is treated as an unordered set of words and the word tokens in such a document are sampled independently with each other. The bag-of-words assumption brings computational convenience, however, it sacrifices the characterization of sequential properties of words in a document and the topic coherence between words belonging to the same language segment (e.g., sentence). As a result, people have observed many negative examples. Just list one for illustration BIBREF8 : the department chair couches offers and the chair department offers couches have very different topics, although they have exactly the same bag of words.

There have been some works trying to solve the aforementioned problems, although still insufficiently. For example, several sentence level topic models BIBREF9 , BIBREF10 , BIBREF11 tackle the topic coherence problem by assuming all the words in a sentence to share the same topic (i.e., every sentence has only one topic). In addition, they model the sequential information by assuming the transition between sentence topics to be Markovian. However, words within the same sentence are still exchangeable in these models, and thus the bag-of-words assumption still holds within a sentence. For another example, in BIBREF12 , the embedding based neural language model BIBREF13 , BIBREF14 , BIBREF15 and topic model are integrated. They assume the generation of a given word in a sentence to depend on its local context (including its preceding words within a fixed window) as well as the topics of the sentence and document it lies in. However, using a fixed window of preceding words, instead of the whole word stream within a sentence, could only introduce limited sequential dependency. Furthermore, there is no explicit coherence constraints on the word topics and sentence topics, since every word can have its own topics in their model.

We propose Sentence Level Recurrent Topic Model (SLRTM) to tackle the limitations of the aforementioned works. In the new model, we assume the words in the same sentence to share the same topic in order to guarantee topic coherence, and we assume the generation of a word to rely on the whole history in the same sentence in order to fully characterize the sequential dependency. Specifically, for a particular word INLINEFORM0 within a sentence INLINEFORM1 , we assume its generation depends on two factors: the first is the whole set of its historical words in the sentence and the second is the sentence topic, which we regard as a pseudo word and has its own distributed representations. We use Recurrent Neural Network (RNN) BIBREF16 , such as Long Short Term Memory (LSTM) BIBREF17 or Gated Recurrent Unit (GRU) network BIBREF18 , to model such a long term dependency.

With the proposed SLRTM, we can not only model the document generation process more accurately, but also construct new natural sentences that are coherent with a given topic (we call it topic2sentence, similar to image2sentece BIBREF19). Topic2sentence has its huge potential for many real world tasks. For example, it can serve as the basis of personalized short text conversation system BIBREF20 , BIBREF21 , in which once we detect that the user is interested in certain topics, we can let these topics speak for themselves using SLRTM to improve the user satisfactory.

We have conducted experiments to compare SLRTM with several strong topic model baselines on two tasks: generative model evaluation (i.e. test set perplexity) and document classification. The results on several benchmark datasets quantitatively demonstrate SLRTM's advantages in modeling documents. We further provide some qualitative results on topic2sentence, the generated sentences for different topics clearly demonstrate the power of SLRTM in topic-sensitive short text conversations.

Related Work

One of the most representative topic models is Latent Dirichlet Allocation BIBREF2 , in which every word in a document has its topic drawn from document level topic weights. Several variants of LDA have been developed such as hierarchical topic models BIBREF22 and supervised topic models BIBREF3 . With the recent development of deep learning, there are also neural network based topic models such as BIBREF23 , BIBREF24 , BIBREF25 , BIBREF26 , which use distributed representations of words to improve topic semantics.

Most of the aforementioned works take the bag-of-words assumption, which might be too simple according to our discussions in the introduction. That is, it ignores both sequential dependency of words and topic coherence of words.

There are some efforts trying to address the limitations of the bag-of-words assumption. For example, in BIBREF27 , both semantic (i.e., related with topics) and syntactic properties of words were modeled. After that, a hidden Markov transition model for topics was proposed BIBREF9 , in which all the words in a sentence were regarded as having the same topic. Such a one sentence, one topic assumption was also used by some other works, including BIBREF10 , BIBREF11 . Although these works have made some meaningful attempts on topic coherence and sequential dependency across sentences, they have not sufficiently model the sequential dependency of words within a sentence. To address this problem, the authors of BIBREF12 adopted the neural language model technology BIBREF13 to enhance topic model. In particular, they assume that every document, sentence, and word have their own topics and the topical information is conveyed by their embedding vectors through a Gaussian Mixture Model (GMM) as a prior. In the GMM distribution, each topic corresponds to a mixture parameterized by the mean vector and covariance matrix of the Gaussian distribution. The embedding vectors sampled from the GMM are further used to generate words in a sentence according to a feedforward neural network. To be specific, the preceding words in a fixed sized window, together with the sentence and document, act as the context to generate the next word by a softmax conditional distribution, in which the context is represented by embedding vectors. While this work has explicitly modeled the sequential dependency of words, it ignores the topic coherence among adjacent words.

Another line of research related to our model is Recurrent Neural Network (RNN), especially some recently developed effective RNN models such as Long Short Term Memory BIBREF17 and Gated Recurrent Unit BIBREF18 . These new RNN models characterize long range dependencies for a sequence, and has been widely adopted in sequence modeling tasks such as machine translation

BIBREF18 and short text conversation BIBREF20 . In particular, for language modeling tasks, it has been shown that RNN (and its variants such as LSTM) is much more effective than simple feedforward neural networks with fixed window size BIBREF16 given that it can model dependencies with nearly arbitrary length.

Sentence Level Recurrent Topic Model

In this section, we describe the proposed Sentence Level Recurrent Topic Model (SLRTM). First of all, we list three important design factors in SLRTM as below.

With the three points in mind, let us introduce the detailed generative process of SLRTM, as well as the stochastic variational inference and learning algorithm for SLRTM in the following subsections.

The generative process

Suppose we have INLINEFORM0 topics, INLINEFORM1 words contained in dictionary INLINEFORM2 , and INLINEFORM3 documents INLINEFORM4 . For any document INLINEFORM5 , it is composed of INLINEFORM6 sentences and its INLINEFORM7 th sentence INLINEFORM8 consists of INLINEFORM9 words. Similar to LDA, we assume there is a INLINEFORM10 -dimensional Dirichlet prior distribution INLINEFORM11 for topic mixture weights of each document. With these notations, the generative process for document INLINEFORM12 can be written as below:

Sample the multinomial parameter INLINEFORM0 from INLINEFORM1 ;

For the INLINEFORM0 th sentence of document INLINEFORM1 INLINEFORM2 , INLINEFORM3 , where INLINEFORM4 is the INLINEFORM5 th word for INLINEFORM6 :

Draw the topic index INLINEFORM0 of this sentence from INLINEFORM1 ;

For INLINEFORM0 :

Compute LSTM hidden state INLINEFORM0 ;

INLINEFORM0 , draw INLINEFORM1 from DISPLAYFORM0

Here we use bold characters to denote the distributed representations for the corresponding items. For example, INLINEFORM0 and INLINEFORM1 denote the embeddings for word INLINEFORM2 and topic INLINEFORM3 , respectively. INLINEFORM4 is a zero vector and INLINEFORM5 is a fake starting word. Function INLINEFORM6 is the LSTM unit to generate hidden states, for which we omit the details due to space restrictions. Function INLINEFORM7 typically takes the following form: DISPLAYFORM0

where INLINEFORM0 , INLINEFORM1 denotes the output embedding for word INLINEFORM2 . INLINEFORM3 are feedforward weight matrices and INLINEFORM4 is the bias vector.

Then the probability of observing document INLINEFORM0 can be written as: DISPLAYFORM0

where INLINEFORM0 is the probability of generating sentence INLINEFORM1 under topic INLINEFORM2 , and it is decomposed through the probability chain rule; INLINEFORM3 is specified in equation (EQREF11) and (EQREF12); INLINEFORM4 represents all the model parameters, including the distributed representations for all the words and topics, as well as the weight parameters for LSTM.

To sum up, we use Figure FIGREF14 to illustrate the generative process of SLRTM, from which we can see that in SLRTM, the historical words and topic of the sentence jointly affect the LSTM hidden state and the next word.

Stochastic Variational Inference and Learning

As the computation of the true posterior of hidden variables in equation (EQREF13) is untractable, we adopt mean field variational inference to approximate it. Particularly, we use multinomial distribution INLINEFORM0 and Dirichlet distribution INLINEFORM1 as the variational distribution for the hidden variables INLINEFORM2 and INLINEFORM3 , and we denote the variational parameters for document INLINEFORM4 as INLINEFORM5 , with the subscript INLINEFORM6 omitted. Then the variational lower bound of the data likelihood BIBREF2 can be written as: DISPLAYFORM0

where INLINEFORM0 is the true distribution for corresponding variables.

The introduction of LSTM-RNN makes the optimization of (EQREF16) computationally expensive, since we need to update both the model parameters INLINEFORM0 and variational parameters INLINEFORM1 after scanning the whole corpus. Considering that mini-batch (containing several sentences) inference and training are necessary to optimize the neural network, we leverage the stochastic variational inference algorithm developed in BIBREF4 , BIBREF28 to conduct inference and learning in a variational Expectation-Maximization framework. The detailed algorithm is given in Algorithm SECREF15 . The execution of the whole inference and learning process includes several epochs of iteration over all documents INLINEFORM2 with Algorithm SECREF15 (starting with INLINEFORM3).

[ht] Stochastic Variational EM for SLRTM Input: document INLINEFORM0 , variation parameters INLINEFORM1 , and model weights INLINEFORM2 . every sentence minibatch INLINEFORM3 in

INLINEFORM4 INLINEFORM5 E-Step: INLINEFORM6 INLINEFORM7 , i.e., every topic index: Obtain INLINEFORM8 by LSTM forward pass. INLINEFORM9 DISPLAYFORM0

convergence Collect variational parameters INLINEFORM0 . M-Step: Compute the gradient INLINEFORM1 by LSTM backward pass. Use INLINEFORM2 to obtain INLINEFORM3 by stochastic gradient descent methods such as Adagrad BIBREF30 . In Algorithm SECREF15 , INLINEFORM4 is the digamma function. Equation (EQREF18) guarantees the estimate of INLINEFORM5 is unbiased. In equation (), INLINEFORM6 is set as INLINEFORM7 , where INLINEFORM8 , to make sure INLINEFORM9 will converge BIBREF4 . Due to space limit, we omit the derivation details for the updating equations in Algorithm SECREF15 , as well as the forward/backward pass details for LSTM BIBREF17 .

Experiments

We report our experimental results in this section. Our experiments include two parts: (1) quantitative experiments, including a generative document evaluation task and a document classification task, on two datasets; (2) qualitative inspection, including the examination of the sentences generated under each topic, in order to test whether SLRTM performs well in the topic2sentence task.

Quantitative Results

We compare SLRTM with several state-of-the-art topic models on two tasks: generative document evaluation and document classification. The former task is to investigate the generation capability of the models, while the latter is to show the representation ability of the models.

We base our experiments on two benchmark datasets:

20Newsgroup, which contains 18,845 emails categorized into 20 different topical groups such as religion, politics, and sports. The dataset is originally partitioned into 11,314 training documents and 7,531 test documents.

Wiki10+ BIBREF31 , which contains Web documents from Wikipedia, each of which is associated with several tags such as philosophy, software, and music. Following BIBREF25 , we kept the most frequent 25 tags and removed those documents without any of these tags, forming a training set and a test set with 11,164 and 6,161 documents, respectively. The social tags associated with each document are regarded as supervised labels in classification. Wiki10+ contains much more words per document (i.e., 1,704) than 20Newsgroup (i.e., 135).

We followed the practice in many previous works and removed infrequent words. After that, the dictionary contains about INLINEFORM0 unique words for 20Newsgroup and INLINEFORM1 for Wiki10+. We adopted the NLTK sentence tokenizer to split the datasets into sentences if sentence boundaries are needed.

The following baselines were used in our experiments:

LDA BIBREF2 . LDA is the classic topic model, and we used GibbsLDA++ for its implementation.

Doc-NADE BIBREF24 . Doc-NADE is a representative neural network based topic model. We used the open-source code provided by the authors.

HTMM BIBREF9 . HTMM models consider the sentence level Markov transitions. Similar to Doc-NADE, the implementation was provided by the authors.

GMNTM BIBREF12 . GMNTM considers models the order of words within a sentence by a feedforward neural network. We implemented GMNTM according the descriptions in their papers by our own.

For SLRTM, we implemented it in C++ using Eigen and Intel MKL. For the sake of fairness, similar to BIBREF12 , we set the word embedding size, topic embedding size, and LSTM hidden layer size to be 128, 128, and 600 respectively. In the experiment, we tested the performances of SLRTM and the baselines with respect to different number of topics INLINEFORM0 , i.e., INLINEFORM1 . In initialization (values of INLINEFORM2 and INLINEFORM3), the LSTM weight matrices were initialized as orthogonal matrices, the word/topic embeddings were randomly sampled from the uniform distribution INLINEFORM4 and are fined-tuned through the training process, INLINEFORM5 and INLINEFORM6 were both set to INLINEFORM7 . The mini-batch size in Algorithm SECREF15 was set as INLINEFORM8 , and we ran the E-Step of the algorithm for only one iteration for efficiently consideration, which leads to the final convergence after about 6 epochs for both datasets. Gradient clipping with a clip value of 20 was used during the optimization of LSTM weights. Asynchronous stochastic gradient descent BIBREF32 with Adagrad was used to perform multi-thread parallel training.

We measure the performances of different topic models according to the perplexity per word on the test set, defined as INLINEFORM0 , where INLINEFORM1 is the number of words in document INLINEFORM2 . The experimental results are summarized in Table TABREF33 . Based on the table, we have the following discussions:

Our proposed SLRTM consistently outperforms the baseline models by significant margins, showing its outstanding ability in modelling the generative process of documents. In fact, as tested in our further verifications, the perplexity of SLRTM is close to that of standard LSTM language model, with a small gap of about 100 (higher perplexity) on both datasets which we conjecture is due to the margin between the lower bound in equation (EQREF16) and true data likelihood for SLRTM.

Models that consider sequential property within sentences (i.e., GMNTM and SLRTM) are generally better than other models, which verifies the importance of words' sequential information. Furthermore, LSTM-RNN is much better in modelling such a sequential dependency than standard feed-forward networks with fixed words window as input, as verified by the lower perplexity of SLRTM compared with GMNTM.

In this experiment, we fed the document vectors (e.g., the INLINEFORM0 values in SLRTM) learnt by different topic models to supervised classifiers, to compare their representation power. For 20Newsgroup, we used the multi-class logistic regression classifier and used accuracy as the evaluation criterion. For Wiki10+, since multiple labels (tags) might be associated with each document, we used logistic regression for each label and the classification result is measured by Micro- INLINEFORM1 score BIBREF33 . For both datasets, we use INLINEFORM2 of the original training set for validation, and the remaining for training.

All the classification results are shown in Table TABREF37 . From the table, we can see that SLRTM is the best model under each setting on both datasets. We can further find that the embedding based methods (Doc-NADE, GMNTM and SLRTM) generate better document representations than other models, demonstrating the representative power of neural networks based on distributed representations. In addition, when the training data is larger (i.e., with more sentences per document as Wiki10+), GMNTM generates worse topical information than Doc-NADE while our SLRTM outperforms Doc-NADE, showing that with sufficient data, SLRTM is more effective in topic modeling since topic coherence is further constrained for each sentence.

Qualitative Results

In this subsection, we demonstrate the capability of SLRTM in generating reasonable and understandable sentences given particular topics. In the experiment, we trained a larger SLRTM with 128 topics on a randomly sampled INLINEFORM0 Wikipedia documents in the year of 2010 with average 275 words per document. The dictionary is composed of roughly INLINEFORM1 most frequent words including common punctuation marks, with uppercase letters transformed into lowercases. The size of word embedding, topic embedding and RNN hidden layer are set to 512, 1024 and 1024, respectively.

We used two different mechanisms in sentence generating. The first mechanism is random sampling new word INLINEFORM0 at every time step INLINEFORM1 from the probability distribution defined in equation (EQREF13). The second is dynamic programming based beam search BIBREF19 , which seeks to generate sentences by globally maximized likelihood. We set the beam size as 30. The generating process terminates until a predefined maximum sentence length is reached (set as 25) or an EOS token is met. Such an EOS is also appended after every training sentence.

The generating results are shown in Table TABREF40 . In the table, the sentences generated by random sampling and beam search are shown in the second and the third columns respectively. In the fourth column, we show the most representative words for each topics generated by SLRTM. For this purpose, we constrained the maximum sentence length to 1 in beam search, and removed stop words that are frequently used to start a sentence such as the, he, and there.

From the table we have the following observations:

Most of the sentences generated by both mechanisms are natural and semantically correlated with particular topics that are summarized in the first column of the table.

The random sampling mechanism usually produces diverse sentences, whereas some grammar errors may happen (e.g., the last sampled sentence for Topic 4; re-ranking the randomly sampled words by a standalone language model might further improve the correctness of the sentence). In contrast, sentences outputted by beam search are safer in matching grammar rules, but are not diverse enough. This is consistent with the observations in BIBREF21 .

In addition to topic2sentece, SLRTM maintains the capability of generating words for topics (shown in the last column of the table), similar to conventional topic models.

Conclusion

In this paper, we proposed a novel topic model called Sentence Level Recurrent Topic Model (SLRTM), which models the sequential dependency of words and topic coherence within a sentence using Recurrent Neural Networks, and shows superior performance in both predictive document modeling and document classification. In addition, it makes topic2sentence possible, which can benefit many real world tasks such as personalized short text conversation (STC).

In the future, we plan to integrate SLRTM into RNN-based STC systems BIBREF20 to make the dialogue more topic sensitive. We would also like to conduct large scale SLRTM training on bigger corpus with more topics by specially designed scalable algorithms and computational platforms.