Journal of Computers Vol. 30 No. 6, 2019, pp. 60-72
doi:10.3966/199115992019123006005

A Semantic Enhanced Topic Model Based on Bi-directional LSTM Networks

Wang Gao[1,2*], Zhi-feng Yang[1], Hai Wang[1], Fan Zhang[3], Yuan Fang[4]

[1] College of Sports Science and Technology, Wuhan Sports University, Wuhan 430079, China
{gaowang2000, yangzhif, wanghai}@foxmail.com

[2] Computer School, Wuhan University, Wuhan 430072, China

[3] College of Computer Science, Wuhan Donghu University, Wuhan 430212, China
whzhangfan@126.com

[4] School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China
fangyuan@foxmail.com

Received 3 January 2018; Revised 14 May 2018; Accepted 4 July 2018

Abstract. Topic modeling techniques are widely used for text modeling and analysis. However, they suffer from the sparseness problem and the complex inference process, which can be alleviated by deep learning techniques such as bi-directional long short-term memory (LSTM) networks. To explore the combination of topic modeling and bi-directional LSTM, we propose a new probabilistic topic model, named GPU-LDA-LSTM. Differently from existing approaches, we first design a document semantic coding framework based on bi-directional LSTM (DSC-LSTM) to learn the representation of documents. Then, we utilize the document-topic and word-word dual-generalized Polya urn (GPU) mechanism to enhance semantics. Furthermore, a LSTM network is also used to improve the contextual consistency in the parameter inference process. Experimental results on two real-world datasets show that our model significantly outperforms state-of-the-art models on several evaluation metrics, suggesting that it can extract more meaningful topics.

Keywords: bi-directional LSTM, semantic enhancement, topic model, word embedding

1 Introduction

Probabilistic topic models, such as latent Dirichlet allocation (LDA) [4], have been proven to be useful for documents modeling and analysis. Topic modeling regards documents as a mixture of latent topics, where a topic is represented by a probability distribution over words. Given a huge volume of documents available, effective and efficient models to extract the coherent topics from documents become fundamental to several applications that require semantic understanding of textual content, such as text categorization [6], emerging event detection [7], topic evolution analysis [1] and recommendation systems [2].

Although topic models have shown great success in automatic topic extraction, there are still some limitations in the modeling process. A key limitation is the scalability of traditional topic models. They are often functionally enhanced by adding extra random variables to the models, which in turn leads to higher time complexity. More importantly, conventional topic models lack a mechanism for combining related semantic reinforcement, which results in the poor semantic coherence of generated topics.

Several heuristic strategies have been adopted to solve the above limitations. Cao et al. proposed a neural topic model (NTM) with supervised extension to reduce the computation complexity of topic models [8]. However, NTM neglects the long-distance dependencies between words in the process of

[*] Corresponding Author

generating document representations. It therefore cannot capture comparatively more distant patterns in a document. Based on the Dirichlet multinomial mixture (DMM) model, Li et al. proposed a topic model which promotes the semantically related words under the same topic during the sampling process by using the GPU model [22]. This model measures the semantic relatedness between two words by their word embeddings, which has not considered the semantic relations between words and documents.

In this paper, we propose a novel topic model to address the above challenges. The main idea comes from the answers of the following two questions: (1) How to find an effective solution for learning document-level embeddings? (2) How to improve topic modeling by both document-topic and word-word semantic reinforcement?

Specifically, we design a new topic model, named GPU-LDA-LSTM. The proposed model leverages both document-topic and word-word semantic reinforcement to improve topic modeling. We first propose a document semantic encoding framework based on bi-directional LSTM (DSC-LSTM) to learning high-quality embeddings for documents. In the second phase, GPU-LDA-LSTM integrates GPU [3] into LDA [4] to discover more coherent topics by the semantic reinforcement of document-topic and word-word respectively. Furthermore, a new sampling algorithm is designed to inference the parameters of the proposed model.

The main contributions of this paper are summarized as follows:

- This paper presents a document-level semantic coding framework DSC-LSTM. Unlike previous methods, DSC-LSTM can effectively capture the document semantic information by using a bi-directional LSTM network, and provide a new deep learning method for generating the semantic representation of documents.
- In this paper, we propose a novel topic model, named GPU-LDA-LSTM. GPU-LDA-LSTM integrates the dual-GPU model into LDA to enhance both document-topic and word-word semantics. To the best of our knowledge, this is the first work for a topic model to incorporate the document-topic and word-word semantic reinforcement with the dual-GPU model.
- A LSTM network is introduced into the Gibbs sampling process of the GPU-LDA-LSTM model, and thus the context consistency in the parameter inference process can be guaranteed.
- The performance of our model is evaluated on two real-world datasets against a few state-of-the-art methods. Experimental results demonstrate our model outperforms the baseline models on several evaluation metrics.

2 Related Work

2.1 Recurrent Neural Network

Recurrent neural network (RNN) is a deep neural network that has been proven powerful in document modeling tasks. In RNN, the output is not only related to the input, but also related to the output of the previous moment. However, RNN cannot effectively solve the problem of long-range dependencies. To solve this problem, Hochreiter et al. proposed a new RNN network, long and short-term memory (LSTM) network [11]. LSTM addresses the problem with an extra memory “cell” that is constructed as a linear combination of the previous state and the input signal. Graves et al. employed the bi-directional LSTM to simultaneously capture forward and backward semantic information [12]. The bi-directional LSTM network has two LSTM hidden layers. Each pair of forward and backward layers are connected to the same output unit, which can provide more context information for each moment in the output layer. Qian et al. discovered that modeling the linguistic role can enhance sentence-level sentiment classification by using LSTM models [23]. Song et al. proposed a new abstractive text summarization method based on a LSTM-CNN neural network [24]. Nevertheless, there has been limited research on utilizing LSTM to improve topic modeling.

2.2 Topic Modeling Based on Semantic Enhancement

Topic modeling based on semantic enhancement aims to integrate domain knowledge into topic modeling, which can significantly improve topic coherence. One method is to impose constraints in the process of word generation. For instance, Andrzejewski et al. encoded the Must-Links and Cannot-Links between words over the topic-word multinomials. Words with Must-Links are encouraged to have high probabilities to share the same topic label, while those with Cannot-Links are disallowed to be in the same topic [13]. Hu et al. proposed a topic model based on an entity taxonomy from a knowledge base [14]. Each topic is generated with a random walk over the entity hierarchy to extract semantically meaningful topic. However, in some complex scenarios, domain knowledge itself needs to be continuously updated during the modeling process, which is very time-consuming. In contrast, the current work incorporates the document-topic and word-word semantic reinforcement based on pre-trained word embeddings that are learned only once in the whole process. Lifelong Topic Model (LTM) is a multi-domain life-long learning topic model that can automatically retrieve knowledge patterns from historical data to improve topic modeling by mining frequent itemsets [15]. Chen et al. introduced the thinking way of "learning like a person" based on LTM [16]. Although the algorithm of frequent itemsets mining is also used to acquire knowledge, it emphasizes the integration and adjustment of knowledge. As a result, it not only detects false domain knowledge, but also can be applied in large data scenarios. Unlike these approaches, this paper combines deep learning techniques with topic modeling to enhance semantics.

2.3 Topic Modeling Based on Neural Network

Topic modeling based on neural networks is designed to combine directed probability graphs and neural networks. This combination can on the one hand avoid the complex inference process and on the other hand extract the high-quality representation of documents. Traditional neural network topic models are often based on the restricted Boltzmann machine [17]. However, the training process of the restricted Boltzmann machine is complicated, and it is not suitable for the text serialization modeling. Cao et al. proposed a topic model NTM based on a feedforward neural network [8]. The generation processes of document-topic and topic-word distributions are represented by two hidden layers respectively. The document-word generation probability is calculated by a dot product computation. Tian et al. proposed a topic model based on RNN to generate thematic sentences [18]. In this model, the generation of each word is related not only to the thematic sentence, but also to all its previous words. With the advent of the encoder-decoder framework, topic modeling based on the end-to-end approach and attention mechanism has emerged in recent years. Xing et al. incorporated Twitter-LDA into an encoder-decoder framework, combining the text attention and topic attention to generate conversations [19]. Li et al. proposed a recurrent attentional topic model to integrate the attentional mechanism into the generation of text sequences [20]. Our model differs significantly from these studies. GPU-LDA-LSTM exploits the semantic reinforcement of document-topic and word-word by using the dual-GPU model. Furthermore, a new sampling algorithm based on LSTM is introduced to inference the parameters of the proposed model. To the best of our knowledge, GPU-LDA-LSTM is the first attempt to combine LSTM and the dual-GPU model for topic modeling based on semantic enhancement.

3 GPU-LDA-LSTM Topic Modeling

3.1 Document Semantic Coding

Although word embeddings are useful to represent lexical semantic features of words, they cannot directly provide high-granularity semantic information at sentence or document level [18]. Probabilistic topic models are built on document collections. Therefore, we should generate more reasonable document representation to achieve better semantic enhancement.

Traditional methods of semantic information extraction usually need to extract the entity relationship of documents, and then semantic templates or constraints are constructed to mine the semantic information. However, this paper builds a topic model based on a bi-directional LSTM neural network, and all words are represented as word embeddings. Accordingly, documents need to be represented as semantic codes that can be integrated into the model.

In recent years, based on deep neural networks, the methods of learning document-level representations usually rely on the features of entities and relationships. These methods employ the extension of word embeddings or the encoding of deep neural networks to generate the semantic embeddings of documents.

However, the semantics of a document is closely related to the topics it contains. We therefore propose an encoding framework, named DSC-LSTM (Document Semantic Coding based on Bi-directional LSTM), to learn document-level embeddings. In order to capture both forward and backward contexts, DSC-LSTM takes advantage of a bi-directional LSTM neural network to encode the semantics of documents.

The proposed DSC-LSTM framework is shown in Fig. 1. The framework is divided into five layers, taking the word embeddings of all words in the document as input and the document semantic code as output. These layers are described in detail below.

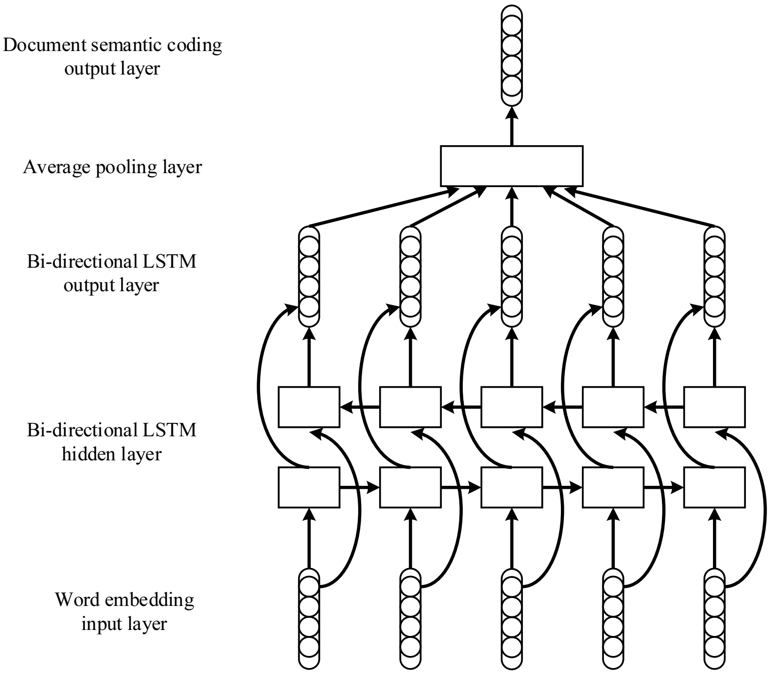
Document semantic coding output layer
Average pooling layer
Bi-directional LSTM output layer
Bi-directional LSTM hidden layer
Word embedding input layer

Fig. 1. The framework of DSC-LSTM

Word embedding input layer. Because the number of words in some documents is very large, and unimportant words have no direct influence on the representation of topics, we only take the entity-relational words in each document as input. The entity-relational words, which appear in the triples of the DBpedia (http://wiki.dbpedia.org/) knowledge base, can be found through the entity relationship link operation. After obtaining entity-relational words, a Skip-Gram model proposed by Le and Mikolov [21] is used to learn the word embeddings of these words.

Bi-directional LSTM hidden layer. This layer contains two LSTM hidden layers, forward and backward. Each input word embedding is inserted into the forward and backward LSTM hidden layers, and the two hidden layers are connected to the same output. The input word embedding under the current moment t is E_t. The output of the forward LSTM hidden layer is h_{t-1}^f and the output of the backward LSTM hidden layer is h_{t+1}^b. The output of both hidden layers under the current moment is:

$$\begin{aligned} h_t^f &= H\left(E_t, h_{t-1}^f, c_{t-1}, b_{t-1}\right) \\ h_t^b &= H\left(E_t, h_{t+1}^b, c_{t-1}, b_{t-1}\right) \end{aligned}, \quad (1)$$

where $H(\cdot)$ represents the hidden layer function, c_{t-1} represents the state value of the cell at the previous moment, and b_{t-1} refers to the bias at the previous moment.

Bi-directional LSTM output layer. Each output unit is connected to the forward and backward LSTM hidden layers.

$$o_t = \sigma\left(W_{ho}^f h_t^f + W_{ho}^b h_t^b + b_o\right), \quad (2)$$

where W_{ho}^{f} and W_{ho}^{b} are the forward and the backward weights between the hidden layers and the bidirectional LSTM output layer respectively, b_o denotes the bias. The output of this layer is a vector and the dimension of each vector is consistent with the input vector.

Average pooling layer. There are two commonly used pooling operations to extract robust features: maximum pooling and average pooling. In this paper, we choose average pooling because the semantics of a document is closely related to each entity-relational word in the document. It is, therefore, reasonable to perform down-sampling by dividing the input into rectangular pooling regions and computing the average values of each region. This pooling scheme can be represented as

$$pool(o)=\sum_{t=1}^{T}\frac{o_t}{L}, \tag{3}$$

where L is the length of the input word sequence.

Document semantic coding output layer. We learn the final document semantic codes by Equation (4).

$$s=\sigma\left(pool(o)\right), \tag{4}$$

where $\sigma(\cdot)$ denotes the activation function. The dimension of document semantic codes is the same as the dimension of input word embeddings. As a result, the similarity between documents and words can be calculated by the cosine similarity. DSC-LSTM is able to generate document semantic codes that extract the semantic information at the document level, which lays the foundation for the topic modeling based on document semantics.

3.2 Latent Dirichlet Allocation

This paper builds a topic model based on LDA proposed by Blei et al. [4], which is the basis of most probabilistic topic models. LDA is a typical three-level Bayesian generative model, consisting of documents, topics and words from top to bottom. The model contains two Dirichlet-polynomial conjugate structures: document-topic and topic-word. The document-topic distribution θ and topic-word distribution ϕ both are the polynomial distributions. The parameters of θ are Dirichlet distributions with α as the prior parameter, while the parameters of ϕ are Dirichlet distributions with β as the prior parameter. The generative process of LDA is described as follows:

- Choose the number of documents to be generated and the corresponding number of words per document (i.e., the length of each document).
- For each word of each document: 1) From document-topic distribution θ_m, sample a topic assignment for the current word; 2) According to the topic assignment, choose the corresponding topic-word distribution $\varphi_{z_{m,n}}$; 3) Sample a word according to the topic-word distribution $w_{m,n} \sim Mult(\varphi_{z_{m,n}})$.

3.3 Semantic Enhancement Based on GPU Model

Document semantic codes generated by DSC-LSTM reflect the semantic elements of documents, and thus can be used as a constraint on topics. For each word in the current document, if the word embedding is close to the semantic document code, the word becomes a representative word of the document. The representative word should be increased the probability that it is selected by the corresponding topic of the document.

In this paper, the GPU model is used to enhance semantics, which has been widely used for lexical enhancement in probabilistic topic models. However, these previous works only promote the semantically related words under the same topic [22], which neglect the semantic relations between words and documents. On the contrary, this paper considers the influence of document semantics on topic modeling. The enhancement of topic semantics is not only reflected in the semantic association between words, but also the semantic association between words and documents. The dual-GPU model adopted in this paper has the following two meanings:

- When a word w is sampled by a topic z_w, if the word is related to the semantic code of a document d, the number of co-occurrences between z_w and d will increase.

- When a word w is sampled by a topic z_w, all the words $w^* \in R_w$ that are semantically correlated to the word w should be semantic enhanced. As a result, the probability that w^* is sampled by the topic z_w will increase.

The semantic relatedness between document-topic can be calculated by the cosine similarity between the word embedding and the semantic coding vector. If the cosine similarity is greater than a certain threshold ξ, the number of co-occurrences between the corresponding topic z and the document d increases by $a(0 < a < 1)$ in a document-topic enhanced matrix M, that is:

$$M_{d,z} = \begin{cases} 0, & dist(s_d, e_{w_z}) < \xi \\ a, & dist(s_d, e_{w_z}) \geq \xi \end{cases}, \tag{5}$$

where $dist(s_d, w_z)$ represents the cosine similarity between the word embedding e_{w_z} of the word w_z and the semantic code s_d of the document d.

Semantic relatedness between words can be calculated by the cosine similarity between their word embeddings. For the sampled word w, all words w^* whose cosine similarities are greater than a threshold ψ form a related word collection R_w. Let N be a word-word enhanced matrix. For the word w, the related words $w^* \in R_w$ need to be enhanced. The number of co-occurrences will increase by $b(0 < b < 1)$. In addition, for the enhancement of the current word w, the number of co-occurrences is increased by 1.

$$N_{w,w^*} = \begin{cases} 1, & w = w^* \\ b, & w^* \in R_w \text{且} w \neq w^* \\ 0, & \text{otherwise} \end{cases}. \tag{6}$$

Through the document-topic and word-word semantic reinforcement by using the dual-GPU model, our model not only increases the proportion of the related topics in the documents, but also makes the semantic related words co-occur in a topic with a higher probability.

3.4 Model Structure

The model structure of GPU-LDA-LSTM is shown in Fig. 2. The gray part indicates the document-topic and word-word enhancement. The former relies on both document semantic codes and word embeddings, while the latter depends on word embeddings only. Compared with LDA, the proposed model does not introduce additional random variables and prior distribution, which guarantees the simplicity of the parameter inference process.

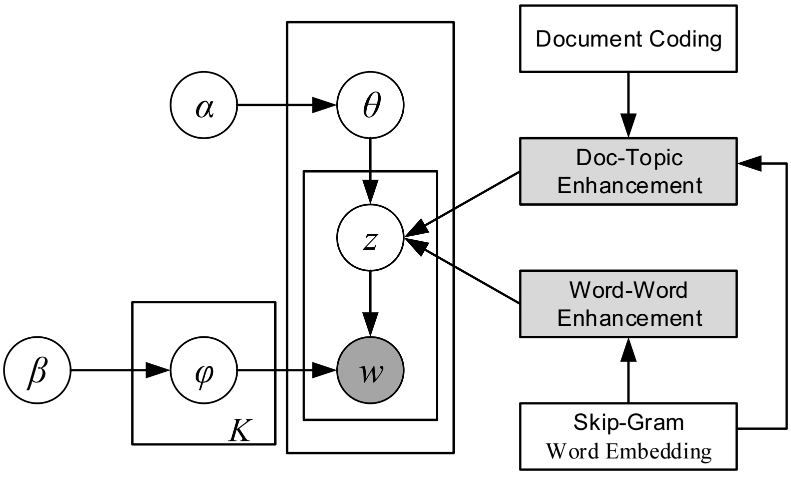

Fig. 2. The model structural graph of GPU-LDA-LSTM

3.5 Model Inference

Collapsed Gibbs sampling is a stochastic algorithm commonly used in the parameters inference process of topic models. It is widely used because of its high computational efficiency and simple operation based on the Markov hypothesis [5]. This paper also employs the collapsed Gibbs sampling to infer the parameters of our model.

The proposed model uses the document-topic and word-word enhancement by dual-GPU to constrain the topic assignments of words. Therefore, the process of sampling needs to consider the integration of the dual-GPU model. In addition, traditional collapsed Gibbs sampling is based on the Markov assumption that the topic assignment at the current moment is only related to the previous topic assignment. However, the dependency between the topic distribution sequences at different moments remains to be studied. In this paper, the parameter inference process of GPU-LDA-LSTM utilizes a deep neural network to enhance the relevance between topics at different time slices. Combined with the dual-GPU model and a LSTM neural network, this paper proposes a new parameter inference algorithm for our model.

According to the collapsed Gibbs sampling framework, the sampling probability which merges the dual-GPU enhancement model of the sampled word w is:

$$\begin{aligned} & p\left(z_{d,n}=ct \mid z_{-(d,n)}, w, \alpha, \beta\right) \\ & \propto \frac{C_{d,ct}^{-(d,n)} M_{d,ct}+\alpha}{\sum_{k=1}^{K}\left(C_{d,k}^{-(d,n)} M_{d,k}+\alpha\right)} \frac{\sum_{u=1}^{V} C_{ct,u}^{-(d,n)} N_{u,n}+\beta}{\sum_{v=1}^{V}\left(\sum_{u=1}^{V} C_{ct,u}^{-(d,n)} N_{u,v}+\beta\right)}, \end{aligned} \quad \textbf{(7)}$$

where ct represents the topic assignment of the current word, $C_{d,ct}^{-(d,n)}$ represents the number of times the word in the document d is assigned to ct excluding the current word, $C_{ct,u}^{-(d,n)}$ represents the number of times the word u is assigned to ct excluding the current word.

Furthermore, this paper also considers the influence of the historical topic assignment of words on the final topic assignment sequence, and thus we construct a topic distribution dependent network based on LSTM. The network consists of a forward LSTM hidden layer, which is used to predict the next topic assignments. The input layer of the network is the topic assignment sequence at the current time slice. The output layer is a softmax function, corresponding to the probability of assigning topics at the next time slice. All the topic distributions are represented as one-hot vectors.

Since the process of topic modeling is unsupervised, each word needs to be topic-tagged to train network weights. In this paper, we use the LDA model to annotate the topics of words in document collections. The topic distribution sequence after Gibbs sampling convergence is regarded as the topic annotation sequence of document collections. With the dual-GPU collapsed Gibbs sampling and topic distribution dependent network, the topic distribution under the current word is:

$$\begin{aligned} p\left(z_{d,n}=ct\right) &= \lambda \cdot p\left(z_{d,n}=ct \mid z_{-(d,n)}, w, \alpha, \beta\right) \\ &\quad +(1-\lambda) \cdot LSTM\left(z_{d,n}\right), \end{aligned} \quad \textbf{(8)}$$

where $LSTM\left(z_{d,n}\right)$ represents the output of the topic distribution dependent network, λ represents the balance factor of two parts.

According to the sampling algorithm, it is effective to obtain information from the historical topic assignment, which can be used to improve the sampling of the current topic. Therefore, the parameter inference process improves the context consistency by using a LSTM neural network.

4 Experiment

In this section we conduct extensive experiments to evaluate our proposed model GPU-LDA-LSTM on two real world corpora against the state-of-the-art baselines. The experimental results demonstrate that our topic model provides promising performance in terms of topic coherence, topic words and text classification.

4.1 Data and Setting

We use two datasets in the experiments: 20-Newsgroups (http://qwone.com/jason/20Newsgroups/) and 163News. 163News dataset is a collection of 40,000 news crawled from a popular Chinese news website (http://www.163.com/), including sports, cars and other 6 categories. Their statistics are summarized in Table 1.

Table 1. Dataset statistics

Dataset	20-Newsgroups	163News
# documents	18846	40000
# categories	20	6

We compare our model with two baseline methods: LDA and NTM. The introduction of two models and related parameters are as follows:

- LDA is the most widely used topic model. However, this model is unable to incorporate external knowledge. We use the jGibbLDA package (http://jgibblda.sourceforge.net) with collapsed Gibbs sampling to implement the LDA model, which is provided online.
- NTM is a neural topic model where the representation of words and documents are combined into a deep learning framework [8]. For this model, we use the implementation (https://github.com/elbamos/NeuralTopicModels) released by the authors

For the 163News dataset, we train 300-dimensional word embeddings from 3 million Chinese news crawled from the 163 website using Skip-gram algorithm. For the 20-Newsgroups dataset, we use the pre-trained 300-dimensional word embeddings (https://code.google.com/p/word2vec). If a word has no embedding, the word is considered as having no word semantic correlation knowledge.

For the baselines, we choose the parameters according to their original papers. For the document-topic enhancement of our model, $\xi = 0.3$ and $a = 0.1$. For the word-word enhancement, $\varphi = 0.4$ and $b = 0.2$. The balance factor λ in the network is set to 0.8. In addition, the vocabulary size is set to 4,500 for all models.

4.2 Evaluation by Topic Coherence

The most commonly used automatic evaluation for topic coherence is Point-wise Mutual Information (PMI) [9]. Related research shows that PMI are often highly consistent with human evaluation, and the higher the score of PMI, the better the semantic coherence of topics [9]. Therefore, this paper uses PMI to calculate topic coherence. Given a topic-word distribution φ_k, the PMI-Score of φ_k is:

$$PMI\left(\varphi_k\right) = \frac{2}{V\left(V-1\right)} \sum_{1 \le i < j \le V} \log \frac{p\left(w_i, w_j\right)}{p\left(w_i\right) p\left(w_j\right)}, \tag{9}$$

where $p\left(w_i\right)$ denotes the probability of word w_i in the document sets, $p\left(w_i, w_j\right)$ denotes the joint probability of word w_i and word w_j in the document sets, and V denotes the vocabulary size.

Fig. 3 and Fig. 4 show the PMI values of the three topic models under different settings on the number of topics $K = \{40, 60, 80\}$.

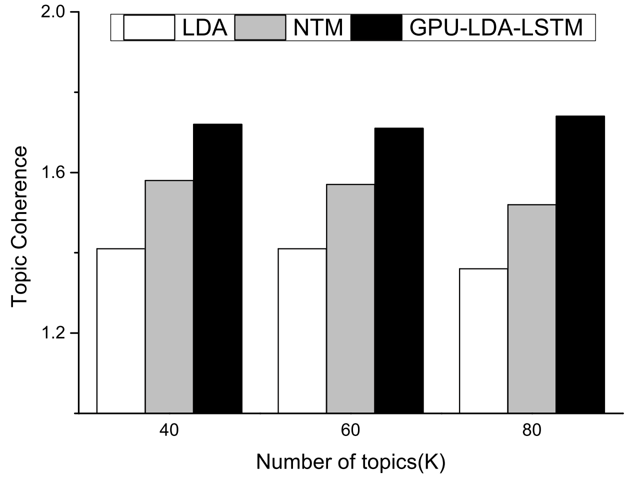

Fig. 3. PMI on 163News dataset

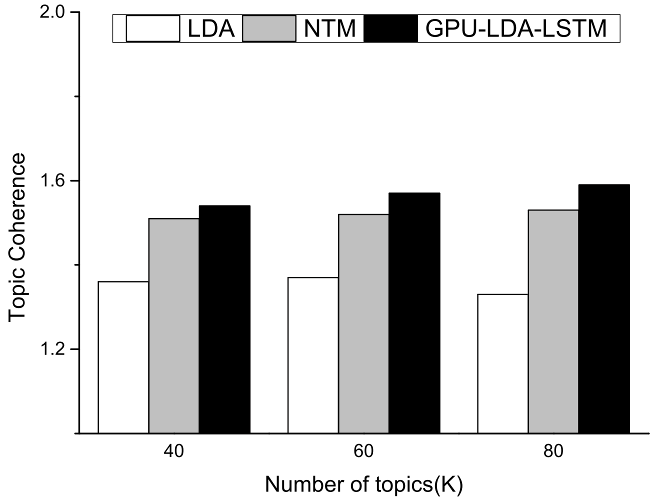

Fig. 4. PMI on 20-Newsgroups dataset

As shown in Fig. 3, on the 163News dataset, the PMI value of GPU-LDA-LSTM is the highest, indicating that its extracted topics have strong semantic coherence. The LDA model has the lowest PMI score because it does not take into account the semantic reinforcement of documents and words. NTM is a neural network reconstruction of LDA. However, the model neglects the semantic reinforcement of words, and thus the PMI value is lower than the proposed model.

As described in Fig. 4, though the differences in the PMI values of each model are not as obvious as on the 163News dataset, the above conclusions still hold for the 20-Newsgroups dataset. Furthermore, as the number of topics increases, the PMI value remains essentially unchanged. This may be due to the difference between topic modeling on a Chinese dataset and an English dataset. A longer vocabulary list is required under English datasets [10].

3.3 Evaluation by Topic Words

In order to demonstrate the semantic coherence of topics more intuitively, we analyze the representative words of the topics extracted from each topic model. For each topic, we choose 5 words with the highest probability in the topic distribution. For NTM, we use bigrams words because of its intrinsic characteristics. Due to space limitations, Table 2 only lists topic words under the three categories “Automobile”, “Finance” and “Sports” on the 163News dataset. Words that are noisy and lack representativeness are highlighted with bold font. The topic number of all topic models is set to 40.

Table 2. Topic words learned from 163News dataset

	LDA	NTM	GPU-LDA-LSTM
Automobile	marketing **acid** **rate** high-grade **standard**	**discount underweight** **south about** BMW brake consumer groups **wet method**	car speed consumption vehicle wheelbase
The number of noise words	3	3	0
Finance	**paragraph** coal **special** cost economic	newspaper reporter industry hotel limited company **version phenomenon** **scope down**	IPO impact equity market **qualified**
The number of noise words	2	2	1
Sports	**cost** team income interaction **standard**	**baby participate** technical level players tour **euro rose** yahoo sports	sports foul qualified final reporter
The number of noise words	2	2	1

Table 2 shows that LDA generates the worst coherent topics with more noise words than the other models. For instance, words like “standard”, “acid” and “rate” could be irrelevant to automobile related topics. Although NTM can enhance the understandability of topics by bigram phrases, it does not explicitly consider lexical semantic reinforcement. Therefore, there are many meaningless phrases in topics. By contrast, our model outperforms other models and almost all topic words are related to the topic. The reason is that our method learns topics with both semantic reinforcement of document-topic and word-word, which significantly enhance the quality of topics.

Table 3 shows some topics learned from the 20-Newsgroups dataset. Three topics correspond to health, crime and sports respectively. From the table, we observe that the topics learned by the propose model are better in coherence than those learned from baseline methods, which demonstrates the effectiveness of our model again.

Table 3. Topic words learned from 20-Newsgroups dataset

	LDA	NTM	GPU-LDA-LSTM
Health	money **Columbia** health insurance **private**	tax program **year month** public companies consumer groups care insurance	company health insurance tax costs
The number of noise words	2	1	0
Crime	gun weapon firearms **used** police	gun weapons criminal killed **men child** police control **com today**	**people** crime criminal killed deaths
The number of noise words	1	2	1
Sports	**will** games hockey **year** teams	play games **people men** game player league season hockey ball	ball league game games season
The number of noise words	2	1	0

3.3 Evaluation by Text Classification

Text categorization is an effective method for the evaluation of topic models. Better classification accuracy means that the extracted topics are more discriminative and representative. Fig. 5 and Fig. 6 report the classification accuracy on the two datasets by using the three models with different settings on the number of topics $K=\{40,60,80\}$.

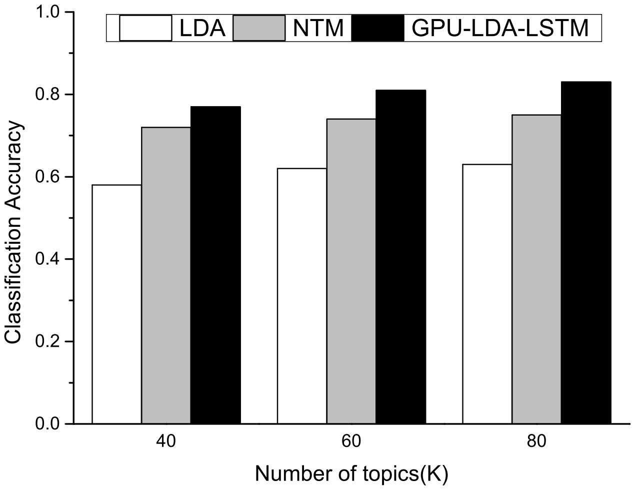

Fig. 5. Classification accuracy on 163News dataset

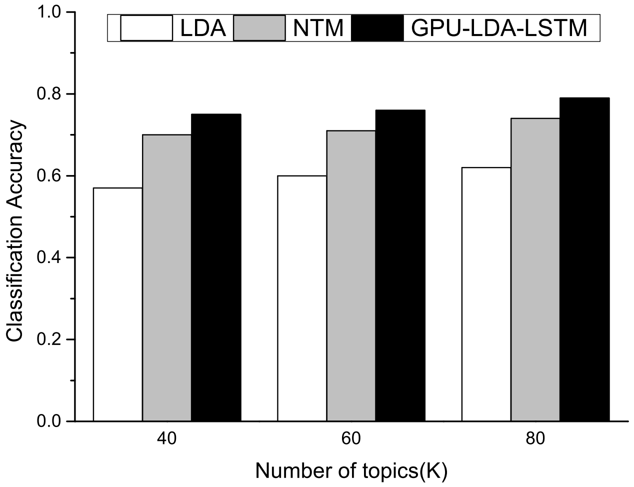

Fig. 6. Classification accuracy on 20-Newsgroups dataset

As shown in Fig. 5 and Fig. 6, our GPU-LDA-LSTM model significantly outperforms other state-of-the-art models on both datasets. Particularly, the performance gains of GPU-LDA-LSTM with respect to NTM are achieved on all $K=\{40,60,80\}$ settings. This validates that the proposed model based on semantic reinforcement with bi-directional LSTM has strong characterization in document feature representation.

4 Conclusion

In this paper, we propose a novel topic model GPU-LDA-LSTM based on bi-directional LSTM networks. This paper first emphasizes the importance of document semantics. We therefore design a document semantic coding framework to learn the semantic codes of documents. Secondly, the document semantic codes and word embeddings are used respectively for the document-topic enhancement and word-word enhancement during the Gibbs sampling process. Furthermore, the iterative process of the Gibbs sampling is implemented by a LSTM network. Experiments exhibit the high quality of the topics generated by our topic model and competitive performance on text classification tasks compared to state-of-the-art approaches. In the future, we will study how to apply our model on various data mining tasks such as tracing topic evolutions of text streams or text retrieval.

References

[1] P. Zhang, B. Li, R. Yang, Research on the topic evolution of microblog based on BTM-LPA, in: Proc. 2017 Conference on Computer Science and Technology, 2017.

[2] Y. Chen, C. Wu, M. Xie, X. Guo, Solving the sparsity problem in recommender systems using association retrieval, Journal of Computers 6(9)(2011) 1896-1902.

[3] F. Caron, M. Davy, A. Doucet, Generalized polya urn for time-varying Dirichlet process mixtures, in: Proc. 2007 Conference on the 13th Uncertainty in Artificial Intelligence, 2007.

[4] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research (3)(2003) 993-1022.

[5] P. Xie, D. Yang, E.P. Xing, Incorporating word correlation knowledge into topic modeling, in: Proc. 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015.

[6] B. Xu, X. Guo, Y. Ye, J. Cheng, An improved random forest classifier for text categorization, Journal of Computers 7(12)(2012) 2913-2920.

[7] U. Sayan, X. Li, S.A. Mohamed, Emerging event detection in social networks with location sensitivity, World Wide Web-Internet & Web Information Systems 18(5)(2015) 1393-1417.

[8] Z. Cao, S. Li, Y. Liu, W. Li, H. Ji, A novel neural topic model and its supervised extension, in: Proc. 2015 Conference on the 29th AAAI Conference on Artificial Intelligence, 2015.

[9] D. Mimno, H.M. Wallach, E. Talley, M. Leenders, A. McCallum, Optimizing semantic coherence in topic models, in: Proc. 2011 Conference on Empirical Methods in Natural Language Processing, 2011.

[10] H. Xu, F. Zhang, W. Wang, Implicit feature identification in Chinese reviews using explicit topic mining model, Knowledge-based Systems 76(1)(2015) 166-175.

[11] S. Hochreiter, J. Schmidhuber, Long short-term memory neural computation, Neural Computation 9(8)(1997) 1735-1780.

[12] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Networks 18(5-6)(2005) 602-610.

[13] D. Andrzejewski, X. Zhu, M. Craven, Incorporating domain knowledge into topic modeling via Dirichlet forest priors, in: Proc. 2009 the 26th International Conference on Machine Learning, 2009.

[14] Z. Hu, G. Luo, M. Sachan, E. Xing, Z. Nie, Grounding topic models with knowledge bases, in: Proc. 2016 the 25th International Joint Conference on Artificial Intelligence, 2016.

[15] Z. Chen, B. Liu, Topic modeling using topics from many domains, lifelong learning and big data, in: Proc. 2014 the 31st International Conference on Machine Learning, 2014.

[16] Z. Chen, B. Liu, Mining topics in documents: standing on the shoulders of big data, in: Proc. 2014 the 20th International Conference on Knowledge Discovery and Data Mining, 2014.

[17] G.E. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets, Neural Computation 18(7)(2006) 1527-1554.

[18] F. Tian, B. Gao, D. He, T. Liu, Sentence level recurrent topic model: letting topics speak for themselves. <https://arxiv.org/abs/1604.02038>, 2016.

[19] C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, W. Ma, Topic aware neural response generation, in: Proc. 2017 the 31st AAAI Conference on Artificial Intelligence, 2017.

[20] S. Li, Y. Zhang, R. Pan, M. Mao, Y. Yang, Recurrent attentional topic model, in: Proc. 2017 the 31st AAAI Conference on Artificial Intelligence, 2017.

[21] Q.V. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proc. 2014 the 31st International Conference on Machine Learning, 2014.

[22] C. Li, H. Wang, Z. Zhang, A. Sun, Z. Ma, Topic modeling for short texts with auxiliary word embeddings, in: Proc. 2016 the 39th International Conference on Research and Development in Information Retrieval, 2016.

[23] Q. Qian, M. Huang1, J. Lei, X. Zhu, Linguistically regularized LSTM for sentiment classification, in: Proc. 2017 the 55th Annual Meeting of the Association for Computational Linguistics, 2017.

[24] S. Song, H. Huang, T. Ruan, Abstractive text summarization using LSTM-CNN based deep learning, Multimedia Tools and Applications (10)(2018) 1-19.