Knowledge-Based Systems 174 (2019) 194–204

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Learning document representation via topic-enhanced LSTM model

Wenyue Zhang [a], Yang Li [a], Suge Wang [a,b,*]

[a] School of Computer and Information Technology, Shanxi University, China
[b] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, China

ARTICLE INFO

Article history:
Received 4 June 2018
Received in revised form 4 March 2019
Accepted 10 March 2019
Available online 14 March 2019

Keywords:
Document representation
Deep learning
Long–short term memory
Topic modeling

ABSTRACT

Document representation plays an important role in the fields of text mining, natural language processing, and information retrieval. Traditional approaches to document representation may suffer from the disregard of the correlations or order of words in a document, due to unrealistic assumption of word independence or exchangeability. Recently, long–short-term memory (LSTM) based recurrent neural networks have been shown effective in preserving local contextual sequential patterns of words in a document, but using the LSTM model alone may not be adequate to capture global topical semantics for learning document representation. In this work, we propose a new topic-enhanced LSTM model to deal with the document representation problem. We first employ an attention-based LSTM model to generate hidden representation of word sequence in a given document. Then, we introduce a latent topic modeling layer with similarity constraint on the local hidden representation, and build a tree-structured LSTM on top of the topic layer for generating semantic representation of the document. We evaluate our model in typical text mining applications, i.e., document classification, topic detection, information retrieval, and document clustering. Experimental results on real-world datasets show the benefit of our innovations over state-of-the-art baseline methods.

1. Introduction

Document representation is mainly concerned with how an unstructured document should be transformed into a machine-understandable representation that captures the semantic meaning of the document. It is a fundamental element in various applications in text mining, natural language processing, and information retrieval. For example, to classify a textual document into a category, one key step would be to represent the document as a meaningful fixed-length vector.

The most popular approach to document representation is the bag-of-words (BoW) model [1], also known as vector space model [2], where a textual document is typically represented as a bag (vector) of individual words or n-grams. However, BoW tends to produce sparse representation and may result in reduced performance for the downstream tasks. By exploiting co-occurrence patterns among words in a document, semantic analysis models, such as probabilistic latent semantic analysis (PLSA) [3] and latent Dirichlet allocation (LDA) [4] are able to detect global semantic topical structure for deriving dense document representation. These models tend to suffer from an unrealistic word exchangeability assumption. Recently, long–short-term memory (LSTM)-based [5] recurrent neural networks have been shown to be effective in modeling contextual sequential patterns among words in a given document widely employed to derive distributed vector representations of the text [6,7]. However, employing LSTM alone may not be adequate to capture global thematic semantics when generating document representation.

In this work, we propose a new Topic-Enhanced LSTM neural network model with topic similarity constraint (TE-LSTM+SC) to learn the distributed vector representations of documents. In particular, given a textual sequence of words in a document, we first employ an attention-based LSTM model to generate contextual hidden representation of the sequence. Then, we develop a topic modeling layer on the attentive representation to capture the high-level latent thematic structure of the data. Note that we design a Similarity Constraint (SC) strategy in this layer and enable the inferred latent semantic topics to be as diverse as possible. Next, a tree-structured LSTM module is introduced on top of the topic modeling layer to generate the representation of the document.

We have made the following main contributions in this work: (1) We develop a unified learning framework that incorporates the latent topic model into the LSTM-based network for document representation. (2) The proposed TE-LSTM+SC allows us not only to capture local sequential contexts but also to infer global latent topical semantics from a given text. (3) We introduce a similarity constraint strategy in the topic modeling layer to

* Corresponding author at: School of Computer and Information Technology, Shanxi University, China.
E-mail address: wsg@sxu.edu.cn (S. Wang).

https://doi.org/10.1016/j.knosys.2019.03.007
0950-7051/

encourage diversity among the learned underlying topics. (4) We conduct quantitative and qualitative evaluations to validate the effectiveness of TE-LSTM+SC for the document representation problem.

The rest of this paper is structured as follows. Section 2 presents related the work related to document representation. Section 3 introduces the proposed TE-LSTM+SC model in detail. Next, we evaluate the new method for typical text mining applications on the real-world data, and present experimental results in Section 4. We conclude this paper with a brief discussion in Section 5.

2. Related work

Document representation has been extensively investigated for many years. In this section, we present the major existing approaches to document representation. We roughly organize them into three categories, i.e., bag-of-words models, semantic analysis models, and deep neural network models.

2.1. Bag-of-words models

One of the most common approaches to document representation is the bag-of-words (BoW) model [1]. Briefly, BoW model represents an unstructured textual document as a bag or vector of individual words that appear in the document. It primarily focuses on word-level occurrence patterns and pays no attention to the order or semantic structure of the sequence of the words within the document. Though it is simple, the BoW model may fail to model word correlations or contextual sense about the semantics of the words, for example, synonymous and polysemous words occurring in a document.

A straightforward extension on BoW, i.e., the bag-of-n-grams model [8], may partially consider the word order or correlations in short context, but it typically leads to sparser document representation, which may harm the performance for downstream text processing or mining tasks. Thus, bag-of-words model and its variants show very little understanding about the semantics of the sequence of words in a document while generating the representation.

2.2. Semantic analysis models

Generally, semantic analysis approaches to document representation attempt to rotate vector space and then project documents onto latent principal factor space via discovering underlying semantic structure of the text data. To represent textual documents and terms (words) in high-level conceptual thematic space, Deerwester et al. [9] proposed a latent semantic analysis method (LSA), which basically leverages singular value decomposition technique onto original term-document co-occurrence association data. Hofmann [3] presented a statistical view on LSA that resulted in a more principled model named probabilistic latent semantic analysis (PLSA). In contrast to LSA, the probabilistic variant shows a sound statistical foundation and provides a proper generative framework for deriving low-dimensional representations of documents. Blei et al. [4] further generalized PLSA and developed a statistical model named latent Dirichlet allocation (LDA) to discover hidden semantic structure of text data. In LDA, semantic topic distributions are generated through Dirichlet process.

By leveraging statistical co-occurrence patterns among words in documents, semantic topic models and their variants are good at mining the global semantic topical structure of the text data, but they normally tend to suffer from a disregard of the order of words in documents, owing to unrealistic assumption of word exchangeability in the models.

2.3. Deep neural network models

Various deep neural network models have been developed to learn distributed vector representations of text at various levels of granularity, such as word-level [10–12], sentence-level [13,14], or document-level [15–17]. In this part, we mainly focus on previous work related to deriving document representation.

Wong et al. [15] developed a recursive autoencoder model which can be trained efficiently and produces compact representations of documents. Sun et al. [16] proposed a computationally economical algorithm for evolving unsupervised deep neural networks to efficiently learn meaningful representations. Zhao and Mao [17] presented a document representation learning framework, where each document is represented through clustering embeddings of words in the document. Though they are able to learn the distributed representation of a document, none of the aforementioned models are not good at preserving local contextual patterns over word sequences. Sun et al. [18] a hierarchical attention-based classification model (known as HAN). It has two levels of attention mechanisms, i.e., word-level and sentence-level. However, in the HAN model, there is no topic information involved, i.e., latent global semantic information that can be identified across documents. Dieng et al. [19] proposed a variational autoencoder (VAE) based method, which uses topics as latent parameters among all RNN nodes, and then represents documents by concatenating the output of inference network and the last state of the RNN. However, simply concatenating the two parts may limit the performance for the generation of document representation.

Recently, LSTM [5], a type of recurrent neural network models, has proven effective in a variety of sequential learning tasks. Li et al. [20] extended basic LSTM, and introduced a hierarchical LSTM model with attention, which sequentially builds an embedding for a document from the learned embeddings at the sentence and word levels. Tai et al. [21] generalized the LSTM to tree-structured network topology and introduced a neural network model named tree-LSTM to improve the semantic representation of each document (paragraph). In addition, Jo et al. [22] proposed the use of latent topic information in an LSTM network for mortality prediction based on time-series multidocuments.

In addition, there are some other kind of representation methods, i.e., Trace Ratio Relevance Feedback (TRRF) proposed by Yang et al. [23], and Discriminative Locally Document Embedding (Disc -LDE) proposed by Wei et al. [24]. The aim of TRRF focuses on refining representations of multimedia data, which means that all information processed through the method is kept in the same granularity. Disc-LDE generates document embeddings by a generative probabilistic model, which is trained through a discrimination process.

In this work, we propose a new topic-enhanced LSTM model for deriving dense distributed document representation, which is related to but different from all the aforementioned models. TE-LSTM+SC is basically a unified learning framework that mainly consists of a sequential LSTM layer, topic modeling module with similarity constraint, and tree-structured LSTM layer. To learn the representation from documents, the proposed model cannot only preserve local contextual semantic information but can also leverage the global high-level latent thematic structure of the textual data.

3. Methodology

In this section, we present the topic-enhanced LSTM model for learning of document representation.

Table 1
Notations used in TE-LSTM+SC.

Symbols	Desciption
x	Input word embedding generated by word2vec.[a]
I, F, O, G, C	Input gate, forget gate, output gate, candidate state, and memory cell state.
σ, tanh	Sigmoid and hyperbolic tangent activation functions.
W, b, B	Parameters of a LSTM model.
seq, att, tr, r, s	Sequence module, attention module, tree-LSTM module, representation dimension variation and similarity constraint strategy, respectively.
h, a	Hidden state vector and attention strength vector.
t	Index of a word within a given text.
K, k	Number of latent topics and index of a topic.
T, TC	State vectors of a latent topic unit and its corresponding memory cell.
Ts, TCs	Matrix formed by all T or TC vectors, respectively.
L, v, s	Topic label, projection vector to the topic, and their similarity score.
f	Forget gate of a tree-LSTM node.
Rep, p, g	Document representation, prediction vector generated by TE-LSTM+SC, and ground truth.

[a] http://radimrehurek.com/gensim/models/word2vec.html.

3.1. Overview

We aim to deal with the document representation problem, i.e., given a text document, we focus on learning its distributed representation. We introduce a latent semantic topic modeling layer on top of the attentive LSTM network, and then build a tree-structured LSTM on the topic layer. In this way, we develop a new unified learning framework named topic-enhanced LSTM model. Fig. 1 shows the overview of TE-LSTM+SC, which mainly consists of four modules, i.e., sequence modeling layer, attention modeling layer, semantic topic modeling layer, and tree-structured representation layer. Table 1 lists the notations used in TE-LSTM+SC.

In particular, given a document that consists of a sequence of words, the sequence modeling layer first transforms the input word embedding data into hidden contextual representations. The attention modeling layer takes as input the hidden representations and evaluates which parts of the input are responsible for deriving meaningful higher-level hidden representations. Then, a latent topic modeling layer is introduced on top of the attention layer to further explore the semantic topical structure of the data. A specially designed similarity constraint strategy is integrated in the layer to encourage the latent topics to be as diverse as possible. Based on the topic layer, a tree-structured LSTM model is employed to derive the semantic representation of the text.

In the following sections, we will describe each of the modeling layers in detail.

3.2. Sequence modeling layer

Given an input sequence of word vectors generated via word2vec, the LSTM-based sequence modeling module produces the hidden representations that preserve local contextual semantics of the word sequence. One key advantage of the LSTM model over other kinds of neural networks is that it can exploit the internal memory to cope with arbitrary sequences of input and capture dynamic temporal behavior and dependency among the data.

Formally, given a word embedding x_t at the time step t, we can derive corresponding hidden state h_t by using the LSTM-based sequence modeling module, as shown in Eqs. (1)–(3):

$$\begin{bmatrix} I_t \\ F_t \\ O_t \\ G_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(W_{seq} \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} + B_{seq} \right) \tag{1}$$

$$C_t = I_t \cdot G_t + F_t \cdot C_{t-1} \tag{2}$$

$$h_t = O_t \cdot \tanh(C_t) \tag{3}$$

Actually, the hidden state of each word in the sequence has summarized local contextual semantics. This is because, in addition to the current observed word, the historical hidden state is exploited by the sequential learning model to derive the hidden state of the current word. Possibly the same words may have different hidden representations if their contexts are diverse in the sequences. It is worth noting that the propagation of contextual semantics is controlled through the input gate I_t and output gate O_t.

We collect all the hidden states and updated internal memory cell states into $\boldsymbol{Hs}=\{h_1, h_2, \ldots, h_n\}$ and $\boldsymbol{Cs}=\{C_1, C_2, \ldots, C_n\}$, which serve as input to the following attention modeling layer.

3.3. Attention modeling layer

The attention layer bridges the sequence modeling layer and latent topic layer. It can evaluate which parts of the hidden states from previous layer are most important for detecting underlying semantic topics in the following layer. Generally, a latent topic is expressed in a high-level cluster of semantically related words, and thus each observed word may show different associative relationships with various topics. For example, in the sentence *"As the war went on, oil prices reached record highs, which raised concern in the auto industry."*, the word "war" is much more relevant to the topic "military" compared to other topics such as "economy" and "technology". The attention modeling layer can automatically estimate the attentive weights of the hidden states of individual words with regard to latent topics.

Formally, we generate a K-dimensional attention strength vector a_t via Eq. (4) as follows (K: number of latent topics):

$$a_t = \sigma\left(W_{att} h_t + b_{att}\right) \tag{4}$$

Each component of a_t indicates the relation strength between contextual hidden state h_t and the corresponding topic. Given the attention vectors, the attention strength matrix is then generated as follows:

$$\widetilde{A} = [a_1, a_2, a_3, \ldots, a_n]^T$$

Note that each row of the attention matrix $\widetilde{A}$ is not normalized, and only contains relative strength values. To facilitate the following semantic topic modeling process, we need to make a slight change to the attention matrix as follows:

$$A = softmax_{rows}(\widetilde{A}). \tag{5}$$

In Eq. (5), we apply the softmax function to each row of the raw matrix $\widetilde{A}$, and yield the desired normalized attention strength matrix A.

3.4. Latent topic modeling layer

A latent topic often refers to a cluster of semantically related words that are observed in textual documents. For instance, the semantic topic *"space"* may group a list of words such as *"space"*, *"orbit"* and *"shuttle"*, etc. The topic modeling layer aims to discover the underlying semantic structure of text data and infer high-level latent topics from the input sequence.

Fig. 2 shows the process of inferring latent semantic topics. Taking as input the hidden states of observed words and normalized attentive weight vectors, the topic modeling layer generates the latent topic matrix *Ts* via Eq. (6). During the process, the layer has embedded the co-occurrence semantics of words in a sequence and have also taken into account the weights of hidden states for detection of the topics. In addition, the topic layer also

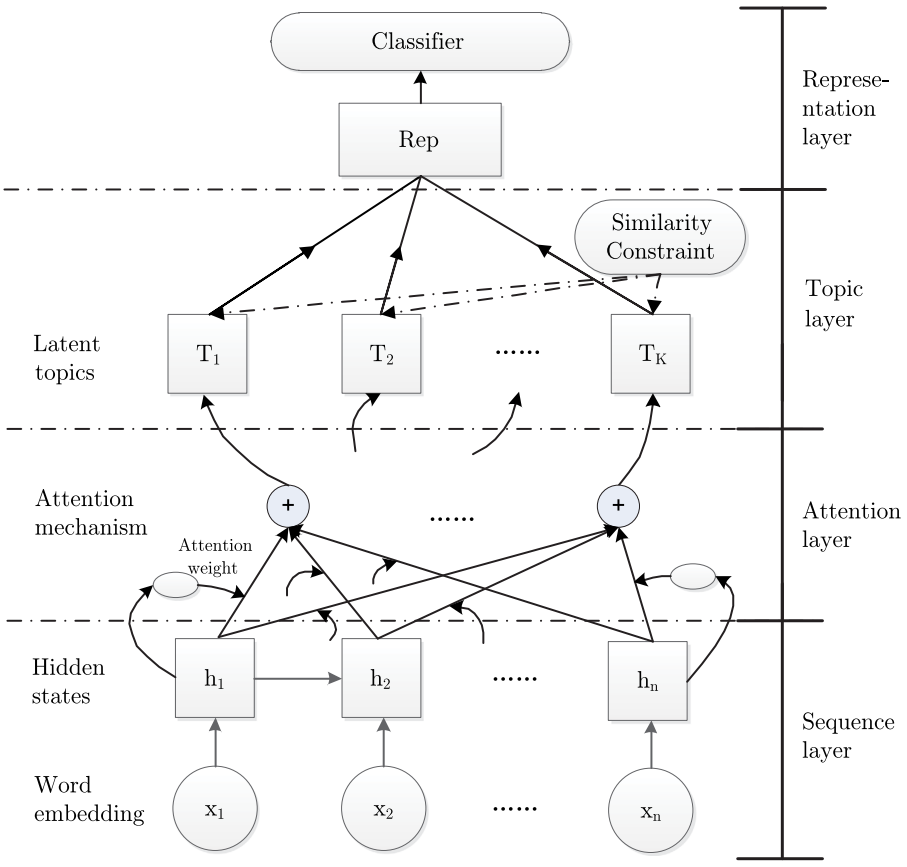


Fig. 1. The proposed TE-LSTM+SC model for document representation.

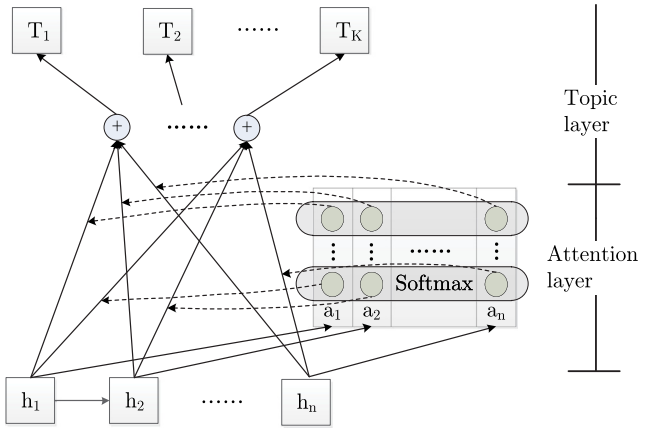


Fig. 2. Latent topic modeling layer. Taking as input hidden states and normalized attentive relation vectors, the layer then produces latent semantic topics (Ts and TCs).

produces the topic-dependent memory cell matrix *TCs* based on input hidden memory states.

$$\begin{bmatrix} Ts \\ TCs \end{bmatrix} = A \begin{bmatrix} Hs \\ Cs \end{bmatrix} \tag{6}$$

Then, the *Ts* and *TCs* are used as input to the topic similarity constraint module and tree-LSTM representation layer.

3.5. Topic similarity constraint

The constraint on the similarity among latent topics has been largely overlooked in previous studies. As a consequence, there is perhaps redundancy in the latent topic detection result. For example, in a practical application, the expected set of latent topics are *"military"*, *"economy"*, and *"health"*, etc., but without using similarity constraint, the generated set of topics could be *"national defense"*, *"weapons"*, *"economy"*, and *"health"*, etc. It is thus important and necessary to introduce the similarity constraint strategy into the latent topic detection layer.

The similarity constraint mechanism is designed based on the assumption that: a corpus is generated on a hyperspace, where each document is characterized by its semantic latent topic information, and the information of each topic is represented by a hypersurface in the space. In this paper, the topics are represented in one-hot vectors (L in Table 1) and can be treated as the basis of the space. As a result, the comprehensive latent topic information of a document can be mathematically calculated through multinomial summation of all topics. According to the assumption, it may be necessary to learn the latent semantic topics that are distinctive and diverse from each other.

We propose a constraint measure for topic similarity upon *Ts*, which is viewed as outcomes of projection from the document context semantic to corresponding topic basis of the hyperspace. The main idea of this measure is to obtain a similarity score s_k between topic projection vector v_k and topic label L_k, where each v_k is generated based on the corresponding T. All L_k are manually generated and are orthogonal to each other. In this paper, each L_k represents a row of an identity matrix and corresponds to a T unit. Note that the topic label vectors are seen as the basis that constitutes a latent semantic space. As a result, all semantic topics

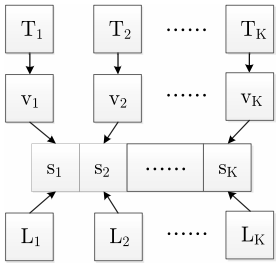


Fig. 3. Structure of topic similarity constraint mechanism.

can be represented by a linear combination of the label vectors. In this constraint measure, we leverage the average similarity score as the similarity among all topics. Eqs. (7)–(9) compute the topic similarity score. Fig. 3 explicitly shows the structure of the constraint mechanism.

$$v_k = \sigma\left(W_s T_k + B_s\right) \tag{7}$$

$$s_k = -\sum_{j=1}^{K} [L_k \cdot \log(v_k)]_j \tag{8}$$

$$S = \sum_{k=1}^{K} s_k \Big/ K \tag{9}$$

We note that reducing $\boldsymbol{S}$ results in larger differences among the T vectors. This constraint is optimized through an objective function (15).

3.6. Tree-LSTM representation layer

The algorithm used for generating *Rep* is similar to an N-ary Tree-LSTM network [21]. The reason we choose the Tree-structure LSTM is that it is able to automatically select the information of subnodes as opposed to manual assignment. Almost all the components of the Tree-LSTM network are employed to combine the information from *Ts* into a single outcome, with the exception of forget gates. The gates control the information flow from *Ts* to *Rep*. In addition, an additional processing step described by Eq. (14) allows the dimension of *Rep* to be variable.

$$\begin{bmatrix} I \\ O \\ G \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(W_{tr} Ts + B_{tr}\right) \tag{10}$$

$$f_k = \sigma(W^f_{tr(k)} T_k + b^f_{tr(k)}) \tag{11}$$

$$C = I \cdot G + \sum_{k=1}^{K} f_k \cdot TC_k \tag{12}$$

$$h = O \cdot \tanh(C) \tag{13}$$

$$Rep = \sigma(W_r h + b_r) \tag{14}$$

As shown in the equations, each child node has a unique forget-gate status, which reflects the variable effects of latent topic semantic on the *Rep*. Next, a distributed dense representation for the given textual document is generated, which leverages the benefits from inferred latent topics of the document.

Finally, we introduce a softmax classifier on top of *Rep* during the training process. The loss function described in Eq. (15) is used to fit classification labels g (c categories in total) of a document through updating model parameters. In Eq. (15), the parameter λ is the tradeoff coefficient between fitting ground truth data and constraining topics similarity, which can be determined via cross-validation in practice.

$$\epsilon = -\sum_{i=1}^{c} [g \cdot log(p)]_i + \lambda S \tag{15}$$

Table 2
Statistics of datasets.

Datasets	# of documents	Avg length of doc	# of classes
20Newsgroup	18,848	93	20
Wiki10+	17,325	936	25
Amazon	8000	108	2
SemEval	1250	6	6

4. Experiments

In this section, we evaluate the proposed TE-LSTM+SC for document representation in some typical text mining applications, namely, document classification, topic detection, information retrieval, and document clustering. In the classification application, three specific tasks are evaluated, i.e., topic, sentiment [25,26], and emotion classification, respectively. For topic detection, a topic is recognized by finding semantically related words according to their attention strength with regard to the topic.

For the information retrieval experiment, we test our model for the answer selection (AS) task given a question and a candidate set of answer sentences. In other words, the goal of the task is to recommend the most likely answer sentences for each given question. One typical approach to dealing with AS may consist of the following steps: (1) Transforming each question and candidate answer into fixed-length vector representations; (2) Estimating correlation score between the question and answer pair; and (3) Ranking all the candidate pairs according to correlation scores. In addition, for the document clustering experiment, we group the derived document vectors into multiple clusters, where the number of clusters is simply the number of categories in the given corpus.

4.1. Datasets

We used four publicly available datasets for evaluation, i.e., 20*Newsgroup*,[1] [27,28] *wiki*10+,[2] [29,30] *amazon reviews*[3] and *SemEval*2007 *task* 14[4] [31]. Specifically, the amazon reviews dataset consists of 4 different domains, where each document is classified into 1 out of 2 sentiment polarities. The SemEval2007 dataset contains 1250 documents that are labeled with the following emotion categories, i.e., *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*. The statistics of datasets are listed in Table 2.

In preprocessing for the datasets, following the work [32], we removed the documents that contain less than 6 words. Then, we conducted 10-fold cross validation experiments on each processed dataset, where for each run we used 90% of the dataset to train the model and the remainder of the dataset for testing.

To evaluate the proposed method for information retrieval task, two extract datasets were used, namely, WikiQA [33] and TrecQA [34]. The Table 3 lists some basic statistics about the datasets. Note that we removed all questions that do not have any correct answers or have negative answers.

1 http://qwone.com/~jason/20Newsgroups.
2 http://nlp.uned.es/social-tagging/wiki10+/.
3 http://www.cs.jhu.edu/~mdredze/datasets/sentiment/.
4 https://aclanthology.coli.uni-saarland.de/papers/S07-1013/s07-1013.

Table 3
Statistics of the IR datasets.

Datasets	# of Q		# of A		Avg-len(Q)	Avg-len(A)
	Train	Test	Train	Test		
WikiQA	873	126	20,360	2733	7	25
TrecQA	78	65	5919	1117	9	28

Table 4
Classification results of all the evaluated methods on the four datasets.

	20Newsgroup	Wiki10+	Amazon	SemEval07
LDA	0.657	0.351	0.526	0.208
Doc-NADE	0.67	0.462	0.532	0.407
HTMM	0.665	0.389	0.550	0.328
TopicRNN	0.632	0.301	0.51	0.335
Seq-LSTM	0.663	0.486	0.715	**0.424**
GMNTM	0.731	0.425	0.601	0.347
SLRTM	0.739	0.489	0.57	0.352
TE-LSTM	0.737	0.522	0.73	0.397
TE-LSTM+SC	**0.75**	**0.534**	**0.736**	0.412

4.2. Experimental setup

We relied on well-known tool Theano[5] to implement our proposed TE-LSTM+SC model. The dimensions of input word vector, hidden layer, and document representation were set to 50, 100, and 50, respectively. TE-LSTM+SC were trained by using Adadelta method with an initial learning rate of 0.01, and the model parameters were randomly initialized. After several tries upon small dataset from the corpus, the value of λ was set as 0.2.

We compared the proposed TE-LSTM+SC model with its variants for document representation in the three problems:

- ***seq − LSTM*** is just the plain LSTM.
- ***TE − LSTM***(***without similarity constraint***) does not consider the topic similarity constraint. Then, we removed the similarity term λS from the objective function in Eq. (15). The variant model is used to test the effect of the new topic similarity constraint strategy on document representation for each task.

For contrast, we also compare the proposed TE-LSTM+SC with the following state-of-the-art baseline methods:

- ***LDA*** [4], or latent Dirichlet allocation: It is one of the most popular topic models and is widely used in studies.
- ***Doc − NADE*** [35], or Document Neural Autoregressive Distribution Estimator: It is a representative topic model based on neural networks.
- ***HTMM*** [36], or Hidden Topic Markov Model: It relies on the Markov transitions to process topics.
- ***GMNTM*** [32]: It combines a feedforward neural network with a Gaussian mixture model.
- ***SLRTM*** [37], or Sentence Level Recurrent Topic Model: It incorporates the LDA process into a plain LSTM.
- ***TopicRNN*** [19], an unsupervised model that can capture latent global topic semantics.

4.3. Document classification

Intuitively, better vector representations can lead to higher classification accuracy. Hence, for evaluating the effectiveness of our model, we employ classification experiments and the results of all the aforementioned baseline methods are listed in Table 4.

[5] http://deeplearning.net/software/theano.

Table 5
T-test value of pairs which are between TE-LSTM+SC and baseline methods on different corpus.

Pairs	20Newsgroup	Wiki10+	Amazon
TE-LSTM+SC, LDA	10.9	13.7	10.1
TE-LSTM+SC, seq-LSTM	8.6	3.0	3.6
TE-LSTM+SC, SLRTM	1.5	2.0	11.7

In Table 4, we show the average classification accuracy of the proposed methods and baselines on 20Newsgroup, amazon reviews, and SemEval2007, while for wiki10+ the measure is micro-F_1 score because the corpus is multilabeled. It is clear that, compared with baselines, TE-LSTM+SC/TE-LSTM achieves considerably improved performance for the classification problem. The proposed TE-LSTM+SC/TE-LSTM achieves significant improvement on the 20Newsgroup and amazon reviews datasets. This is because the length of documents in the corpuses are usually shorter than that in the wiki10+ dataset on average. In addition, the imbalanced nature of wiki10+ also incurs difficulties in learning process. However, our proposed model significantly improves its micro-F_1 score, because TE-LSTM+SC/TE-LSTM is capable of overcoming the difficulties. On the amazon reviews, representations generated by the TE-LSTM+SC/TE-LSTM can be more easily mapped to their sentiment polarities; this phenomenon may indicate that our model is more capable of encoding semantic topic information in sentiment classification. However, sequence structured models have higher average accuracy than our model in the SemEval2007 task #14. One explanation may be that documents of the corpus are too short to carry enough latent topic information.

The comparison between seq-LSTM and TE-LSTM+SC/TE-LSTM indicates that document representations generated by using attention mechanism lead to better performance of classification. Moreover, comparing with seq-LSTM, the improvement may be attributed to the fact that the mechanism makes it possible to reveal correlations between words and topics.

Then, we conducted a *T-test* on the 10-fold cross validation results to further evaluate the significance of the benefits of TE-LSTM+SC over baselines.

Table 5 lists the T-test scores between TE-LSTM+SC and corresponding baselines. The higher the *T-test* score is, the more superior the performance of the proposed model is. The results of *T-test* indicate that our model significantly outperforms the baselines for document classification. Specifically, TE-LSTM+SC classifies amazon reviews considerably better than the SLRTM, perhaps owing to that more diverse themes of corpus making its topic information ambiguous. In addition, topic information is the most distinct in the 20Newsgroup, and the classification result of the sequence structure LSTM is not as good as that on another corpus. By contrast, hybrid-structured TE-LSTM+SC performance is remarkably superior in the dataset.

Furthermore, to validate effectiveness of topic similarity constraint mechanism, we sorted testing datasets by their topic similarity score in descending order and divided them into 15 batches of the same size. The average classification accuracy of each batch is shown in Fig. 4.

As shown in the figure, both curves indicate an ascending tendency. In other words, the less similar the topics are, the better performance the new model achieves. The correlation is a strong support for the effectiveness of the constraint. Given the small size of the wiki10+ testing dataset, the performance curve on the dataset is much fluctuant than another ones.

Table 6
Performance of topic detection.

Models	Topics	Words	C_v
lda2vec	Space	Astronomical, astronomy, satellite, planetary, telescope	0.556
	Encryption	Encryption, wiretap, encrypt, escrow, clipper	0.572
	X Windows	Mydisplay, xlib, window, cursor, pixmap	0.472
	Mid-East	Armenian, Lebanese, Muslim, Turk, Sy	0.200
	All 20 Topics		0.567(avg)
TE-LSTM	Space	Space, orbit, mission, astro, NASA	0.593
	Encryption	Encryption, cryptography, key, escrow, chip	0.610
	X Windows	Motif, faq, widget, window, windows	0.602
	Mid-East	Israel, Israeli, Lebanon, Armenia, Arab	**0.671**
	All 20 Topics		0.588(avg)
TE-LSTM+SC	Space	Orbit, space, shuttle, launch, mars	**0.600**
	Encryption	Cryptography, key, encryption, nsa, chip	**0.618**
	X Windows	Motif, faq, sunos, windows, widget	**0.621**
	Mid-East	Israel, Israeli, Armenia, Armenian, Jew	0.630
	All 20 Topics		**0.611**(avg)

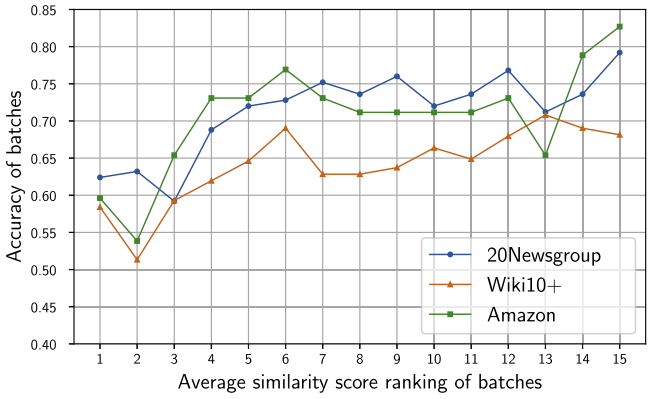


Fig. 4. Correlation between the ranking of mean similarity score and accuracy of the batch.

Table 7
Performance of rest detected topics detected by the TE-LSTM+SC.

Topics	Words	C_v
Atheism	Atheism, god, murder, atheist, car	0.547
Graphics	Graphics, polygon, algorithm, image, windows	0.498
PC.hardware	Windows, isa, gateway, bios, vlb	0.611
Mac.hardware	Mac, centris, apple, quadra, iisi	0.664
Forsale	Sale, shipping, offer, sell, obo	0.607
Autos	Car, dealer, honda, toyota, wheel	0.589
Motorcycles	Bike, dod, ride, motorcycle, rider	0.723
Baseball	Baseball, season, team, hitter, player	0.641
Hockey	Hockey, nhl, lindros, selanne, team	0.659
Electronics	Car, electronics, circuit, project, joystick	0.494
Med	Drug, med, treatment, medical, patient	0.623
Christian	Christian, Jesus, Christ, God, Christianity	0.703
Guns	Gun, handgun, firearm, weapon, baseball	0.616
Politics	Homosexual, sexual, gay, windows, economic	0.641
Religion	Christian, gay, Christianity, God, religion	0.584
Ms-windows	Windows, Indiana, Microsoft, version, site	0.551

4.4. Topic detection

For the latent topic detection task, topics are represented by using clusters of semantically related words. Unlike other methods, the topic enhanced model is able to reveal coherence between words and topics.

In the lda2vec [38], four topics from the 20Newsgroup were showen with their highly related words and were then submitted to an online system *Palmetto*[6] for measuring coherence of the words. We followed the settings in the lda2vec, i.e., for each of the four topics, coherence of the top 5 strongest attention words was evaluated.

The coherence between words and topics indicates the correlation among word cluster and topic. For fair quantitative comparison, we adopted the coherence measurement introduced in lda2vec [38]. The closer the relation between words and topic is, the higher the C_v achieved. We tested the coherence performance of lda2vec and TE-LSTM+SC/TE-LSTM, and the results are listed in Table 6.

Compared to the values of 4 selected topics of lda2vec, the performance of our model is remarkable. Moreover, the average score of the 20 words groups is notably improved by TE-LSTM+SC/TE-LSTM. This suggests that the word clusters detected by the new model are much more relevant to their corresponding topics. In addition, better results of the proposed model indicate that the similarity constraint strategy works very well for the task.

Furthermore, the results of the rest 16 topics detected via TE-LSTM+SC are listed in Table 7 for further comparison.

4.5. Information retrieval

In information retrieval, the Answer Selection task can be considered as a pairwise comparison [39] problem, where models are trained with binary classifier to distinguish correct question answer pairs and negative pairs [40]. In practice, for each question answer pair, we combine their *REP*s with a binary classifier, whose answer is regarded as the score of correction.

Furthermore, we chose the Mean Reciprocal Rank (MRR) and the Mean Average Precision (MAP) as evaluation metrics, which are calculated as follow:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{Rank_i} \tag{16}$$

$$MAP = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{m_i}\sum_{j=1}^{m_i}\frac{j}{Rank_{ij}} \tag{17}$$

where Q is a set of given questions, $Rank_i$ refers to the rank position of the first correct answer in the candidate list for the ith question, whose correct candidate set is m_i, and the $Rank_{ij}$ represents the jth correct answer of the question. Obviously, the higher the values of *MRR* and *MAP* are, the higher the model prefers ranking the correct documents.

6 http://palmetto.aksw.org/palmetto-webapp/. It records the coherence value (C_v), which is the Normalized Pointwise Mutual Information (NPMI) of each pair of words within a sliding window upon an external corpus, and return the NPMI mean of the submitted words.

Table 8
Answer selection results of models.

Models	TrecQA		WikiQA	
	MRR	MAP	MRR	MAP
TopicRNN	0.414	0.394	0.463	0.453
SLRTM	0.518	0.438	0.441	0.431
Seq-LSTM	0.466	0.407	0.432	0.426
TE-LSTM	0.524	0.467	0.473	0.467
TE-LSTM+SC	**0.535**	**0.472**	**0.494**	**0.486**

Table 9
Document clustering performance of methods.

	20Newsgroup	Wiki10+
LDA	0.249	0.107
Doc-NADE	0.128	0.091
HTMM	0.175	0.123
GMNTM	0.100	0.096
SLRTM	0.369	0.216
Seq-LSTM	0.503	0.319
TE-LSTM	0.623	0.418
TE-LSTM+SC	**0.782**	**0.477**

Table 8 lists the results of our model and baselines in answer selection task.

In Table 8, the TE-LSTM+SC with topic similarity constraint mechanism achieves the best results. In particular, the model with similarity constraint performs better than the one having no extra topic processing, and the performance was improved through the applying of attention mechanism. Again, the contrast experiments validate the effectiveness of attention mechanism constraint and topic similarity strategy of the proposed method.

4.6. Document clustering

By employing our proposed model, textual documents can be represented by real-value semantic vectors, which may indicate semantic similarities of documents. To validate the effectiveness of the representations derived by our method, we conducted document clustering experiments via K-means algorithm on the 20Newsgroup and wiki10+ dataset. Generally, if a pair of documents are semantically correlative, the Euclidean distance between their vectors would be small, and then they are more likely to be gathered into the same cluster.

In our experiments, the numbers of clusters were set to 20 and 25 on the 20Newsgroup and wiki10+, respectively. Then, for quantitative analysis, we recorded the percentage of categories in each cluster, and then all categories are assigned to clusters one by one with their highest percentages scores. We report the average percentage scores of all clusters, as listed in Table 9.

Clearly, TE-LSTM+SC/TE-LSTM achieve the highest average score, especially for the proposed model with topic similarity constraint mechanism, which performs best in this test. In addition, the results of models based on neural networks are significantly better than others.

Next, Tables 10 and 11 list the detailed results of document clustering based on the representations via the proposed models on the 20Newsgroups dataset, where the color depth of a cell represents the percentage score of its corresponding category in the cluster.

Category distributions of all clusters generated on the 20Newsgroup are detailed in Tables 10 and 11. It is obvious that document vectors derived by TE-LSTM+SC have more clear correspondence between category and clusters. Furthermore, the semantic relativity among categories are also more reasonable in Table 10. For example, category distribution of cluster *"C"* indicate that categories *"religion.misc"* and *"atheism"* are correlative, and the *"religion.misc"* row reflects the close relationship among *"atheism"*, *"religion.christian"*, *"politics.guns"*, and *"religion.misc"*.

Moreover, Tables 12 and 13 list the detailed document clustering results based on the representations via the proposed models on the wiki10+ dataset. It is clear that category and clusters in Table 12 present the strongest correspondence, which means that topical semantic information distance of document representations generated by the new model are more likely to be revealed through linear calculation.

In addition, in the column *"K"* of Table 12, documents of both *"web"* and *"web2.0"* categories are extensively contained, which also imply that there is a close relationship between those two categories. In fact, this is reasonable and intuitive, and additionally, the phenomenon in Table 12 is more notable than other baseline models.

From the tables we can see that TE-LSTM+SC/TE-LSTM obtain the best performance on both 20Newsgroups and wiki10+, compared to all baselines. The results indicate that document representations generated by the new model are more effective in encoding semantic information for document clustering.

5. Conclusions

In this paper, we propose a TE-LSTM+SC model for learning textual document representations. When given a document, TE-LSTM+SC first employs the attentive LSTM model to generate local contextual hidden representation. Then, the model relies on a topic modeling layer with similarity constraint to capture global latent thematic structure of the data. Next, it applies a tree-structured LSTM module to derive the distributed dense representation. The representations generated by the TE-LSTM+SC result in superior performance compared to state-of-the-art models in document classification, topic detection, information retrieval, and document clustering applications.

According to the experimental results of TE-LSTM+SC, TE-LSTM and seq-LSTM in all tasks on each corpus, the effectiveness of the two main contributions, namely, attention and topic similarity constraint mechanisms, has been validated. In particular, special structures of the attention modeling and latent topic mining layer improve the capability of our model in embedding global latent topic semantics and local sequential contexts, which was clearly demonstrated in experiments on the datasets with clear topic information. Relatively, for the corpus, which is tiny-scaled or contains ambiguous topic information, the global latent semantics are less helpful for some tasks and may affect the performance of the TE-LSTM+SC model. In conclusion, the experimental results clearly support the main contributions of this study: (1) It presents an effective unified method with hybrid structure combining topic model and LSTM neural networks; (2) It is able to capture sequential local contexts as well as latent global semantics; (3) It can leverage the topic similarity constraint mechanism to make latent information diverse; and (4) It conducts extensive experiments to validate the model in quantitative as well as qualitative ways.

For future work, we plan to extend the proposed model in an unsupervised learning framework and will also develop new layers in the model, which may work well for the short text representation problem. Then, one of the possible directions is to replace the classification layer with a structure of variational autoencoder (VAE), which makes it capable of training the model using context semantics from the document itself.

Table 10
The document clustering results for TE-LSTM+SC on the 20Newsgroup dataset.

Categories \ Clusters	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
rec.sport.baseball	1																			
comp.os.ms-windows.misc		1																		
alt.atheism			1																	
misc.forsale				1																
soc.religion.christian					1															
talk.politics.guns						1														
sci.electronics							1													
rec.motorcycles								1												
comp.sys.ibm.pc.hardware									1											
talk.politics.misc										1										
sci.crypt											1									
sci.space												1								
rec.sport.hockey													1							
comp.windows.x														1						
talk.politics.mideast															1					
talk.religion.misc																1				
rec.autos																	1			
comp.graphics																		1		
comp.sys.mac.hardware																			1	
sci.med																				1

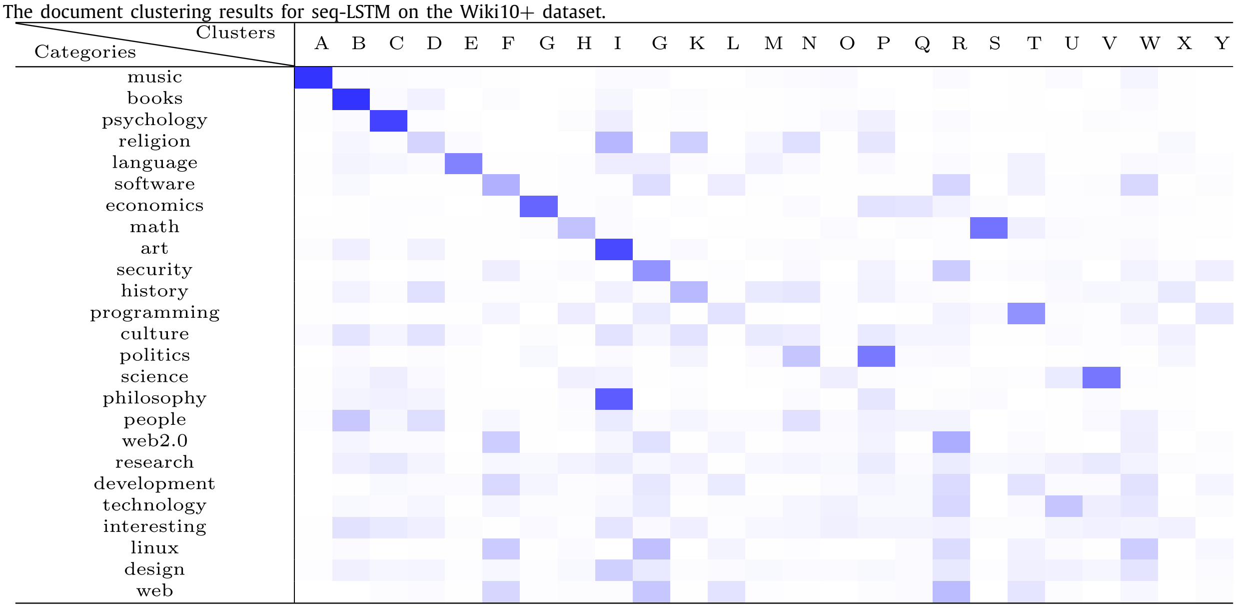
Table 11
The document clustering results for seq-LSTM on the 20Newsgroup dataset.

Categories \ Clusters	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
rec.sport.hockey	1																			
talk.religion.misc		1																		
alt.atheism			1																	
soc.religion.christian				1																
sci.electronics					1															
rec.motorcycles						1														
comp.sys.mac.hardware							1													
comp.os.ms-windows.misc								1												
talk.politics.mideast									1											
comp.graphics										1										
rec.autos											1									
sci.space												1								
misc.forsale													1							
sci.crypt														1						
talk.politics.misc															1					
sci.med																1				
rec.sport.baseball																	1			
comp.windows.x																		1		
comp.sys.ibm.pc.hardware																			1	
talk.politics.guns																				1

Table 12
The document clustering results for TE-LSTM+SC on the Wiki10+ dataset.

Categories \ Clusters	A	B	C	D	E	F	G	H	I	G	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
music	1																								
culture		1																							
psychology			1																						
technology				1																					
books					1																				
language						1																			
software							1																		
economics								1																	
math									1																
design										1															
web											1														
history												1													
programming													1												
politics														1											
philosophy															1										
science																1									
religion																	1								
art																		1							
linux																			1						
development																				1					
security																					1				
interesting																						1			
research																							1		
people																								1	
web2.0																									1

Table 13
The document clustering results for seq-LSTM on the Wiki10+ dataset.



Acknowledgments

The work presented in this paper benefits from the interesting discussion with Zhen Hai.[7] The authors would like to express gratitude to all reviewers, whose valuable comments and suggestions are significantly helpful to improve the quality of this paper. The work is supported by the National Natural Science Foundation of China (Nos. 61632011, 61573231, 61672331, 61432011, 61603229) and the Key research and development projects of Shanxi Province, China (201803D421024).

References

[1] Z.S. Harris, Distributional structure, Word 10 (1981) 146–162.
[2] G. Salton, A. Wong, C.-S. Yang, A vector space model for automatic indexing, Commun. ACM 18 (1975) 613–620.
[3] T. Hofmann, Probabilistic latent semantic analysis, in: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, 1999, pp. 289–296.
[4] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[5] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.
[6] Y. Liu, M. Lapata, Learning structured text representations, Trans. Assoc. Comput. Linguist. 6 (2018) 63–75.
[7] C. De Boom, S. Van Canneyt, T. Demeester, B. Dhoedt, Representation learning for very short texts using weighted word embedding aggregation, Pattern Recognit. Lett. (2016) 150–156.
[8] M. Kekha, A. Khonsari, F. Oroumchian, Rich document representation and classification: An analysis, Knowl.-Based Syst. 22 (2009) 67–71.
[9] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (1990) 391–407.
[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inf. Process. Syst. 26 (2013) 3111–3119.
[11] S. Lai, K. Liu, S. He, J. Zhao, How to generate a good word embedding, IEEE Intell. Syst. 31 (2016) 5–14.
[12] J. Li, J. Li, X. Fu, M.A. Masud, J.Z. Huang, Learning distributed word representation with multi-contextual mixed embedding, Knowl.-Based Syst. 106 (2016) 220–230.
[13] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, R. Ward, Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval, IEEE/ACM Trans. Audio Speech Lang. Process., TASLP 24 (2016) 694–707.
[14] M.J. Er, Y. Zhang, N. Wang, M. Pratama, Attention pooling-based convolutional neural network for sentence modelling, Inform. Sci. 373 (2016) 388–403.
[15] D.F. Wong, Y. Lu, L.S. Chao, Bilingual recursive neural network based data selection for statistical machine translation, Knowl.-Based Syst. 108 (2016) 15–24.
[16] Y. Sun, G.G. Yen, Y. Zhang, Evolving unsupervised deep neural networks for learning meaningful representations, IEEE Trans. Evol. Comput. PP (2018) 1–1.
[17] R. Zhao, K. Mao, Fuzzy bag-of-words model for document representation, IEEE Trans. Fuzzy Syst. 26 (2018) 794–804.
[18] Y. Sun, G.G. Yen, Z. Yi, Evolving unsupervised deep neural networks for learning meaningful representations, IEEE Trans. Evol. Comput. (2018) 1–1.
[19] A.B. Dieng, C. Wang, J. Gao, J.W. Paisley, Topicrnn: A Recurrent Neural Network with Long-Range Semantic Dependency, CoRR, 2016, abs/1611.01702.
[20] J. Li, T. Luong, D. Jurafsky, A hierarchical neural autoencoder for paragraphs and documents, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, vol. 1, 2015, pp. 1106–1115.
[21] K.S. Tai, R. Socher, C.D. Manning, Improved semantic representations from tree-structured long short-term memory networks, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, vol. 1, 2015, pp. 1556–1566.
[22] Y. Jo, L. Lee, S. Palaskar, Combining LSTM And Latent Topic Modeling for Mortality Prediction, CoRR, 2017, abs/1709.02842.
[23] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, Y. Pan, A multimedia retrieval framework based on semi-supervised ranking and relevance feedback, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 723–742.
[24] C. Wei, S. Luo, J. Guo, Z. Wu, L. Pan, Discriminative locally document embedding: Learning a smooth affine map by approximation of the probabilistic generative structure of subspace, Knowl.-Based Syst. 121 (2017) 41–57.
[25] C. Zhao, S. Wang, D. Li, Determining fuzzy membership for sentiment classification: A three-layer sentiment propagation model, PLoS One 11 (2016) e0165560.
[26] C. Zhao, S. Wang, D. Li, Exploiting social and local contexts propagation for inducing chinese microblog-specific sentiment lexicons, Comput. Speech Lang. 55 (2019) 57–81.
[27] B. Altinel, M.C. Ganiz, B. Diri, Instance labeling in semi-supervised learning with meaning values of words, Eng. Appl. Artif. Intell. 62 (2017) 152–163.
[28] D. Cai, X. He, Manifold adaptive experimental design for text categorization, IEEE Trans. Knowl. Data Eng. 24 (2012) 707–719.
[29] A. Zubiaga, Enhancing Navigation on Wikipedia with Social Tags, CoRR, 2012, abs/1202.5469.
[30] M. Gupta, R. Li, Z. Yin, J. Han, Survey on social tagging techniques, ACM Sigkdd Explor. Newsl. 12 (2010) 58–72.
[31] C. Strapparava, R. Mihalcea, Semeval-2007 task 14: Affective text, in: Proceedings of the 4th International Workshop on Semantic Evaluations, Association for Computational Linguistics, 2007, pp. 70–74.

[7] Institute for Infocomm Research, A*STAR, Singapore.

[32] M. Yang, T. Cui, W. Tu, Ordering-sensitive and semantic-aware topic modeling, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015, pp. 2353–2359.
[33] Y. Yang, W.-t. Yih, C. Meek, Wikiqa: A challenge dataset for open-domain question answering, in: Proceedings of the 2015 Conference on Empirical Methods, in: Natural Language Processing, 2015, pp. 2013–2018.
[34] M. Wang, N.A. Smith, T. Mitamura, What is the jeopardy model? A quasi-synchronous grammar for QA, in: Proceedings of the 2007 Joint Conference on Empirical Methods, in: Natural Language Processing and Computational Natural Language Learning, 2007, pp. 22–32.
[35] H. Larochelle, S. Lauly, A neural autoregressive topic model, in: Proceedings of the 25th International Conference on Neural Information Processing Systems, vol. 2, 2012, pp. 2708–2716.
[36] A. Gruber, Y. Weiss, M. Rosen-Zvi, Hidden topic Markov models, in: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, vol. 2, 2007, pp. 163–170.
[37] F. Tian, B. Gao, D. He, T.-Y. Liu, Sentence Level Recurrent Topic Model: Letting Topics Speak for Themselves, CoRR, 2016, abs/1604.02038.
[38] C.E. Moody, Mixing Dirichlet Topic Models and Word Embeddings to Make Lda2vec, CoRR, 2016, abs/1605.02019.
[39] L. Nie, X. Wei, D. Zhang, X. Wang, Z. Gao, Y. Yang, Data-driven answer selection in community QA systems, IEEE Trans. Knowl. Data Eng. 29 (2017) 1186–1198.
[40] Y. Xiang, Q. Chen, X. Wang, Y. Qin, Answer selection in community question answering via attentive neural networks, IEEE Signal Process. Lett. 24 (2017) 505–509.