Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems

School of Information Systems

2-2019

Topical co-attention networks for hashtag recommendation on microblogs

Yang LI

Ting LIU

Jingwen HU

Jing JIANG
Singapore Management University, jingjiang@smu.edu.sg

Citation

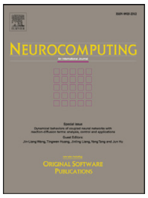
LI, Yang; LIU, Ting; HU, Jingwen; and JIANG, Jing. Topical co-attention networks for hashtag recommendation on microblogs. (2019). *Neurocomputing*. 331, 356-365. Research Collection School Of Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/4900

Neurocomputing 331 (2019) 356–365

Contents lists available at ScienceDirect

Neurocomputing

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



Topical Co-Attention Networks for hashtag recommendation on microblogs

Yang Li[a,*], Ting Liu[b], Jingwen Hu[b], Jing Jiang[c]

[a] College of Information and Computer Engineering, Northeast Forestry University, China
[b] Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China
[c] School of Information System, Singapore Management University, Singapore

ARTICLE INFO

Article history:
Received 19 May 2018
Revised 30 September 2018
Accepted 16 November 2018
Available online 22 November 2018

Communicated by Dr. Weike Pan

Keywords:
Hashtag recommendation
Long short-term memory
Co-attention
Topic model

ABSTRACT

Hashtags provide a simple and natural way of organizing content in microblog services. Along with the fast growing of microblog services, the task of recommending hashtags for microblogs has been given increasing attention in recent years. However, much of the research depends on hand-crafted features. Motivated by the successful use of neural models for many natural language processing tasks, in this paper, we adopt an attention based neural network to learn the representation of a microblog post. Unlike previous works, which only focus on content attention of microblogs, we propose a novel Topical Co-Attention Network (TCAN) that jointly models content attention and topic attention simultaneously, in the sense that the content representation(s) are used to guide the topic attention and the topic representation is used to guide content attention. We conduct experiments and test with different settings of TCAN on a large real-world dataset. Experimental results show that our model significantly outperforms various competitive baseline methods. Furthermore, the incorporation of topical co-attention mechanism gives more than 13.6% improvement in F1 score compared with the standard LSTM based methods.

1. Introduction

In recent years, as a social network and news media, microblog has achieved great success and become very important. There is a large amount of information produced every day. To facilitate navigation in the deluge of information, microblogging services allow users to insert hashtags starting with the "#" symbol (e.g., #HarryPotter) into their posts to indicate the context or key idea. Consisting of freely chosen keywords assigned to posts by users, hashtags help bring together relevant microblogs on a particular topic or event. In this way, hashtags provide a simple and natural way of organizing content and enhance information diffusion in microblog services. It has also been proven that hashtags are important for many applications in microblogs such as microblog retrieval [11], query expansion [2] and sentiment analysis [7,27,43]. However, not all microblog posts have hashtags created by their authors. Reported in a recent study, only about 11% of tweets were annotated with one or more hashtags [20]. Hence, the task of recommending hashtags for microblogs has become an important research topic and attracted much attention in recent years.

Existing approaches to hashtag recommendation range from classification and collaborative filtering to probabilistic models such as Naive Bayes and topic models. Most of these methods depend on sparse lexical features including bag-of-word (BoW) models and exquisitely designed patterns. However, feature engineering is labor-intensive and the *sparse* and *discrete* features cannot effectively encode semantic and syntactic information of words. On the other hand, neural models have shown great potential in learning effective representations recently, and have achieved state-of-the-art performance on various natural language processing tasks [6,39,41]. Among these methods, the long short-term memory (LSTM), a variant of recurrent neural network (RNN), is widely used because of its capability of capturing long-term dependencies in learning sequential representations [12,19,37].

We model the hashtag recommendation task as a multi-class classification problem. A typical approach is to adopt LSTM to learn the representation of a microblog post and then perform text classification based on this representation. However, a potential issue with this approach is that all the necessary information of the input post has to be compressed into a fixed-length vector. This may make it difficult to cope with long sentences [1]. One possible solution is to perform an average pooling operation over the hidden vectors of LSTM [5], but not all words in a microblog post contribute equally for hashtag recommendation. Inspired by the success of attention mechanism in computer vision and natural

* Corresponding author.
E-mail addresses: yli@nefu.edu.cn (Y. Li), tliu@ir.hit.edu.cn (T. Liu), jwhu@ir.hit.edu.cn (J. Hu), jingjiang@smu.edu.sg (J. Jiang).

https://doi.org/10.1016/j.neucom.2018.11.057
0925-2312/© 2018 Elsevier B.V. All rights reserved.

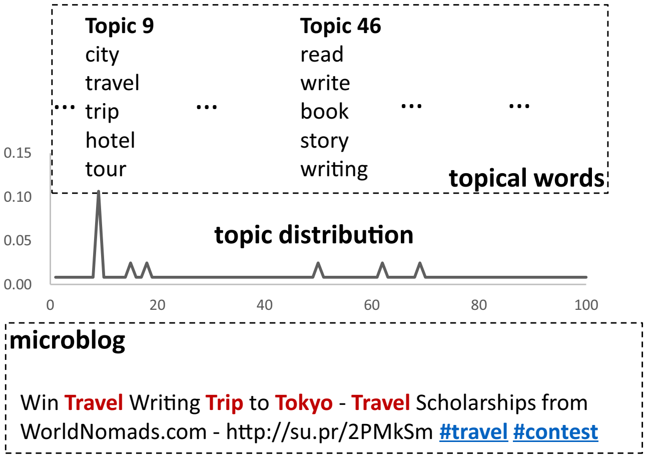


Fig. 1. Illustration of hashtags of a microblog post and its topic information.

language processing [1,31,36], we investigate the use of attention mechanism to automatically capture the most relevant words in a microblog to the recommendation task. In addition, it has been observed that most hashtags indicate the topics of a microblog post [10,13], as illustrated in Fig. 1. The example post has a high probability in the travel topic (Topic 9). With the topical word information, we can predict its hashtag #travel. Similarly, if we have the semantic information of the post, the topical word “travel” and “trip” will be emphasized which can also help in the hashtag recommendation task.

To solve the above problems, we propose a topical co-attention network (TCAN) that jointly models the content and the topic information of a microblog post simultaneously. TCAN has a natural symmetry between the content and the topic, in the sense that the content representation(s) are used to guide the topic attention and the topic representation is used to guide content attention. By incorporating the co-attention mechanism, our model captures the deep interactions between the local content representation and the global topic representation, and is able to learn effective representations of microblogs for hashtag recommendation. Experimental results on a large real microblogging dataset show that our model significantly outperforms various competitive baseline methods. Furthermore, the incorporation of topical co-attention mechanism gives more than 13.6% improvement in F1 score compared with standard LSTM method.

The main contributions of this paper can be summarized as follows:

- We propose a novel topical co-attention mechanism that jointly performs content-guided topic attention and topic-guided content attention simultaneously. To the best of our knowledge, this is the first work combines both topic and content attention with deep neural network into an integrated framework for this task.
- We test with different settings of the TCAN and find that both topic-guided content attention and content-guided topic attention can help the recommendation task, TCAN further improves the recommendation performance by combining them together.
- We thoroughly investigate several baseline methods including recent neural attention-based models for comparison. Experiment results show that our model outperforms various state-of-the-art methods.

2. Related work

We compare and relate our work with a few lines of recent works including hashtag recommendation, attention models and hashtag recommendation with attention based models in literature.

2.1. Hashtag recommendation

Hashtag recommendation has been given a lot of attention from academic in the past few years. The proposed methods can be roughly divided into two categories: content-based methods and collaborative filtering methods.

Content-based methods take different techniques to build semantic bridges between hashtags and messages, such as the TFIDF scheme [21,40,48], Bayes rules [34], word similarity information from WordNet [26] and topic translation methods [9,10,29]. Zangerle et al. [48] recommend hashtags based on tweets’ similarity. For a given tweet, they first retrieve its similar tweets and then rank the hashtags by their usage on the most similar tweets. Sedhai and Sun [40] represent each candidate hashtag as a feature vector and use pairwise learning to rank method to find the top ranked hashtags from the candidate set. Mazzia and Juett [34] apply a Naive Bayes model to estimate the maximum a posteriori probability of each hashtag class given the words of the tweet. Furthermore, Godin et al. [13] propose to incorporate topic models to learn the underlying topic assignment of language classified tweets, and suggest hashtags to a tweet based on the topic distribution. Under the assumption “hashtags and tweets are parallel description of a resource” that proposed by Liu et al. [29] and Ding et al. [10] try to integrate latent topical information into translation model. The model uses topic-specific word trigger to bridge the vocabulary gap between the words in tweets and hashtags [9,10].

Kywe et al. [24] propose a collaborative filtering model to incorporate user preferences in hashtag recommendation. Inspired by previous work, Wang et al. [44] propose a joint model based on topic modeling and collaborative filtering to take advantages of both local (the current microblog content and the user) and global (hashtag-related content and likeminded users’ usage preference) information. Zhao et al. [51] propose a hashtag-LDA recommendation approach that combines user profile-based collaborative and LDA-based collaborative filtering. They jointly model the relations between users, hashtags and words through latent topics.

Most of the above work is only based on text information. There have also been some attempts that combine text with other types of data. Zhang et al. [49] and Ma et al. [33] try to incorporate temporal information. Gong et al. [15] propose to model type of hashtag as a hidden variable into their DPMM (Dirichlet Process Mixture Models) based method. Li et al. [25] use a learning to rank algorithm to incorporate features built from topic enhanced embedding, tweet entity data, hashtag frequency, hashtag temporal data and tweet URL domain information. Gong et al. [16] combine textual and visual information together to recommend hashtags for multimodal microblog posts.

2.2. Attention-based models

Attention-based models have demonstrated success in a wide range of NLP tasks including sentence summarization [39], reading comprehension [18] and text entailment [38,42]. The basic idea of the attention mechanism is that it assigns a weight to each position in a lower-level of the neural network when computing an upper-level representation [1,31]. Bahdanau et al. [1] made the first attempt to use an attention-based neural machine translation (NMT) approach to jointly translate and align words. The model is based on the basic encoder-decoder model [6]. Differently, it encodes the input sentence into a sequence of vectors and chooses a subset of these vectors adaptively through the attention mechanism while generating the translation.

Specifically, Yin et al. [47] propose a two-way attention mechanism to project the paired inputs into a common representation space. Xiong et al. [45] introduce a Dynamic Co-attention Network (DCN) for question answering. The model consists of a co-attentive encoder that captures the interactions between the question and the document, as well as a dynamic pointing decoder that alternates between estimating the start and end of the answer spans. Recently, in the field of Visual Question Answering (VQA), a number of recent works have proposed attention models that generate spatial maps highlighting image regions relevant to answering the question. Lu et al. [30] present a novel co-attention model for VQA that jointly reasons about image and question attention.

2.3. Hashtag recommendation with attention-based models

More recently, there have been some attempts to use attention-based models for hashtag recommendation [14,32,50]. Gong and Zhang [14] propose an attention-based convolutional neural network, which incorporates a local attention channel and a global channel for hashtag recommendation. Zhang et al. [50] propose a co-attention network incorporating textual and visual information to recommend hashtags for multimodal tweets.

Motivated by the previous work [1,30,50], we propose a Topical Co-Attention Network to capture the deep interactions between the local content representations and the global topic of microblogs. The co-attention mechanism allows our model to attend to different position of content representations as well as different topical word representation. To the best of our knowledge, there is no work yet on employing both combines both topic and content attention with deep neural network into an integrated framework for this task.

3. Methodology

In this section, we will present the Topical Co-Attention Network for hashtag recommendation. We formulate the task of hashtag recommendation as a multi-class classification problem. Our model is mainly based on an LSTM neural network, given a microblog post, to predict its hashtags, we would like to process it sequentially and learn hidden representation at each position. Then perform text classification based on the representation of the microblog post.

We propose a Topical Co-Attention Network (TCAN) that jointly models content and topic information simultaneously. The model learns the content representations based on a bidirectional LSTM model and constructs a topical word matrix to represent the topic representation, and combine them through the co-attention mechanism. TCAN has a natural symmetry between the content and topic, in the sense that the content representations are used to guide the topic attention and the topic representation is used to guide content attention. We believe that, in this way, our model can capture the deep interactions of local content representations and the global topic representation of a microblog post. The overall model of TCAN is illustrated in Fig. 2. The model mainly consists of three parts, namely, *LSTM based sequence encoder, topic modeling*, and *topical co-attention*.

In the rest of this section, we will present each of these three parts in detail. A basis of all three parts is that each word is represented as a low dimensional, continuous and real-valued vector, also known as word embedding [3,35]. All the word vectors are stacked in a word embedding matrix $L_w \in \mathbb{R}^{d_{emb} \times |V|}$, where d_{emb} is the dimension of word vector and $|V|$ is vocabulary size. We pretrain the values of word vectors from text corpus with embedding learning algorithms to make better use of semantic and grammatical associations of words [35]. Given an input microblog s, we take the embeddings $\mathbf{x}_t \in \mathbb{R}^{d_{emb} \times 1}$ for each word in the microblog to obtain the first layer. Hence, a microblog post of length N is represented with a sequence of word vectors $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$. The notations and descriptions are shown in Table 1.

Table 1
Notations and descriptions.

	Description
V	Total number of unique words
C	Total number of hashtags
T	Total number of topics
N	Number of word in a post
M	Number of topical words for each topic
d_{hidden}	Dimension of LSTM hidden layer
d_{emb}	Dimension of word embedding
L_w	$\mathbb{R}^{d_{emb} \times \vert V \vert}$, word embedding matrix
$\mathbf{x}_t$	$\mathbb{R}^{d_{emb} \times 1}$, embedding of tth word in a specific post
$\mathbf{h}_t$	$\mathbb{R}^{d_{hidden} \times 1}$, hidden state of tth word in a specific post
$\mathbf{b}_k$	$\mathbb{R}^{d_{emb} \times 1}$, embedding of kth topical word in a specific topic
e_{tk}	Attention weight between $\mathbf{h}_t$ and $\mathbf{b}_k$ in a specific post
$\tilde{\mathbf{b}}_k$	Weighted summation of $\{\mathbf{h}_t\}_{t=1}^{N}$ that is relevant to $\mathbf{b}_k$
$\tilde{\mathbf{h}}_t$	Weighted summation of $\{\mathbf{b}_k\}_{k=1}^{M}$ that is relevant to $\mathbf{h}_t$
$\mathbf{a}^h$	The content attention weight vectors with length N
$\mathbf{a}^b$	The topic attention weight vectors with length M

3.1. LSTM based sequence encoder

LSTM is a special form of recurrent neural networks (RNNs), and is widely used to model sequence data. LSTM uses input gate, forget gate and output gate vectors at each position to control the passing of information along the sequence and thus improves the modeling of long-range dependencies [19].

Given a microblog $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, LSTM processes it sequentially. For each position $\mathbf{x}_t$, given the previous output $\mathbf{h}_{t-1}$ and cell state $\mathbf{c}_{t-1}$, an LSTM cell use the input gate $\mathbf{i}_t$, the forget gate $\mathbf{f}_t$ and the output gate $\mathbf{o}_t$ together to generate the next output $\mathbf{h}_t$ and cell state $\mathbf{c}_t$. The transition equations of LSTM are defined as follows:

$$\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W}^i\mathbf{x}_t + \mathbf{U}^i\mathbf{h}_{t-1} + \mathbf{b}^i) \\
\mathbf{f}_t &= \sigma(\mathbf{W}^f\mathbf{x}_t + \mathbf{U}^f\mathbf{h}_{t-1} + \mathbf{b}^f) \\
\mathbf{o}_t &= \sigma(\mathbf{W}^o\mathbf{x}_t + \mathbf{U}^o\mathbf{h}_{t-1} + \mathbf{b}^o) \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}^c\mathbf{x}_t + \mathbf{U}^c\mathbf{h}_{t-1} + \mathbf{b}^c) \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \qquad (1)
\end{aligned}$$

where $\odot$ stands for element-wise multiplication, σ is the sigmoid function, all $\mathbf{W} \in \mathbb{R}^{d_{hidden} \times l}$ and $\mathbf{U} \in \mathbb{R}^{d_{hidden} \times d_{hidden}}$ are weight matrices, all $\mathbf{b} \in \mathbb{R}^{d_{hidden}}$ are bias vectors.

Bidirectional LSTM is an extension of traditional LSTM that enables the hidden states to capture both historical and future context information. In problems where all time steps of the input sequence are available, Bidirectional LSTM models text semantics both from forward and backward. For a microblog $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, the forward LSTM reads sequence from $\mathbf{x}_1$ to $\mathbf{x}_N$ and the backward LSTM reads sequence from $\mathbf{x}_N$ to $\mathbf{x}_1$, and similarly processes the sequence according to Eq. (1). Then we concatenate the forward hidden state $\overrightarrow{\mathbf{h}_t}$ and backward hidden state $\overleftarrow{\mathbf{h}_t}$, i.e., $\mathbf{h}_t = [\overrightarrow{\mathbf{h}_t}; \overleftarrow{\mathbf{h}_t}]$, where the $[\cdot; \cdot]$ denotes concatenation operation. Finally, the $\mathbf{h}_t$ summarizes the information of the whole sequence centered around $\mathbf{x}_t$.

3.2. Topic modeling

Topic models have been a powerful technique for finding useful structures in a collection of documents. In topic models, it is assumed that a document is generated by a mixture of topics, each of which is a distribution over words in the vocabulary. By fitting

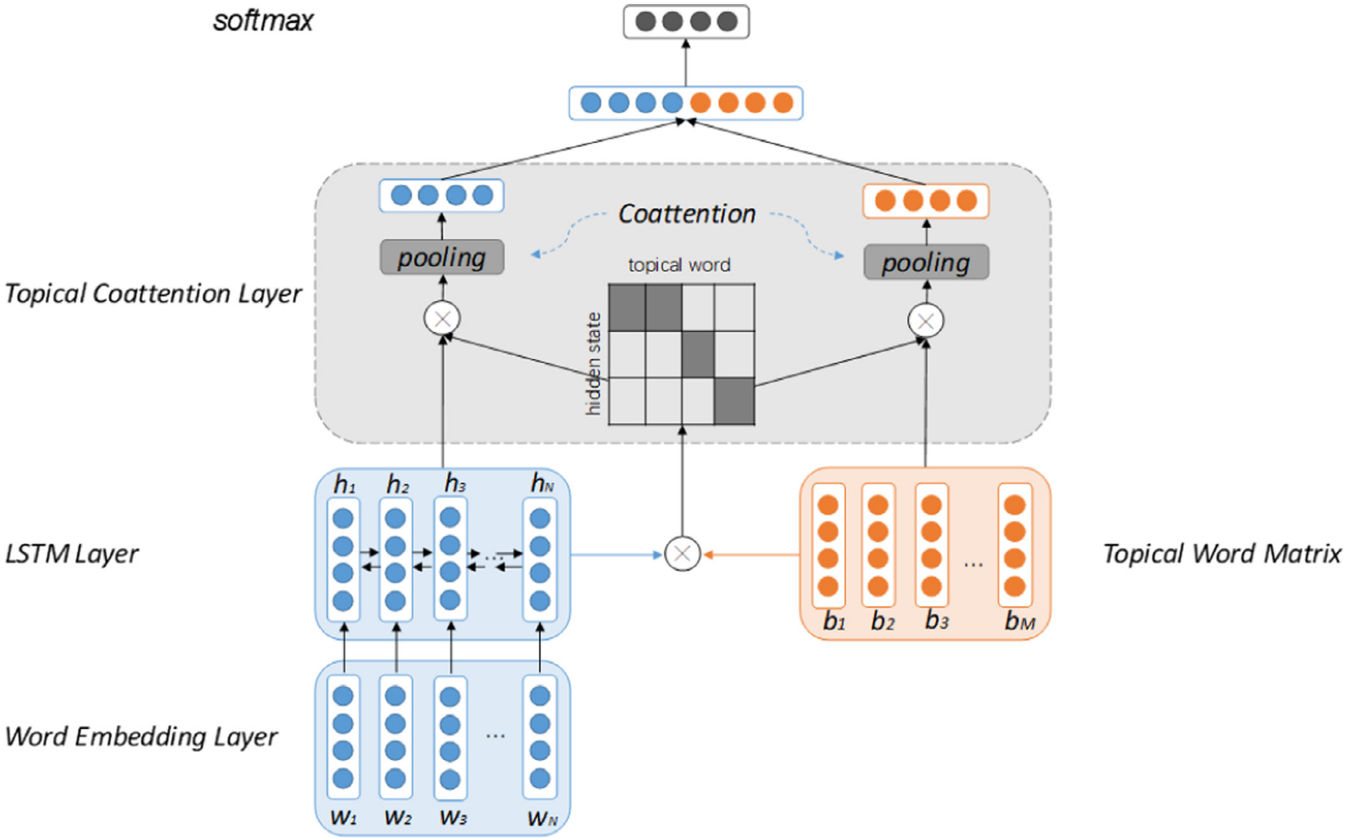


Fig. 2. The graphical illustration of the proposed Topical Co-Attention Network (TCAN).

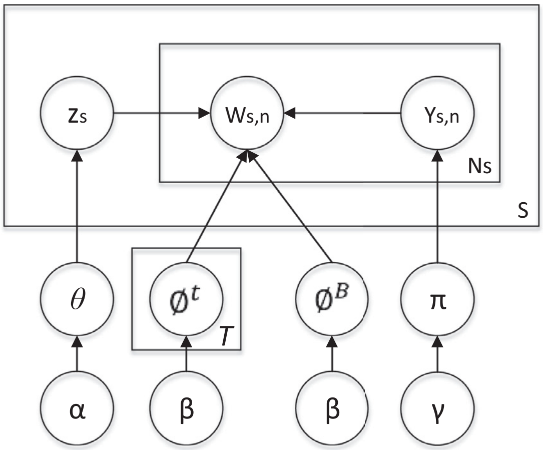


Fig. 3. Graphical model of Twitter LDA.

the models, we can represent each document through the learned topics as well as understand topics in the corpus through the most probable words of each topic. These topical words represent the main semantic information of the topic. Therefore, we propose to incorporate such information of microblogs as prior knowledge into the LSTM based neural network.

To model the topics of microblogs, we choose to use Twitter LDA [52], which is the state-of-the-art topic model for short texts. Unlike the original setting in standard LDA [4], where each word has a topic label, in Twitter LDA, a single microblog post is more likely to talk about one topic. Fig. 3 gives the graphical model of the Twitter LDA. Assuming that there are T topics, each topic is represented by a word distribution. Let ϕ^t denotes the word distribution for topic t and ϕ^B the word distribution for background words. π is a Bernoulli distribution that governs the choice between background words and topic words. To generate a microblog post, the model first chooses a topic, then chooses a bag of words one by one based on the chosen topic or the background topic. The parameters of Twitter LDA can be estimated by the collapsed Gibbs sampling algorithm [52].

After topic modeling, the model assigns a topic z to each microblog post s. Then we can extract M most probable words as the semantic information of topic z. Hence, the topic information of the post s is represented by a sequence of embedding vectors of topical words $[\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_M]$. M is the number of topical words we choose for each topic, which is pre-defined.

3.3. Topical co-attention

The co-attention allows our model to attend to different position of the content representation as well as different topical word representations. Given all hidden states $[\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N]$ and the topical word embedding vectors $[\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_M]$ learned from topic modeling, the topical co-attention layer outputs a continuous context vector *vec* for each microblog post s.

Now we introduce the topical co-attention layer in detail. First, let $\mathbf{E} \in \mathbb{R}^{N \times M}$ denote an affinity matrix. For each row $t \in \{1, 2, \dots, N\}$ and each column $k \in \{1, 2, \dots, M\}$, we define each entry e_{tk} in matrix $\mathbf{E} \in \mathbb{R}^{N \times M}$ as follows:

$$e_{tk} = \mathbf{h}_t^\top \mathbf{W}^{hb} \mathbf{b}_k \tag{2}$$

where $\mathbf{W}^{hb} \in \mathbb{R}^{d_{hidden} \times d_{emb}}$ is a trainable weight matrix. e_{tk} is the attention weight between the hidden state $\mathbf{h}_t$ and the topical word embedding $\mathbf{b}_k$, standing for the similarity between $\mathbf{h}_t$ and $\mathbf{b}_k$.

3.3.1. Content-guided topic attention

For the hidden state of a word $\mathbf{h}_t$, the relevant semantics in the global topic information is identified and computed with Eq. (3).

$$\tilde{\mathbf{h}}_t = \sum_{k=1}^{M} \mathbf{a}_k^b \mathbf{b}_k \tag{3}$$

where $\tilde{\mathbf{h}}_t$ is a weighted summation of $\{\mathbf{b}_k\}_{k=1}^M$, $\mathbf{a}_k^b$ is the attention weight of $\mathbf{b}_k$ and can be computed as follows:

$$\mathbf{a}_k^b = \frac{exp(e_{tk})}{\sum_{j=1}^M exp(e_{tj})} \tag{4}$$

In this step, the content representation(s) are used to guide the topic attention vectors $\mathbf{a}^b$ and learn a new topic representation. Intuitively, the information in $\{\mathbf{b}_k\}_{k=1}^M$ that is relevant to $\mathbf{h}_t$ will be selected and represented as $\tilde{\mathbf{h}}_t$.

3.3.2. Topic-guided content attention

Similarly, the content in $\{\mathbf{h}_t\}_{t=1}^N$ that is relevant to $\mathbf{b}_k$ will be selected and represented as $\tilde{\mathbf{b}}_k$:

$$\tilde{\mathbf{b}}_k = \sum_{t=1}^{N} \mathbf{a}_t^h \mathbf{h}_t \tag{5}$$

where $\mathbf{a}_t^h$ is the attention weight of $\mathbf{h}_t$ and can be computed as follows:

$$\mathbf{a}_t^h = \frac{exp(e_{tk})}{\sum_{i=1}^N exp(e_{ik})} \tag{6}$$

In this step, the topic representation is used to guide the content attention vectors $\mathbf{a}^h$ and learn a new content representation $\tilde{\mathbf{b}}_k$.

Next, our model converts the resulting vectors obtained above (Eqs. (3)–(6)) to a fixed length vector *vec* with pooling and feeds it to the final classifier. Specifically, we perform both average and max pooling, and concatenate all these vectors to form the final fixed length vector *vec, vec* is calculated as follows:

$$\begin{aligned} \mathbf{v}_{average_h} &= \textstyle\sum_{t=1}^N \frac{\tilde{\mathbf{h}}_t}{N}, \mathbf{v}_{max_h} = \max_{t=1}^N \tilde{\mathbf{h}}_t \\ \mathbf{v}_{average_b} &= \textstyle\sum_{k=1}^M \frac{\tilde{\mathbf{b}}_k}{M}, \mathbf{v}_{max_b} = \max_{k=1}^M \tilde{\mathbf{b}}_k \end{aligned} \tag{7}$$

$$vec = [\mathbf{v}_{average_h}; \mathbf{v}_{average_b}; \mathbf{v}_{max_h}; \mathbf{v}_{max_b}] \tag{8}$$

We then feed the output vector *vec* to a linear layer whose output length is the number of hashtags. Then a softmax layer is added to output the probability distributions of all candidate hashtags. The softmax function is calculated as follows, where C is the number of hashtag categories:

$$softmax(c_i) = \frac{exp(c_i)}{\sum_{i'=1}^C exp(c_{i'})} \tag{9}$$

3.4. Model training

We train our model in a supervised manner by minimizing the cross-entropy error of the hashtag classification. The loss function is given below:

$$\mathcal{J} = -\sum_{s \in S} \sum_{t \in tags(s)} \log p(t|s) \tag{10}$$

where S stands for all training instances, $tags(s)$ is the hashtag collection for microblog s.

Table 2
Statistics of the dataset, Nt (avg) is the average number of hashtags in the dataset.

# Tweets	# Hashtags	Vocabulary size	Nt (avg)
600,000	27,720	337,245	1.308

4. Experiments

We apply the proposed method to the task of hashtag recommendation to evaluate the performance. In this section, we design experiments to answer the following research questions: (*i*) How much can neural network help for hashtag recommendation compared with traditional baseline methods? (*ii*) Does attention mechanism help on top of neural network for this task? (*iii*) Does the Topical Co-Attention Network perform better than other attention based neural networks?

4.1. Dataset

Our dataset is constructed from a large Twitter dataset spanning the second half of 2009 [46]. We collect a dataset with 185,391,742 tweets from October to December. Among them, there are 16,744,189 tweets including hashtags annotated by users. We randomly select 500,000 tweets as training set, 50,000 tweets as development and test set respectively. Finally, we get 337,245 unique words and 27,720 hashtags. The statistics of our dataset is shown in Table 2.

4.2. Experimental settings

4.2.1. Baseline methods

For comparison, we consider the following baseline methods:

- *LDA:* We use the LDA based method proposed by Krestel et al. [23] to recommend hashtags.
- *SVM:* We build a multi-class SVM classification model [17] with LibSVM. The feature we use are word embedding features with 300 dimension. We believe that comparing to Bag-of-words, word embedding features can capture deep semantic information of the microblog posts. SVM parameters are chosen by grid search on the development set.
- *TTM:* The topical translation model is proposed by Ding et al. [9] for hashtag extraction. We implement their method for evaluating it on the corpus constructed in this work.
- *LSTM:* We regard the last hidden vector from LSTM as the microblog representation. Then we feed it to a linear layer whose output length is the number of hashtags. Finally, a softmax layer is added to output the probability distributions of all candidate hashtags.
- *BLSTM:* BLSTM is similar to LSTM, except that we adopt the Bidirectional LSTM to learn the representation of a microblog.
- *AVG-BLSTM:* We perform an average pooling operation on the hidden vectors at each position of LSTM that processes a post, and use the result as the representation of that post.
- *TAB-BLSTM:* The topical attention-based LSTM model is proposed by Li et al. [28]. This method bears similarity to our method in that it also incorporates topic information (topic distribution) into the neural network. The difference is that it only focuses on content attention.
- *VAB-BLSTM:* In this model, we use the last hidden vector from the LSTM as the global representation of that post and incorporate attentions to measure the interactions between each word and the global representation. This method is a degenerate version of TAB-BLSTM [28], we refer to it as vanilla attention based BLSTM, or VAB-BLSTM for short.

We refer to our proposed model as the Topical Co-Attention Network (Fig. 2), or TCAN for short. To evaluate the effectiveness of the co-attention mechanism, we also compare two degenerate version of TCAN: TCAN$_{topic}$ and TCAN$_{content}$.

- *TCAN$_{topic}$*: TCAN with content-guided topic attention only, i.e. $vec = [\mathbf{v}_{average_h}; \mathbf{v}_{max_h}]$ in Eq. (8).
- TCAN$_{content}$: TCAN with topic-guided content attention only, i.e. $vec = [\mathbf{v}_{average_b}; \mathbf{v}_{max_b}]$ in Eq. (8).

4.2.2. Evaluation metrics

We use hashtags annotated by users as the golden set. To evaluate the performance, we use precision (P), recall (R), and F1-score (F) as the evaluation metrics. Precision means the percentage of “tags truly assigned” among “tags assigned by system”. Recall denotes that “tags truly assigned” among “tags manually assigned”. F1-score is the average of Precision and Recall. The same settings are adopted by previous work [9,10,15].

$$P = \frac{\text{number of tags truly assigned}}{\text{number of tags assigned by system}},$$

$$R = \frac{\text{number of tags truly assigned}}{\text{number of tags manually assigned}},$$

$$F = \frac{2 \times P \times R}{P + R}. \tag{11}$$

4.2.3. Experimental setup

We perform hashtag recommendation as follows. Suppose given an unlabeled dataset, we first train our model on training data, and save the model which has the best performance on the validate dataset. For the microblog of the unlabeled data, we will encode the microblog post through our proposed model and then perform the softmax classification. We train all the neural models and our proposed model TCAN with the sentences of length up to 50 words. For each of the above models, the dimension of all the hidden states in the LSTMs is set to 500 and the dimension of word embeddings is 300, unless otherwise noted. We use a minibatch stochastic gradient descent (SGD) algorithm together with the Adam method to train each model [22]. The hyperparameters $\beta 1$ is set to 0.9 and $\beta 2$ set to 0.999 for optimization. The learning rate is set to be 0.001. The batch size is set to be 100. The network was used for training for 20 epochs with early stopping. For our models, we tested with different numbers of LDA topic size T and different number of topical words M, we found $T = 200$ and $M = 30$ is an optimal setting for TCAN.

For both our models and the baseline methods, we use the validation data to tune the hyperparameters, and report the results of the test data in the same setting of hyperparameters. Furthermore, the word embeddings used in all methods are pre-trained from the original twitter data released by Yang and Leskovec [46] with the word2vec toolkit [35].

4.3. Comparison to other methods

In Table 3, we compare the results of our method and the state-of-the-art discriminative and generative methods on the dataset. To summarize, we find the following:

First, considering the comparison between the SVM and LDA methods, we observe that SVM performs much better than LDA. This indicates that the embedding features capture more semantic information than bag-of-words (BoW).

Secondly, in comparing the traditional baseline methods such as LDA, SVM and TTM to the neural models, we observe that the neural models achieve a more than 50% relative improvement of the F1-score. For the task of hashtag recommendation, the key ingredient is learning the representation of the microblog. This indicates that neural networks are effective in learning the semantic information of microblog posts and can improve the performance considerably.

Table 3
Evaluation results of different methods for hashtag recommendation. All improvements obtained by TCAN over other methods are statistically significant within a 0.99 confidence interval using the *t*-test.

Methods	Precision	Recall	F1-score
LDA	0.098	0.078	0.087
SVM	0.238	0.203	0.219
TTM	0.324	0.280	0.300
LSTM	0.470	0.404	0.434
BLSTM	0.478	0.411	0.442
AVG-BLSTM	0.475	0.408	0.439
VAB-BLSTM	0.492	0.423	0.455
TAB-BLSTM	0.506	0.437	0.469
TCAN	**0.532**	**0.458**	**0.493**

Table 4
Evaluation results of TCAN and its two degenerate models for hashtag recommendation. TCAN$_{topic}$ stands for TCAN with content-guided topic attention only, TCAN$_{content}$ stands for TCAN with topic-guided content attention only.

Methods	Precision	Recall	F1-score
TCAN$_{topic}$	0.486	0.418	0.449
TCAN$_{content}$	0.522	0.450	0.484
TCAN	**0.532**	**0.458**	**0.493**

Table 5
The influence of number of training data of TCAN.

Training data	Precision	Recall	F1-score
100 K (20%)	0.377	0.322	0.347
200 K (40%)	0.438	0.376	0.405
300 K (60%)	0.472	0.405	0.436
400 K (80%)	0.500	0.431	0.463
500 K (100%)	0.532	0.458	0.493

Thirdly, observing the comparisons of the LSTM, BLSTM, AVG-BLSTM and VAB-BLSTM, it is clear that the attention mechanism is useful to learn the representation of the microblog post.

Finally and most importantly, both TAB-BLSTM and TCAN outperform VAB-BLSTM significantly, which shows the topic information is useful for this task. Moreover, our model TCAN achieves better results than TAB-BLSTM. TAB-BLSTM bears similarity to our method in that it also incorporates topic information into the neural network. The difference is that TCAN constructs a topical word matrix to represent the topic information and has a natural symmetry between the content and topic. The properties of this aspect of our proposed model have been proven to be effective by observing the results in Table 3.

We compare TCAN with its two degenerate models TCAN$_{topic}$ and TCAN$_{content}$ in Table 4. We observe that TCAN$_{content}$ outperforms TCAN$_{topic}$ significantly. We hypothesize that this is because TCAN$_{topic}$ model predicts the hashtags mainly based on the topical word representation and TCAN$_{content}$ mainly based on the microblog representation. TCAN is able to improve the recommendation performance over TCAN$_{content}$ and TCAN$_{topic}$ by incorporating the topical co-attention mechanism.

Table 5 shows the influence of the number of training data elements. Based on the results, we observe the performance of TCAN increases when larger training data sets are used. The results also demonstrate that our proposed method achieves significant better performance than the traditional baseline methods even with only 20% of the training data.

Many microblog posts have more than one hashtags. Therefore, we also evaluate the top k recommendation results of different

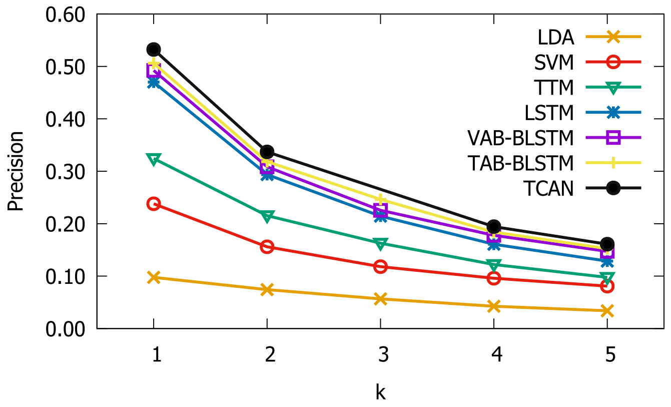


Fig. 4. Precision with recommended hashtags range from 1 to 5.

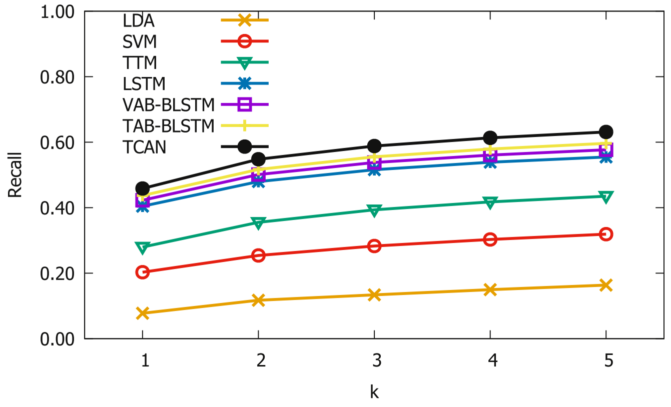


Fig. 5. Recall values with recommended hashtags range from 1 to 5.

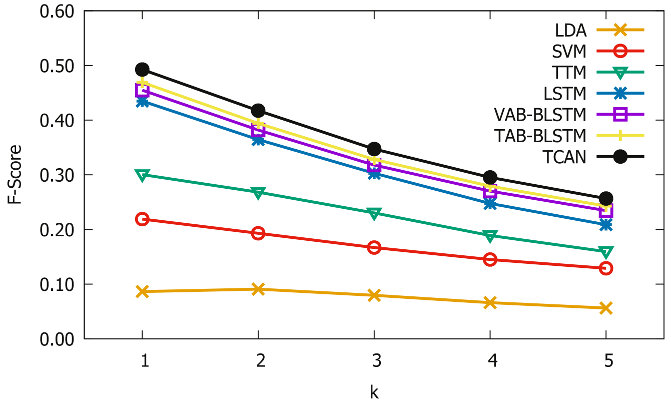


Fig. 6. F1 values with recommended hashtags range from 1 to 5.

methods. Figs 4–6 show the precision, recall, and F1 curves of LDA, SVM, TTM, LSTM, VAB-BLSTM, TAB-BLSTM and TCAN on the test data respectively. Each point of a curve represents the extraction of a different number of hashtags, ranging from 1 to 5. We observe that although the precision and F1-score of TCAN decreases when the number of hashtags is larger, TCAN still outperforms the other methods. In addition, the relative improvement on extracting only one hashtag is higher than that on more than one hashtags, showing that it is more difficult to recommend hashtags for a microblog post with more than one hashtags.

4.4. Parameter sensitive analysis

We further investigate the effect of hyperparameters to the performance. It is well accepted that a good word embedding is crucial to composing a powerful text representation at a higher level. First, we would like to study the effects of different word embeddings. Table 6 shows the precision, recall and F1-score when we vary the dimension of word embeddings in TCAN. This indicates that a larger dimension of word embedding is more effective for this task.

Table 6
Precision, recall and F1 of TCAN with different dimension of word embeddings when the number of topics is 200 and number of topical words is 30.

Methods	Precision	Recall	F1-score
Emb50	0.479	0.412	0.443
Emb100	0.483	0.415	0.447
Emb200	0.502	0.432	0.464
Emb300	0.532	0.458	0.493

Table 7
Precision, recall and F1 of TCAN with different number of topics when the dimension of word vectors is set to be 300 and number of topical words is 30.

# Topics	Precision	Recall	F1-score
50	0.512	0.441	0.474
100	0.524	0.452	0.486
150	0.527	0.455	0.489
200	0.532	0.458	0.493
250	0.529	0.456	0.490

Table 8
Precision, recall and F1 of TAB-LSTM with different number of topical words selected when the dimension of word vectors is set to be 300 and the number of topics is 200.

#Topical words	Precision	Recall	F1-score
10	0.509	0.439	0.472
20	0.520	0.448	0.482
30	0.532	0.458	0.493
40	0.523	0.452	0.485
50	0.515	0.443	0.477

Recall that in the part of topic modelling, there are two hyperparameters control the topic information incorporated, the number of topics and the number of topical words for each topic. Next, we vary the number of topics K from 50 to 250 with a gap of 50 while fixing the other parameters. Results in Table 7 show that the performance of the precision, recall and F1-score improves when K increases. The results do not change much when K is between 100 and 250. When K is equal to 200, TCAN achieves the best results.

Table 8 shows the results when we vary the number of topical words M from 10 to 50. We can observe that an optimal setting of M is 30. We find a larger M does not help in this task. This is because a large number of topical words may introduce more topic information as well as noise information into the model.

4.5. Qualitative analysis

We also conduct qualitative analysis of our results through case studies.

Example 1: *A 10-year-old battling cancer recently scored her dream.* **#cancer #beatcancer**

In Fig. 7, we show the attention heat maps learned by TCAN and TAB-BLSTM of an example microblog post. In this example, the hashtag #cancer is correctly recommended by both TCAN and TAB-BLSTM. We observe that the word “cancer” is not only a word in microblog post, but also a topical word. The topical co-attention mechanism gives the word “cancer” a high weight over the topic representations and the topical word “cancer” a high weight over

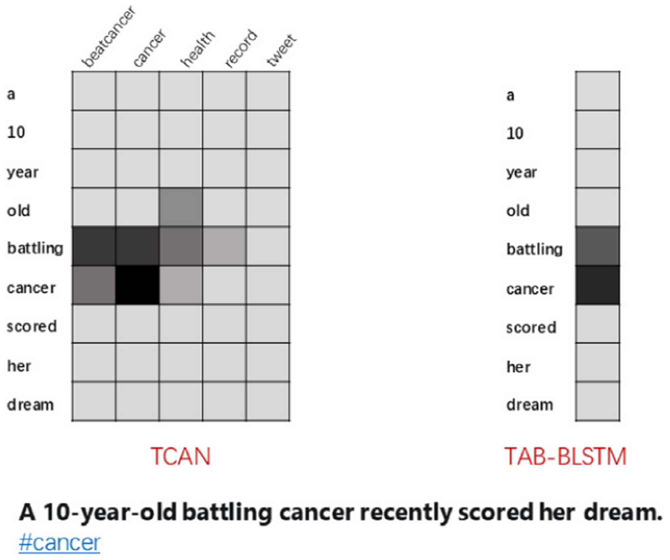


Fig. 7. Attention heat maps from TCAN and TAB-BLSTM for Example 1.

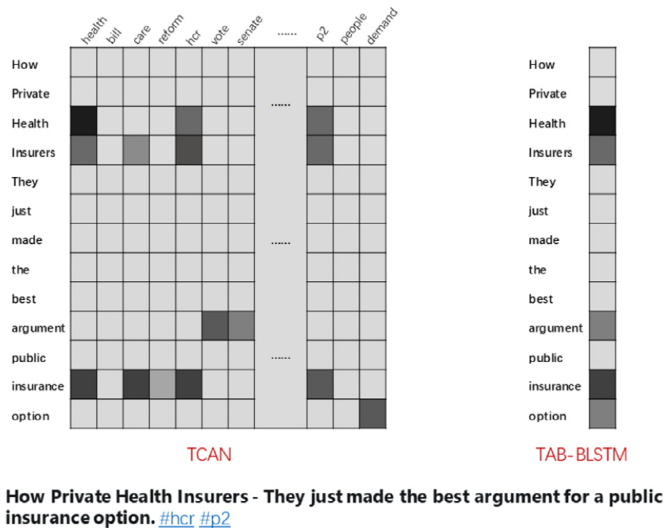


Fig. 8. Attention heat maps from TCAN and TAB-BLSTM for Example 2.

Table 9
Top 30 topical words for the topic of the microblog post in Example 2.

Topical words
health, bill, care, reform, hcr, vote, senate, healthcare, insurance, public, tcot, obamaoption, house, gop, support, congress, publicoption, abortion, rep, sign, reid, plan, hc, Americans, obamacare, tax, gd, p2, people, demand

the content representations. While TAB-BLSTM also gives the word “cancer” a high weight based on its topic distribution.

Example 2: *How Private Health Insurers – They just made the best argument for a public insurance option.* **#hcr #p2**

In the second example, the attention heat maps learned by TCAN and TAB-BLSTM are shown in Fig. 8. We observe that, unlike the first example, the hashtag “hcr” and “p2” do not appear in the microblog post. The hashtag #hcr and #p2 are both correctly predicted by TCAN but #p2 is not recommended by TAB-BLSTM. Although “p2” is not a high probable topical word in Table 9, the co-attention mechanism of TCAN gives “p2” a high weight. However, in the case of TAB-BLSTM, only the hashtag #hcr is recommended, as it is highly related to the topic. This demonstrates that the topic information incorporated in TCAN is richer than that in TAB-BLSTM.

5. Conclusion

In this article, we propose a Topical Co-Attention Network (TCAN) for the task of hashtag recommendation. The Topical Co-Attention Network incorporates topic modeling into the LSTM architecture through a co-attention mechanism and takes over the advantages of the both. We design experiments to evaluate our model against several state-of-the-art models. By comparing with traditional baseline methods including LDA, SVM and TTM, we found that our neural models significantly helped in this task. In addition, we found it beneficial to incorporate the co-attention mechanism by comparing our model with a recent work topical attention-based LSTM. Finally, we also tested with different settings of TCAN and found that both topic-guided content attention and content-guided topic attention can help the recommendation task, $\text{TCAN}_{content}$ (TCAN with topic-guided content attention only) outperforms TCAN_{topic} (TCAN with content-guided topic attention only) significantly. TCAN is able to further improve the recommendation performance by combining them together in the topical co-attention mechanism. The overall experimental results show that our model outperforms competitive baseline methods effectively.

There are a few directions we would like to explore in the future. First, the present work does not consider the use of other types of data in microblogs for hashtag recommendation. In the future, other types of data such as user information and background knowledge can be incorporated into the model. Second, previous work [8] demonstrated posts that published around the same time are more likely to have the same topic. We will consider the temporal information of posts in the future. All these issues will be left as our future works.

Acknowledgment

We thank the anonymous reviewers for their constructive comments, and gratefully acknowledge the support of the National Natural Science Foundation of China (NSFC) via Grant 61806049, the special funds of Central College Basic Scientific Research Bursary via Grant 2572018BH04 and the Natural Science Foundation of Heilongjiang Province of China via Grant F2018001.

References

[1] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: International Conference on Learning Representations, 2015.
[2] A. Bandyopadhyay, K. Ghosh, P. Majumder, M. Mitra, Query expansion for microblog retrieval, Int. J. Web Sci. 1 (4) (2012) 368–380.
[3] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, J. Mach. Learn. Res. 3 (2003) (Feb) 1137–1155.
[4] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) (Jan) 993–1022.
[5] Y.-L. Boureau, N. Le Roux, F. Bach, J. Ponce, Y. LeCun, Ask the locals: multi--way local pooling for image recognition, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2651–2658.
[6] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2014, Doha, Qatar pp. 1724–1734. http://dl.acm.org/citation.cfm?id=2390524.2390599.
[7] D. Davidov, O. Tsur, A. Rappoport, Enhanced sentiment learning using twitter hashtags and smileys, in: Proceedings of the 23rd international conference on computational linguistics: posters, Association for Computational Linguistics, 2010, pp. 241–249.
[8] Q. Diao, J. Jiang, F. Zhu, E.-P. Lim, Finding bursty topics from microblogs, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, 1, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 536–544.

[9] Z. Ding, X. Qiu, Q. Zhang, X. Huang, Learning topical translation model for microblog hashtag suggestion, in: Proceeding of the IJCAI, Citeseer, 2013.

[10] Z. Ding, Q. Zhang, X. Huang, Automatic hashtag recommendation for microblogs using topic-specific translation model, in: Proceedings of COLING 2012: Posters, The COLING 2012 Organizing Committee, Mumbai, India, 2012, pp. 265–274. http://www.aclweb.org/anthology/C12-2027.

[11] M. Efron, Hashtag retrieval in a microblogging environment, in: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2010, pp. 787–788.

[12] F.A. Gers, J. Schmidhuber, F.A. Cummins, Learning to forget: continual prediction with LSTM., Neural Comput. 12 (10) (2000) 2451–2471. http://dblp.uni-trier.de/db/journals/neco/neco12.html#GersSC00.

[13] F. Godin, V. Slavkovikj, W. De Neve, B. Schrauwen, R. Van de Walle, Using topic models for Twitter hashtag recommendation, in: Proceedings of the 22nd International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, 2013, pp. 593–596.

[14] Y. Gong, Q. Zhang, Hashtag recommendation using attention-based convolutional neural network, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI, New York, NY, USA, 2016, pp. 2782–2788. http://www.ijcai.org/Abstract/16/395. 9–15 July 2016.

[15] Y. Gong, Q. Zhang, X. Huang, Hashtag recommendation using dirichlet process mixture models incorporating types of hashtags, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 401–410. http://aclweb.org/anthology/D15-1046.

[16] Y. Gong, Q. Zhang, X. Huang, Hashtag recommendation for multimodal microblog posts, Neurocomputing 272 (2018) 170–177, doi:10.1016/j.neucom.2017.06.056.

[17] M.A. Hearst, S.T. Dumais, E. Osman, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Appl. 13 (4) (1998) 18–28.

[18] K.M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, in: Proceedings of the Advances in Neural Information Processing Systems, 2015, pp. 1693–1701.

[19] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[20] L. Hong, A. Ahmed, S. Gurumurthy, A.J. Smola, K. Tsioutsiouliklis, Discovering geographical topics in the twitter stream, in: Proceedings of the 21st International Conference on World Wide Web, ACM, New York, NY, USA, 2012, pp. 769–778, doi:10.1145/2187836.2187940.

[21] M. Jeon, S. Jun, E. Hwang, Hashtag recommendation based on user tweet and hashtag classification on twitter, in: Y. Chen, W.-T. Balke, J. Xu, W. Xu, P. Jin, X. Lin, T. Tang, E. Hwang (Eds.), Web-Age Information Management, Springer International Publishing, Cham, 2014, pp. 325–336.

[22] D. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015. arXiv: 1412.6980.

[23] R. Krestel, P. Fankhauser, W. Nejdl, Latent dirichlet allocation for tag recommendation, in: Proceedings of the Third ACM Conference on Recommender Systems, ACM, 2009, pp. 61–68.

[24] S.M. Kywe, T.-A. Hoang, E.-P. Lim, F. Zhu, On recommending hashtags in Twitter networks, in: Social Informatics, Springer, 2012, pp. 337–350.

[25] Q. Li, S. Shah, A. Nourbakhsh, X. Liu, R. Fang, Hashtag recommendation based on topic enhanced embedding, tweet entity data and learning to rank, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2016, pp. 2085–2088, doi:10.1145/2983323.2983915.

[26] T. Li, Y. Wu, Y. Zhang, 2012. Twitter Hash Tag Prediction Algorithm. Icomp.

[27] Y. Li, J. Jiang, T. Liu, X. Sun, Personalized microtopic recommendation with rich information, in: Proceedings of the Social Media Processing: 4th National Conference, SMP, Springer, Guangzhou, China, 2015, pp. 1–14.

[28] Y. Li, T. Liu, J. Jiang, L. Zhang, Hashtag recommendation with topical attention-based lstm., in: N. Calzolari, Y. Matsumoto, R. Prasad (Eds.), COLING, ACL, 2016, pp. 3019–3029. http://dblp.uni-trier.de/db/conf/coling/coling2016.html#LiLJZ16.

[29] Z. Liu, X. Chen, M. Sun, A simple word trigger method for social tag suggestion, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Edinburgh, Scotland, UK, 2011, pp. 1577–1588. http://www.aclweb.org/anthology/D11-1146.

[30] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical Question-image Co-attention for Visual Question Answering, in: NIPS'16 Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016, pp. 289–297.

[31] T. Luong, H. Pham, C.D. Manning, 2015. Effective Approaches to Attention-based Neural Machine Translation, 1412–1421. http://aclweb.org/anthology/D15-1166.

[32] J. Ma, C. Feng, G. Shi, X. Shi, H. Huang, Temporal enhanced sentence-level attention model for hashtag recommendation, CAAI Trans. Intell. Technol. 3 (2) (2018) 95–100, doi:10.1049/trit.2018.0012.

[33] Z. Ma, A. Sun, Q. Yuan, G. Cong, Tagging your tweets: a probabilistic modeling of hashtag annotation in twitter, in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, ACM, 2014, pp. 999–1008.

[34] A. Mazzia, J. Juett, Suggesting hashtags on twitter, in: EECS 545m, Machine Learning, Computer Science and Engineering, University of Michigan, 2009.

[35] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.

[36] V. Mnih, N. Heess, A. Graves, et al., Recurrent models of visual attention, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 2204–2212.

[37] 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. 24 (4) (2016) 694–707.

[38] T. Rocktäschel, E. Grefenstette, K.M. Hermann, T. Kočiskỳ, P. Blunsom, Reasoning about entailment with neural attention, in: Proceedings of ICLR, 2016.

[39] A.M. Rush, S. Chopra, J. Weston, 2015. A Neural Attention Model for Abstractive Sentence Summarization, 379–389. http://aclweb.org/anthology/D15-1044.

[40] S. Sedhai, A. Sun, Hashtag recommendation for hyperlinked tweets, in: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM, 2014, pp. 831–834.

[41] D. Tang, B. Qin, T. Liu, Document modeling with gated recurrent neural network for sentiment classification, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1422–1432.

[42] S. Wang, J. Jiang, Learning natural language inference with lstm, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 1442–1451. http://www.aclweb.org/anthology/N16-1170.

[43] X. Wang, F. Wei, X. Liu, M. Zhou, M. Zhang, Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach, in: Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, 2011, pp. 1031–1040.

[44] Y. Wang, J. Qu, J. Liu, J. Chen, Y. Huang, What to tag your microblog: hashtag recommendation based on topic analysis and collaborative filtering, in: L. Chen, Y. Jia, T. Sellis, G. Liu (Eds.), Web Technologies and Applications, Springer International Publishing, Cham, 2014, pp. 610–618.

[45] C. Xiong, V. Zhong, R. Socher, 2016. Dynamic Coattention Networks for Question Answering. CoRR abs/1611.01604.1611.01604. http://arxiv.org/abs/1611.01604.

[46] J. Yang, J. Leskovec, Patterns of temporal variation in online media, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, ACM, 2011, pp. 177–186.

[47] W. Yin, H. Schütze, B. Xiang, B. Zhou, ABCNN: Attention-based convolutional neural network for modeling sentence pairs, Computer Science (2015).

[48] E. Zangerle, W. Gassler, G. Specht, Recommending#-tags in twitter, in: Proceedings of the Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop, 730, 2011, pp. 67–78.

[49] Q. Zhang, Y. Gong, X. Sun, X. Huang, Time-aware personalized hashtag recommendation on social media, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 203–212. http://www.aclweb.org/anthology/C14-1021.

[50] Q. Zhang, J. Wang, H. Huang, X. Huang, Y. Gong, Hashtag recommendation for multimodal microblog using co-attention network, in: Proceedings of IJCAI, 2017.

[51] F. Zhao, Y. Zhu, H. Jin, L.T. Yang, A personalized hashtag recommendation approach using LDA-based topic model in microblog environment, Fut. Gener. Comput. Syst. 65 (C) (2016) 196–206.

[52] W.X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, X. Li, Comparing twitter and traditional media using topic models, in: Proceedings of the 33rd European Conference on Advances in Information Retrieval, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 338–349.

Yang Li received her Ph.D. degree in July 2017 from the Department of Computer Science, Harbin Institute of Technology, Harbin, China. She is an assistant professor in the College of Information and Computer Engineering at the Northeast Forestry University. Her current research interests include natural language processing, information retrieval, and social media analysis.

Ting Liu received his Ph.D. degree in 1998 from the Department of Computer Science, Harbin Institute of Technology, Harbin, China. He is a Full Professor in the Department of Computer Science, and the Director of the Research Center for Social Computing and Information Retrieval (HIT-SCIR) from Harbin Institute of Technology. His research interests include information retrieval, natural language processing, and social media analysis.

Jingwen Hu is an undergraduate student in the Department of Computer Science, Harbin Institute of Technology, Harbin, China. Her current research interests include natural language processing, information retrieval.

Jing Jiang received her Ph.D. degree in 2008 from the Department of Computer Science, University of Illinois at Urbana-Champaign, Illinois, USA. She is an associate professor in the School of Information Systems at the Singapore Management University. Her research interests include natural language processing, information extraction and social media analysis.