Hashtag Recommendation with Topical Attention-Based LSTM

Yang Li†, Ting Liu†, Jing Jiang ‡, Liang Zhang †
†Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China
‡School of Information Systems, Singapore Management University, Singapore
{yli,tliu,lzhang}@ir.hit.edu.cn
jingjiang@smu.edu.sg

Abstract

Microblogging services allow users to create hashtags to categorize their posts. In recent years, the task of recommending hashtags for microblogs has been given increasing attention. However, most of existing methods depend on hand-crafted features. Motivated by the successful use of long short-term memory (LSTM) for many natural language processing tasks, in this paper, we adopt LSTM to learn the representation of a microblog post. Observing that hashtags indicate the primary topics of microblog posts, we propose a novel attention-based LSTM model which incorporates topic modeling into the LSTM architecture through an attention mechanism. We evaluate our model using a large real-world dataset. Experimental results show that our model significantly outperforms various competitive baseline methods. Furthermore, the incorporation of topical attention mechanism gives more than 7.4% improvement in F1 score compared with standard LSTM method.

1 Introduction

Over the past few years, microblogging has experienced tremendous success and become very important as both a social network and a news media. There is a significant amount of information generated every day. To facilitate the navigation in the deluge of information, microblogging services allow users to insert hashtags starting with the "#" symbol (e.g., #followfriday) into their posts to indicate the context or the core idea. In this way, hashtags help bring together relevant microblogs on a particular topic or event and enhance information diffusion in microblog services. It has been proven that hashtags are important for many applications in microblogs (Efron, 2010; Bandyopadhyay et al., 2012; Davidov et al., 2010; Wang et al., 2011; Li et al., 2015). 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 (Hong et al., 2012). 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 recently have shown great potential for learning effective representations and delivered state-of-the-art performance on various natural language processing tasks (Cho et al., 2014; Tang et al., 2015; Rush et al., 2015). Among these methods, the long short-term memory (LSTM), a variant of recurrent neural network (RNN), is widely adopted due to its capability of capturing long-term dependencies in learning sequential representations (Hochreiter and Schmidhuber, 1997; Gers et al., 2000; Palangi et al., 2016).

In this work, 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

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3019–3029, Osaka, Japan, December 11-17 2016.

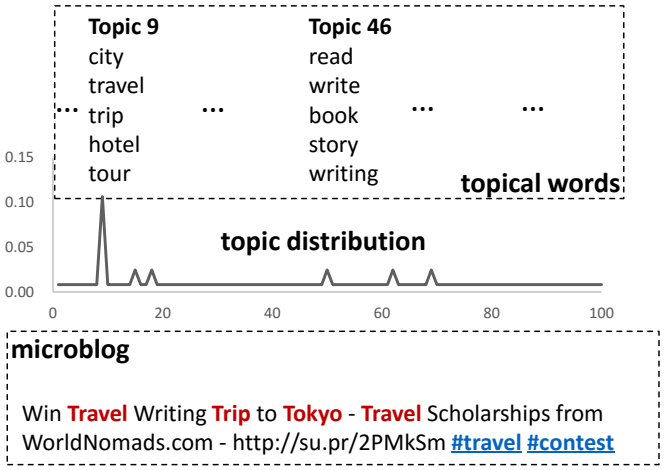

Figure 1: Illustration of hashtags of a microblog post and its topic distribution. The example post has a high probability in a travel topic (Topic 9). Words that are related to the topic (marked in red bold) can be selected through the topic distribution of the post. Using these words, we can predict its hashtag #travel.

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 (Bahdanau et al., 2015). One possible solution is to perform an average pooling operation over the hidden vectors of LSTM (Boureau et al., 2011), 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 language processing (Mnih et al., 2014; Bahdanau et al., 2015; Luong et al., 2015), we investigate the use of attention mechanism to automatically capture the most relevant words in a microblog to the recommendation task. Furthermore, it has been observed that most hashtags indicate the topics of a microblog (Ding et al., 2012; Godin et al., 2013), as illustrated in Figure 1. To this end, we propose a novel attention-based LSTM model which incorporates LDA topics of microblogs into the LSTM architecture through an attention mechanism. By modeling the interactions between the words and the global topics, our model can 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 attention mechanism gives more than 7.4% improvement in F1 score compared with standard LSTM method.

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

- We thoroughly investigate several neural attention-based models for hashtag recommendation.
- We propose a novel attention-based LSTM model that incorporates topics of microblog posts into the LSTM architecture through an attention mechanism. Experiments on data from a real microblogging service show that our model achieves significantly better performance than various state-of-the-art methods.

2 Background

Before going to the details of our method, we provide some background on two topics relevant to our work: the attention mechanism and Latent Dirichlet Allocation (LDA).

2.1 Attention Mechanism

Attention-based models have demonstrated success in a wide range of NLP tasks including sentence summarization (Rush et al., 2015), reading comprehension (Hermann et al., 2015) and text entailment (Rocktäschel et al., 2016; Wang and Jiang, 2016). 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 (Bahdanau et al., 2015; Luong et al., 2015). Bahdanau et al. (2015) 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 (Cho et al., 2014). 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.

In this work, we adopt the attention mechanism to scan input microblog posts and select key words to hashtag recommendtaion. Motivated by Bahdanau et al. (2015), we first investigate a vanilla attention-based LSTM model, which is referred to as VAB-LSTM in Section 4.2.1. In VAB-LSTM, we use the last hidden vector from the LSTM that processes a post as the global representation of that post and incorporate attentions to measure the interactions between each word and the global representation. Then we further compare it with our proposed topical attention-based LSTM model.

2.2 Latent Dirichlet Allocation (LDA)

Topic models have been a powerful technique for finding useful structures in a collection of documents. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a well-developed and widely-used topic model for inferring the semantic meaning of documents through a set of representative words (topics). It models a document as a mixture of latent topics. In LDA, each document of the corpus is assumed to have a distribution over K topics, where the discrete topic distributions are drawn from a symmetric Dirichlet distribution. The high probability words in each distribution gives us a way of understanding the contents of the corpus at a very high level.

Given a collection of microblog posts, LDA is able to learn a sparse topic representation for each post. The topic representation is viewed as a kind of global semantic information of a post, which we can utilize to learn the interactions between each words and the whole microblog post.

3 The Approach

In this section, we will present our proposed model for hashtag recommendation. We formulate the hashtag recommendation task as a multi-class classification problem. It has been observed that hashtags indicate the primary topics of microblog posts (Ding et al., 2012; Godin et al., 2013). To incorporate the topics of microblogs, we take into account the attention mechanism and develop a novel Topical Attention-Based LSTM model, or TAB-LSTM for short. The basic idea of TAB-LSTM is to combine local hidden representations with global topic vectors through an attention mechanism. We believe that in this way our model can capture the importance of different local words according to the global topics of a microblog post.

Our overall model is illustrated in Figure 2. The model mainly consists of three parts, namely, LSTM based sequence encoder, topic modeling, and topical 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 (Bengio et al., 2003; Mikolov et al., 2013). All the word vectors are stacked in a word embedding matrix $L_w \in \mathbb{R}^{dim \times |V|}$, where dim is the dimension of word vector and $|V|$ is vocabulary size. We pre-train the values of word vectors from text corpus with embedding learning algorithms to make better use of semantic and grammatical associations of words (Mikolov et al., 2013). Given an input microblog s, we take the embeddings $\mathbf{x}_t \in \mathbb{R}^{dim}$ for each word in the microblog to obtain the first layer. Hence, a microblog post of length N is represented with $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)$.

3.1 LSTM based Sequence Encoder

LSTM is special form of recurrent neural networks (RNNs), 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 (Hochreiter and Schmidhuber, 1997).

Given a microblog $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)$, 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

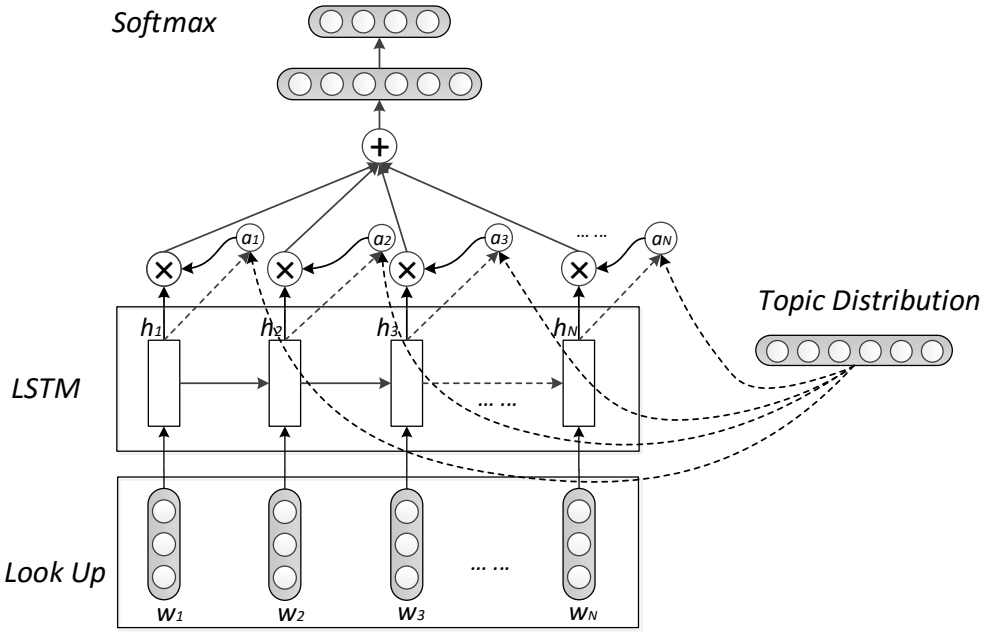

Figure 2: The graphical illustration of the proposed topical attention-based LSTM model (TAB-LSTM).

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:

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

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

The output of LSTM layer is a sequence of hidden vectors $[\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N]$. Each annotation $\mathbf{h}_t$ contains information about the whole input microblog with a strong focus on the parts surrounding the t-th word of the input microblog.

3.2 Topic Modeling

In TAB-LSTM, we propose an topical attention to introduce a series of attention-weighted combinations of these hidden vectors using the external topical distribution. We use LDA to learn the topic structures of microblogs and the model is trained offline. Specifically, given a set of microblog posts $\mathcal{S}$, where each post $s \in \mathcal{S}$ contains N_s words $\{w_{s,1}, w_{s,2}, ..., w_{s,N_s}\}$, LDA makes the following assumptions. There exist K topics, each associated with a multinomial word distribution φ_k. Each post has a topic distribution θ_s in the K-dimensional topic space. Each word in a microblog post has a hidden topic label drawn from the post's topic distribution. Formally, the generative process of LDA is described as follows:

- For each topic $k = 1, ..., K$, draw $\varphi_k \sim Dir(\beta)$
- For each microblog post $s \in \mathcal{S}$, draw $\theta_s \sim Dir(\alpha)$
  - For each word $w_{s,n}$, draw $z_{s,n} \sim Multi(\theta_s)$ and $w_{s,n} \sim Multi(\phi_{z_{s,n}})$

where α and β are parameters of the Dirichlet priors, θ_s is the topic distribution we will incorporate in our model.

3.3 Topical Attention

Taking all hidden states $[\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N]$ and the external topic vector $\theta_\mathbf{s} \in \mathbb{R}^{K\times 1}$, the topical attention layer outputs a continuous context vector $vec \in \mathbb{R}^{d\times 1}$ for each microblog post s. The output vector is computed as a weighted sum of each hidden state $\mathbf{h}_j$:

$$vec = \sum_{j=1}^{N} a_j \mathbf{h}_j \tag{2}$$

where d is the hidden dimension of LSTM, $a_j \in [0, 1]$ is the attention weight of $\mathbf{h}_j$ and $\sum_j a_j = 1$.

Next, we will introduce how we obtain $[a_1, a_2, ..., a_N]$ in detail. Specifically, for each $\mathbf{h}_j$, we use the following equation to compute scores on how well the inputs around position j and the topic distribution θ_s match:

$$g_j = v_a^\top \tanh(\mathbf{W^a}\theta_\mathbf{s} + \mathbf{U^a}\mathbf{h}_j) \tag{3}$$

where K is the number of topics, $\mathbf{W^a} \in \mathbb{R}^{d\times K}$, $\mathbf{U^a} \in \mathbb{R}^{d\times d}$ and $v_a \in \mathbb{R}^{d\times 1}$ are the weight matrices. After obtaining $[g_1, g_2, ..., g_N]$, we feed them to a $softmax$ function to calculate the final weight scores $[a_1, a_2, ..., a_N]$.

Finally, we use the output from the topical attention layer as the embedding of the microblog from our deep neural network. We 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 M is the number of hashtag categories:

$$softmax(m_i) = \frac{exp(m_i)}{\sum_{i'=1}^{M} exp(m_{i'})} \tag{4}$$

3.4 Model Training

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

$$\mathcal{J} = -\sum_{s\in S} \sum_{t\in tags(s)} \log p(t|s) \tag{5}$$

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

4 Experiments

We apply the proposed method to the task of hashtag recommendation to evaluate the performance. In this section, we first describe our dataset and experimental settings, then the results and analysis.

4.1 Dataset

Our dataset is constructed from a large Twitter dataset which spans the second half of 2009 (Yang and Leskovec, 2011). We collect a dataset with 185,391,742 tweets from October to December 2009. 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. The statistics of our dataset is shown in Table 1.

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

Table 1: Statistics of the dataset, Nt(avg) is the average number of hashtags in the dataset.

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. (2009) to recommend hashtags.

• **SVM**: We build a multi-class SVM classification model (Hearst et al., 1998) 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. (2013) 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.

We also compare two degenerate versions of our model TAB-LSTM as follows.

• **AVG-LSTM**: 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.

• **VAB-LSTM**: In this model, we use the last hidden vector from the LSTM that processes a post as the global representation of that post and incorporate attentions to measure the interactions between each word and the global representation. This method is similar to our model except that we replace the topic distribution θ_s with the last hidden vector h_N in Equation (3).

4.2.2 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. We train four types of neural models including LSTM, AVG-LSTM, VAB-LSTM and our proposed model TAB-LSTM. For each of the above models, the sentences of length is up to 40 words. We set the dimension of all the hidden states of the LSTMs to be 500. We use a minibatch stochastic gradient descent (SGD) algorithm together with the Adam method to train each model (Kingma and Ba, 2014). 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. For TAB-LSTM, we tested with different numbers of LDA topic size K and found $K = 100$ is an optimal setting.

For both our models and the baseline methods, we use the validation data to tune the hyperparameters, we 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, 2011) with the word2vec toolkit (Mikolov et al., 2013).

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. The same settings are adopted by previous work (Ding et al., 2012; Ding et al., 2013; Gong et al., 2015).

4.3 Comparison to Other Methods

In Table 2, we compare the results of our method and the state-of-the-art discriminative and generative methods on the dataset. TAB-LSTM denotes our proposed model. We have the following observations. (1) First of all, SVM performs much better than LDA, showing that the embedding features capture more semantic information than bag-of-words (BoW). (2) Both TAB-LSTM and the degenerate models significantly outperform the baseline methods LDA, SVM and TTM. The results demonstrate that the neural network can achieve better performance on this task. (3) If we compare TAB-LSTM with LSTM and AVG-LSTM, we can see that TAB-LSTM improves the F1-score by nearly 7.4% in the same setting of parameters, showing that incorporating attention mechanism is useful. (4) Using topical attention, TAB-LSTM outperforms VAB-LSTM, which shows the effectiveness of topic information for this task.

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
AVG-LSTM	0.472	0.405	0.436
VAB-LSTM	0.489	0.419	0.452
TAB-LSTM	**0.503**	**0.435**	**0.467**

Table 2: Evaluation results of different methods for hashtag recommendation. The dimension of word embeddings is set to be 300 for all methods. All improvements obtained by TAB-LSTM over other methods are statistically significant within a 0.99 confidence interval using the t-test.

Considering that many microblog posts have more than one hashtags, we also evaluate the top k results of different methods. Figure 3 shows the precision, recall, and F1 curves of LDA, SVM, TTM, LSTM and TAB-LSTM on the test data. Each point of a curve represents the extraction of a different number of hashtags, ranging from 1 to 5. From Figure 3, we can see although the precision and F1-score of TAB-LSTM decreases when the number of hashtags is larger, the performance of TAB-LSTM 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.

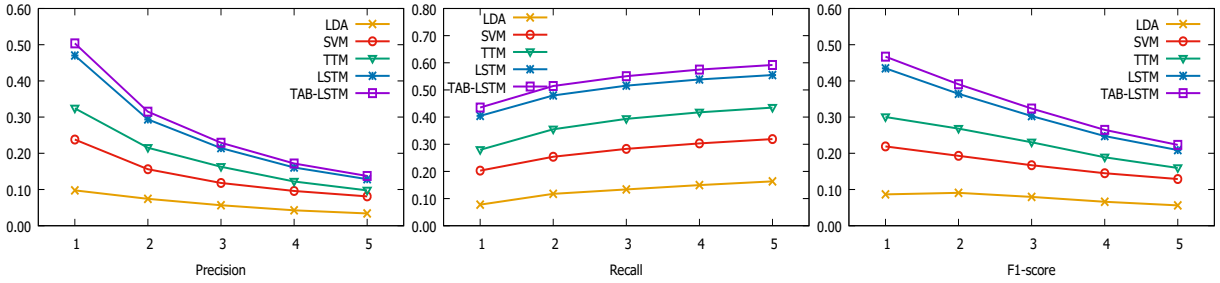

Figure 3: Precision, Recall and F1 with recommended hashtags range from 1 to 5.

4.4 Parameter Sensitive Analysis

We further investigate the effect of hyperparameters to the performance. First, we vary the values of topic size K while fixing the other parameters. We've tried various settings with $K = 50, 100, 150, 200$ when training the topic models with LDA. Results in Table 3 show that the best results are achieved when K is larger than 100.

Methods	Precision	Recall	F1-score
Attn50	0.492	0.422	0.454
Attn100	0.503	0.435	0.467
Attn150	0.501	0.432	0.464
Attn200	0.499	0.431	0.463

Table 3: Precision, Recall and F1 of TAB-LSTM with different number of topics when the dimension of word vectors is set to be 300.

It is well accepted that a good word embedding is crucial to composing a powerful text representation at a higher level. Next, we would like to study the effects of different word embeddings. Table 4 shows

the precision, recall and F1-score when we vary the dimension of word embeddings. We find a larger dimension of word embedding is more effective for this task.

Methods	Precision	Recall	F1-score
Emb50	0.470	0.403	0.434
Emb100	0.487	0.419	0.450
Emb200	0.495	0.425	0.457
Emb300	0.503	0.435	0.467

Table 4: Precision, Recall and F1 of TAB-LSTM with different dimension of word embeddings when the number of topics is 100.

4.5 Qualitative Analysis

We also perform qualitative analysis of our results. In Figure 4, we compare the attention heat maps learned by TAB-LSTM and VAB-LSTM of two example microblog posts. In the first example, hashtag #H1N1 is correctly recommended by TAB-LSTM because the word *H1N1* is selected by the topic of this post, while in the case of VAB-LSTM, H1N1 is not selected. In the second example, both hashtags are correctly predicted by TAB-LSTM, while VAB-LSTM missed the word "ff", which is short for #followfriday.

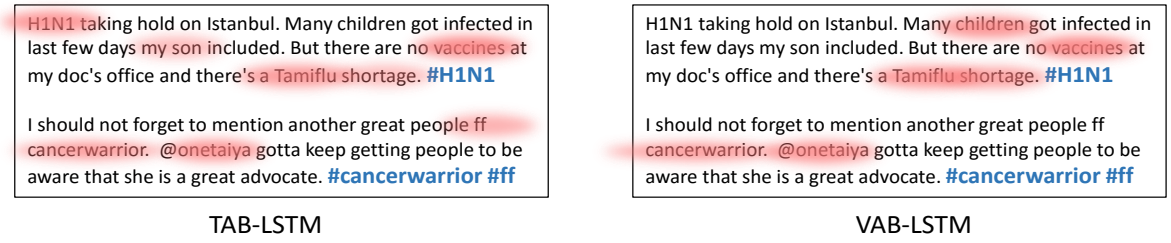

Figure 4: Attention heat maps for two example microblog posts.

5 Related Work

There has been a variety of work proposed for hashtag recommendation in the past few years. Zangerle et al. (2011) exploit the similarity between tweets. 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 (2014) formulate hashtag recommendation task as a learning to rank problem. They 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 (2009) 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. (2013) 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. (2011), Ding et al. 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 (Ding et al., 2012; Ding et al., 2013).

Most of the works mentioned above are based on textual information. There have also been some attempts that combine text with other types of data. Kywe et al. propose a collaborative filtering model to incorporate user preferences in hashtag recommendation (Kywe et al., 2012). Besides that, Zhang et al. (2014) and Ma et al. (2014) try to incorporate temporal information. Gong et al. (2015) propose to model type of hashtag as a hidden variable into their DPMM (Dirichlet Process Mixture Models) based method.

More recently, Gong and Zhang (2016) propose an attention-based convolutional neural network, which incorporates a local attention channel and global channel for hashtag recommendation. However, to the best of our knowledge, there is no work yet on employing both topic models and deep neural networks for this task.

6 Conclusion

In this paper, we investigated a novel topical attention-based LSTM model for the task of hashtag recommendation. We adopted the architecture of LSTM to avoid hand-crafted features. Our model incorporates topic modeling into the LSTM architecture through an attention mechanism and takes over the advantages of the both. Through evaluations run on a large dataset from Twitter, we have demonstrated that the proposed method outperforms competitive baseline methods effectively.

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 temporal information can be incorporated into the model. We will also consider using alternative topic models which are particularly designed for short microblog texts such as Twitter-LDA (Zhao et al., 2011).

7 Acknowledgements

We thank the anonymous reviewers for their constructive comments, and gratefully acknowledge the support of the National Basic Research Program (973 Program) of China via Grant 2014CB340503, the National Natural Science Foundation of China (NSFC) via Grant 61472107 and 71532004. Corresponding author: Ting Liu, E-mail: tliu@ir.hit.edu.cn.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations*.

Ayan Bandyopadhyay, Kripabandhu Ghosh, Prasenjit Majumder, and Mandar Mitra. 2012. Query expansion for microblog retrieval. *International Journal of Web Science*, 1(4):368–380.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *journal of machine learning research*, 3(Feb):1137–1155.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Y-Lan Boureau, Nicolas Le Roux, Francis Bach, Jean Ponce, and Yann LeCun. 2011. Ask the locals: multi-way local pooling for image recognition. In *2011 International Conference on Computer Vision*, pages 2651–2658. IEEE.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. 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)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics: posters*, pages 241–249. Association for Computational Linguistics.

Zhuoye Ding, Qi Zhang, and Xuanjing Huang. 2012. Automatic hashtag recommendation for microblogs using topic-specific translation model. In *Proceedings of COLING 2012: Posters*, pages 265–274, Mumbai, India, December. The COLING 2012 Organizing Committee.

Zhuoye Ding, Xipeng Qiu, Qi Zhang, and Xuanjing Huang. 2013. Learning topical translation model for microblog hashtag suggestion. In *IJCAI*. Citeseer.

Miles Efron. 2010. Hashtag retrieval in a microblogging environment. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788. ACM.

Felix A. Gers, Jürgen Schmidhuber, and Fred A. Cummins. 2000. Learning to forget: Continual prediction with lstm. *Neural Computation*, 12(10):2451–2471.

Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. 2013. Using topic models for twitter hashtag recommendation. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 593–596. International World Wide Web Conferences Steering Committee.

Yuyun Gong and Qi Zhang. 2016. Hashtag recommendation using attention-based convolutional neural network. In *IJCAI*.

Yeyun Gong, Qi Zhang, and Xuanjing Huang. 2015. Hashtag recommendation using dirichlet process mixture models incorporating types of hashtags. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 401–410, Lisbon, Portugal, September. Association for Computational Linguistics.

Marti A. Hearst, Susan T Dumais, Edgar Osman, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 769–778, New York, NY, USA. ACM.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. 2009. Latent dirichlet allocation for tag recommendation. In *Proceedings of the third ACM conference on Recommender systems*, pages 61–68. ACM.

Su Mon Kywe, Tuan-Anh Hoang, Ee-Peng Lim, and Feida Zhu. 2012. On recommending hashtags in twitter networks. In *Social Informatics*, pages 337–350. Springer.

Yang Li, Jing Jiang, Ting Liu, and Xiaofei Sun. 2015. Personalized microtopic recommendation with rich information. In *Social Media Processing: 4th National Conference, SMP 2015, Guangzhou, China*, pages 1–14. Springer.

Zhiyuan Liu, Xinxiong Chen, and Maosong Sun. 2011. A simple word trigger method for social tag suggestion. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1577–1588, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. pages 1412–1421, September.

Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. 2014. 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*, pages 999–1008. ACM.

Allie Mazzia and James Juett. 2009. Suggesting hashtags on twitter. *EECS 545m, Machine Learning, Computer Science and Engineering, University of Michigan*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, pages 2204–2212.

Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):694–707.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. *ICLR*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. pages 379–389, September.

Surendra Sedhai and Aixin Sun. 2014. Hashtag recommendation for hyperlinked tweets. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 831–834. ACM.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432.

Shuohang Wang and Jing Jiang. 2016. 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*, pages 1442–1451, San Diego, California, June. Association for Computational Linguistics.

Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. 2011. 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*, pages 1031–1040. ACM.

Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM.

Eva Zangerle, Wolfgang Gassler, and Gunther Specht. 2011. Recommending#-tags in twitter. In *Proceedings of the Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop Proceedings*, volume 730, pages 67–78.

Qi Zhang, Yeyun Gong, Xuyang Sun, and Xuanjing Huang. 2014. Time-aware personalized hashtag recommendation on social media. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 203–212, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 338–349, Berlin, Heidelberg. Springer-Verlag.