Hashtag Recommendation using Attention-based Convolutional Neural Network

Yuyun Gong, Qi Zhang
School of Computer Science, Fudan University
Shanghai Key Laboratory of Intelligent Information Processing
825 Zhangheng Road, Shanghai, P.R. China
{yygong12, qz}@fudan.edu.cn

Abstract

Along with the increasing requirements, the hashtag recommendation task for microblogs has been receiving considerable attention in recent years. Various researchers have studied the problem from different aspects. However, most of these methods usually need handcrafted features. Motivated by the successful use of convolutional neural networks (CNNs) for many natural language processing tasks, in this paper, we adopt CNNs to perform the hashtag recommendation problem. To incorporate the trigger words whose effectiveness have been experimentally evaluated in several previous works, we propose a novel architecture with an attention mechanism. The results of experiments on the data collected from a real world microblogging service demonstrated that the proposed model outperforms state-of-the-art methods. By incorporating trigger words into the consideration, the relative improvement of the proposed method over the state-of-the-art method is around 9.4% in the F1-score.

Introduction

On social networks and microblogging services, users usually use hashtags, which are types of labels or metadata tags, to make it easier to find messages with a specific theme or content. In most microblogging services, users can place the hash character # in front of words or unspaced phrases to create and use hashtags. Hashtags can occur anywhere in a microblog: at the beginning, middle, or end. Moreover, along with the increase in social media services, microblogs have also been widely used as data sources for public opinion analyses [Bermingham and Smeaton, 2010; Jiang *et al.*,], prediction [Asur and Huberman, 2010; Bollen *et al.*, 2011], reputation management [Pang and Lee, 2008; Otsuka *et al.*, 2012], and many other applications [Sakaki *et al.*, 2010; Becker *et al.*, 2010; Guy *et al.*, 2010; 2013]. Many approaches for various applications have also demonstrated the usefulness of hashtags, including microblog retrieval [Efron, 2010], query expansion [A.Bandyopadhyay *et al.*, 2011], and sentiment analysis [Davidov *et al.*, 2010; Wang *et al.*, 2011].

However, only a portion of microblogs contain hashtags created by their authors. Hence, the task of automatically recommending hashtags for microblogs has received considerable attention in recent years. Existing works have studied discriminative models with various kinds of features and models [Heymann *et al.*, 2008; Liu *et al.*,], collaborative filtering [Kywe *et al.*, 2012], and generative models [Krestel *et al.*, 2009; Ding *et al.*, 2013; Godin *et al.*,] based on textual and social information. Most of these methods are commonly based on lexical level features, including the bag-of-words (BoW) model and exquisitely designed patterns. In addition to these methods, the effectiveness of word trigger assumption [Liu *et al.*, ; Ding *et al.*, 2013] has also been demonstrated. This means that substance of a given sentence can be realized by some important words in it.

In recent years, the rapid development of deep neural networks with word embedding has made it possible to perform various NLP tasks and achieve remarkable results. Among these methods, convolutional neural networks (CNNs)[LeCun *et al.*, 1998], which were originally invented for computer vision, have shown their effectiveness for various NLP tasks, including semantic parsing [Yih *et al.*, 2014], machine translation [Meng *et al.*,], sentence modeling [Kalchbrenner *et al.*,], and a variety of traditional NLP tasks [Dos Santos *et al.*, 2015; Chen *et al.*,]. Instead of building hand-crafted features, these methods utilize layers with convolving filters that are applied on top of pre-trained word embeddings. Moreover, compared to standard feedforward neural networks, CNNs have far fewer parameters and thus are easier to train.

Inspired by the advantages of CNNs in processing NLP tasks, we propose to use it to solve the hashtag recommendation task. Previous methods for hashtag and keyphrase recommendation [Liu *et al.*, ; Ding *et al.*, 2013] have also demonstrated the effectiveness of the trigger word mechanism, for example, in the tweet "#ipad #iphone How to share calendar events on iPhone and iPad", the trigger words are iPhone and iPad. However, standard CNNs cannot handle it. Motivated by the attention mechanism, which has been used in speech recognition [Chorowski *et al.*, 2014] and machine translation [Luong *et al.*, 2015], in this work, we introduce a novel CNN that incorporates an attention mechanism. The hashtag recommendation task is modelled as a classification problem. We employ a attention layer to produce a weight for a word with its surrounding context. Then, trigger words selected based on

the attention layer and whole microblogs are transformed into fixed length vectors respectively. In the final step, a fully connected layer with softmax outputs is constructed. Through experiments using a dataset collected from real online services, we demonstrated the effectiveness of the CNNs and attention mechanism. The CNN method could achieve a performance that was better than those of state-of-the-arts methods. The proposed attention-based CNNs yielded additional improvement compared to standard CNNs.

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

- To take advantages of deep neural networks, we adopted CNNs to perform the hashtag recommendation task.
- To incorporate a trigger word mechanism, we proposed a novel attention-based CNN architecture, which incorporates a local attention channel and global channel.
- Experimental results using a dataset collected from a real microblogging service showed that the proposed method can achieve significantly better performance than the state-of-the-arts methods.

Related Works

Hashtag recommendation

Due to the requirements of hashtag recommendation, a variety of methods have been proposed from different perspectives [Zangerle *et al.*, 2011; Ding *et al.*, 2013; Sedhai and Sun, 2014; Gong *et al.*,]. Zangerle et al. [2011] introduced a similarity based method to achieve the task. Firstly, they tried to retrieve a set of hashtags used within these most similar messages. Then, heuristics ranking methods are used to select hashtags from these selected candidates. Ding et al. [2013] proposed to convert the hashtag recommendation task as a translation process. They assume that hashtag and the trigger words of the tweets are two different language and have the same meaning. They integrate topical based translation model to perform this task.

Most of the works mentioned above are based on textual information. Besides these methods, Zhang et al. [2014] observed that when picking hashtags users have different perspectives, which are impacted by their own interests or the global topic trend. To model these factors, they proposed a topical model based method to incorporate the temporal and personal information. Since may tweets contain not only textual information but also hyperlinks, Sedhai and Sun [2014] studied the problem of hashtag recommendation for hyperlinked tweets.

From the brief descriptions given above, we can observe that deep neural networks have not been implemented to perform hashtag suggestion tasks. In this work, we propose to use convolution network with attentional mechanism to achieve the hashtag recommendation task.

Attention Mechanism

In recent years, *attention*-based neural network architectures, which learn to focus their "attention" to specific parts of the input, have shown promising results on various tasks, such as speech recognition [Bahdanau *et al.*, 2015b; Chorowski *et al.*, 2015], machine translation [Bahdanau *et al.*, 2015a], visual object classification [Mnih *et al.*, 2014], caption generation [Xu *et al.*, 2015] and so on.

Bahdanau et al. [2015b] proposed an attention-based recurrent sequence generators for sequence prediction. They performed the method on speech recognition task at the character level. The attention mechanism scanned the input sequence and chooses relevant frames. Chorowski [2015] also used attention-based RNN on a phoneme recognition task. On image classification tasks, Mnih [2014] introduced a recurrent neural network model, which can select a sequence of regions or locations from an image or video. Only the selected regions were incorporated into further processing.

Motived from these successful usages in these tasks, in this work, we adopt attentional mechanism to scan input microblogs and select trigger words. We further combine both the selected words and the whole microblog together to achieve the task.

The Proposed Models

In this work, we formulate a hashtag recommendation task as a multi-class classification problem. Networks handle input microblogs of varying length. Each dimension of the output layer represents the probability of a hashtag recommended. As discussed in the introduction section, we also follow the trigger words assumption, which has been successfully used in previous studies [Liu *et al.*, 2012; Ding *et al.*, 2013]. Hence, the proposed model incorporates two channels: a local attention channel and global channel. Figure shows the architecture of the proposed model.

In the global channel, all the words will be encoded, while the local attention channel will only encode a few trigger words. Common to these two types of channels is the fact that for each input microblog, both channels first operate the input. The goal is then to derive a feature vector that captures relevant information. However, these channels differ in how the feature vector is derived. The global channel has to attend to all the words for each tag, while the tag may only have relations with some trigger words. A local attention mechanism chooses to focus only on a small subset of the words for each tag which is depended on gate scores. Specifically, we employ a simple convolutional layer to combine the information from both vectors as follows:

$$\widehat{\mathbf{h}} = tanh(\mathbf{M} * \mathbf{v}[\mathbf{h_g}; \mathbf{h_l}] + b), \quad (1)$$

where $\mathbf{h_g}$ is the feature vector extracted from global channel, $\mathbf{h_l}$ is the feature vector of the local attention channel. $\mathbf{M}$ is a filter matrix for the convolutional operation. b is a bias. Each filter will produce one feature, we use multiple filters to produce a feature vector.

In our model, we use CNN to encode a sentence for the global channel, whereas we construct a local attention network for the local attention channel. The parameters in the global and local attention channels are learned jointly with our final objective function instead of being trained separately.

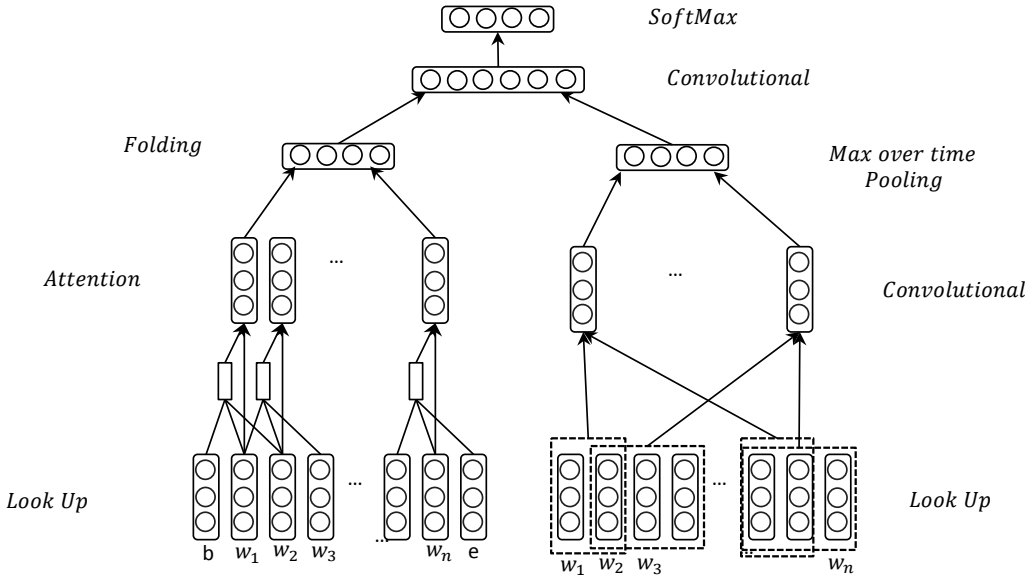


Figure 1: The architecture of the attention-based Convolutional Neural Network

Local Attention Channel

In the local attention channel, we consider the attention problem as a decision process. Given an input microblog m, we take the embeddings $w_i \in R^d$ for each word in the microblog to obtain the first layer, where d is the dimension of the word vector. A microblog of length n is represented with $w_{1:n}$ which is the concatenation of words $w_1, w_2, ..., w_n$. In general, let $w_{i:i+j}$ refer to the concatenation of words $w_i, w_{i+1}, ..., w_{i+j}$.

After converting words into embeddings, the next step is the local attention layer. Given a threshold value η and an input microblog m, the attention layer generates a sequence of trigger words $(w_i, ..., w_j)$. Each word is extracted from a small window. In general, given an attention window size h, we define $\mathbf{M^l} \in R^{h \times d}$ to be the parameter matrix. At the i-th step, the local attention layer generates the score of the i-th word in the microblog by focusing on the words in the window. The score of the central word in the window is obtained as follow:

$$s_{(2i+h-1)/2} = g(\mathbf{M^l} * \mathbf{w}_{i:i+h} + b), \tag{2}$$

where $s_{(2i+h-1)/2}$ is the score of the word $w_{(2i+h-1)/2}$. b is a bias, g is a non-linear function. We extract the words depend on their scores, if the score of the word greater than the threshold η, it will be extracted as a trigger word. The operation has been defined as follows:

$$\widehat{w_i} = \begin{cases} w_i & if \quad w_i > \eta \\ 0 & if \quad w_i \leq \eta \end{cases} \quad 0 \leq i < n, \tag{3}$$

where η is the threshold, we let η to be a function of the length of the microblog. Although many functions are possible, we simply model the threshold as follows:

$$\eta = \delta \cdot \min\{\mathbf{s}\} + (1 - \delta) \cdot \max\{\mathbf{s}\} \tag{4}$$

$\mathbf{s}$ is a sequence of scores for each word in the microblog. $min\{\mathbf{s}\}$ is the minimum score of the words, and $max\{\mathbf{s}\}$ is the maximum score. $\delta \in [0, 1]$ is a parameter to balance the minimum and maximum.

The local attention layer makes it possible to extract the most important words in the microblog. We apply the local attention layer as the first layer in the attention channel which insure that the following layer will only operate on the trigger words.

After local attention layer, the trigger words extracted will input the folding layer which can operate on different number of words. The folding layer is to abstract the features of the trigger words by the following operation.

$$z = g(\mathbf{M^l} * folding(\widehat{\mathbf{w}}) + \mathbf{b}), \tag{5}$$

where $\widehat{\mathbf{w}}$ are the trigger words. $\mathbf{M^l} \in R^{d \times r}$ is the parameter matrix, d is the dimension of the word vector, r is dimension of the output vector. $b \in R^r$ is a bias, g is a non-linear function. $folding$ is the sum operation for each dimension of all the trigger words, $f_i = \sum_j \widehat{w}_{j,i}$, where $\widehat{\mathbf{w}}_{\mathbf{j,i}}$ is the value in the ith position of the embedding of the jth trigger word.

The final output of the local attention channel is a fixed-length vector, which represents the embeddings of the trigger words $\widehat{\mathbf{w}}$.

Global Channel

In the global channel, we use a CNN architecture to model the whole microblog. In the convolutional layer, we use a filter which is a weight matrix $\mathbf{M^g} \in R^{l \times d}$ to produce a feature. l is the window size which means the filter operates on l words, d is the dimension of the word. For example, the feature z_i generated from a filter operated on l words $w_{i:i+l-1}$ can be

calculated as follow:

$$z_i = g(\mathbf{M^g} \cdot \mathbf{w}_{i:i+l-1} + b), \quad (6)$$

where g is a non-linear function and $b \in R$ is a bias term. We operate this filter on all the combinations of the words in the microblog $\{w_{1:l}, w_{2:l+1}, ..., w_{n-l+1:n}\}$ to produce a map of feature as follow:

$$\mathbf{z} = [z_1, z_2, ..., z_{n-l+1}], \quad (7)$$

In the pooling layer, a max-overtime pooling operation is applied over the feature map $\mathbf{z}$ to produce a feature for this filter. Using this operation we can extract the most important feature for each feature map. And it can deal with different microblog lengths.

From the process described above, we can see that one feature is extracted from one filter. To obtain multiple features, we use multiple filters with varying window sizes in the model. After max-overtime pooling operation, a bias and a non-linear function $tanh$ are applied to the pooled matrix.

The final output of our global channel is also a fixed-length vector, which represents the embeddings of the input microblog m.

After the local attention channel and global channel, we use a convolutional layer with multiple feature maps to combine the outputs of the local attention channel and the global channel. We regard the output from the convolutional layer as the embedding of the microblog from our deep neural network.

Training

In this work, we joined learning the parameters Θ in local attention and global channels.

$$\Theta = \{\mathbf{W}, \mathbf{M^l}, \mathbf{M^g}\}, \quad (8)$$

where $\mathbf{W}$ are words embeddings; $\mathbf{M^l}$ and $\mathbf{M^g}$ are the parameters of the local attention channel and global channel respectively; the rest parameters belong to the fully connected layer. Our training objective function is formulated as follows:

$$J = \sum_{(m,a)\in D} -logp(a|m), \quad (9)$$

where D is the training corpus, a is the hashtag for microblog m.

To minimize the objective function, we use the robust learning method AdaDelta [Zeiler, 2012].

Hashtag Recommendation

We perform hashtag recommendation as follows. Suppose given an unlabelled 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 unlabelled data, we will encode the microblog through the local attention channel and global channel by the saved model.

After the encoded processes, we combine the features generated from both channels. Then we get the scores of the hashtags for the dth microblog in unlabelled data by the fully connected layer:

$$P(y^d = a|\mathbf{h}^d; \beta) = \frac{exp(\beta^{(a)T}\mathbf{h}^d)}{\sum_{j\in A} exp(\beta^{(j)T}\mathbf{h}^d)}, \quad (10)$$

where β are the parameters. $\mathbf{h}$ is the feature vector connected from the channels. A is a set of candidate hashtags.

According to the scores output from fully connected layer, we can rank the hashtags for each microblog and recommending the top-ranked hashtags to users.

Experiments

Dataset and Setup

We use the microblog dataset collected by Ding et al. [2013] for evaluation. There are 110,000 microblogs in the dataset which contain hashtags annotated by users. The vocabulary of words is 106,323 in the dataset, the vocabulary of hashtags is 37,224, and the average number of words and hashtags in each microblog is 20.45 and 1.20 respectively. The dataset has been splitted into training set (100,000 microblogs) and test set (10,000 microblogs). In our experiment, we randomly select 10% of the training set as the development set.

To evaluate the performance, we use Precision(P), Recall(R), and F-score(F_1):

$$P = \frac{N_r}{N_s}, R = \frac{N_r}{N_m}, F_1 = \frac{2PR}{P+R}, \quad (11)$$

N_r is the right number recommended, N_s is the total number recommended by system, N_m is the total number of the hashtags assigned in the corpus.

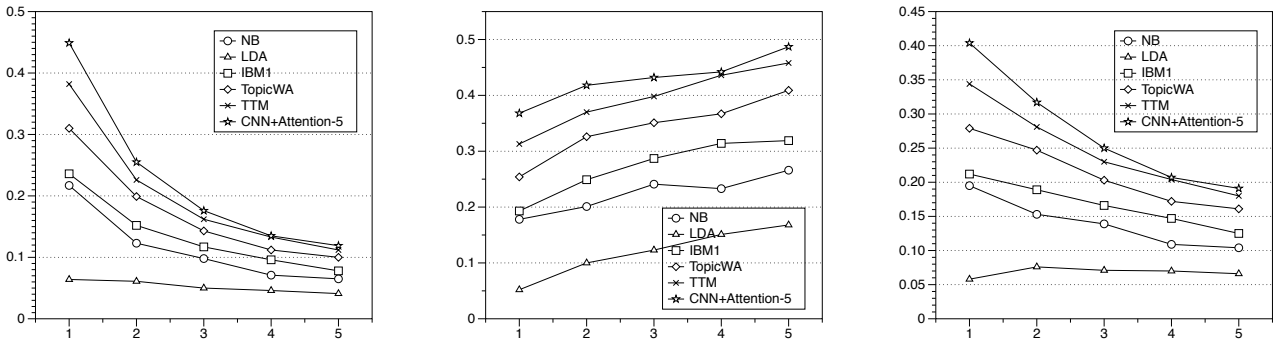
The parameters in our model include both of parameters in global channel and local attention channel. For the global channel, after trying different single window sizes and multi window sizes, we empirically set filter windows to multi window 1, 2, 3, and we use 100 feature maps for each window size. For the local attention channel, we set the width of attention matrix to 5, and δ to 0.8. Both the global and local attention channel we use hyperbolic tangent as non-linear function and set mini-batch size to 200.

We trained the word vectors on 10 million words from Sina Weibo. The vectors have dimensionality of 100 and were trained using the architecture proposed in [Mikolov *et al.*, 2013]. New words are initialized randomly.

Baseline

In this section, we compare our attention-based model with some baseline models and two degeneration models. We consider the following methods

- **Naive Bayes (NB)**: We used NB to model the hashtag recommendation as a classification task. Given the textual of the microblogs, we can estimate the posterior probability of each hashtag.
- **LDA**: we used the LDA based method proposed in [Krestel *et al.*, 2009] to recommend hashtags.
- **Translation model (IBM-1)**: IBM model 1 is proposed by [Liu *et al.*,] which is directly applied to obtain the alignment probability between the word and the hashtag.



Precision Recall F_1

Figure 2: Precision, Recall and F_1 with recommended Hashtags range from 1 to 5

Table 1: Evaluation results of different methods on the evaluation collection.

Methods	Precision	Recall	F_1
NB	0.217	0.197	0.203
LDA	0.064	0.060	0.062
IBM1	0.236	0.214	0.220
TopicWA	0.310	0.285	0.292
TTM	0.382	0.357	0.364
CNN	0.416	0.338	0.373
Attention-5	0.410	0.335	0.369
CNN+Attention-5	**0.443**	**0.362**	**0.398**

- **TopicWA**: TopicWA is a topical word alignment model proposed in [Ding *et al.*,]. In this method, the standard LDA is employed to discover the latent topic and combine the alignment method for this task.
- **TTM**: TTM was proposed by [Ding *et al.*, 2013] for hashtag recommendation.
- **CNN**: CNN was proposed by [Kim, 2014] for sentence classification. We use the publish code to accomplish the hashtag recommendation task.
- **Attention-5**: Attention-5 was the local attention channel in our model with the window size equals to 5.

The first method NB is traditional method. LDA, IBM-1, TopicWA and TTM are the methods based on topic model and translation model which are attracted a lot of attention in recent years. CNN and Attention-5 are the variant methods of our method.

Results and Discussion

Table shows comparisons of the results of the proposed method with those of the state-of-the-art discriminative and generative methods on the evaluation dataset. "CNN+Attention-5" denotes the method proposed in this paper. "CNN" and "Attention-5" represent the convolutional neural network and attention model, respectively. The results show that the proposed method is significantly better than the other methods. "TTM", "TopicWA", and "IBM1" are the models based on trigger words. From the results of "TTM", "TopicWA", "IBM1", and "NB", we can observe that the trigger words methods could improve the performance of the hashtag recommendation task. A comparison of the "CNN" and "TTM" results shows that "CNN" achieves a significantly better F_1 than "TTM". The results demonstrate that the neural network can achieve better performance on this task. From the results of "Attention-5", "TTM", "TopicWA", and "IBM1", we can observe that our local attention channel can improve the performance of the trigger words for this task. From the results of CNN and CNN+Attention-5, we observe that the attention model can benefit the task, and the multiple channels can achieve better performance than a single channel. The relative improvement of the proposed CNN-Attention-5 over TTM is around 9.4% in the F_1 score, which demonstrates the effectiveness of our method.

Figure shows the precision, recall, and F_1 curves of NB, LDA, IBM1, TopicWA, TTM, and CNN+Attention-5 on the test data. Each point of a curve represents the extraction of a different number of hashtags, ranging from 1 to 5. The curve that is the highest of the graphs indicates the best performance. Based on the results, we can observe that the performance of CNN+Attention-5 is the highest in all the curves. This indicates that the proposed method was significantly better than the other methods, and when we recommended the top 1 hashtag for each microblog, we obtained the highest F_1 score.

Table lists the static and non-static results. These results show that the non-static model could achieve better results for all the models. The non-static method tuned the word vector more specifically to the task-at-hand. For randomly initialized tokens which are not in the set of pre-trained vectors, we can learn more meaningful representations from fine-tuning.

Parameters sensitive analysis

From the description of the proposed model, we can know that there are several hyperparameters in the proposed model. To evaluate their impacts, we evaluated two crucial ones, window size l of the global channel and the width of the attention size h.

Table 2: Evaluation results of static and non-static models.

Methods	Non-static			Static		
	Precision	**Recall**	$\mathbf{F}_1$	**Precision**	**Recall**	$\mathbf{F}_1$
CNN	0.416	0.338	0.373	0.412	0.337	0.371
Attention-5	0.410	0.335	0.369	0.382	0.313	0.344
CNN+Attention-5	0.443	0.362	0.398	0.425	0.350	0.384

Table 3: Performance on various window size of global channel.

window sizes	Precision	Recall	$\mathbf{F}_1$
1	0.422	0.346	0.380
2	0.428	0.351	0.386
3	0.425	0.348	0.383
1-2	0.440	0.361	0.397
1-2-3	0.443	0.362	0.398

Table 4: Evaluation results of different attention size.

Methods	Precision	Recall	$\mathbf{F}_1$
Attention-1	0.326	0.267	0.294
Attention-3	0.406	0.332	0.365
Attention-5	0.410	0.335	0.369
Attention-7	0.405	0.331	0.364
CNN-Attention-1	0.395	0.323	0.355
CNN-Attention-3	0.435	0.356	0.392
CNN-Attention-5	0.443	0.362	0.398
CNN-Attention-7	0.438	0.358	0.394

Table lists the results of different window sizes for the global channel. We modeled the global channel on window sizes of 1, 2, and 3. To show the performance of the model with multiple window sizes, we use the window size with the the following combinations: (1,2), (1,2,3), and we obtained the best performance when the window sizes were (1,2,3). From the results of window sizes equal to 1, 2, or 3, we observe that the best window size is 2 on hashtag recommendation. The reason is that when the window size equals to 1, the convolutional operation will extract the unigram information while ignore the context information. The results of window sizes equal to 2 and 3 show that the bigram information is more important than the trigram for our task. From the results of multiple window sizes, we can observe that the multiple can achieve better performance. This demonstrated the advantages of the models with multiple window sizes over the single window size models. Comparing the results of window sizes equal to (1,2) and (1,2,3), we can observe that the performance between the multiple window sizes were similar, so we can choose the multiple window size conveniently.

Table lists the comparisons of different attention sizes on the constructed evaluation dataset. "Attention-h" represent the methods that only considered the local attention channel. "CNN+Attention-h" represent the methods proposed with the multiple window sizes (1,2,3) in the global channel ($h \in \{1, 3, 5, 7\}$ is the attention size). Based on the results, we can see that the performance will increase with the number of words considered, and we will obtain the best performance when the window size is set to 5. "Attention-1" achieved a bad performance. Because in this model, the importance of the central word only depended on itself while ignoring the context. From the results of "Attention-5" and "Attention-7", we can observe that when the distance between the words and the central word greater than 2, we can ignore the influence of these words to the central word.

Conclusions

In this paper, we investigated a novel attention based CNNs for performing the hashtag recommendation task. We adopted the architecture of CNNs to avoid hand-craft features and take other advantages of CNNs. To incorporate a trigger word mechanism, we proposed a novel attention-based CNN architecture, which consists of a local attention channel and global channel. Experimental results on the data collected from a real world microblogging service demonstrated that the proposed method outperforms the methods which take only global or local information into consideration and state-of-the-art methods.

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No. 61532011, 61473092, and 61472088), the National High Technology Research and Development Program of China (No. 2015AA015408).

References

[A.Bandyopadhyay *et al.*, 2011] A.Bandyopadhyay, M. Mitra, and P. Majumder. Query expansion for microblog retrieval. In *Proceedings of TREC*, 2011.

[Asur and Huberman, 2010] S. Asur and B.A. Huberman. Predicting the future with social media. In *WI-IAT'10*, volume 1, pages 492–499, 2010.

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

[Bahdanau *et al.*, 2015b] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. *arXiv preprint arXiv:1508.04395*, 2015.

[Becker *et al.*, 2010] Hila Becker, Mor Naaman, and Luis Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of WSDM '10*, 2010.

[Bermingham and Smeaton, 2010] Adam Bermingham and Alan F. Smeaton. Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of CIKM'10*, 2010.

[Bollen *et al.*, 2011] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8, 2011.

[Chen *et al.*,] Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of ACL-IJCNLP 2015*.

[Chorowski *et al.*, 2014] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. End-to-end continuous speech recognition using attention-based recurrent nn: First results. In *arXiv preprint arXiv:1412.1602*, 2014.

[Chorowski *et al.*, 2015] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. *arXiv preprint arXiv:1506.07503*, 2015.

[Davidov *et al.*, 2010] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *COLING*, 2010.

[Ding *et al.*,] Zhuoye Ding, Qi Zhang, and XuanJing Huang. Automatic hashtag recommendation for microblogs using topic-specific translation model. In *Proceedings of COLING 2012*.

[Ding *et al.*, 2013] Zhuoye Ding, Xipeng Qiu, Qi Zhang, and Xuanjing Huang. Learning topical translation model for microblog hashtag suggestion. In *Proceedings of IJCAI 2013*, 2013.

[Dos Santos *et al.*, 2015] Cicero Dos Santos, Bing Xiang, and Bowen Zhou. Classifying relations by ranking with convolutional neural networks. In *Proceedings of ACL-IJCNLP*, 2015.

[Efron, 2010] Miles Efron. Hashtag retrieval in a microblogging environment. In *Proceedings of SIGIR '10*, 2010.

[Godin *et al.*,] Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. Using topic models for twitter hashtag recommendation. In *Proceedings of WWW '13 Companion*.

[Gong *et al.*,] Yeyun Gong, Qi Zhang, and Xuanjing Huang. Hashtag recommendation using dirichlet process mixture models incorporating types of hashtags.

[Guy *et al.*, 2010] Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, and Erel Uziel. Social media recommendation based on people and tags. In *Proceedings of SIGIR*, 2010.

[Guy *et al.*, 2013] Ido Guy, Uri Avraham, David Carmel, Sigalit Ur, Michal Jacovi, and Inbal Ronen. Mining expertise and interests from social media. In *Proceedings of WWW*, 2013.

[Heymann *et al.*, 2008] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. Social tag prediction. In *SIGIR*, 2008.

[Jiang *et al.*,] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In *Proceedings of ACL 2011*.

[Kalchbrenner *et al.*,] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of ACL 2014*.

[Kim, 2014] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[Krestel *et al.*, 2009] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. Latent dirichlet allocation for tag recommendation. In *Proceedings of RecSys '09*, 2009.

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

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Liu *et al.*,] Zhiyuan Liu, Xinxiong Chen, and Maosong Sun. A simple word trigger method for social tag suggestion. In *Proceedings of EMNLP 2011*.

[Liu *et al.*, 2012] Zhiyuan Liu, Chen Liang, and Maosong Sun. Topical word trigger model for keyphrase extraction. In *Proceedings of COLING*, 2012.

[Luong *et al.*, 2015] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*, 2015.

[Meng *et al.*,] Fandong Meng, Zhengdong Lu, Mingxuan Wang, Hang Li, Wenbin Jiang, and Qun Liu. Encoding source language with convolutional neural network for machine translation. In *Proceedings of ACL-IJCNLP 2015*.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[Mnih *et al.*, 2014] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *NIPS*, 2014.

[Otsuka *et al.*, 2012] Takanobu Otsuka, Takuya Yoshimura, and Takayuki Ito. Evaluation of the reputation network using realistic distance between facebook data. In *WI-IAT*, 2012.

[Pang and Lee, 2008] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January 2008.

[Sakaki *et al.*, 2010] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of WWW '10*, 2010.

[Sedhai and Sun, 2014] Surendra Sedhai and Aixin Sun. Hashtag recommendation for hyperlinked tweets. In *Proceedings of SIGIR*, 2014.

[Wang *et al.*, 2011] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of CIKM '11*, 2011.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.

[Yih *et al.*, 2014] Wen-tau Yih, Xiaodong He, and Christopher Meek. Semantic parsing for single-relation question answering. In *Proceedings of ACL*, 2014.

[Zangerle *et al.*, 2011] Eva Zangerle, Wolfgang Gassler, and Gunther Specht. Recommending#-tags in twitter. In *Proceedings of SASWeb 2011*, 2011.

[Zeiler, 2012] Matthew D Zeiler. Adadelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[Zhang *et al.*, 2014] Qi Zhang, Yeyun Gong, Xuyang Sun, and Xuanjing Huang. Time-aware personalized hashtag recommendation on social media. In *Proceedings of COLING 2014*, Dublin, Ireland, 2014.