Computers, Materials & Continua CMC, vol.61, no.3, pp.1261-1269, 2019

Hashtag Recommendation Using LSTM Networks with Self-Attention

Yatian Shen[1], Yan Li[1], Jun Sun[1, *], Wenke Ding[1], Xianjin Shi[1], Lei Zhang[1], Xiajiong Shen[1] and Jing He[2]

Abstract: On Twitter, people often use hashtags to mark the subject of a tweet. Tweets have specific themes or content that are easy for people to manage. With the increase in the number of tweets, how to automatically recommend hashtags for tweets has received wide attention. The previous hashtag recommendation methods were to convert the task into a multi-class classification problem. However, these methods can only recommend hashtags that appeared in historical information, and cannot recommend the new ones. In this work, we extend the self-attention mechanism to turn the hashtag recommendation task into a sequence labeling task. To train and evaluate the proposed method, we used the real tweet data which is collected from Twitter. Experimental results show that the proposed method can be significantly better than the most advanced method. Compared with the state-of-the-art methods, the accuracy of our method has been increased 4%.

Keywords: Hashtags recommendation, self-attention, neural networks, sequence labeling.

1 Introduction

With the rapid development of social media, people are more and more likely to record their moods and lives on the Internet with short words. To facilitate the management of thousands of tweets posted each day, some tweets will contain a symbol "#". The symbol "#" is called hashtag and is usually used to mark keywords or topics in a tweet. Originally, users of social media created hashtag to classify messages. Now, various studies have also shown that hashtags can provide valuable information on different issues such as twitter spam detection [Benevenuto, Magno, Rodrigues et al. (2010)], sentiment analysis [Wang, Wei, Liu et al. (2011)] and so on.

As the different demand has increased, the hashtag recommendation tasks have received widespread attention in recent years. Scholars have proposed different methods, such as discriminant models [Heymann, Ramage and Garcia-Molina (2008)], collaborative filtering [Kywe, Hoang, Lim et al. (2012)] and generating models [Ding, Qiu and Zhang (2013); Godin, Slavkovikj, Neve et al. (2013)]. Most of these methods treat hashtag recommendation tasks as a multi-class classification problem. In recent years, with the development of deep learning, deep neural networks have been well applied to various

[1] School of Computer and Information Engineering, Henan University, Kaifeng, 475000, China.

[2] The Corporate and Investment, Bank Technology, Canary Wharf, London, E145JP, UK.

[*] Corresponding Author: Jun Sun. Email: sunjunjerry@outlook.com.

natural language processing tasks [Xiong, Shen, Wang et al. (2018); Shen and Huang (2016)]; Scholars began to study the application of deep neural networks to hashtag recommendation tasks. The convolutional neural networks are applied to hashtag recommendation tasks [Gong and Zhang (2016)]. Some researchers apply additional information to the hashtag recommendation task. Huang et al. [Huang, Zhang, Gong et al. (2016)] use the user's historical information to obtain the user's interest characteristics, and design an end-to-end neural network model for hashtag recommendation. Zhang et al. [Zhang, Wang, Huang et al. (2017)] combines the image information and text information of the tweet and uses the attention mechanism to recommend the hashtag.

However, these methods considered that the hashtag recommendation task is a multi-classification task, and they can only recommend hashtag that have already appeared for new tweets. When a new tweet appears, the hashtag it needs may never be appeared. Thus, being able to recommend new tags for tweets has become an important research topic. To solve this problem, we propose a novel network structure. Unlike traditional multi-classification models, we consider the hashtag recommendation task as a sequence labeling task, which is very simple and effective. We believe that any word in the tweet may be a hashtag, a model is designed to calculate the probability that each word in the tweet is hashtag. The word with a higher probability value is more likely to become the hashtag of the tweet. Most importantly, our approach can recommend tags that have never appeared before, while previous methods can only recommend similar tags in historical information. In general, our main contributions are as follows:

1. Unlike previous methods, we consider the hashtag recommendation task as a sequence labeling task. We predicted hashtags of tweets by calculating the probability that each word in the tweet becomes a hashtag.
2. We propose a novel network structure to solve the hashtag recommendation task. The model can easily obtain the long-distance dependence information in the sentence and the internal structure of the sentence without considering the distance of the words in the sentence. Moreover, the model can automatically recommend new tags.
3. Experimental results from data set collected on Twitter show that the performance of this method is significantly better than current state-of-the-art methods, which can get the result of precision about 0.557.

2 Related works

With the increase in demand, the hashtag recommendation task has received extensive attention in recent years, and scholars have conducted various studies on this task. Heymann et al. [Heymann, Ramage and Garciamolina (2008)] used data collected from the social bookmarking system to investigate tag recommendation issues. They introduced an entropy-based indicator to capture the versatility of tags. Krestel et al. [Krestel, Fankhauser and Nejdl (2009)] introduced latent Dirichlet allocation that leads to a shared theme structure from the collaboration markup work of multiple users to recommend hashtags. Kywe et al. [Kywe, Hoang, lim et al. (2012)] used a similarity-based approach to solve this problem. They recommend hashtags by combining tags from similar tweets and similar users. Lu et al. [Lu, Yu, Chang et al. (2009)] found that similar web pages often have the same tags and proposed a method that takes into account the tag

information and page content to achieve the task.

There are also some approaches to focus on topic modeling. Based on the hypothesis that hashtags and triggers are from two different languages but with the same meaning, Ding et al. [Ding, Qiu, Zhang et al. (2013)] proposed a translation process to simulate this task. Godin et al. [Godin, Slavkovikj, De Neve et al. (2013)] established a theme model to tag recommendations for tweets. However, most of these works are based on tweet text information. Other studies have found that it is helpful to have different types of information. Zhang et al. [Zhang, Gong, Sun et al. (2014)] proposed a topic model-based approach that combines time information with personal information. Sedhai et al. [Sedhai and Sun (2014)] combine text information and hyperlink information to recommend tags. In addition, unlike these methods, Shi et al. [Shi, Ifrim and Hurley (2016)] proposed a label correlation modeling method. With the development of deep neural networks, CNN has been applied to this task [Gong and Zhang (2016)]. Huang et al. [Huang, Zhang, Gong et al. (2016)] used end-to-end memory networks with hierarchical attention to perform the hashtag recommendation task.

Most of the above work is based on textual information. In addition to these methods, the information brought by the image data has also received extensive attention. These works focus on tags that users annotate through social media services such as Flickr, Zooomr. These studies mainly recommend tags that sever as good descriptors of the photo itself. Sigurbjörnsson et al. [Sigurbjörnsson and Van Zwol (2008)] studied the tag recommendation task for images. When a user uploads a photo and enters some tags, an ordered list of candidate tags is derived for each of the entered tags. In Garg et al. [Garg and Weber (2008)], based on the statistics of tags co-occurrence, the issue of personalized, interactive tags recommendation was also studied. In Li et al. [Li and Snoek (2013)], tags correlation is classified using a set of support vector machines for each tag. However, most studies treat tag recommendation tasks as multi-category task.

Different from these methods, we propose a hashtag recommendation model that combines self-attention mechanism. New tags can be recommended for tweets where the tags have never been appeared.

3 Approach

This section mainly introduces the methods we proposed. Firstly, we introduced the basic self-attention mechanism. Then we introduced self-attention tag recommendation model and training methods that we proposed. Finally, we present the methods of how do we apply the learned model to achieve the best result of the hash tag recommendation.

3.1 Self-attention

Self-attention is a special case of the attention mechanism, and only one sequence is needed to calculate its representation. Self-attention has been successfully applied to many tasks, including reading comprehension [Cheng, Dong and Lapata (2016)], abstract summarization [Parikh, Täckström, Das et al. (2016)] and machine translation [Vaswani, Shazeer, Parmar et al. (2017)].

The main application of our model is multi-head attention [Vaswani, Shazeer, Parmar et al. (2017)]. Multi-attention is a variant of scaled dot-product attention. Commonly, the

used attention functions have additive attention and dot product attention. Compared with the additive attention, dot product attention is faster and more space-saving in practical application. Scaled dot-product attention refers to the addition of a large-scale dot-product function on the basis of dot product attention. Given queries, keys and values as input, and the calculation formula for scaled dot-product attention is as follows:

$$\text{Att}ention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d^k}})V \tag{1}$$

where Q is a matrix consisting of a set of queries, K and V are matrices made up of keys and values.

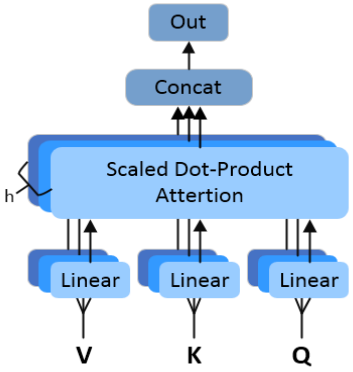

Figure 1: Computational diagram of the multi-head attention mechanism

Fig. 1 depicts a computational diagram of the multi-head attention mechanism. The Q, K, V are passed into the linear transformation and then enter into the scaled dot-product attention. This process needs to execute h times in parallel, with different parameters for each linear transformation. Connect h output values together, and then perform a linear transformation to get the output of multi-head attention. The specific calculation formula is as followings:

$$\text{head}_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2}$$

$$\text{Multi}Head(Q,K,V) = Conect(head_1,\cdots,head_h)W^O \tag{3}$$

where $W_i^Q \in R^{\frac{n\times d}{h}}$, $W_i^K \in R^{\frac{n\times d}{h}}$, $W_i^V \in R^{\frac{n\times d}{h}}$ and $W^O \in R^{d\times d}$.

The self-attention mechanism has many advantages. Self-attention can capture some syntactic or semantic features between words in the same sentence. It is easier to capture the interdependent features of long distance in sentences after introducing self -attention, and there is no gradient disappearing problem compared with RNN.

3.2 The proposed methods

In order to solve the hashtags recommendation issue better, we proposed a novel network architecture, where is shown in Fig. 2.

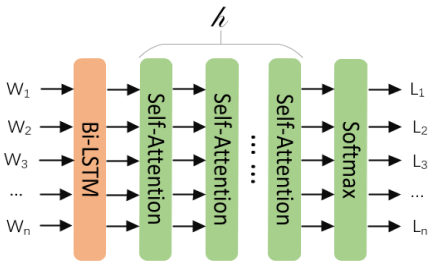

Figure 2: The self-attention network architecture

Unlike other methods, our model does not considered the hashtag recommendation task as a multi-classification task, but as hashtags prediction task. Our goal is to predict which word in a tweet is appropriate as the hashtag for this tweet. For each word in the tweet, we predict the probability that it will become the hashtag for this tweet. Our approach is very simple, similar to the sequence labeling issue. For each word in a tweet, we assign it a label with a value of 0 or 1 (1 is a hashtag, 0 is not).

Firstly, each word of a given tweet t is represented as a vector x_i of vocabulary. Then, we sum the embedded vectors to get a sentence-level tweet representation: $t=x_1, x_2, x_3, \ldots, x_T$. The maximum length of T is 30. In our work, if the length of sentences is less than T, the remainder is filled with zero.

Since the LSTM has shown good performance in text understanding task and has been widely used in recent years, we use it to generate text features. At each step, the LSTM unit accepts an input vector x_t and outputs a hidden state h_t. The details are as follows:

$$i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i) \tag{4}$$

$$f_t = \sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f) \tag{5}$$

$$o_t = \sigma(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o) \tag{6}$$

$$g_t = \tanh(W_g \cdot x_t + U_g \cdot h_{t-1} + b_g) \tag{7}$$

$$c_t = i_t \otimes g_t + f_t \otimes c_{t-1} \tag{8}$$

$$h_t = o_t \otimes \tanh c_t \tag{9}$$

where i_t is the input gate, f_t is the forget gate, c_t is the memory unit, o_t is the output gate, and g_t is the extracted feature vector.

In our model we use Bi-LSTM. The Bi-LSTM includes a forward LSTM and a backward LSTM. The forward LSTM encodes the sentence from left to right, and the backward LSTM encodes the sentence from right to left:

$$\vec{h}_t = LSTM(x_t, \vec{h}_{t-1}) \tag{10}$$

$$\overleftarrow{h}_t = LSTM(x_t, \overleftarrow{h}_{t-1}) \tag{11}$$

Then, the hidden layer states obtained by the forward network and the backward network are concatenated together to obtain the representation of sentence:

$$h_t = \begin{bmatrix} \overrightarrow{h}_t \\ \overleftarrow{h}_t \end{bmatrix} \tag{12}$$

After that, the representation of sentence is input into the stacked self-attention network to capture the potential dependencies between the nested structure of the sentence and the tag. Finally, we apply the softmax layer to get the probability that each word has a label of 1. The word with a higher probability value is the hashtag recommended for the tweet.

4 Experiments

4.1 Data sets and parameters

Our dataset collects public tweets from randomly selected users. First, we randomly selected 30,000 users to crawl the tweets. Second, we filtered out tweets that are not in English and tags that are not used very frequently. According to statistics, there are 27,000 tweets in our evaluation. The number of #tags in the corpus is 3885, and the average number of #tags in each tweet is 1.287. The list of labels annotated by the users is considered as a basic fact. The specific statistics are shown in Tab. 1. All the tweets are processed by deleting stop words and special characters. In our experiments, we divided the data set into a training set, a validation set, and a test set. There were 232,378 tweets in the training set, and 28,092 tweets in the verification set, and the remaining 20,875 tweets were in the test set.

We experimentally studied the effects of different parameters in our proposed method: word embedding size, learning rate, small batch size, Bi-LSTM hidden size and so on. For the initialization of the word embedding used in our model, we use the publicly available word2vec vectors, which are 100 billion words trained from Google News. The final hyperparameters are shown in Tab. 1.

Table 1: Hyperparameters of our model

Parameter	values
Minibatch size	10
Word embedding size	300
Bi-LSTM hidden size	192
Rate of regularization	0.01
Self-attention size	64
Attention heads	6
Learning rate	0.01

4.2 Baseline

In this section, in order to compare with our model, we choose some effective methods as the baseline, as described below:

1. **Naive Bayes (NB):** In order to complete the task, we converted the hashtag recommendation into a classification problem. We use naive Bayes methods to model the posterior probability of each label.
2. **Support Vector Machine (SVM):** This method is proposed by Chen et al. [Chen, Chang, Cahang et al. (2008)] to use SVM to solve the label recommendation problem.
3. **TTM:** TTM is a hashtag recommendation using only text information proposed by Godin et al. [Godin, Slavkovikj, Neve et al. (2013)]. The topic translation model is used to recommend hashtags.
4. **CNN-Attention:** CNN-attention by Gong et al. [Gong and Zhang (2016)]. It is a convolutional neural network structure with attention mechanism.
5. **HMemN2N:** To incorporate the user history information, Huang et al. [Huang, Zhang, Gong et al. (2016)] to extend the end-to-end memory networks to perform the hashtag recommendation task. It was the state-of- the-art method for this task. In this paper, we compare our method with it.

4.3 Experiment and analysis

In this section, we compare the proposed method with existing methods using real data. The experimental results are shown in Tab. 2. It can be observed that our method works better than the state-of-the-art method.

Table 2: Comparison of the proposed method with existing methods

Methods	Precision
NB	0.102
SVM	0.211
TTM	0.231
CNN-Attention	0.311
HMemN2N	0.518
SA-LSTM	0.557

We experimented with the existing methods on real data, and we can see that our method works best. Compared to these methods, our model has been significantly improved. HMemN2N is the state-of-the-art method. Compared with it, our method has increased by 4%. It can be seen that HMemN2N is not a good solution to treat the hashtag recommendation task as a traditional multi-classification problem.

NB and SVM are the basic methods which is used to solve multi-classification problems. It can be seen from Tab. 2 that their results are very low, which indicates that treating tag recommendation task as classification task is not an optimal solution. TTM uses the topic translation model for tag recommendations, and the results are not very good. CNN-Attention uses convolutional neural networks and uses attention mechanisms to recommend hashtags, yet it still considers hashtag recommendations as a multi-class

classification problem. HMemN2N is currently the state-of-the-art method. It adds user's historical interest information while using tweet text information, thus achieving a good result. Compared to these methods, our method achieves the best results. We propose a novel self-attention network to calculate the probability that each word in the tweet becomes a hashtag. This way, our model can recommend new hashtags to tweets even in the absence of similar tags. This clearly demonstrates the effectiveness of our approach.

5 Conclusion

This paper proposes a novel self-attention network architecture that translates hashtag recommendation issues into sequence labeling task. The model can calculate the probability that each word in the tweet becomes a hashtag. The greater the probability value, the more likely it is to become a hashtag. This is the biggest contribution of our work. Unlike other solutions, our proposed method can recommend new hashtags for tweets. This is something that the previous method could not do. To evaluate the proposed approach, we collect data from real twitter services. The experimental results on the evaluation data set show that the proposed method achieves better results in this task than the state-of- the-art methods.

References

Benevenuto, F.; Magno, G.; Rodrigues, T.; Almeida, V. (2010): Detecting spammers on twitter. *Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, vol. 6, no. 2010, pp. 12.

Cheng, J.; Dong, L.; Lapata. M. (2016): Long short-term memory-networks for machine reading. arXiv:1601.06733.

Chen, H. M.; Chang, M. H.; Chang, P. C.; Tien, M. C.; Hsu, W. H. et al. (2008): Sheepdog: group and tag recommendation for flickr photos by automatic search-based learning. *Proceedings of the 16th ACM international conference on Multimedia*, pp. 737-740.

Ding, Z.; Qiu, X.; Zhang, Q.; Huang, X. (2013): Learning topical translation model for microblog hashtag suggestion. *International Joint Conference on Artificial Intelligence*, pp. 2078-2084.

Garg, N.; Weber, I. (2008): Personalized, interactive tag recommendation for flickr. *Proceedings of the 2008 ACM Conference on Recommender Systems*, pp. 67-74.

Godin, F.; Slavkovikj, V.; De Neve, W.; Schrauwen, B.; Van de Walle, R. (2013): Using topic models for twitter hashtag recommendation. *Proceedings of the 22nd International Conference on World Wide Web*, pp. 593-596.

Gong, Y.; Zhang, Q. (2016): Hashtag recommendation using attention-based convolutional neural network. *Proceedings of the 26rd International Joint Conference on Artificial Intelligence*, pp. 2782-2788.

Heymann, P.; Ramage, D.; Garciamolina, H. (2008): Social tag prediction. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 531-538.

Huang, H.; Zhang, Q.; Gong, Y.; Huang, X. (2016): Hashtag recommendation using

end-to-end memory networks with hierarchical attention. *26th International Conference on Computational Linguistics*, pp. 943-952.

Krestel, R.; Fankhauser, P.; Nejdl, W. (2009)**:** Latent dirichlet allocation for tag recommendation. *ACM Conference on Recommender Systems*, pp. 61-68.

Li, X.; Snoek, C. G. (2013): Classifying tag relevance with relevant positive and negative examples. *Proceedings of the 21st ACM International Conference on Multimedia*, pp. 485-488.

Liu, Z.; Chen, X.; Sun, M. (2011): A simple word trigger method for social tag suggestion. *Conference on Empirical Methods in Natural Language Processing*, pp. 1577-1588.

Lu, Y. T.; Yu, S. I.; Chang, T. C.; Hsu, Y. J. (2009): A content-based method to enhance tag recommendation. *International Jont Conference on Artifical Intelligence*, pp. 2064-2069.

Kyew, S. M.; Hang, T. A.; Lim, E. P.; Zhu, F. (2012): On recommending hashtags in twitter networks. *Social Informatics*, pp. 337-350.

Parikh, A. P.; Täckström, O.; Das, D.; Uszkoreit, J. (2016): A decomposable attention model for natural language inference. arXiv:1606.01933.

Sedhai, S.; Sun, A. (2014): Hashtag recommendation for hyperlinked tweets. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 831-834.

Shen, Y.; Huang, X. (2016): Attention-based convolutional neural network for semantic relation extraction. *26th International Conference on Computational Linguistics*, pp. 2526-2536.

Shi, B.; Ifrim, G.; Hurley, N. (2016): Learning-to-rank for real-time high-precision hashtag recommendation for streaming news. *International Conference on World Wide Web*, pp. 1191-1202.

Sigurbjörnsson, B.; Van Zwol, R. (2008): Flickr tag recommendation based on collective knowledge. *Proceedings of the 17th International Conference on World Wide Web*, pp. 327-336.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L. et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, pp. 5998-6008.

Wang, X.; Wei, F.; Liu, X.; Zhou, M.; Zhang, M. (2010): Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pp. 1031-1040.

Xiong, Z.; Shen, Q.; Wang, Y.; Zhu, C. (2018): Paragraph vector representation based on word to vector and CNN learning. *Computers, Materials & Continua*, vol. 55, no. 1, pp. 1.

Zhang, Q.; Gong, Y.; Sun, X.; Huang, X. (2014): Time-aware personalized hashtag recommendation on social media. *25th International Conference on Computational Linguistics*, pp. 203-212.

Zhang, Q.; Wang, J.; Huang, H.; Huang, X.; Gong, Y. (2017): Hashtag recommendation for multimodal microblog using co-attention network. *Proceedings of the Twenty-sixth International Joint Conference on Artificial Intelligence*, pp. 3420-3426.