CAAI Transactions on Intelligence Technology

Special Issue: Internet of Things and Intelligent Devices and Services

Temporal enhanced sentence-level attention model for hashtag recommendation

ISSN 2468-2322
Received on 30th December 2017
Revised on 13th January 2018
Accepted on 21st January 2018
doi: 10.1049/trit.2018.0012
www.ietdl.org

Jun Ma ✉, *Chong Feng, Ge Shi, Xuewen Shi, Heyang Huang*

Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, College of Computer Science, Beijing Institute of Technology University, 5 South Zhong Guan Cun Street of Haidian District, Beijing, People's Republic of China
✉ *E-mail: jmaBIT@163.com*

Abstract: Hashtags of microblogs can provide valuable information for many natural language processing tasks. How to recommend reliable hashtags automatically has attracted considerable attention. However, existing studies assumed that all the training corpus crawled from social networks are labelled correctly, while large sample statistics on real social media shows that there are 8.9% of microblogs with hashtags having wrong labels. The notable influence of noisy data to the classifier is ignored before. Meanwhile, recency also plays an important role in microblog hashtag, but the information is not used in the existing studies. Some temporal hashtags such as *World Cup* will ignite at a particular time, but at other times, the number of people talking about it will sharply decrease. To address the twofold shortcomings above, the authors propose an long short-term memory-based model, which uses temporal enhanced selective sentence-level attention to reduce the influence of wrong labelled microblogs to the classifier. Experimental results using a dataset of 1.7 million microblogs collected from SINA Weibo microblogs demonstrated that the proposed method could achieve significantly better performance than the state-of-the-art methods.

1 Introduction

Microblogs, such as Twitter and SINA Weibo, allow users to insert a relevant keyword or phrase starting with the hash '#' symbol (e.g. #AI), which called hashtag. Hashtag indicates the core idea of microblogs and can bring the microblogs with the same topic or event together to strengthen the information dissemination. It has been proved that hashtags can provide valuable information for many natural language processing (NLP) tasks such as microblog retrieval [1], spammer detection [2], sentiment analysis [3–5], query expansion [6], and popularity prediction [7].

Unfortunately, relatively few microblogs contain hashtags manually annotated by their authors. Therefore, recommending reliable hashtags for microblogs automatically has attracted considerable attention in recent years. Most existing methods are based either on probabilistic models or on deep neural network models. Probabilistic model methods range from discriminative models with handcrafted features [8, 9] and generative models [10–12, 23] to collaborative filtering [24]. Deep neural network methods model the hashtag recommendation task as a multi-class classification problem [13] and some work incorporated attention mechanism [14–17] to gain better performance.

All existing studies directly used crawled microblogs with hashtags as training corpus, which assume that the user-generated instances with hashtags are labelled correctly. However, there are lots of noisy instances marked with irrelevant hashtags in real crawled data. To show empirical evidence of the noisy instance, we estimate the ratio of noise instances by randomly sampling the data.

- First, we gained over five million microblogs set range from January 2013 to June 2014 using our crawler after several pre-processing steps.
- Second, constructing statistical data by randomly selecting five groups which each contains 1000 microblogs from the whole collected dataset.
- Then, we employed six graduate students majoring in Chinese to annotate each microblog as correct labelled or wrongly labelled. They have read relevant instructions before annotating, and each microblog is labelled by at least three people.
- Finally, the majority opinion is considered as the final annotation result.

The statistic results show that 8.9% of microblogs with hashtags are intentionally wrongly labelled. By further investigation and analysis, we sum up those improperly labelled microblogs into three major error categories: for product promotion (intentional error), for blogger hype (intentional error) and for weak relevance (unintentional error). Table 1 gives examples of these kinds of errors. The first microblog is about the product advantages of teeth whitening powder. It is a product promotion strategy that attracts users reading by tagging popular hashtags. The second type can be attributed to the popularity of 'instant internet celebrity'. As more and more people want to be this online star, they will label the microblog with hot topic tags despite the content when they post their articles, photos or videos. The microblog content of these two types mentioned above is not related to the hashtag labelled. The third example is about Bieber's song, but it has nothing to do with Bieber and Selena. The third type of error annotation is that the content of the microblog has weak relevance to the hashtag. If we train the classifier on such noisy training corpus directly, the performance will be affected by these wrong labelled microblogs.

Meanwhile, recency also plays an important role in microblog hashtags. Some hashtags such as *World Cup* and *Grammy Award* will ignite at a particular time, but at other times, the number of people talking about it will sharply decrease. However, existing neural network based methods to solve the hashtag recommendation task only consider the textual information of microblogs and not consider the temporal information.

To address the shortcomings of existing methods, we propose a long short-term memory (LSTM)-based model which uses selective sentence-level attention to reduce the influence of wrong labelled microblogs to the classifier and introduce temporal information to expand the attention model. To evaluate the effectiveness of our model, we carry out experiments on a dataset of 1.7 million microblogs collected from SINA Weibo. Experimental results illustrate

Table 1 Types and examples of error labelled microblog

Types	Examples	Proportion, %
for product promotion (intentional error)	#Bieber shopping with Selena# The price of teeth whitening powder is [13.9] after using the coupon. It can not only whiten your yellow teeth but also remove bad breath, smoke stains, and tea stains.	5.1
for blogger hype (intentional error)	#Bieber shopping with Selena# Start playing! Recorded a funny short video, please pay attention to me.	2.6
for weak relevance (unintentional error)	#Bieber shopping with Selena# Justin Bieber's 'Love Yourself' live version, the magnetic voice is very comfortable to hear!	1.2

that the proposed model can achieve 2.6% improvement in F1-score than the state-of-the-art model by training with noisy data directly. It is also better than only using the textual information of microblogs.

The main contributions of our work can be summarised as follows.

- To model the notable errors and reduce its influence in microblogs with the labelled hashtag, we introduce selective sentence-level attention to assign different weights to each sentence.
- We incorporate the temporal information of microblogs into the sentence-level attention model to further improve hashtag recommendation.
- Experimental results demonstrate that the proposed method can achieve significantly better performance than the state-of-the-art methods on the large-scale dataset.

2 Related work

With the growing demands for hashtag recommendation, many methods have been proposed from different perspectives. There are two major types of existing approaches: probabilistic model methods and deep neural network methods. We will also introduce some effective noise reduction methods based on the distant supervision strategy and compare their merits and demerits. Besides, how to use LSTM networks in the NLP field will be explained in this section.

2.1 Probabilistic model methods for hashtag recommendation

A probabilistic model is often used for recommendation tasks. Among the probabilistic model based methods, Mazzia and Juett [23] used the Naive Bayes (NB) model to recommend hashtag for microblogs, which generates a list of top 20 recommended hashtags. Ding *et al.* [10] focused on the topic model and assume that the content and hashtags of the tweet are talking about the same themes but written in different languages. They converted hashtag suggestion into a translation process from content to hashtags. Sedhai and Sun [18] formulated the task as learning to the rank problem and adopt RankSVM to aggregate and rank the candidate hashtags.

Most of these methods use word-level features, including term frequency–inverse document frequency (TF-IDF) and exquisitely designed patterns to perform the task. However, feature engineering is labour-intensive, as well as the sparse and discrete features it created, could ignore the semantic information in microblogs. Unlike long texts, a microblog is a kind of short text. It is impossible to understand the content without semantic information only based on only word frequency.

2.2 Neural network methods for hashtag recommendation

In recent years, the deep neural network has been widely applied in computer vision, NLP and other fields. This technology has also been introduced into the hashtag recommendation task. Li *et al.* [13] proposed a recurrent neural network model to learn vector-based tweet representations to recommend hashtags. Gong *et al.* [14] proposed an attention-based convolutional neural network (CNN) architecture, which consists of a local attention channel and global channel. Yang Li *et al.* [15] proposed an attention-based LSTM model which incorporates topic modelling into the LSTM architecture through an attention mechanism. Apart from these methods focused on textual information of microblog only, there are some works utilising other types of information. Huang *et al.* [16] incorporated the histories of users into the external memory. Zhang *et al.* [17] proposed a co-attention network incorporating textual and visual information to recommend hashtags for multi-modal tweets.

All these neural network methods assumed that the microblogs in the training corpus are labelled correctly. They did not consider the impact of noisy data on the classifier. Recency played an essential role in microblog hashtag but was not taken into account either.

2.3 Noise reduction methods

Noise is common in the dataset, especially in distant supervision relation extraction. Zeng *et al.* [19] modelled the task as a multi-instance problem in which the uncertainty of instance labels is taken into account to solve the wrong label problem. Lin *et al.* [20] built sentence-level attention over multiple instances, which is expected to reduce the weights of those noisy instances dynamically. Luo *et al.* [21] designed a dynamic transition matrix structure to characterise the noise and a curriculum learning based framework to guide the training procedure to learn with noise adaptively.

In our work, we learn the noise reduction methods from distant supervision relation extraction. We adopt a temporal enhanced sentence-level attention model to reduce the influence of wrong labelled microblogs to the classifier.

2.4 LSTM networks

LSTM is a special form of recurrent neural networks, which has shown good performance in understanding text and has been widely used to model sequence data in recent years. 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 modelling of long-range dependencies [25].

At each time step, the LSTM unit takes an input vector $\boldsymbol{e}_t$ and outputs a hidden state h_t, using input gate i_t, memory cell c_t, forget gate f_t, and output gate o_t. The details are defined as follows:

$$i_t = \sigma\left(\boldsymbol{W}^i \cdot \left[h_{t-1}, e_t\right] + \boldsymbol{b}_i\right), \tag{1}$$

$$f_t = \sigma\left(\boldsymbol{W}^f \cdot \left[h_{t-1}, e_t\right] + \boldsymbol{b}_f\right), \tag{2}$$

$$o_t = \sigma\left(\boldsymbol{W}^{\mathrm{o}} \cdot \left[h_{t-1}, e_t\right] + \boldsymbol{b}_{\mathrm{o}}\right), \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \tanh \odot \tanh\left(\boldsymbol{W}^c \cdot \left[h_{t-1}, e_t\right] + \boldsymbol{b}_c\right), \tag{4}$$

$$h_t = o_t \tanh\left(c_t\right), \tag{5}$$

where σ stands for the sigmoid function, $\boldsymbol{W}$ is the weight matrix, $\boldsymbol{b} \in \mathbb{R}^d$ are bias vectors, $\odot$ is element-wise multiplication. The output of the LSTM layer is a sequence of hidden vectors $\left[h_1, h_2, \ldots, h_M\right]$. Each annotation h_t contains information about the whole input microblog with a strong focus on the parts surrounding the *t*th word.

3 Proposed models

We model the hashtag recommendation task as a multi-class classification problem. To solve the noisy data in the training corpus, selective sentence-level attention is introduced to assign different weights to each sentence. Furthermore, the model is

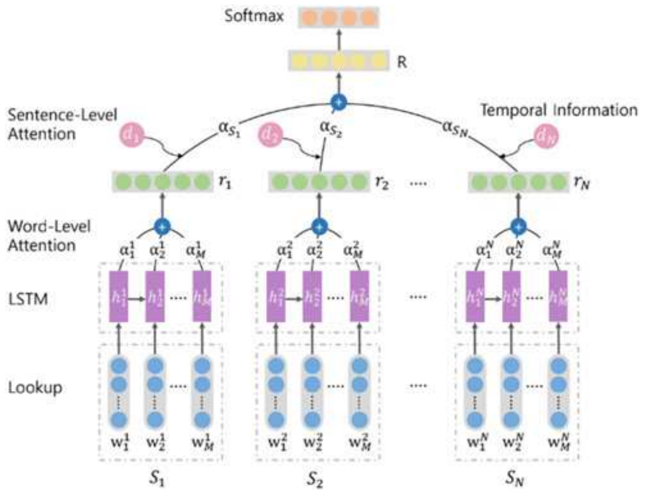

Fig. 1 *Architecture of temporal enhanced sentence-level attention model based on LSTM*

expanded and enhanced with temporal information. Fig. 1 describes the architecture of the neural network in the training phase used for hashtag recommendation. Since short texts, especially microblogs, have a non-standard syntax structure and contain a large number of named entities. They are not suitable for vectorisation at the text level. Word vectors can preserve more entity's information. Besides, word vectors are more expansible than sentence vectors and can be trained with more different fields of data. For example, sentence vectors trained by news data are completely unsuitable for microblogs, but word vectors can relatively alleviate this problem. Therefore, our model uses word vectors to generate the sentence vector rather than directly using doc2vec.

The model proposed in this study contains five layers of components:

(1) Lookup layer: considers each microblog as an integral sentence (for instance) and maps each word of microblog into a low-dimensional vector. The word embedding is pertained by gensim.
(2) LSTM layer: utilises LSTM unit to get high-level semantic features rather than statistical properties from step (1).
(3) Word-level attention layer: produces a weight vector for each word, and merge word-level features into a sentence-level vector. This layer can help focus on more important words in the microblog, rather than considering all words equally.
(4) Sentence-level attention layer: incorporates the output vector of word-level attention layer and the temporal information of every microblog, to produce a weight vector for each instance with the same hashtag labelled to select the microblogs which really related to the hashtag.
(5) Softmax layer: obtains the hashtag associated with the input vector from step (4).

During the training phase, we randomly select a certain number of microblogs with the same hashtag labelled into a bag and then give a different weight vector to each instance through a temporal enhanced sentence-level attention model to reduce the influence of wrong labelled microblogs to the classifier. During the test phase, the input bag consists of only one instance to predict, and other structures are the same as the training stage.

3.1 LSTM-based sentence encoder

The original inputs of the first part are bags, each of which contains N microblogs with the same hashtag. The word representation of every microblog is $S_N = \{w_1, w_2, \ldots, w_M\}$, where M is the maximum length of the microblogs. In our work, sentences with length less than M are padded with zeros. Every word w_i is converted into a real-valued vectored e_i first. Then, for each word in S_N, we look up the embedding matrix $\boldsymbol{W} \in \mathbb{R}^{d^{\omega}|V|}$, where V is a fixed-sized vocabulary and d^{ω} is the size of the word embedding. The matrix $\boldsymbol{W}$ is pre-trained by word2vec and d^{ω} is a hyper-parameter which can be chosen by a user. Every word w_i is transformed into its embedding e_i by using the matrix-vector product

$$e_i = \boldsymbol{W}\boldsymbol{v}^i, \tag{6}$$

where v^i is a vector of size $|V|$ which has value 1 at index e_i and 0 in all other positions.

Then the real-valued vector $\text{emb}_N = \{e_1, e_2, \ldots, e_M\}$ is feed into the LSTM layer. The output of the LSTM layer is a sequence of hidden vectors $[h_1, h_2, \ldots, h_M]$. Each annotation h_t contains information about the whole input microblog with a strong focus on the parts surrounding the tth word.

3.2 Attention mechanism incorporating temporal information

The attention mechanism has been demonstrated successfully in a wide range of NLP tasks. In this hashtag recommendation task, not only the word-level attention was used but also introduced selective sentence-level attention incorporating temporal information of microblogs to reduce the influence of noisy data in the training corpus.

First, the word-level attention mechanism is introduced for a hashtag recommendation task because all words of microblog should not be equally important. $\boldsymbol{H} = [h_1, h_2, \ldots, h_M]$ is a matrix consisting of output vectors that the LSTM layer produced. The representation r of the microblog is computed as a weighted sum of each hidden state h_j. Just as these following equations demonstrate:

$$\alpha_W = \text{softmax}\big(\boldsymbol{\omega}^{\mathrm{T}} \tanh(H)\big), \tag{7}$$

$$r = H\alpha_W^{\mathrm{T}}, \tag{8}$$

where $\boldsymbol{\omega}$ is a trained parameter vector and $\boldsymbol{\omega}^{\mathrm{T}}$ is its transpose.

In the large training corpus, the wrong labelled problem inevitably occurs. If every microblog is considered equally, the wrong labelled microblogs will bring in lots of noises during the training phase. Therefore, a selective sentence-level attention model over multiple instances with the same hashtag is introduced, which is expected to reduce the weights of those noisy microblogs dynamically. Besides, some temporal hashtags can be considered as topics which have a limited lifespan. They have high frequencies in a specific period and low frequencies in other times. Temporal information has a particular indicative effect on a hashtag recommendation task. So the model is expanded and enhanced with temporal information to improve the performance. Above all, a two-dimensional matrix $\boldsymbol{B} \in \mathbb{R}^{|\text{time}|\times|\text{hashtag}|}$ is defined, where |time| means the number of time nodes and |hashtag| means the number of hashtags. Time nodes are like yyyy-mm-dd-hh which contain year, month, day and hour. This two-dimensional matrix $\boldsymbol{B}$ is a parameter that needs to be trained. Thus, given a tuple like <microblog S_i, hashtag h>, we can look up an element (denoted as d_i) from the matrix $\boldsymbol{B}$ according to the time and the hashtag of the microblog S_i.

The output of the word-level attention layer is a set $S = \{r_1, r_2, \ldots, r_N\}$. Our model represents the set S with a real-valued vector $\boldsymbol{R}$ when predicting hashtag. The set vector $\boldsymbol{R}$ is computed as a weighted sum of these sentence vector $\boldsymbol{r}_i$

$$m_i = \boldsymbol{r}_i\boldsymbol{A}t, \tag{9}$$

$$\alpha_{Si} = \frac{\exp(m_i \times d_i)}{\sum_k \exp(m_k \times d_k)}, \tag{10}$$

$$\boldsymbol{R} = \sum \alpha_{Si}\boldsymbol{r}_i, \tag{11}$$

where $\boldsymbol{A}$ is a weighted diagonal matrix, and $\boldsymbol{t}$ is the query vector associated with the hashtag, m_i is referred as a query-based function which scores how well the input sentence representation $\boldsymbol{r}_i$ and the predict hashtag t matches, d_i is the temporal information and α_{Si} is the weight of each sentence vector $\boldsymbol{r}_i$.

After having the output vector $\boldsymbol{R}$ of the sentence-level attention layer, the final output which corresponds to the scores associated with all hashtags is defined as follows:

$$O = \boldsymbol{M}R + \boldsymbol{b} \tag{12}$$

where $\boldsymbol{b}$ is a bias vector and $\boldsymbol{M}$ is the representation matrix of hashtags.

Then a softmax layer is added to output the probability distributions of all candidate hashtags whose length is the number of hashtags. The softmax function is defined as follows:

$$\hat{p}(t|S,\theta) = \frac{\exp(O_i)}{\sum_{i'}^{K}\exp(O_{i'})}, \tag{13}$$

$$\hat{t} = \arg\max_{t}\ \hat{p}(t|S,\theta), \tag{14}$$

where K is the total number of hashtag categories.

3.3 Training

The training objective function by minimising the cross-entropy error of the hashtag recommendation in our model is defined as follows:

$$J(\theta) = \sum_{i=1}^{N}\log\hat{p}(t_i|S_i,\theta) \tag{15}$$

where N indicates the number of sentence sets and θ indicates all parameters of our model.

To solve the optimisation problem, stochastic gradient descent with the Adam is used to minimise the objective function. For learning, a mini-batch from the training set is selected randomly to iterate until converging. In addition, dropout regularisation has been proved to be an effective method for reducing the overfitting in deep neural networks with millions of parameters. In this work, L2-norm regularisation terms are added as the parameters of the network to augment the objective function.

4 Experiments

4.1 Dataset and evaluation metrics

The 28.4G microblogs set was collected range from January 2013 to June 2014 using our crawler. After observing the raw data, several preprocessing steps were elaborately designed.

- Firstly, to simplify the hashtag recommendation task, we only consider the microblogs with the single hashtag. So, we deleted the microblogs with several hashtags, and there were 5,732,360 microblogs left.
- Then, to get pure text contents, we got rid of the duplicated microblogs and retweets.
- Consequently, jieba was used for Chinese word segmentation. Punctuations were not meaningful in semantic coding, so they were also removed.
- Finally, the top 2000 popular hashtags were selected, there were 1,692,507 microblogs left.

We randomly selected 160,000 microblogs as a testing set and other 1,532,507 microblogs as a training set. The vocabulary of words is 159,427 in the dataset, and the max length of microblog after the word segmentation is 65. The overall statistics of our dataset is shown in Table 2.

Table 2 Overall statistics of the dataset

#Microblogs	#Hashtags	Vocabulary size	Max length
1,692,507	2000	159,427	65

To evaluate the performance, we use precision (P), recall (R), and F1-score (F1)

$$P = \frac{N_\mathrm{r}}{N_\mathrm{c}}, \tag{16}$$

$$R = \frac{N_\mathrm{r}}{N_\mathrm{d}}, \tag{17}$$

$$F_1 = \frac{2PR}{P+R}, \tag{18}$$

where N_r is the right number recommended, N_c is the total number recommended by the classifier, and N_d is the total number of the hashtags assigned to the dataset.

4.2 Baselines and experimental settings

In this section, to evaluate the proposed model which incorporates selective sentence-level attention and temporal information, we compared with the following methods, which contain probabilistic model methods and neural network methods:

- NB: NB is applied to model the posterior probability of each hashtag only using the textual information of the microblogs. This method is a traditional machine learning algorithm and easy to implement, which only focus on the text and ignore the temporal information.
- CNN: the CNN is proposed for sentence classification by Kim [22]. We modify the public source code to complete the hashtag recommendation task. This method is a basic neural network method without attention mechanism. It is a primary method based on the semantic representation
- CNN-LSTM: This method is proposed by Jia Li [13]. They apply a CNN to learn semantic sentence vectors and then make use of the sentence vectors to train an LSTM network. Compared with the traditional CNN, this method can get better semantic representation. Compared with the simple CNN, the effect has been improved. However, it does not use temporal information and attention mechanism, either.
- Topical attention model (TAM)-LSTM: this method is proposed by Yang Li [15]. It is an attention-based LSTM model which incorporates topic modelling into the LSTM architecture through an attention mechanism. Moreover, it is the state-of-the-art method for this task. Compared with this method, our model takes into account the fact that the training set contains noise data and utilises the sentence level attention mechanism to reduce their influence to the classifier. Meanwhile, the temporal information is introduced.
- Sentence attention model (SAM)-LSTM: This is a variant of our proposed model, which only uses textual information with selective sentence-level attention without using temporal information of microblogs.

The first method NB is a probabilistic model. Others are all models based on neural networks. For the baselines and our models, 1,692,507 original microblogs data crawled from SINA Weibo are used to train the embedding by word2vec toolkit. The embedding has a dimensionality of 200. During the training phase, five microblogs with the same hashtag are randomly selected into a bag to train the classifier. A minibatch stochastic gradient descent (SGD) algorithm together with the Adam method is used to train the model. The number of training epoch is 4. The dimension of hidden states is 300. The value of batch size, learning rate and dropout rate is 100, 0.001 and 0.5, respectively. Besides, the multi-classification confidence threshold is 0.2.

Table 3 Evaluation results of different models for hashtag recommendation

Methods	Precision	Recall	F1
NB	0.519	0.502	0.510
CNN	0.598	0.561	0.579
CNN-LSTM	0.601	0.596	0.598
TAM-LSTM	0.649	0.637	0.643
SAM-LSTM	0.653	0.646	0.650
tSAM-LSTM	0.673	0.665	0.669

4.3 Results and analysis

Table 3 shows a comparison of the proposed model to the state-of-the-art methods on the constructed evaluation collection. Based on the results, we have the following observations:

(1) First of all, methods based on neural networks such as CNN performs better than the probabilistic model method NB, showing that the embedding feature can capture more semantic information than the static feature such as TF-IDF.
(2) Among methods based on the neural network, TAM-LSTM performs better than CNN and CNN-LSTM, showing that word-level attention mechanism is comparatively useful which can improve precision, recall and F1-score by >4%.
(3) SAM-LSTM, the variant of our proposed model, outperforms neural methods, TAM-LSTM, which only uses word-level attention. A reasonable explanation is that the effectiveness of selective sentence-level attention to reducing the influence of the noisy date in training corpus to the classifier. Therefore, it is necessary to denoise the training set which contains noise.
(4) Our model temporal enhanced sentence attention model (tSAM)-LSTM obtained the best result on all three metrics, showing that the temporal information of microblogs can further enhance the performance of sentence-level attention. Both sentence-level attention and temporal information have played a role in improving the classification results.

The hashtag recommendation task is considered as a multi-class classification problem. The classifier we trained produces a real-valued confidence score for its decision, rather than just a class label. With different classification confidence thresholds, the performance of the classifier is different. Fig. 2 shows the precision curves, Fig. 3 shows the recall curves and Fig. 4 shows the F1 curves which contain CNN, CNN-LSTM, TAM-LSTM, SAM-LSTM and tSAM-LSTM on the test data. These curves show the performance of different classifiers under different thresholds.

We set five confidence thresholds, which are 0.05, 0.2, 0.5, 0.8 and 0.98, respectively. From the figures, we can see that tSAM-LSTM outperforms all of the models in three metric curves with varying confidence threshold. It also indicates that incorporating temporal information and sentence-level attention based on LSTM encoding to reduce the influence of noisy data to the classifier is effective and robust in hashtag recommendation.

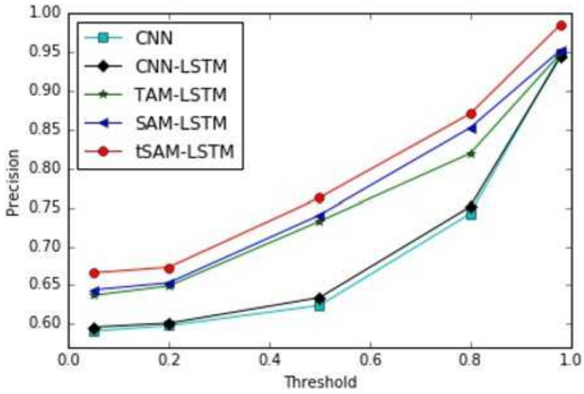

Fig. 2 *Precision with a different threshold*

4.4 Parameter influence

Proper embedding is crucial to forming a powerful textual representation at a higher level. We study the effects of embedding with a different dimension. Table 4 shows the precision, recall, and F1-score when the dimension of embedding is changed. From the results, we can see that with a certain range, as the size of the embedding dimension represents the expression ability of each word, a higher embedding dimension results in a better performance. However, the dimension cannot be increased without limitation. When the embedding dimension is varied from 200 to 300, the improvement of performance is not significant, but the network structure could be more complicated because of this. Therefore, 200 is the most suitable dimension to our model.

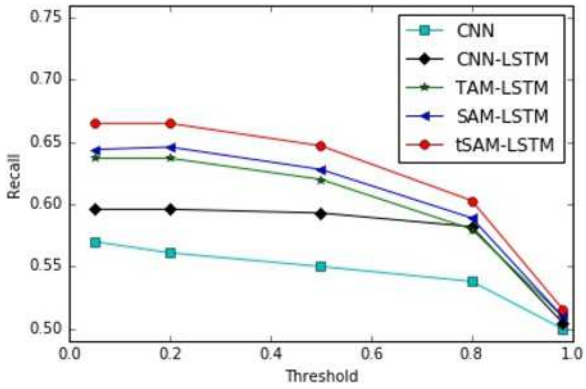

Fig. 3 *Recall with a different threshold*

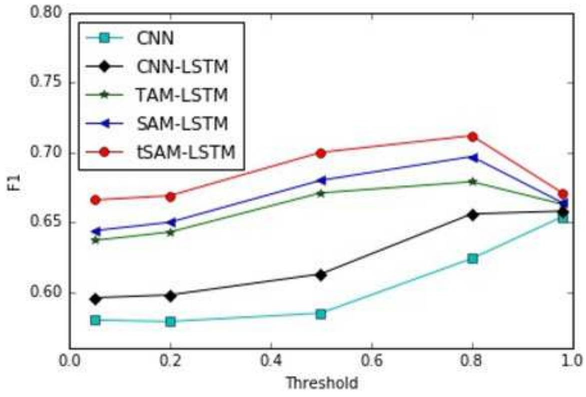

Fig. 4 *F1 with a different threshold*

Table 4 Evaluation results with a different dimension of embedding

Methods	Dim	Precision	Recall	F1
SAM-LSTM	50	0.632	0.623	0.627
	100	0.642	0.635	0.638
	200	0.653	0.646	0.650
	300	0.654	0.647	0.650
tSAM-LSTM	50	0.655	0.649	0.652
	100	0.668	0.662	0.665
	200	0.673	0.665	0.669
	300	0.673	0.666	0.669

5 Conclusion

To solve the problem of noisy microblogs in hashtag recommendation, we proposed a novel LSTM-based method, which could incorporate selective sentence-level attention to reducing the influence of noisy data to the classifier. The attention model was further improved by introducing the temporal information. Previous studies based on neural networks utilised the textual information only without considering the time factor. The evaluation of a large-scale dataset collected from SINA Weibo microblog demonstrates the effectiveness of our model. In the future work, more methods to reduce the noise will be tried.

6 Acknowledgments

The work was mainly supported by the National Key Research and Development Program of China (no. 2017YFB1002101) and the National Natural Science Foundation of China (no. U1636203).

7 References

[1] Efron, M.: 'Hashtag retrieval in a microblogging environment'. Proc. SIGIR'10, Geneva, Switzerland, 2010

[2] Benevenuto, F., Magno, G., Rodrigues, T., *et al.*: 'Detecting spammers on twitter'. Collaboration, Electronic Messaging, Anti-abuse and Spam Conf. (CEAS), Stockholm, Sweden, 2010, vol. 6, p. 12

[3] Davidov, D., Tsur, O., Rappoport, A.: 'Enhanced sentiment learning using twitter hashtags and smileys'. COLING, Beijing, China, 2010

[4] Wang, X., Wei, F., Liu, X., *et al.*: 'Topic sentiment analysis in twitter:a graph-based hashtag sentiment classification approach'. Proc. CIKM'11, Glasgow, Scotland, UK, 2011

[5] Mohammad, S.M., Kiritchenko, S., Zhu, X.: 'NRC-Canada: building the state-of-the-art in sentiment analysis of tweets'. Proc. Int. Workshop on Semantic Evaluation, Atlanta, USA, 2013

[6] Bandyopadhyay, A., Mitra, M., Majumder, P.: 'Query expansion for microblog retrieval'. Proc. TREC, Gaithersburg, MD, USA, 2011

[7] Tsur, O., Rappoport, A.: 'What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities'. Proc. Fifth ACM int. Conf. on Web search and data mining, Seattle, WA, USA, 2012, pp. 643–652

[8] Heymann, P., Ramage, D., Garcia-Molina, H.: 'Social tag prediction'. SIGIR, Singapore, 2008

[9] Liu, Z., Chen, X., Sun, M.: 'A simple word trigger method for social tag suggestion'. Proc. EMNLP, Edinburgh, UK, 2011, pp. 1577–1588

[10] Ding, Z., Qiu, X., Zhang, Q., *et al.*: 'Learning topical translation model for microblog hashtag suggestion'. Proc. Int. Joint Conf. on Artificial Intelligence, Beijing, China, 2013

[11] Godin, F., Slavkovikj, V., De Neve, W., *et al.*: 'Using topic models for twitter hashtag recommendation'. Proc. 22nd Int. Conf. on World Wide Web, Rio de Janeiro, Brazil, 2013, pp. 593–596

[12] She, J., Chen, L.: 'Tomoha: topic model-based hashtag recommendation on twitter'. Proc. 23rd Int. Conf. on World Wide Web, Seoul, Korea, 2014, pp. 371–372

[13] Li, J., Xu, H., He, X., *et al.*: 'Tweet modeling with LSTM recurrent neural networks for hashtag recommendation'. IEEE Int. Joint Conf. on Neural Networks (IJCNN), Vancouver, Canada, 2016, pp. 1570–1577

[14] Gong, Y., Zhang, Q.: 'Hashtag recommendation using attention-based convolutional neural network'. Int. Joint Conf. on Artificial Intelligence, New York, NY, USA, 2016

[15] Li, Y., Liu, T., Jiang, J., *et al.*: 'Hashtag recommendation with topical attention-based LSTM'. Proc. COLING 2016, the 26th Int. Conf. on Computational Linguistics: Technical Papers, Osaka, Japan, 11–17 December 2016, pp. 3019–3029

[16] Huang, H., Zhang, Q., Gong, Y., *et al.*: 'Hashtag recommendation using end-to-end memory networks with hierarchical attention'. Proc. COLING 2016, the 26th Int. Conf. on Computational Linguistics: Technical Papers, Osaka, Japan, 11–17 December 2016, pp. 943–952

[17] Zhang, Q., Wang, J., Huang, H., *et al.*: 'Hashtag recommendation for multimodal microblog using Co-attention network'. Proc. 26th Int. Joint Conf. on Artificial Intelligence, Melbourne, Australia, 2017

[18] Sedhai, S., Sun, A.: 'Hashtag recommendation for hyperlinked tweets'. Proc. SIGIR, Gold Coast , QLD, Australia, 2014

[19] Zeng, D., Liu, K., Chen, Y., *et al.*: 'Distant supervision for relation extraction via piecewise convolutional neural networks'. Proc. EMNLP, Lisbon, Portugal, 2015

[20] Lin, Y., Shen, S., Liu, Z., *et al.*: 'Neural relation extraction with selective attention over instances'. Proc. 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, August 2016, pp. 2124–2133

[21] Luo, B., Feng, Y., Wang, Z., *et al.*: 'Learning with noise: enhance distantly supervised relation extraction with dynamic transition matrix'. Meeting of the Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 430–439

[22] Kim, Y.: 'Convolutional neural networks for sentence classification'. Proc. EMNLP, Doha, Qatar, 2014

[23] Mazzia, A., Juett, J.: 'Suggesting hashtags on twitter, EECS 545m, machine learning', Computer Science and Engineering, University of Michigan, 2009

[24] Kywe, S.M., Hoang, T.-A., Lim, E.-P., *et al.*: 'On recommending hashtags in twitter networks'. Social Inf., 2012, 7710, pp. 337–350

[25] Hochreiter, S., Schmidhuber, J.: 'Long short-term memory', *Neural Comput.*, 1997, **9**, (8), pp. 1735–1780