Multi-label Community-based Question Classification via Personalized Sequence Memory Network Learning

Xinyu Duan[1]**, Shengyu Zhang**[2]**, Zhou Zhao**[1]**, Fei Wu**[1]
[1]{duanxinyu, zhaozhou, wufei}@zju.edu.cn, [2]light.e.gal@gmail.com
[1]College of Computer Science, Zhejiang University
[2]School of Information management, Wuhan University
No.38 Zheda Road, Hangzhou, Zhejiang, China, 310027

Abstract

Multi-label community-based question classification is a challenging problem in Community-based Question Answering (CQA), arising in many real applications such as question navigation and expert finding. Most of the existing approaches consider the problem as content-based tag suggestion task, which suffers from the textual sparsity issue. In this paper, we consider the problem from the viewpoint of personalized sequence learning. We introduce the personalized sequence memory network that leverages not only the semantics of questions but also the personalized information of askers to provide the sequence tag learning function to capture the high-order tag dependency. The experiment on real-world dataset shows the effectiveness of our method.

Introduction

The benefits of CQA have been well recognized today (Zhao et al. 2015). Most existing methods consider the problem of multi-label community-based question classification as text categorization task, which trains the discriminative classifier model, and then ranks the tags to the given question based on their semantic relevance. Although existing methods have achieved excellent performance, they mainly use the textual content, which still suffer from the textual sparsity issue. On the other hand, these methods ignore the important influences of question askers. Let us take the question ”what is the price of apple?” as an example. In this case, the businessman tends to know the stock price of the apple company while the teenagers may be aware of the price of apple product. Therefore, the role of question askers is important for tackling the textual sparsity issue of question classification, where the above question can be classified into the category of company or product based on the askers role.

In this paper, we consider the problem of multi-label community-based question classification from the viewpoint of Personalized Sequence Memory Networks (PSMN) learning. Our proposed model leverages not only the semantics of questions but also the peculiar interest of each user. Specifically, we introduce the tag sequence learning method with sequence memory networks to exploit the semantic tag dependency. We jointly learn the embedding of questions, personalized information (which can be extracted from user context) and tags through PSMN for multi-label community-based question classification. When a certain user posts a new question, our method can recursively generate the candidate tag sequences with the trained PSMN.

Methodology

We denote the input question by $Q = \{q_1, q_2, q_3, ..., q_n\}$, where n is the number of questions in question set Q. q_i is an one-hot vector representing the textual content of the i-th input question using bag-of-word (BOW) feature. We then denote the set of user context by $U = \{u_1, u_2, u_3, ..., u_m\}$ for personalized context representation, where m is the number of users. u_i is also an one-hot vector representing the K most frequent words in a decreasing order based on the TF-IDF scores from the user's posted questions. Using TF-IDF scores means that we hope to neglect the general terms that users commonly use, since they are not helpful for personalization. The set of tags are denoted by $L = \{l_1, l_2, l_3, ..., l_k\}$, where k is the number of distinct tags.

Inspired by some works of image caption, we construct three memories to store three types of context information: 1. question memory that stores the semantics of the original question; 2. user context memory for TF-IDF weighted K frequent words from questions one user post; 3. tag memory for the generated tags. The whole architecture of our model is outlined in Figure 1. The internal memory vector for each memory at time step t is computed as follows:

$$a_{*i} = W_{*a}x_i; \quad c_{*i} = W_{*c}x_i$$

$$q_t = W_l l_t; \quad p_{*i} = a_{*i}q_t; \quad o_{*i} = p_{*i}c_{*i}$$

where $*$ represents one of the three memory defined above. W represents the embedding matrix. x_i is the i-th stored feature in the corresponding memory. a, b and c are the internal vector of the memory. p is a probability vector over the input and o is the output vector of the memory.

Then we use a convolutional neural network (CNN) with one convolutional layer and one pooling layer to obtain a more powerful representation. The output vectors of CNN from three memories are concatenated and form m_t which is then passed through a final weight matrix W_g and a softmax to produce the predicted tag probability vector h_t:

$$h_t = Softmax(W_g m_t)$$

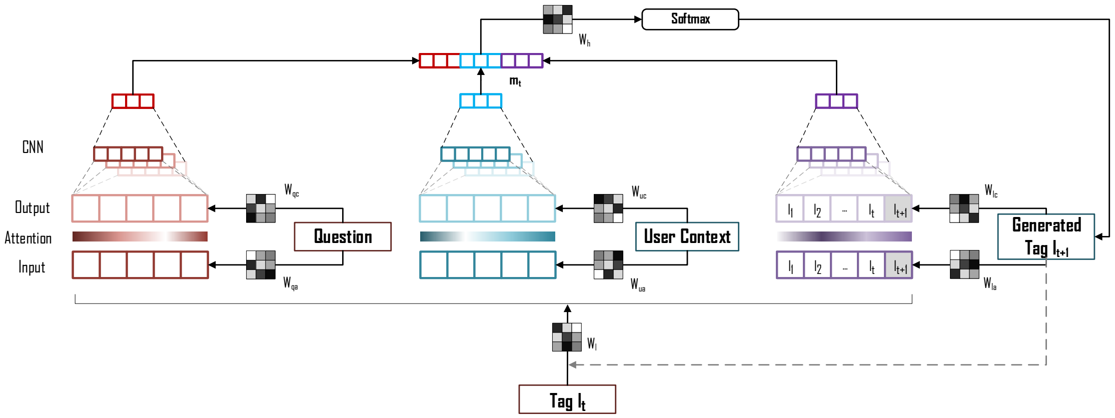

Figure 1: The overview of Personalized Sequence Memory Network.

Tag with the highest probability in h_t is considered as the generated tag l_{t+1} which is then inserted into the tag memory and treated as the input of the model at time $t + 1$ to generate next tag. Denoting all the parameters as θ, the objective function for training the model is given by:

$$\min_{\theta} L(\theta) = -\sum_{Q}\sum_{L} y_i * \log h_{ti} + \lambda \left\|\theta\right\|_2^2$$

where y is the true label of the answer. $\lambda > 0$ is a hyper parameter to trade-off the training loss and regularization. To optimization the objective function, we employ the stochastic gradient descent (SGD).

Experiments

We evaluate our method using the dataset from Stack Exchange. The dataset contains 116,072 questions, 216,522 words, 7,359 users and 9,358 tags from 26 different communities of Stack Exchange. Each question owns 5 tags at most and the average number of tags for a single question is 1.8522. The average length of the question is 192.18 and the longest question possesses 3,320 words. We use the first 70% posted questions as training set and the remaining ones for testing. We evaluate the performance using *Precision@K (K = 1, 2, 3)*. The compared four baselines are as follows:

TSTM (Huang 2012) is a topic specific translation model based on latent topic allocation for automatic tag suggestion.

WTM (Liu, Chen, and Sun 2011) is a word trigger method that suggests tags according to the words in a resource description based on the translation model.

TTM (Ding et al. 2013) is a topic translation model which combines both topic model and translation model.

SRW (Wu et al. 2016) is the supervised random walk that leverages similar questions and similar tags to boost the recommendation of tail tags and learn question similarity, tag similarity, and tag importance in a unified framework.

The experimental results in Table 1 demonstrate that our proposed PSMN achieves the best performance in all cases, which illustrates that the performance of multi-label question classification can be further improved by employing both the semantics of the questions and personalized information of question askers.

Methods	TSTM	WTM	TTM	SRW	Ours
P@1 (%)	12.94	24.87	25.11	25.42	26.14
P@2 (%)	10.03	18.22	18.45	19.98	21.51
P@3 (%)	6.85	10.21	13.48	15.83	19.34

Table 1: Experimental results on Precision@K (K = 1, 2, 3).

Conclusion

In this paper, we consider the problem of multi-label community-based question classification from the viewpoint of personalized sequence memory network learning. Our proposed model leverages not only the semantics of questions but also the personalized information of each user to capture the high-order tag dependency.

Acknowledgement

This work was supported by NSFC (U1611461), Key Program of Zhejiang Province (2015C01027), China Knowledge Center for Engineering and Microsoft Research Asia.

References

Ding, Z.; Qiu, X.; Zhang, Q.; and Huang, X. 2013. Learning topical translation model for microblog hashtag suggestion. In *IJCAI*, 2078–2084.

Huang, Z. D. Q. Z. X. 2012. Automatic hashtag recommendation for microblogs using topic-specific translation model. In *ICCL*, 265.

Liu, Z.; Chen, X.; and Sun, M. 2011. A simple word trigger method for social tag suggestion. In *EMNLP*, 1577–1588.

Wu, Y.; Wu, W.; Li, Z.; and Zhou, M. 2016. Improving recommendation of tail tags for questions in community question answering. In *AAAI*, 3066–3072.

Zhao, Z.; Zhang, L.; He, X.; and Ng, W. 2015. Expert finding for question answering via graph regularized matrix completion. *IEEE Trans. Knowl. Data Eng.* 27(4):993–1004.