Overview of the NLPCC 2018 Shared Task: Automatic Tagging of Zhihu Questions

Bo Huang(✉) and Zhenyu Zhao(✉)

Zhihu Institute, No. 5 Xueyuan Road, Beijing, China
{huangbo,zhaozhenyu}@zhihu.com

Abstract. In this paper, we give an overview for the shared task at the CCF Conference on Natural Language Processing & Chinese Computing (NLPCC 2018): Automatic Tagging of *Zhihu* Questions. The dataset is collected from the Chinese question-answering web site *Zhihu*, which consists 25551 tags and 721608 training samples in this shared task. This is a multi-label text classification task, and each question can have as much as five relevant tags. The dataset can be assessed at http://tcci.ccf.org.cn/conference/2018/taskdata.php.

Keywords: Automatic tagging · Multi-label classification
Text classification

1 Introduction

The task aims to tag questions in *Zhihu* with relevant tags from a collection of predefined ones. This is a multi-label classification problem, several tags can be relevant to a given question. With the rise of social media, the text data on the web is growing exponentially. Furthermore, the label space is relatively huge compared to traditional text classification tasks. Make it is impractical for a human being to accurately assign tags to all those data. Machine learning methods are quite suitable for this task, and accurate tags can benefit several downstream applications such as recommendation and search.

Formally, the task is defined as follows: given a question with its title $x_t = (x_{t_1}, x_{t_2}, \cdots, x_{t_n})$ and description $x_d = (x_{d_1}, x_{d_2}, \cdots, x_{d_m})$, where x_{t_j} denotes the jth word in the title. The objective is to find its possible relevant tags in the predefined tag set. More specifically, given a specific tag tag_i, we need to find a function to predict whether tag_i is relevant to the current question with title x_t and description x_d.

$$p(tag_i|x_t, x_d) = f(x_t, x_d, tag_i, \theta) \tag{1}$$

where θ is the parameter of the function.

M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11109, pp. 434–438, 2018.
https://doi.org/10.1007/978-3-319-99501-4_40

2 Data

In this task, we provide training, development, and test data. Each question in the dataset contains a title, an unique id and an additional description. The labels are tagged collaboratively by users from the community question answering web site *Zhihu*. To improve the quality of the data, we removed infrequency tags, and relabeled manually to build development and test dataset.

There are 25551 tags and 721608 training samples in training data, 8947 samples in development data and 20597 samples in test data. Some samples from training dataset are shown in Table 1.

The dataset is different from widely used text classification datasets. Firstly, the label space is relatively huge and there is a data imbalance problem. Table 2 shows the statistics of the numbers of training samples for each label, we can see that almost 30% labels only have 5 to 10 training samples, while there still are some labels may have more than 5000 training samples. Secondly, the task is a multi-label problem and the number of labeled tags is not fixed for each question with a range from 1 to 5, Table 3 shows the statistics of the numbers of labeled tags for each question. Thirdly, since the dataset is collected from *Zhihu* whose contents are all generated by users, the text styles vary from user to user, Fig. 1 shows the length distributions of titles and descriptions respectively.

Table 1. Training samples from the dataset.

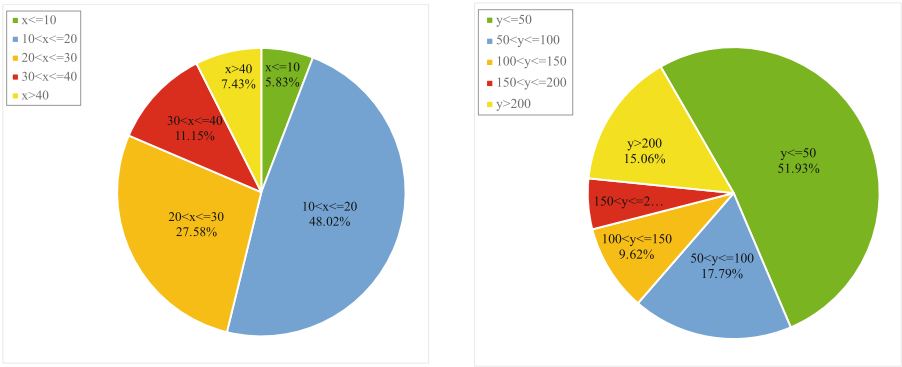
question_title	question_description	tags
如何做到不让别人影响自己的心情? How to keep others from affecting your mood?		情绪, 情绪管理 mood, mood control
怎样写商业计划书? How does one create a business plan?	我有些想法，怎么获得投资 I have some ideas, how can I get investment?	商业计划书，风投 business plan, VC

Table 2. Statistics of the numbers of training samples for each label.

Number of training samples	5 to 10	10 to 50	50 to 500	500 to 1000	1000 to 5000	5000+
Count of labels	7651	11988	5158	422	296	36
Percentage of labels(%)	29.94	46.92	20.19	1.65	1.16	0.14

Table 3. Statistics of the numbers of labeled tags for each sample in training data.

Number of labeled tags	1	2	3	4	5
Count of samples	134190	123397	151553	143388	169080
Percentage of samples(%)	18.60	17.10	21.00	19.87	23.43



(a) Length distribution of titles.

(b) Length distribution of descriptions.

Fig. 1. Length distributions of training data.

3 Evaluation

For each question in the test set, the model is required to predict as much as five relevant tags, and the tags are sorted by their predicted probabilities. Specifically, the number of predicted tags for a given question can be less than 5 or even be 0 if the model can't find enough relevant tags to the question.

The results are evaluated on the F_1 measure. We compute the positional weighted precision. Let $correct\_num_{p_i}$ denotes the correct count of predicted tags at position i, and $predict\_num_{p_i}$ denotes the count of predicted tags at position i.

The precision, recall and F_1 measure are computed as following formulas:

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{2}$$

$$P = \frac{\sum_{i=1}^{5} correct\_num_{p_i} / \log(i+2)}{\sum_{i=1}^{5} predict\_num_{p_i} / \log(i+2)} \tag{3}$$

$$R = \frac{\sum_{i=1}^{5} correct\_num_{p_i}}{ground\_truth\_num} \tag{4}$$

4 Baseline Implementations

Text classification is an important task in Natural Language processing with many applications, including search query classification, sentiment analysis, news categorization, which have been studied for years. In recent years, Deep

Learning on text classification has gained much attention due to its prominent achievement.

We have implemented several deep learning baseline models on text classification, which are effective and widely used in recent years, including LSTM [1], FastText [2] and CNN [3]. The results of the baselines are listed in Table 4.

Table 4. Results of baselines.

Method	Weighted precision	Recall	F1
LSTM	33.47	46.15	38.80
CNN	34.62	46.87	39.82
FastText	32.79	48.52	39.13

5 Participants Submitted Results

There are total 15 participants actively participate and submit their predictions on the test set. The number of submissions is limit to 5 in total, and we report the best result for each participant. The predictions are evaluated and the results are listed in Table 5.

Table 5. Participants Submitted Results.

Participant	Weighted precision	Recall	F1	Number of submissions
P1	0.5251	0.7783	0.6271	1
P2	0.5384	0.6380	0.5840	3
P3	0.5267	0.5123	0.5194	5
P4	0.5048	0.4692	0.4863	4
P5	0.5031	0.4664	0.4841	1
P6	0.3859	0.5502	0.4536	2
P7	0.3423	0.4759	0.3982	2
P8	0.3743	0.3770	0.3756	4
P9	0.2882	0.4157	0.3404	5
P10	0.2880	0.4100	0.3383	5
P11	0.2894	0.3427	0.3138	1
P12	0.2996	0.3221	0.3104	5
P13	0.2431	0.3477	0.2861	5
P14	0.4333	0.1546	0.2279	2
P15	0.1614	0.1242	0.1404	1

6 Conclusion

Text classification has been studied for years, several text classification datasets have been studied extensively in recent years. In this task we collected a new text classification dataset, which addresses two problems: (1) the label space is relatively huge, (2) the training samples are very imbalance. We contributed the dataset to the research community for further study.

References

1. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
2. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: FastText.zip: compressing text classification models. arXiv preprint arXiv:1612.03651 (2016)
3. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)