Overview of the NLPCC 2017 Shared Task: Chinese News Headline Categorization

Xipeng Qiu(✉), Jingjing Gong, and Xuanjing Huang

School of Computer Science, Fudan University,
825 Zhangheng Road, Shanghai, China
{xpqiu,jjgong15,xjhuang}@fudan.edu.cn

Abstract. In this paper, we give an overview for the shared task at the CCF Conference on Natural Language Processing & Chinese Computing (NLPCC 2017): Chinese News Headline Categorization. The dataset of this shared task consists 18 classes, 12,000 short texts along with corresponded labels for each class. The dataset and example code can be accessed at https://github.com/FudanNLP/nlpcc2017_news_headline_categorization.

1 Task Definition

This task aims to evaluate the automatic classification techniques for very short texts, i.e., Chinese news headlines. Each news headline (i.e., news title) is required to be classified into one or more predefined categories. With the rise of Internet and social media, the text data on the web is growing exponentially. Make a human being to analysis all those data is impractical, while machine learning techniques suits perfectly for this kind of tasks. After all, human brain capacity is too limited and precious for tedious and non-obvious phenomenons.

Formally, the task is defined as follows: given a news headline $x = (x_1, x_2, ..., x_n)$, where x_j represents jth word in x, the object is to find its possible category or label $c \in \mathcal{C}$. More specifically, we need to find a function to predict in which category does x belong to.

$$c^* = \underset{c \in \mathcal{C}}{\operatorname{argmax}} f(x; \theta_c), \tag{1}$$

where θ is the parameter for the function.

2 Data

We collected news headlines (titles) from several Chinese news websites, such as toutiao, sina, and so on.

There are 18 categories in total. The detailed information of each category is shown in Table 1. All the sentences are segmented by using the python Chinese segmentation tool *jieba*.

Some samples from training dataset are shown in Table 2.

X. Huang et al. (Eds.): NLPCC 2017, LNAI 10619, pp. 948–953, 2018.
https://doi.org/10.1007/978-3-319-73618-1_85

Table 1. The information of categories.

Category	Train	Dev	Test
Entertainment	10000	2000	2000
Sports	10000	2000	2000
Car	10000	2000	2000
Society	10000	2000	2000
Tech	10000	2000	2000
World	10000	2000	2000
Finance	10000	2000	2000
Game	10000	2000	2000
Travel	10000	2000	2000
Military	10000	2000	2000
History	10000	2000	2000
Baby	10000	2000	2000
Fashion	10000	2000	2000
Food	10000	2000	2000
Discovery	4000	2000	2000
Story	4000	2000	2000
Regimen	4000	2000	2000
Essay	4000	2000	2000

Table 2. Samples from dataset. The first column is Category and the second column is news headline.

Category	**Title Sentence**
world	首辩在即希拉里特朗普如何备战
society	山东实现城乡环卫一体化全覆盖
finance	除了稀土股，还有哪个方向好戏即将..
travel	独库公路再次爆发第三次泥石流无法...
finance	主力资金净流入9000 万以上28 股...
sports	高洪波：足协眼中的应急郎中
entertainment	世界级十大喜剧之王排行榜

Table 3. Statistical information of the dataset.

Category	Size	Avg. Chars	Avg. Words
Train	156000	22.06	13.08
Dev.	36000	22.05	13.09
Test	36000	22.05	13.08

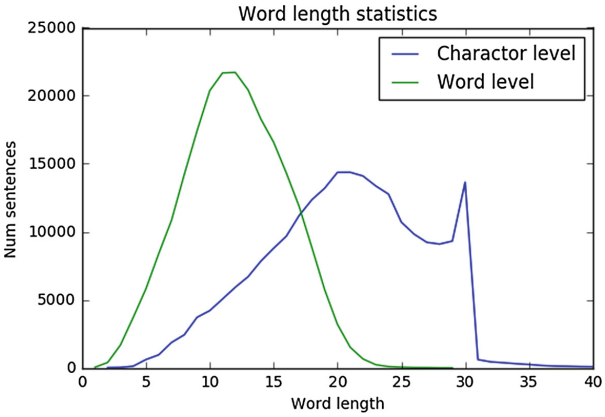

Fig. 1. The blue line is *character length* statistic, and blue line is *word length*. (Color figure online)

Length. Figure 1 shows that most of title sentence character number is less than 40, with a mean of 21.05. Title sentence word length is even shorter, most of which is less than 20 with a mean of 12.07 (Table 3).

The dataset is released on github https://github.com/FudanNLP/nlpcc2017_news_headline_categorization along with code that implement three basic models.

3 Evaluation

We use the macro-averaged precision, recall and F1 to evaluate the performance.

The Macro Avg. is defined as follow:

$$Macro\_avg = \frac{1}{m}\sum_{i=1}^{m}\rho_i$$

And Micro Avg. is defined as:

$$Micro\_avg = \frac{1}{N}\sum_{i=1}^{m}w_i\rho_i$$

where m denotes the number of class, in the case of this dataset is 18. ρ_i is the accuracy of ith category, w_i represents how many test examples reside in ith category, N is total number of examples in the test set.

4 Baseline Implementations

As a branch of machine learning, Deep Learning (DL) has gained much attention in recent years due to its prominent achievement in several domains such as Computer vision and Natural Language processing.

Table 4. Results of the baseline models.

Model	Macro P	Macro R	Macro F	Accuracy
LSTM	0.760	0.747	0.7497	0.747
CNN	0.769	0.763	0.764	0.763
NBoW	0.791	0.783	0.784	0.783

We have implemented some basic DL models such as neural bag-of-words (NBoW), convolutional neural networks (CNN) [3] and Long short-term memory network (LSTM) [2].

Empirically, 2 Gigabytes of GPU Memory should be sufficient for most models, set batch to a smaller number if not.

The results generated from baseline models are shown in Table 4.

5 Participants Submitted Results

There are 32 participants actively participate and submit they predictions on the test set. The predictions are evaluated and the results are shown in Table 5.

6 Some Representative Methods

In this section, we gives three representative methods.

[4] proposed a novel method which enhances the semantic representation of headlines. It first adds some keywords extracted from the most similar news to expand the word features. Then, it uses the corpus in news domain to pre-train the word embedding so as to enhance the word representation. At last, it utilizes Fasttext classifier, which uses a liner method to classify texts with fast speed and high accuracy.

[1] developed a voting system based on convolutional neural networks (CNN), gated recurrent units (GRU), and support vector machine (SVM).

[5] proposed an efficient approach for Chinese news headline classification based on multi-representation mixed model with attention and ensemble learning. It first models the headline semantic both on character and word level via Bi-directional Long Short-Term Memory (BiLSTM), with the concatenation of output states from hidden layer as the semantic representation. Meanwhile, it adopts attention mechanism to highlight the key characters or words related to the classification decision. And lastly it utilizes ensemble learning to determine the final category of the whole test samples by sub-models voting.

Table 5. Results submitted by participants.

Participant	Macro P	Macro R	Macro F	Accu.
P1	0.831	0.829	0.830	0.829
P2	0.828	0.825	0.826	0.825
P3	0.818	0.814	0.816	0.814
P4	0.816	0.809	0.813	0.809
P5	0.812	0.809	0.810	0.809
P6	0.811	0.807	0.809	0.807
P7	0.809	0.804	0.806	0.804
P8	0.806	0.802	0.804	0.802
P9	0.803	0.800	0.802	0.800
P10	0.805	0.800	0.802	0.800
P11	0.799	0.798	0.798	0.798
P12	0.797	0.795	0.796	0.795
P13	0.793	0.789	0.791	0.789
P14	0.791	0.789	0.790	0.789
P15	0.792	0.787	0.789	0.786
P16	0.786	0.783	0.785	0.783
P17	0.778	0.775	0.777	0.775
P18	0.785	0.775	0.780	0.775
P19	0.785	0.775	0.780	0.775
P20	0.766	0.765	0.765	0.765
P21	0.768	0.759	0.764	0.759
P22	0.768	0.748	0.758	0.748
P23	0.744	0.729	0.736	0.729
P24	0.729	0.726	0.728	0.726
P25	0.745	0.700	0.722	0.700
P26	0.734	0.688	0.710	0.688
P27	0.698	0.685	0.691	0.685
P28	0.640	0.633	0.637	0.633
P29	0.645	0.629	0.637	0.629
P30	0.437	0.430	0.433	0.430
P31	0.474	0.399	0.433	0.399
P32	0.053	0.056	0.054	0.056

7 Conclusion

Since large amount of data is required for machine learning techniques like deep learning, we have collected considerable amount of news headline data and

contributed to the research community. We also found that the performance of news headline classification still need be improved. We hope that our dataset provides a valuable training data and a testbed for text classification task.

References

1. Zhu, F., Dong, X., Song, R., Hong, Y., Zhu, Q.: A multiple learning model based voting system for news headline classification. In: Proceedings of the CCF Conference on Natural Language Processing & Chinese Computing (2017)
2. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
3. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
4. Yin, Z., Tang, J., Chengsen, Luo, W., Luo, Z., Ma, X.: A semantic representation enhancement method for Chinese news headline classification. In: Proceedings of the CCF Conference on Natural Language Processing & Chinese Computing (2017)
5. Lu, Z., Liu, W., Zhou, Y., Hu, X., Wang, B.: An effective approach for Chinese news headline classification base on multi-representation mixed model with attention and ensemble learning. In: Proceedings of the CCF Conference on Natural Language Processing & Chinese Computing (2017)