Corpus for Chinese 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.

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

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.

Length The statistical information is also given in Fig. 3.

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.

The dataset is released on github along with a Tensorflow [Abadi *et al.*, 2015] implemented demonstration code.

3 Evaluation

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

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 1: The information of categories.

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.

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

Category	Size	Avg. Chars	Avg. Words
train	156000	22.06	13.08
dev.	36000	22.05	13.09
test	36000	22.05	13.08

Table 3: Statistical information of the dataset.

We have implemented some basic DL models such as neural bag-of-words (NBoW), convolutional neural networks (CNN) [Kim, 2014] and Long short-term memory network (LSTM) [Hochreiter and Schmidhuber, 1997].

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

The results are shown in Table 4.

5 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.

References

[Abadi *et al.*, 2015] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

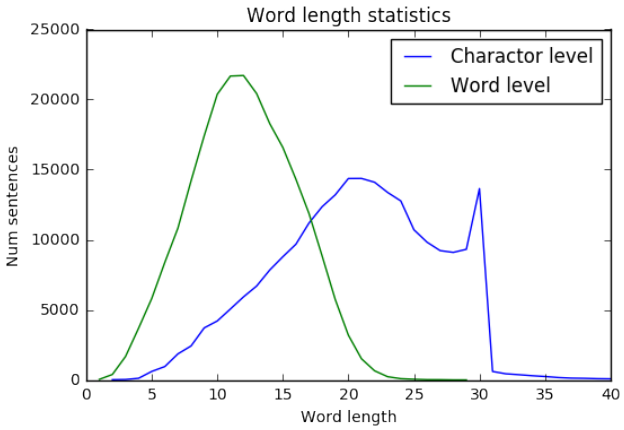

Figure 1: The blue line is *character length* statistic, and blue line is *word length*.

Model	Micro P	Micro R	Micro 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

Table 4: Results of the baseline models.