A Million Chinese News Title Classification

Jingjing Gong, Xipeng Qiu and Xuanjing Huang
Computer Science Department, Fudan University

1 Call for Participation

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.

Deep Learning as a branch of machine learning have gain much attention in recent years due to its prominent achievement in several domains of machine learning such as Computer vision and Natural Language processing. Techniques like Deep Learning requires a lot of training data. There are unlimited amount of unsupervised data on the web, while labeled or supervised data is relatively insufficient compared to the consumption ability of Deep Learning Models. There by we present a large amount of Chinese News Title dataset to alleviate the fact that the amount of labeled Chinese corpus attainable by researchers is relatively small.

In this challenge, participants will be given a fraction of the dataset, such that for each item in the dataset, there consists a news title x^i and a label c^i to that title. Put it simply, what participants need to do is to create a function $y = f(x)$ such that it will map title to the right label correspond to title text.

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

表 1: This table shows the dataset sample, the first column is Category and the second column is title sentence, in the dataset file this two is separated by a tab character

2 Task Definition

This task is defined as follows: Given a Chinese title $x^i = (x^i_1, x^i_2, ..., x^i_n)$, where x^i_j represents jth word in ith sample, there is a label c^i correspond to the title x^i, The data pair sample is shown in Table 1. Participants need to find a model to predict in which category does x^i belong to. More specifically, the goal is to approximate a function $c = f(x; \theta)$ (in which θ is the parameter for the function) so that most of data pair including data that the model have never seen (x^i, c^i) will satisfy the equation.

The evaluation of the quality of the model will be that, for a set of unseen data pairs, how many data pairs will satisfy the equation, the higher the accuracy(the fraction of data that fit to the model) is, the better the model is.

Note that in this challenge, only 30% of the full

Category	**Number**
entertainment	128246
sports	116340
car	104505
society	91276
tech	86118
world	80924
finance	73791
game	63240
travel	35400
military	32350
history	26557
baby	26189
fashion	24776
food	21710
discovery	8990
story	8587
regimen	8564
essay	8100

表 2: This table shows the number of examples that belong to the corresponding category

dataset is given, in which the ratio of training data and development data is 2:1. while the rest data is preserved as test data. A week before deadline, test set *without label* will be released for evaluation. Full dataset will be released after deadline of NLPCC-2017 conference.

3 Dataset Description

The full data consists about 1 Million title-category data pairs all of them are collected from news websites such as toutiao etc., each title belongs to 1 of 18 categories, the category name and numbers of examples to the corresponding category is shown in Table 2. The length statistics is also given in Fig. 1. The blue line is *example character length* statistic, and blue line is *example word length* given that the sentences are segmented using the python Chinese segmentation tool ***jieba***. This Figure have shown 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. Empirically, 2 Gigabytes of GPU Memory should be sufficient for most models, set batch to a smaller number if not.

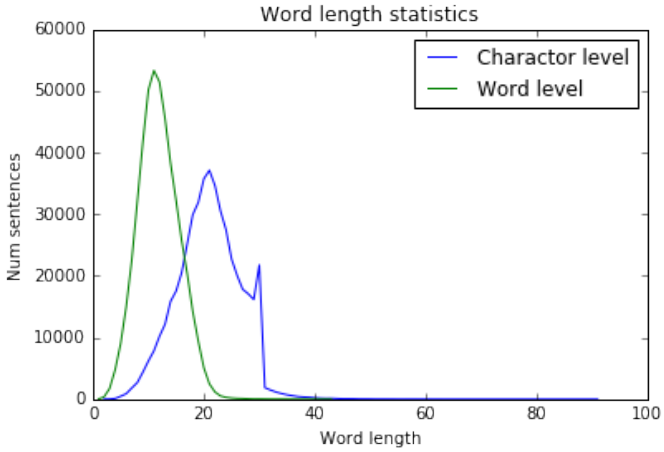

图 1: The blue line is *example character length* statistic, and blue line is *example word length* given that the sentences are segmented using the python Chinese segmentation tool ***jieba***.

Model	**Macro Avg. Acc.**	**Micro Avg. Acc.**
NBoW	0.705	0.745
CNN	0.70	0.760
LSTM	0.763	0.791

表 3: This Table shows basic experimental results. Macro Avg. is the non-weighted average (average accuracy across category), and Micro Avg is weighted average (average accuracy across samples)

4 Basic Experimental Results

We have run some basic 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], and Results are shown in Table 3. 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.

5 Conclusion

Since large amount of data is required for Machine Learning techniques like Deep Learning, we have collected considerable amount of News Title data and contributed to the research community. [1]

参考文献

[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.

[Kim, 2014] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[1] 30% of the dataset is released on github along with a Tensorflow [Abadi *et al.*, 2015] implemented demonstration code. All data will be released after the NLPCC-2017 conference is over.