Text Classification on Large Scale Chinese News Corpus using Character-level Convolutional Neural Network

Xingsu Wu, Jinhui Chen
School of Mathematics and Computer Science, Shangrao Normal University, Shangrao, 34001, China
314108@sru.edu.cn

Abstract. In view of the current problems in the field of Chinese news, such as chaotic news data and lack of automatic and efficient management, most of the existing Chinese text classification methods take word characteristics as the basic unit of text representation, but ignore the beneficial performance of character characteristics. In this paper, an automatic classification method of Chinese news based on character-level convolutional neural networks (char-CNN) is proposed. In order to compare with traditional models and deep learning models, we constructed a large scale Chinese news corpus to show that character-level convolutional neural networks could achieve the most advanced or competitive results.

1. Introduction
Text classification is a basic work in the field of natural language processing (NLP), which is widely used in information retrieval [1], deceptive examination and identification [2][3], emotion analysis [4] and spam E-mail detection [5]. The technical basis of news classification is text classification in natural language processing (NLP). Along with rapidly development of Internet, the information on the Internet has increased explosively. So it is becoming more and more difficult for people to obtain useful information and it’s an urgent need for the classification of the positive and negative of mass news. Fortunately, text classification can solve information disorder for the news data, which can help people to locate information easily in the greater degree. At present, the traditional text classification method and its defects are as follows.

(1) The support vector machine (SVM) [6][7] method is complicated and can not be used to solve nonlinear problems.

(2) Decision tree classification [8] is not suitable for training large sample sets and is prone to sensitive to noises or overfitting in the training set.

(3) Naive bayes classification [9][10] does not match the actual application conditions and assumptions and it also does not satisfy each property independent of each other, so the classification accuracy of naive bayes classification is easily affected.

With the development of neural networks, there are more and more text classification tasks using the idea of deep neural networks. Text classification commonly used neural network comprises a convolutional neural network (CNN), Recursive Neural Networks (RecursiveNNs) Recurrent Neural Networks (RNNs) and more complex network architectures. Compared with the traditional methods, the character-level convolutional neural network is still able to learn without any knowledge on the syntax or semantic structures of a language for text classification.

Based on the above considerations, the convolutional neural network model is given in this paper, there are three contributions in our work:
In this paper, the character-level convolutional neural network is applied to text classification in Chinese corpus, which is seldom studied in the field of NLP. The result shows that Chinese character dataset has generated a better result than its corresponding word-level.

Compared with the traditional models and deep learning models, the character-level convolutional neural network model can be classified more accurate classification results.

Other than existing models, the character-level convolutional neural network has significantly fewer parameters and is more applicable to large scale text classification tasks.

2. Related work

Up to now, most of the previous techniques of text classification are based on word counting, in which simple statistics for some of the ordered word combinations always perform the best [11], which misses deep semantic analysis of texts, so the generated abstract is unsatisfying.

In recent years, research on document classification based on deep learning [12] has made great progress. In general, deep learning models perform well in learning the high level vector representation of words, sentences and documents directly from raw text data and have a great impact on classification results.

CNN were first developed in computer vision, such as image classification, target detection and image segmentation [13-15] etc. With the continuous development of deep learning technology, more and more researchers apply deep learning to the field of natural language processing. Kim [16] uses convolutional neural networks to classify texts and achieve better classification results. However, due to the differences in expression and word-formation between Chinese and English, several words or characters may be needed to determine the category of a Chinese text. Therefore, it is necessary to extract Chinese text by word segmentation. In this paper, CNN is used to classify large scale Chinese news data sets, and a feature extraction model is established to extract advanced features from the network to represent the semantic information of the text.

For Chinese text, most current models first segment the word and then feed the feature to the model for classification [17][18]. Unfortunately, errors due to word segmentation will be propagated to the model. Some researchers have also proposed a character level model. Most of these models use word embedding for Chinese text categorization [19-21]. However, most of the existing Chinese text classification models usually regard text features as the basic unit of text representation, and ignore the beneficial representation of character features.

Compared with word-level and sentence-level convolutional neural network, the character-level convolutional neural network (Char-CNN) has a better applicability for typographical error input and misspellings. Moreover, character-level convolutional neural network can still learn without knowing the syntax or semantic structure of the text classification language. In addition, the character-level can be easily extended to all languages. Due to these, recent researches for text classification mainly focus on character-level convolutional neural network. However, while most of researches employ English corpus for the Char-CNN, the minority of these studies have been done using Chinese corpus, especially large scale Chinese text.

However, to the best of our knowledge, most of these studies use English corpus for character-level text classification, while few studies use Chinese corpus. It is necessary to build a Chinese text dataset to explore Char-CNN for Chinese corpus text classification. We have constructed a large scale Chinese dataset, Compared with the traditional models and deep learning models, the experimental results show that he accuracy of the proposed model is improved.

3. Character-level Convolutional Neural Network

3.1. Model Design

We constructed a deep network CNN for the classification of Chinese News Corpus. Our network architecture is shown in Figure 1. There are four components in our model, namely data preprocessed, embedding layer, CNN layers, fully-connected layers and pooling layers.
Figure 1 The architecture of proposed model

To get the representation of a sentence, it mainly goes through the following steps:

1. Embedding layer: Converting elements in a word sequence into word vectors;
2. CNN unit: Applying a set of convolutional filters;
3. Pooling layer: Abstracting the output characteristics of convolutional layer further;
4. Fully connected layer: Concatenating the results of convolution;
5. Output layer: Using the fully connected softmax layer to complete the classification task.

3.2. Character quantization

There are more than 6,000 commonly used Chinese characters, which means that the size of the dictionary is quite large, but Zhang et al.[22] developed character-level coding scheme is limited to ASCII characters, such as 0-9, a-z and some additional punctuation and ignore other characters. Therefore, a new encoding scheme is proposed to convert any UTF-8 encoded string into an 8-bit sequence, in which each byte is quantized with a one-shot encoding as the input of the model. Our model accepts a sequence of coded characters as input. Encoding is done by specifying Chinese characters of size m for the input language, and then using the one-hot encoding to quantify each character, as shown in Figure 2.

Figure 2 UTF-8 byte encoding for Chinese characters

Before training, the word-level text classifiers are usually preprocessed into different languages using language detection programs. Using the full UTF-8 coded character set, including all possible text characters, and enabling our model to learn patterns based on different languages automatically, including currency symbols, emoticons (smiley faces, and so on), and other non-language-based character sequences.

4. Experimental Evaluation

4.1. Data and Pre-processing

In the analysis below, we use the dataset that our research team collected from a subset of THUCNews. We evaluated the proposed model on a subset of the THUCNews dataset[23], which is based on the filtering of historical data from 2005 to 2011 and the filtering of Sina News RSS subscription channels, including 740000 news documents divided into 14 categories with a total capacity of 2.19 GB. We construct the subset data from the original dataset, namely large scale
Chinese news corpus. The dataset consists of 10 categories, which are divided into training set, validation set and test set, as shown in Table 1.

### Table 1: Training, validation and testing sets

| Data set type | Subset (piece) |
|---------------|----------------|
| Training sets | 50000          |
| Validation sets | 10000        |
| Testing sets  | 5000           |

#### 4.2. Comparison Models

In order to make a fair comparison of the competition model, we conducted a series of deep learning experiments such as RNN and char-RNN, including traditional methods such as LibSVM, CBOW, Skip-Gram, CWE and so on. We do our best to choose models that provide comparable and competitive results, and report the results faithfully without any model selection.

#### 4.3. Results

In general, we present the results of each benchmark data flow independently. Evaluation on precision (P), recall (R), and F1 score (F1) were evaluated for each dataset. We use the macro average, which is the arithmetic mean of the test results of all categories as the evaluation criteria, as shown in Table 2.

### Table 2: Macro average of each model on the THUCNEWS text classification dataset

| Model          | LibSVM | CBOW | Skip-Gram | CWE | CWE+P | CWE+L | CWE+LP | CWE+N | SCWE | SCWE+M | SJL.CWE | SJL.CWE+M | RNN | Char-RNN | Char-CNN |
|----------------|--------|------|-----------|-----|-------|-------|--------|-------|------|--------|---------|----------|-----|-----------|---------|
| P              | 0.8743 | 0.8764 | 0.8797   | 0.8803 | 0.8812 | 0.8842 | 0.8904 | 0.8909 | 0.8923 | 0.8960 | 0.8923 | 0.8976 | 0.9065 | 0.9438 | 0.9693 |
| R              | 0.8606 | 0.8508 | 0.8532   | 0.8536 | 0.8546 | 0.8572 | 0.8624 | 0.8624 | 0.8648 | 0.8699 | 0.8662 | 0.8706 | 0.9238 | 0.9443 | 0.9675 |
| F1             | 0.8625 | 0.8623 | 0.8652   | 0.8656 | 0.8666 | 0.8694 | 0.8749 | 0.8752 | 0.8772 | 0.8816 | 0.8779 | 0.8828 | 0.9147 | 0.9427 | 0.9678 |

Compared with LibSVM, CBOW, Skip-Gram, CWE, CWE+P, CWE+L, CWE+LP, CWE+N, SCWE, SCWE+M, SJL.CWE, SJL.CWE+M, RNN and Char-RNN, our char-CNN approach produces better performance on THUCNews dataset. Our approach performs better with 10.53%, 10.55%, 10.26%, 10.22%, 10.12%, 9.84%, 9.29%, 9.26%, 9.06%, 8.62%, 8.99%, 8.50%, 5.31%, and 2.51% F1-score increases than LibSVM, CBOW, Skip-Gram, CWE, CWE+P, CWE+L, CWE+LP, CWE+N, SCWE, SCWE+M, SJL.CWE, SJL.CWE+M, RNN and Char-RNN respectively.

Figure 3 and figure 4 show the confusion matrix of our char-CNN method and char-RNN method in the THUCNews dataset respectively. In the confusion matrix, the quality of the classifier can be measured in the multi-class accuracy normalized to the main diagonal of the confusion matrix. It can be concluded from Figure 3 that our char-CNN method has achieved excellent performance on Chinese corpus text classification.
For the Char-CNN method, the results show that the accuracy of all classes is higher than 90%. But for char-RNN method, the results show that the accuracy of home class is less than 80%. As the highest accuracy, Real estate news has little relationship with other categories and more unique characteristics, so the classifier achieves the best result in it. And the lowest one home class has many of the same characteristics as fashion, finance and education, so it has many items that can be classified to other classes.

5. Conclusion
In this paper, an automatic classification method of Chinese news based on character-level convolutional neural network is proposed. Experiments demonstrate that our char-CNN approach achieves achieved a classification macro average precision of 96.93% over the Chinese benchmark datasets THUCNews. Compared with the current Chinese text classification method, it has comparability and even better performance.

However, our proposed model only focuses on the global features, without taking into account the local features. To solve this problem, the attention mechanism will be introduced to extract keywords from the text in the future.

To improve the accuracy of the model, we will train the model with more data, so that the model can contain more words. Dealing with words you don't know will be a big challenge.

The parallel algorithm can well meet the needs of large scale text classification, so the text classification model will be implemented in parallel on multi-GPU, and the time cost will be reduced under the premise of ensuring accuracy.

Acknowledgments
This work was supported by the Jiangxi Provincial Department of Education Science and Technology Project Fund: GJJ180870. Thanks for the news corpus provided by NLP Laboratory of Tsinghua University.

References
[1] I. W. B. Croft, D. Metzler, T. Strohman, Search engines: Information retrieval in practice, vol. 283, Addison-Wesley Reading, 2010.
[2] Zhang W, Tang X, Yoshida T. TESC: An approach to Text classification using Semi-supervised Clustering[J]. Knowledge Based Systems, 2015, 75(feb.):152-160.
[3] Zhang W, Du Y, Yoshida T, et al. DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network[J]. Information Processing & Management, 2018, 54(4): 576-592.
[4] Chen H, Sun M, Tu C, et al. Neural Sentiment Classification with User and Product Attention[C]// Conference on Empirical Methods in Natural Language Processing. 2016:1650-1659.
[5] Carreras X, Marquez L. Boosting Trees for Anti-Spam Email Filtering[J]. 2001:58-64.
[6] Wang F, Wang Z, Li Z, et al. Concept-based Short Text Classification and Ranking[C]// the 23rd ACM International Conference. ACM, 2014:1069-1078.
[7] Haddoud M, Mokhtari A, Lecroq T, et al. Combining supervised term-weighting metrics for SVM text classification with extended term representation[J]. Knowledge & Information Systems, 2016, 49(3):1-23.
[8] Islam M Z, Adnan M N, Publications S. ComboSplit: Combining Various Splitting Criteria for Building a Single Decision Tree[C]// The International Conference on Artificial Intelligence and Pattern Recognition (AIPR2014). 2014:1-8.
[9] Baselines and Bigrams: Simple, Good Sentiment and Topic Classification Using bag of features + linear classifiers, in: ACL, 2012: 90-94.
[10] Pratama B Y, Sarno R. Personality classification based on Twitter text using Naive Bayes, KNN and SVM[C]// International Conference on Data & Software Engineering. IEEE, 2016: 170-174.
[11] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning, Springer-Verlag, 1998: 137-142.

[12] Lecun Y, Bengio Y, Hinton G. Deep learning. [J]. Nature, 2015, 521(7553):436-444.

[13] Krizhevsky A; Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems. Neural Information Processing Systems Foundation, Inc.: South Lake Tahoe, NV, USA, 2012:1097-1105.

[14] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// CVPR. IEEE, 2014:580-587.

[15] Girshick R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13-16 December 2015:1440-1448.

[16] Kim Y. Convolutional Neural Networks for Sentence Classification[C]//Proc of Empirical Methods in Natural Language Processing. 2014:1746-1751.

[17] Siwei L, Kang L, Shizhu H, et al. How to Generate a Good Word Embedding?[J]. IEEE Intelligent Systems, 2016, PP(6):1-1.

[18] Lu S H, Chiang D A, Keh H C, et al. Chinese text classification by the Naive Bayes Classifier and the associative classifier with multiple confidence threshold values[J]. Knowledge-Based Systems, 2010, 23(6):598-604.

[19] Huang W, Wang J. Character-level Convolutional Network for Text Classification Applied to Chinese Corpus[J]. 2016.

[20] Li, Yue, Wang X, et al. Chinese Text Classification Model Based on Deep Learning[J]. Future Internet, 2018,10(11):113.

[21] Liu S, Bremer P T, Thiagarajan J J, et al. Visual Exploration of Semantic Relationships in Neural Word Embeddings[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, PP(99):1-1.

[22] Zhang X, Zhao J, Lecun Y. Character-level Convolutional Networks for Text Classification[C]// International Conference on Neural Information Processing Systems. 2015:649-657.

[23] Maosong Sun, Jingyang Li, Zhipeng Guo, Yu Zhao, Yabin Zheng, Xiance Si, Zhiyuan Liu. THUCTC: An Efficient Chinese Text Classifier. 2016.