Headline Expansion for Chinese News Headline Categorization

Maoquan Wang, Weijie An *, and Qin Chen

East China Normal University
{51151201051}@stu.ecnu.edu.cn, {wjan,qchen}@ica.stc.sh.cn

Abstract. The Chinese news headlines are usually short and confusing for eye-catching, which results in lacking enough information to distinguish their category. We propose *headlines expansion* to enhance discriminative information in headlines. We first search the headlines on the Internet to obtain the expanded headlines, and then build an ensemble model combining traditional NLP-based models and deep learning-based neural networks to predict the headline categorization. Experiments show that headline expansion can efficiently improve the performance, and our system obtains 0.887 (Macro-Acc) in development set and 0.816 in test set from NLPCC shared task2.

Keywords: Headlines Categorization, Deep Learning, Expansion

1 Introduction

Text classification is a foundational task in many NLP applications [1, 2]. With the explosion of the social media, natural text (especially news) on the web is growing exponentially. It is urgent to categorize news headlines for different readers to choose their favourite news. However, it is impractical and extremely tedious to manually label a large number of headlines. NLPCC shared task provides *Task 2: News Headline Categorization* [3] to assesses the ability of participant systems to automatically classifier each headlines into one predefined categories.

Various machine learning techniques especially deep learning methods has been widely used in text classification task. Kim [4] introduced a convolution neural network for sentence classification, Lai [5] proposed Recurrent Convolutional Neural Networks(RCNN) to enhance contextual information of words, and Zhang [6] explored character-level convolutional networks for text classification. These methods mainly focus on the labeled data, it is difficult to exploit the existing limited labeled data to achieve higher accuracy when face the unexplored instances. Chinese News headlines is very short and do not provide enough contextual information, the short-text classifier easily be blocked by sparsity and ambiguity problems [7].

* Equal Contribution.

We proposed a headline expansion based classification framework for such the special task. An effective way of news headlines classification is to use some extra knowledge to enrich news headlines [8]. Firstly, we use search engines to expand the headlines and get the expanded headlines which is directly combined with the origin headlines as the final classification instance. After that, we utilize both the traditional classification models and deep learning based method for the textural classification. Finally, we combine the result of traditional classification models and deep learning method to get the final prediction.

We conduct experiments on the released Chinese News Headlines data. The results show that our expansion based classification methods much better than that non-extended.

The rest of this paper is organized as below: Section 2 describe the detail of our model. Section 3 shows the experiments and Section 4 is the result analyses. Then we provided some case study in the Section 5. We conclude our work in the final Section 6.

2 Model Description

Fig. 1 shows the overall architecture of our system. Given the news headlines, we first expand the headlines with the titles searched from the Internet. After that, we build our system based on the expanded headlines, which consists of the following three modules:

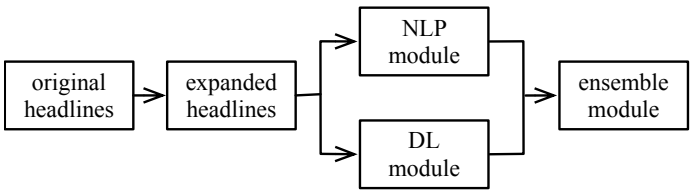

Fig. 1. The system architecture

- **Traditional NLP Module** is to extract two kinds of features. The lexical features are to select the category-related words, and the embedding features to capture the semantic information. All these NLP-based features are used to build classifier to make prediction.
- **Deep Learning Module** is to encode input sentences into fixed-length vector and adopt softmax function to predict the probability of each category.
- **Ensemble Module** is to average the probability of above two modules to get the final predicted category.

Next, we will describe the system in detail.

2.1 Headlines Expansion

The headlines of news are quite short and confused, thus it is helpful to supplement related knowledge to reduce the ambiguity. We first search on the largest Chinese search engine *Baidu.com* to obtain the top 5 related titles as the extra knowledge. Then we directly concatenate the origin headline and the expanded headlines as the input to the following models.

2.2 Traditional NLP Module

In this section, we give the details of feature engineering and learning algorithms.

2.2.1 Lexical Features

- **Unigram Features** Unigram Features are designed to recognize distinguish words for the predicted categories. We represent each input news headline as one-hot representation where each dimension represent corresponding word term frequency. Since Chinese characters also make contributions to predict the truth categories, we use character level unigram features to enhance representations. Specifically, we represent the head lines on both word level and character level. The sentence is represented as a BOW(Bag of Word), where each word is weighted by its term frequency value.
- **Bigram Features** Unigram features only capture the keywords information, in order to contain the sequence information we adopt the bigram features. Sometimes the word groups(two adjacent words) provide more benefits than only one word when explaining the meaning. By this point, we extracted the bigram features of headlines.
- **Word Importance Features** We find that the same word contribution to different categories is not the same. Such as the word "宝宝" is more likely appear in the category of **baby**. Thus we define the Word Importance Features as below:

$$W_{w_i - c_j} = \frac{Count_{ij}}{\sum_{k=1}^{18} Count_{ik}} \tag{1}$$

  where $Count_{ij}$ is the frequency value that the word w_i appear in category c_j. We sum each word importance features W_{wi} as the headlines features W(in this task, it has 18 dimensions).

2.2.2 Word Embedding Features

The unigram and bigram indicates the lexical features in the sentences. In order to better understand the headlines, we need some semantic information. Inspired by the work of [9], we represent each word in headlines as a vector also called word embedding. It built a semantic relationship between words which we thought that contains semantic information. Specifically, we train word embeddings based on our expanded headlines. And we use the average(**AVG**), maxim(**MAX**), and minim(**MIN**) pooling at each word embedding to get the headlines representation.

2.2.3 Learning Algorithms

Due to the large scale of extracted features, we adopt Logistic Regression (LR) algorithm implemented by LIBLINEAR[1] and XGBoost (XGB) algorithm[2] for classification.

2.3 Deep Learning Module

In this module, we adopt Recurrent convolutional neural network (RCNN) [5] for headlines classification. RCNN is effective to capture the contextual information and the most important features for category classification.

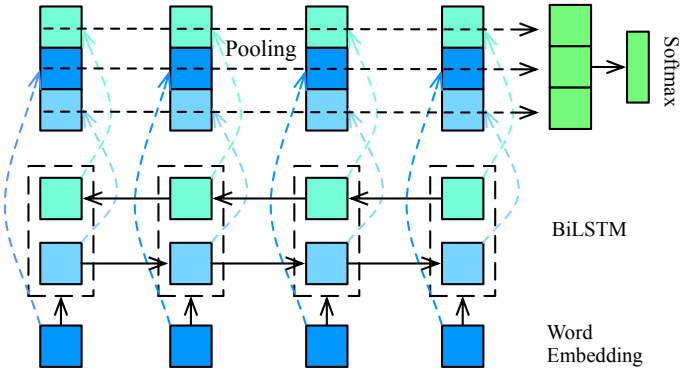

Fig. 2. RCNN

Fig. 2 shows the network structure. The input of the network is a expanded headline sentence S, which is a sequence of words $w_1, w_2, \cdots, w_n$. The output of the network contains class category. We use $p(k|D, \theta)$ to denote the probability of the headline being category k, where θ is the parameters in the network.

First, we use a bidirectional long short-term memory network to capture the contexts information. Then we combine the word embedding and its context as the representation of the word. We define $c_l(w_i)$ as the left context of word w_i and $c_r(w_i)$ as the right context of the word w_i, where $c_l(w_i)$ is the output of the forward LSTM and the backward LSTM respectively. The representation of word w_i is defined as the concatenation of the left-side context vector $c_l(w_i)$, the word embedding $e(w_i)$ and the right-side context vector $c_r(w_i)$:

$$x_i = [c_l(w_i); e(w_i); c_r(w_i)] \tag{2}$$

Second, we apply a max-pooling layer after all of the representations of words are calculated:

$$y = max_{i=1}^{n} x_i \tag{3}$$

[1] `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`
[2] `https://github.com/dmlc/xgboost`

where the max function is element-wise function. The pooling layer converts texts with various lengths into a fixed-length vector as the sentence representation. With the pooling layer, we can capture the information throughout the entire text.

Finally, we use an output layer to convert the sentence representation into probabilities of each category. Similar to traditional neural networks, it is defined as:

$$h = Wy + b \tag{4}$$

$$p = softmax(h) \tag{5}$$

where W is the weight matrix and the b is the bias. The softmax function is to transform the output into a vector in the range $(0, 1]$ that adds up to 1.

2.4 Ensemble Module

To ensemble NLP-based models and DL-based models, we adopt *average strategy,* which means taking the mean of each individual model predictions. We sum up the following three probabilities to give the final predication:

(1) LR algorithms with all NLP-based features;

(2) XGB algorithms, using all NLP-based features;

(3) the softmax probabilities in RCNN model;

After that, we select the maximum summed probability as the final predicted category.

3 Experimental Settings

Datasets: NLPCC 2017 Task 2 collected large-scale news headlines from several Chinese news websites, such as toutiao, sina. Each headline is annotated with a category, which indicates the topic of the news. In total, there are 18 categories, such as entertainment, sports, baby.

To expand the news headlines, we search on Baidu and collected the top 5 related titles. We concatenate the origin headlines and the related titles as the expanded headlines. we segment the headlines using the python Chinese segmentation tool *jieba*. **Table 1** shows the statistics of the original and the expanded datasets.

Table 1. The statistics of the original and expanded datasets.

Category	**Size**	**Original Headlines**		**Expanded Headlines**	
		Avg. Chars	**Avg. Words**	**Avg. Chars**	**Avg. Words**
train	156,000	22.06	13.08	140.79	78.10
dev	36,000	22.05	13.09	136.73	75.73
test	36,000	22.05	13.08	147.02	83.96

Evaluation: We use the macro-averaged precision, recall and F1 to evaulate the performance. The macro-averaged precision is defined as follow:

$$\text{Macro}_{\text{avg}} = \frac{1}{m}\sum_{i=1}^{m} p_i$$

where m denotes the number of class, in the case of this dataset is 18. p_i is the accuracy of ith category.

Pre-trained Word Embedding: We find 41.0% out-of-vocabulary words in the expanded headlines when using the given word embeddings. Thus we train word embeddings based on our expanded headlines, using fastText [10]. We set the dimension of trained embedding to 100. And the embedding is used in Word Embedding Features, as well as the input to the RCNN model.

Parameter Settings in RCNN: We adopt Adam [11] to optimize the cross entropy loss, and set the learning rate to 1e-3, hidden size to 100.

Baselines: The task organizers have implemented three DL baseline models, i.e., neural bag-of-words (NBoW), convolutional neural networks (CNN) [4] and long short-term memory network (LSTM) [12]. And we compare our models with these baseline models.

4 Results and Discussion

Table 2. Our system performance (Macro-Acc) on the original and expanded datasets.

Model		**Original**	**Expanded**
Baselines	NBoW	0.783	-
	CNN	0.763	-
	LSTM	0.747	-
Logistic Regression(LR)	Unigram-Char	0.7348	0.8140
	Unigram-Word	0.7772	0.8425
	Bigram-Char	0.7864	0.8578
	Bigram-Word	0.5805	0.8016
	Word Importance	0.7141	0.5876
	Word Embeddding (AVG)	0.7805	0.8453
	Word Embeddding (MIN/MAX/AVG)	0.7960	0.8535
	All Features	0.8212	0.8741
XGBoost(XGB)	All Features	0.8192	0.8663
Deep Learning(DL)	RCNN	0.8237	0.8751
Ensemble	LR + XGB + DL	0.8312	**0.8871**

Table 2 lists the results of several baselines and our systems. We find:

(1) Comparing different NLP features, Unigram-Word and Bigram-Char archives the better performance than the Unigram-Char and Bigram Features. We think that individual words alone can distinguish the category to classify. Also, we can grant the Bigram-Char as words in Chinese, since the Chinese words is usually constructed by two characters. What's more, the Word Embedding features achieve the better results than the Lexical Features since the word embedding can capture more semantic information. when combine all the features together, the performance increase to 0.8212, which indicates that all features make contributions.

(2) Regarding deep learning models, the RCNN model surface the best the results of NLP methods. We think the RCNN model can capture the important information from word and context information, and when training with the large annotated training data, it surely archive the better results. Also, when comparing the baseline DL models, the RCNN model outperform the NBoW, CNN and LSTM models. We think that RCNN model can capture the most important information by the pooling comparing with the baseline models.

(3) All ensemble methods significantly improved the performance. The ensemble of NLP methods and DL models (LR+XGB+DL) outperforms any single learning algorithm. It suggests that the traditional NLP methods and the deep learning models are complementary to each other and their combination achieves the best performance.

(4) Surprisingly, the headlines expansion is very effective, which increase about 5% comparing with the no-expanded sentences. It shows that our headlines expansion can bring in the extra knowledge which can supply the short origin headlines with more useful information.

5 Case Study

We study two cases in our experiments that show the effectiveness and somewhere to be promoted. As shown in **Table 3**, {“和珅”与“皇上”再聚首} is an entertainment news headline. But it is classified to the wrong category **history** according to the word 皇上 and 和珅. After we expand the headline, we obtain some related titles that contains the words 王刚, 张铁林 and 娱乐八卦 and so on. Thus it is correctly classified to the **entertainment** category.

Generally, it is easily expand some extra irrelevant information. As shown in **Table 4**, we expand the headline {如果穿整套的话，莫名就有一种呆萌喜感！}. We observe that the 4th expanded title {呆萌V!168 潮男&191 双塔刘国梁中网挑边太喜感_乒乓球_新浪竞技...} contains some irrelevant information. Thus it is classified to the wrong category **sports**. From the overall experimental results, we can conclude that the benefit of expansion far greater than the harms it brings.

6 Conclusion

In this paper, we proposed a headlines expansion novel framework for short texts modeling and classification. Extra knowledge which help model to understand the news headlines is obtained by baidu search engine. Experimental results on NLPCC task2 validated the effectiveness of the proposed method. Future works can be focused on how to effectively use the extra knowledge.

Table 3. Case Study I

Origin Headlines	“和珅”与“皇上”再聚首
Expanded Headlines	1. 铁三角和珅皇上再聚首王刚老成这样_奢华_维度女性网
	2. 和珅皇上再聚首68岁的王刚老成这样_13720网
	3. 和珅与皇上再聚首!王刚与张铁林相聊甚欢_网易娱乐
	4. 和珅皇上再聚首68 岁的王刚老成这样张铁林相谈甚欢_娱乐八卦
	5. “和珅”“皇上”再聚首!王刚与张铁林热聊- 新华网
Ground Truth	**entertainment**
Origin Prediction	**history**
Expanded Prediction	**entertainment**

Table 4. Case Study II

Origin Headlines	如果穿整套的话，莫名就有一种呆萌喜感!
Expanded Headlines	1. 看见这货就有莫名的喜感- 糗事百科
	2. 大叔滑稽表演,莫名的喜感!【超级搞笑】-搞笑-高清视频-爱奇艺
	3. 呆萌熊本熊原来这么会跳舞!整个画面充满着莫名喜感…_微信网页版
	4. 呆萌V!168潮男&191双塔刘国梁中网挑边太喜感_乒乓球_新浪竞技...
	5. 【图片】看到这个小黑男嘉宾就想笑怎么破!这是由内而外的喜感!
Ground Truth	**fashion**
Origin Prediction	**fashion**
Expanded Prediction	**sports**

References

1. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. In: Proceedings of the 33rd

international ACM SIGIR conference on Research and development in information retrieval, ACM (2010) 841–842
2. Aggarwal, C.C., Zhai, C.: A survey of text classification algorithms. In: Mining text data. Springer (2012) 163–222
3. Organizer, N..: Corpus for chinese news headline categorization. In: Proceedings of NLPCC 2017
4. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
5. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: AAAI. Volume 333. (2015) 2267–2273
6. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: Advances in neural information processing systems. (2015) 649–657
7. Chen, M., Jin, X., Shen, D.: Short text classification improved by learning multi-granularity topics. In: Twenty-Second International Joint Conference on Artificial Intelligence. (2011)
8. Qiu, Y., Frei, H.P.: Concept based query expansion. In: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, ACM (1993) 160–169
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. (2013) 3111–3119
10. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
11. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8) (1997) 1735–1780