A Semantic Representation Enhancement Method for Chinese News Headline Classification

YIN Zhongbo[1], TANG Jintao[2], RU Chengsen[2], LUO Wei[1]✉, LUO Zhunchen[1], and MA Xiaolei[2]

[1] China National Defense Science and Technology Information Center, Beijing 100142, China
[2] National University of Defense Technology, Changsha 410073, China
lwowen79@gmail.com

Abstract. Recently there has been an increasing research interest in short text such as news headline. Due to the inherent sparsity of short text, the current text classification methods perform badly when applied to the classification of news headlines. To overcome this problem, a novel method which enhances the semantic representation of headlines is proposed in this paper. Firstly, we add some keywords extracted from the most similar news to expand the word features. Secondly, we use the corpus in news domain to pre-train the word embedding so as to enhance the word representation. Moreover, Fasttext classifier, which uses a liner method to classify text with fast speed and high accuracy, is adopted for news headline classification. On the task for Chinese news headline categorization in NLPCC2017, the proposed method achieved 83.1% of the F-measure, which got the first rank in 33 teams.

Keywords: Semantic Representation Enhancement, Short Text Classification, News Headline, Word Embedding.

1 Introduction

With the development of mobile Internet, there are lots of short text such as news headline, microblog and WeChat sent to our mobile everyday. In order to cope with the information exploitation, it is necessary to further process the short texts such as classification. Different from general text classification, short text classification face the problem so-called semantic representation bias because of the lacking of semantic features.

Normally, we use the Vector Space Model (VSM) such as bag of words(BOW) to represent the text semantic. The main idea of VSM is to map text to a vector space which can be used to calculate the semantic similarity between the two text snippets[1]. A simple VSM method is one hot vector, but its vectors usually are too sparse and the vector dimension is too large. In recent years, the more popular method is using neural networks such as word2vec[2] to train a

word embedding model, which can map a text to a more dense and continuous vector. Based on the word2vec's skip-gram representation method, Bojanowski et al. proposed a N-gram approach to add the subword information to the embedding[3] which considers morphological information additionally.

As for the text classification method, the most classic one is Naive Bayesian (Naïve Bayes, NB) algorithm, which is based on Bayesian theorem and feature independent hypothesis[4]. Simultaneously, SVM (support vector machine) and KNN (K-nearest neighbor) also have good performance on Chinese text classification. Compared with these traditional classification methods, the deep learning methods (e.g. CNN and RNN) usually have a better performance but much higher complexity in recent many researches[5][6]. However, Mikolov et al. proves that text classification task is so simple that it does not need the complicated network structure of deep learning methods. Moreover, they put forward a more applicable to large-scale Internet text classification model named Fasttext which is much faster than deep learning approaches[7].

Both text representation and classification methods mentioned above are applied to general texts. However, short texts have the characteristics of sparseness and low-frequency of words. The sparse of words will make it is difficult to count the co-occurrence of features. The low keywords' frequency means that the co-occurrence calculation maybe inaccurate. As a consequence, it would lead to the semantic bias whether using the simplest BOW or the complex embedding. To overcome this problem, this paper focuses on short text semantic representation enhancement.

This paper proposed a novel method for Chinese news headline classification by enhancing semantic representation (CNHCESR), which focuses on the key issues on sparseness and low frequency in short texts. For the sparseness problem, we expand some keywords from the title and snippet parts of the first retrieval result from the search engine. For the bias semantics representation problem caused by low keyword frequency, we build a specific embedding by using the high quality corpus in news domain. Moreover, the Fasttext classification architecture has been used to train a news headline classifier. We do not extend the short texts to general long texts because just several keywords are added to the original texts. Thus, the speed advantage of the short text classification is maintained. Experimental result shows that both the proposed methods to enhance short text representation have a significant improvement in news headline classification.

2 Related Work

In recent years, there are lots of researches in the field of text classification, which is one key task in natural language processing. News headline categorization is one kind of the short text classification, which focuses on short text such as dialog, comment and microblog. As the characteristics of short content, the short text classification is more challenging than traditional long text classification. In order to solve the problem, many researchers have used some external knowledge resources like corpus and thesaurus to improve the existing long text

classification methods and make them suitable for short text classification. The existing short text expansion methods can be divided into two categories: one is using network resources[8] and the other is using domain vocabulary. The network resources (e.g., Wikipedia) is easy to get, but the key point is how to gets high quality resource from lots of network resources. The domain vocabulary is used to build knowledge base or LDA (Latent Dirichlet Allocation) model which is a theme model with probability. LDA model can extract related entities or the theme content to achieve the purpose of expanding corpus[9].

Compared with the English text classification, the first step of Chinese text classification is word segmentation which affects the final classification performance. To improve the segmentation quality, Zhou and Xu et al. constructed a RNN language model with LSTM, which re-integrated neighboring char into the word form in the process of word segmentation [10]. A better semantic representation and classification performance has been obtained while the text is divided into words‘ form but not chars.

In order to further excavate the semantic information from the short text, Wang et al. put forward a method to represent the text‘s apparent semantics and latent aspects [11]. The apparent semantics is caught from Baidubaike by matching vocabulary entry information of the short text; while the latent semantics information is gathered by pLSA method[12]. Finally, the categorization is determined by comparing the specific class‘ correlation coefficient with the computed coefficient based on the apparent semantic information and latent aspect. Wang and Zhou used the hierarchical relationship provided by Baidubaike to identify the semantic topic for the short text[13]. Though this method, they established a convex optimization model to facilitate the short text classification.

In order to represent the semantic links between the few features of the short text more accurately, many researchers used embedding trained under the existing information to improve the classifier's performance[14]. Yao et al. got a 3% accuracy‘s promotion by only using the short text training set to train embedding and added them into the classifier as a preprocessing vector [15]. Furthermore, Ma et al. built some richer embeddings trained under the large sample set and got a more pronounced classification performance[16].

In recent years, some machine learning and neural network models have been developed for short text classification and have been proved effective. For example, Yin et al. used the most basic machine learning algorithm SVM in short text classification[17]. Peng et al. developed a CNN model for short text classification, which could develop more semantic information from short text information[18]. Similarly, the short text classification model based on recursive neural network (RNN) and convolution neural network (CNN) designed by Young et al. achieved a pleasurable result in dialog records classification[19].

Although the deep learning methods have achieved a good performance in text classification, they are still facing the problems of huge resource requirement and training time cost. In contrast, Fasttext[7] used a liner approach in text classification and achieved a similar performance as deep learning methods in terms of a relatively small resource and time. Since the Fasttext architecture is

used for general long text classification, we expand the short text representation by Internet resources.

3 Fasttext

Fasttext is a new text classification tool developed by Facebook. It provides a simple but efficient method for text representation[3] and text classification[2, 7]. For the text representation part, this algorithm train a word embedding, which is similar to the word2vec method. What's different from the word2vec is that the Fasttext representation approach considers additional N-gram in the process of computing embedding. For the text classification part, it only has one hidden layer in the architecture so that the classification process is relatively fast.

Fasttext classification function is similar to word2vec's continue bag of words (CBOW) algorithm[2]. Firstly, the feature vector combined with word sequence is linearly projected to middle hidden layer. Secondly, there is a non-linear activation function which projects middle hidden layer to the categorization label. The difference between Fasttext and CBOW is that Fasttext predicts labels while CBOW predicts middle terms.

Fasttext's official website currently provides the word vector representation of 294 languages. The advantage of Fasttext for text categorization is efficient and fast. Mikolov et al. proved that its performance is on par with deep learning(DL) algorithm, and many orders of magnitude faster than DL algorithm.

4 Semantic Representation Enhancement

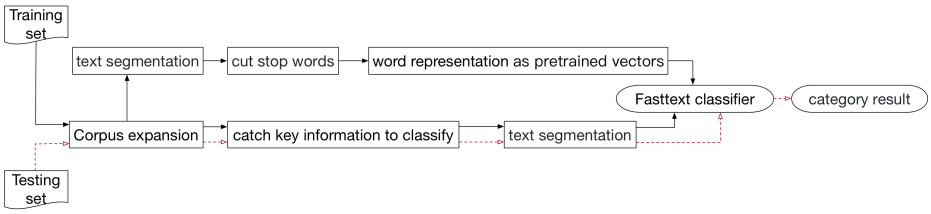

Fig. 1. It is a flow chart of the headline's semantic representation enhancing procedure, where the black solid arrows represent the training process, while the red dotted lines indicate the testing part.

4.1 Feature Expansion

Compared with the general long text classification, news headline classification is characterized by sparse features and low keyword frequency. Based on this aspect, we need to expand the news headlines'features before classification. In

contrast with the LDA based expanding method[20], we use a more simple but reliable approach. With the help of the search engine, the first retrieval's snippet information is gathered for expanding corpus and its keywords are gathered as an additional input corpus to train the classifier.

4.2 Pretreatment

Before classification, the text should be preprocessed such as word segmentation and cutting stop words. In this paper, Jieba is used to segment words for the expanded corpus[21].As the expanded corpus has a large number of functional words, prepositions, punctuation and other noise which are useless for classification. Therefore, we use the stop word list[22] to filter out the stop words in the expanded corpus.

4.3 Pre-train

The semantic distribution is domain related. The same word may have different meaning in different domains. For instance, the word band means a small group of musicians who play popular music or musical instrument in the domain of entertainment, while it means frequency range in the communication domain. Thus, we crawl the corpus in the news domain to enhance the semantic representation by using the headlines in training dataset as the input of search engine. The expanded corpus, which contains the most similar headlines, descriptions, and snippets, is gathered from the first retrieved result. Then we use the expanded corpus to train a word embedding.

Word embedding is a text representation method that maps text semantics to vector spaces[23]. Previous studies have shown that the text classifier with pre-trained word vector pattern has a better classification performance. The literature [2] proved that using skip-gram method to train embedding and negative-sampling method to optimize the trained embedding would obtain a better vocabulary similarity performance. Simultaneously, the literature [24] proposed that adjusting parameters appropriately was beneficial to improve the performance of word embedding representation. In our experiment, we use the skip-gram + negative-sampling method to train a word embedding and optimize the parameters later.

4.4 Keyword Expansion

Short text are characterized by the feature sparseness problem which makes the calculation of feature co-occurrence difficult and inaccurate. Inaccurate co-occurrence represents inaccurate semantic similarity which is the key point to classify. Therefore, it is beneficial to add some keywords for expanding features.

In this paper, we use TF-IDF algorithm to extract keywords from the search snippets. The experiment shows that the best representation performance is gained by adding 13 keywords into original news headlines.

5 Experiments and Results Analysis

5.1 Dataset Sources

The experimental corpus comes from the NLPCC2017 public evaluation: Chinese news headline categorization. This corpus includes 18 news categories. The categorization of discovery, story, regimen and essay has 4000 headlines in each training set. Other 14 categorizations have 10000 headlines. Each categorization of developing set and testing set contains 2000 news headlines.

The specific headline sample are enumerated in Table1, where the first column is category and the second displays some specific headline samples. As the headline samples, these news headlines are typical short text, which has a small amount of vocabulary and few features related to categorization. Therefore, it is necessary to enhance the semantic before classification.

Table 1. Samples for dataset.

Category	Title sentence
entertainment	台媒预测周冬雨金马奖封后，大气的倪妮却佳作难出
food	农村就是好，能吃到纯天然无添加的野生蜂蜜，营养又健康
fashion	14 款知性美装，时尚惊艳搁浅的阳光轻熟的优雅
society	新京报动新闻：高铁断电千人“汗蒸”为啥不能开门窗?
history	红军长征在中国革命史上的地位
story	奇闻录：苗族蛊毒到底是真的么？一则真实的中蛊故事
car	轿车型皮卡在中国会有市场么?

5.2 Performance Evaluation Indicators

The classification performance is evaluated by the following indicators: Macro P, Macro R and Macro F1.

$$\text{Marco P} = \frac{1}{m}\sum_{i=1}^{m}\frac{\text{number of true results to } i \text{ category}}{\text{number of result to } i \text{ category}} \tag{1}$$

$$\text{Marco R} = \frac{1}{m}\sum_{i=1}^{m}\frac{\text{number of true results to } i \text{ category}}{\text{number of } i \text{ category in testing set}} \tag{2}$$

$$\text{Marco F1} = \frac{2 \times \text{Marco P} \times \text{Marco R}}{\text{Marco P} + \text{Marco R}} \tag{3}$$

5.3 Baseline

In this paper, we compared our CNHCESR method with three basic deep learning algorithms which were offered by NLPCC2017[25]: long short-term network (LSTM)[26], neural bag-of-words (NBOW) and convolutional neural networks (CNN)[27].

5.4 Results

Experiment 1: Before Expansion. Table2 lists the performance of the three DL algorithms and Fasttext approach without expansion. Among the four algorithms, Fasttext achieved the best performance with the least training time. In details, the DL algorithms ran on a server node with eight 3.7 GHz Intel (R) Xeon (R) E5-1620 v2 CPUs and Fasttext approach ran on a 2.4 GHz Intel Core i7 CPU with 2 cores. Although the DL algorithms were trained on a more powerful computing server, the training time was still much more than Fasttext.

Table 2. Performance of classification with original dataset.

Model	Macro P%	Macro R%	Macro F%	Accuracy %	Training time
LSTM	70.2	69.1	69.2	69.1	201min12s
CNN	76.2	75.5	75.8	75.5	32min46s
NBOW	77.8	77.1	77.4	77.1	41min03s
Fasttext	78.0	77.3	77.7	77.3	9s

Experiment 2: Enhancing Representation with Embedding. Table3 lists the classification performance of Fasttext and DL algorithms while the 100-dimensional embedding (from baseline or our enhancing method) added into the original short text. Our enhancing embedding was obtained by using the full-scale expanded corpus to train a more elaborate embedding by Fasttext's word representation function. From the table 2 and 3, it can be concluded that each classification accuracy will rise at least 1% after adding an pre-train embedding. As a result, it is reasonable to believe that adding the embedding is helpful to improve the classification performance.

Table 3. Performance of classification with original dataset and embedding.

Model	Macro P%	Macro R%	Macro F%	Accuracy %	Training time	Embedding source
LSTM	77.5	76.8	77.1	76.8	97min36s	Baseline
CNN	79.0	78.4	78.7	78.4	30min23s	Baseline
NBOW	79.7	79.0	79.3	79.0	37min13s	Baseline
Fasttext	79.0	78.3	78.7	78.4	49s	Baseline
LSTM	79.6	79.2	79.4	79.2	99min56s	Enhancing
CNN	79.9	79.3	79.6	79.3	32min54s	Enhancing
NBOW	81.0	80.5	80.8	80.5	30min47s	Enhancing
Fasttext	81.0	80.5	80.8	80.5	23s	Enhancing

Table3 shows the performance over a period between the baseline embedding and enhancing embedding. As can be seen in Table3, each approach had a 2% promotion after replacing baseline embedding with our enhancing embedding,

which indicated that using domain specific resource to train embedding could more accurately represent the semantics of words in this domain.

According to Table 2 and 3, though the classification performance of Fasttext was almost the same as NBOW which had the best performance in the baselines, Fasttext had a significantly advantage over NBOW in efficiency. Therefore, this paper selected Fasttext method with comprehensive consideration below.

Experiment 3: Enhancing Representation with Keywords. This experiment was based on Experiment 2 with a softmax loss function. This experiment's expanded the original news headline with 13 keywords from the first search snippet. Simultaneously, we used different loss function to promote the performance. As can be seen from Table4, the classification accuracy was promoted by 1% after changing the loss function from softmax to negative sampling. And it can be promoted at least 2.4% by using the expanded keywords.

Table 4. Performance of classification with keywords expansion.

Corpus	Loss function	MacroP%	MacroR%	MacroF%	Accuracy%
orginal	softmax	79.0	78.3	78.7	78.4
	negative sampling	79.9	79.4	79.6	79.4
	hierarchical softmax	76.2	75.5	75.8	75.5
expand	softmax	81.2	80.8	81.0	80.8
	negative sampling	82.2	82.0	82.1	82.0
	hierarchical softmax	79.2	78.7	79.0	78.7

Experiment 4: Enhancing Representation with Keywords and embedding. Based on the excellent performance of Fasttext in the experiment 2 and 3, we combined the enhanced embedding and keywords in this experiment. Additionally, we set the n-grams to 2 and negative samples to 10 in this experiment. The fourth experiment results list in the Table5.

Table 5. Performance of classification with keywords expansion and embedding.

Corpus	Loss function	embedding	Macro P%	Macro R%	Macro F%	Accuracy %	Training time
expand	negative-sampling	baseline	82.7	82.6	82.6	82.6	57s
		Enhancing	83.2	83.1	83.1	83.1	26s

5.5 Results Analysis

Figure2 shows the specific category accuracy and recall. It can be seen from the classification precision and recall index of world, society, travel, entertainment and story is less than 80%. The main reason is that the above five categories' news covers a widely range and are easy to be confused with other categories.

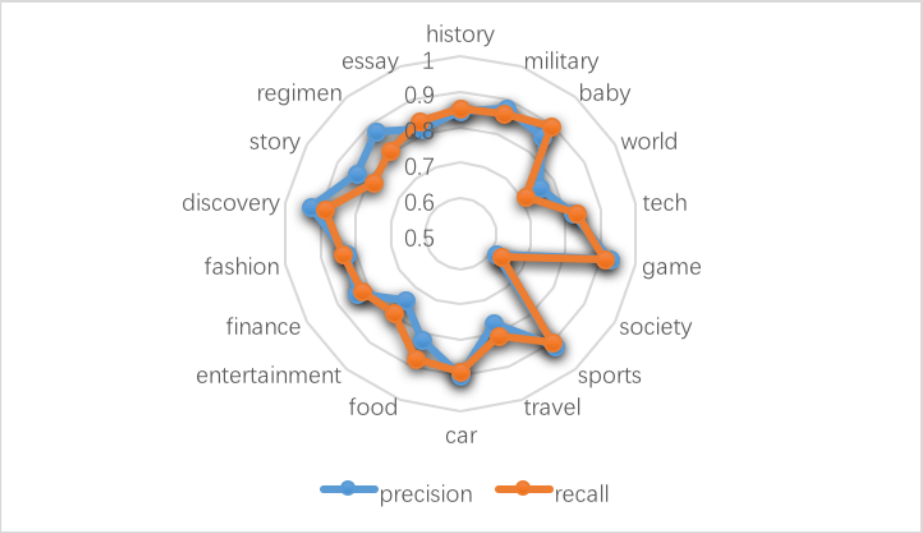

Fig. 2. Performance for specific categorization.

For example, society and world have a high feature coincidence degree which results in the confusion in the classification process.

6 Conclusions

In this paper, we enhanced short text's semantic representation by adding additional keywords of the search snippet and using a more accurate embedding trained with domain-related corpus. Since we just added serval keywords to enhance semantic representation, the expanded input is still a short text which is faster in classification. The experiment results proved that, our Chinese news headline enhancing semantic representation method outperform the art-of-states both on performance and efficiency.

Acknowledgements

Firstly, we would like to thank Jintao Tang and Ting Wang for their valuable suggestions on the initial version of this paper, which have helped a lot to improve the paper. Secondly, we also want to express gratitudes to the anonymous reviewers for their hard work and kind comments, which will further improve our work in the future. This work was supported by the National Natural Science Foundation of China (No. 61602490).

Reference

1. TANG Qi GUO Qing-lin, LI Yan-mei. Similarity computing of documents based on vsmj. *Application Research of Computers*, 25(11):3256–3258, 2008.
2. Greg Corrado Tomas Mikolov, Kai Chen and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
3. P.BojanowskiE.GraveA.JoulinT.Mikolov. Enriching word vectors with subword information. arXiv: 1607.04606, 2016.
4. N Lachiche PA Flach. Naive bayesian classification of structured data. *Machine Learning*, 57(3):233–269, 2004.
5. D Sontag AM Rush Y Kim, Y Jernite. Character-aware neural language models. *Computer Science*, pages 2741–2749, 2015.
6. Yann LeCun Xiang Zhang, Junbo Zhao. Character-level convolutional networks for text classification. arXiv:1509.01626, 2015.
7. P. Bojanowski T. Mikolov A. Joulin, E. Grave. Bag of tricks for efficient text classification. arXiv:1607. 04606, 2016.
8. S Horiguchi XH Phan, LM Nguyen. Learning to classify short and sparse text and web with hidden topics from large-scale data collections. *WWW 2008 Refereed Track: Data Mining – Learning*, pages 91–100, 2008.
9. H Hu X Fan. A new model for chinese short-text classification considering feature expansion. *International Conference on Artificial Intelligence and Computational Intelligence*, 2:7–11, 2010.
10. J Xu L Yang C Li Y Zhou, B Xu. Compositional recurrent neural networks for chinese short text classification. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 137–144, 2016.
11. Cai YQ et al Chen YW, Wang JL. A method for chinese text classification based on apparent semantics and latent aspects. *Journal of Ambient Intelligence and Humanized Computing*, 6(4):473–480, 2015.
12. Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. *Probabilistic latent semantic analysis*, number 289–296, Stockholm, Sweden, 1999.
13. W Luo JX Du YW Chen, Q Zhou. Classification of chinese text based on recognition of semantic topics. *Cognitive Computation*, 8(1):114–124, 2016.
14. X Liu X Wu L Sang, F Xie. Wefest: Word embedding feature expansion for short text classification. In *IEEE International Conference on Data Mining Workshops*, 2017.
15. Jianhui Huang Jin Zhu Di Yao, ingping Bi. A word distributed representation based framework for large-scale short text classification. In *International Joint Conference on Neural Networks*, pages 1–7, 2015.
16. Zhen Zhang Taisong Li Yan Zhang Chenglong Ma, Xin Wan. Short text classification based on semantics. In *International Conference on Intelligent Computing*, volume 9227, pages 463–470, 2015.
17. Hui Zhang Chunyong Yin, Jun Xiang. A new svm method for short text classification based on semi-supervised learning. *Advanced Information Technology and Sensor Application (AITS)*, pages 100–103, 2016.
18. Jiaming Xu ect Peng Wanga, Bo Xua. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing*, 174(PB):806–814, 2016.
19. Proceedings of NAACL-HLT 2016. *Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks*, number 515-520, 2016.

20. Chang Huiyou Hu Yongjun, Jiang Jiaxin. A new method of keywords extraction for chinese short - text classification. *New Technology of Library and Information Service*, 234(6):42–48, 2013.
21. Jieba chinese text segmentation, 6 2017.
22. Stop word list, 6 2017.
23. JS Senécal F Morin JL Gauvain Y Bengio, H Schwenk. Neural probabilistic language models. *Springer Berlin Heidelberg*, 3(6):1137–1155, 2006.
24. I Dagan O Levy, Y Goldberg. Improving distributional similarity with lessons learned from word embeddings. *Bulletin De La Société Botanique De France*, 75(3):552–555, 2015.
25. Corpus for chinese news headline categorization, 6 2017.
26. Jürgen Schmidhuber Sepp Hochreiter. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
27. Yoon Kim. Convolutional neural networks for sentence classfication. arXiv:1408.5882, 2014.