J. Parallel Distrib. Comput. 116 (2018) 18–26

Contents lists available at ScienceDirect

J. Parallel Distrib. Comput.

journal homepage: www.elsevier.com/locate/jpdc

Using convolution control block for Chinese sentiment analysis

Zheng Xiao [a], Xiong Li [b], Le Wang [a], Qiuwei Yang [a,*], Jiayi Du [c], Arun Kumar Sangaiah [d,*]

[a] College of Computer Science and Information Engineering, Hunan University, 410082, Changsha, China
[b] School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China
[c] College of Computer and Information Engineering, Central South Forest Technology University, 410004, Changsha, China
[d] School of Computer science and Engineering, VIT University, Vellore-632014, Tamil Nadu, India

HIGHLIGHTS

- We generate a word vector query library, i.e. the lookup table, trained by the skip-gram algorithm on a large scale unlabeled dataset from Chinese Wikipedia of size 1.3G.
- Based on convolution neural network, we propose a Chinese sentiment classification model on the concept of convolution control block.
- We experiment with the real-world millions of Chinese hotel reviews dataset to compare the performance of our model with LR_all and DCN.

ARTICLE INFO

Article history:
Received 19 July 2017
Received in revised form 26 October 2017
Accepted 30 October 2017
Available online 16 November 2017

Keywords:
Natural language processing
Deep learning
Convolutional neural network
Sentiment analysis

ABSTRACT

Convolutional neural network (CNN) has lately received great attention because of its good performance in the field of computer vision and speech recognition. It has also been widely used in natural language processing. But those methods for English cannot be transplanted due to phrase segmentation. Those for Chinese are not good enough for poorly semantic retrieving. We propose a Chinese sentiment classification model on the concept of convolution control block (CCB). It aims at classifying Chinese sentences into the positive or the negative. CCB based model considers short and long context dependencies. Parallel convolution of different kernel sizes is designed for phrase segmentation, gate convolution for merging and filtering abstract features, and tiering 5 layers of CCBs for word connection in sentence. Our model is evaluated on Million Chinese Hotel Review dataset. Its positive emotion accuracy reaches 92.58%, which outperforms LR_all and DCN by 2.89% and 4.03%, respectively. Model depth and sentence length are positively related to the accuracy. Gate convolution indeed improves model accuracy.

1. Introduction

Nowadays, the Internet have evolved from a static and one-way information carrier to a dynamic and interactive media, where human beings can post comments on news or products, send instant message, etc. The massive interactive information exposes people's emotion, which is users' psychological footprint.

Those posted texts are subjective information reflecting opinion and attitude of people on social events or commodities, and of great value. They help enterpriser to set a good brand image or investigate users' experience. They help consumers to make purchase decision. They help government to know the public opinion on policy or hot social issues. Even they are helpful for medical institutions to evaluate people's healthy status and take counter measures.

Sentiment analysis is widely applied in internet public opinion analysis, product or service review mining, etc. Text sentiment analysis is a classic research topic in the field of natural language processing (NLP). It plays an important role in intelligent network or society.

Machine learning [31,9] and pattern recognition [24,32] are two important technologies in computer field, which can be used in many areas, such as wireless sensor networks [32] and wireless communication networks [24]. Traditional methods of sentiment analysis mainly include the emotional dictionary based classification method [6] and the traditional machine learning based method. Although these methods perform well in classification accuracy, they confront many difficulties. The former relies on a large number of manual operations, such as manual tagging of emotional parts of speech, manual development of search rules, etc. Moreover, such kind of methods is incapable of dealing with new words and unknown words. The latter cannot distinguish the

* Corresponding authors.
E-mail addresses: zxiao@hnu.edu.cn (Z. Xiao), lixiongzhq@163.com (X. Li), wanghnu@hnu.edu.cn (L. Wang), yangqiuwei@hnu.edu.cn (Q. Yang), maxdujiayi@hnu.edu.cn (J. Du), sarunkumar@vit.ac.in, arunkumarsangaiah@gmail.com (A.K. Sangaiah).

https://doi.org/10.1016/j.jpdc.2017.10.018
0743-7315/

semantics of sentences because it ignores the order of words in a sentence. This leads to the problem of sentiment misclassification. Take Bag of Words (BOW) for example, which is a feature model frequently used in machine learning based methods. BOW model represents a text (such as a sentence or document) as a collection of words, but the collection ignores the syntax of the statement and the order in which the words appear. As a result, it cannot catch the context-sensitive information between the words and underneath the sequence. The polarity change sometimes happens in real world cases [23]. The above-mentioned pitfall leads to the potential misclassification.

Words connection has an important impact on classification performance. Due to neglecting it, traditional machine learning exists serious misclassification. For example, given two Chinese sentences, "这个手机很贵但很好看 (the phone is very expensive but good-looking)" and "这个手机很好看但很贵 (the phone is good-looking but very expensive)". The BOW model thinks those sentences identical without noticing the subtle difference, leading to the wrong judgment that those two sentences have the same sentiment. In fact, those two sentences emphasize different parts, and the emotion gets different from each other. The former emphasizes that the phone looks good, and its emotion is positive; while the latter emphasizes that the phone is expensive, and its emotion is negative. BOW model ignores the word order of statement, and thus cannot understand the emotional implication.

The example illustrates the word connection in sentence. There is another problem, the word connection in phrase, which is exclusive in Chinese. For example, given two phrases in Chinese "对不起 (sorry)" and "了不起 (great)", which have opposite sentiment. In English, there is a single word for each phrase. But in Chinese, three characters are contained in each phrase. Furthermore, the character "不 (no)" is negative and challenges the judgment. Therefore, phrase segmentation is a special feature in Chinese, which impedes the transplantation of some methods for English sentiment analysis.

Recently deep learning methods have started to be exploited to deal with NLP tasks, such as word-based ConvNet, long-short term memory (LSTM) [12], recurrent neural network (RNN) model, and convolutional neural network (CNN) model. Section 2 gives a comprehensive analysis and comparison of those methods. This paper focused on CNN for Chinese sentiment analysis task, because of its advantages on training speed and efficiency, parallel implementation.

As to the word connection in phrase and sentence, we propose a Convolution Control Block (CCB) construct and CCB based model. Instead of the bag of words, an unsupervised pre-training algorithm is used to generate word vectors, making the word relation measurable. In a sentence, several words form a semantic unit, which has a unified contribution to sentiment. The number of words in a semantic unit reflects a semantic distance. The larger the distance is, the more words are thought to be interdependent. According to the feature of Chinese, 3 and 5 are believed to be the short and long semantic distances. Based on that, two parallel convolutional operations of different sizes are simultaneously executed. And then gate convolution is used to merge and filter the high-level abstract features. The structure is called Convolution Control Block (CCB) construct, which attempts to deal with the phrase segmentation problem.

A large network of CCBs is organized by tiering and pooling, expecting the word connection to be extended to the sentence level. So a CCB based model is designed using CCB as a basic construct. This model has a good understanding of the dependency and semantics of the Chinese context. The model considers larger context dependency and extracts more abstract hierarchical features by CCB networking.

In contrast to dictionary based and traditional machine learning based methods, our model does not need to manually develop emotional dictionaries and classification rules; partially immune to the new words and the unknown words through training; capable of semantic retrieving damaged in BOW model. To evaluate the performance of our model, two recent deep learning methods, LR_all [3] and DCN [28], are compared. As far as F1-score is concerned, our model outperforms them by about 2%. Model depth and sentence length are positively proportional to accuracy. Gate convolution is testified to work in improving accuracy.

The following summarizes our contributions:

- Based on CNN, we propose a Chinese sentiment classification model on the concept of convolution control block. It is realized by the deep learning library Keras. Its main component is 5 layered CCBs.
- The performance of the model above is assessed on the real millions of Chinese hotel review dataset. The experimental results show that its accuracy is higher than that of LR_all and DCN.
- The depth of convolutional layer, the length of training statement, and the existence of gate convolution are studied, so as to obtain some better Hyper-parameters.

The rest of this paper is organized as follows: In Section 2, we briefly describe some of the relevant work in the field of sentiment analysis. In Section 3, we review the concept of convolution and develop the construct of convolution control block. Sections 4 and 5 describe the model design and the training algorithm. Section 6 shows the experimental results, and finally in Section 7 we summarize the full paper.

2. Related works

2.1. Emotional dictionary based methods

Dictionary based approach is the simplest method for sentimental analysis. Words or phrases are annotated with their sentiment tendency. The emotional intensity of each word or phrase is then aggregated to get the emotional orientation of the whole text. Riloff and Shepherd et al. [25] constructed semantic dictionary based on corpus. Hatzivassiloglou and McKeown et al. [10] studied the emotional tendencies of English words on large-scale corpus, with impact of conjunction on sentiment of adjectives. Kamps and Marx [14] research on emotional tendencies by use of the wellknown emotional dictionary WordNet.

2.2. Traditional machine learning based methods

Traditional machine learning methods include maximum entropy, decision tree, support vector machine, hidden Markov model, and conditional random field. These methods are unnecessary to build dictionary, instead construct feature template for text classification and sentiment recognition. They are based on annotated dataset from which learning is automated. Wang D et al. [33] proposed a new approach to detect the sentiments of Chinese microblogs using layered feature. Semi-supervised learning is a good choice when we have far more unlabeled data than labeled data for training. Li J et al. [19] presented a semi-supervised bootstrapping algorithm for analyzing China's foreign relations from the People's Daily. Yu N and Kübler S [35] investigated the use of Semi-Supervised Learning in opinion detection both in sparse data situations and for domain adaptation and showed that co-training reaches the best results in an in-domain setting with small labeled datasets.