A New Text Sentiment Analysis Method Based on Chinese Morphological Features and HowNet

Yan Cheng¹ Chunjia Liu¹ Yunhong Li¹ Linhui Zhong¹ and Yue Feng¹
¹School of Computer Information Engineering, Jiangxi Normal University, Nanchang, Jiangxi 330022, China
Email: 786564119@qq.com

Abstract. Traditional deep learning methods have two problems when using vectorized words as input. One is that they only consider the overall semantic information of the vocabulary, but ignore the morphological features of the Chinese vocabulary and the prior knowledge of the Chinese external knowledge base. Second, the word vector corresponding to each word will be limited to a single word vector training model. Aiming at these problems, we propose a double-channel convolutional neural network model based on Chinese morphological features and HowNet. First, the cw2vec model and the SAT model (Sememe Attention over Target Model) are used to train the word vectors. Second, the two different word vectors are used as the input of the two channels of the model. Finally, the convolutional neural network are used to extract the characteristics of the two channels to complete the sentiment analysis task. The comparative experimental results on the two data sets show that the proposed model achieves significantly better classification performance than traditional sentiment analysis methods.

1. Introduction
The main research content in text sentiment analysis is to classify texts with sentiment orientation. In recent years, deep learning methods have achieved good results on text sentiment analysis tasks. However, most traditional deep learning methods use a single word vector to train the model, which causes the model to rely heavily on the initial value of the input vector. In addition, these methods cannot make full use of the morphology of the word and external knowledge base information. Aiming at the above problems, and taking into account the particularity of the Chinese language, this paper uses two models of cw2vec [1] and SAT[2] to train two different word vectors. cw2vec can capture the morphological characteristics of Chinese characters, and SAT can make full use of the existing Chinese HowNet knowledge base. Both models can better extract the semantic information of Chinese words. These two different word vectors are then used to input a double-channel convolutional neural network for training.

2. Related Work

2.1. Sentiment Analysis
Text sentiment analysis is to judge the emotional polarity of text by learning the context content of text, and it is an important branch of natural language processing. In 2014, Kim proposed a classical convolutional neural network model and applied it to the English sentiment classification task [3]. Based on this, Kalchbrenner et al. proposed a wide convolution and K-max pooling method [4]. Conneau et al. proposed a VDCNN model and adopted a deep convolutional network method [5]. In order to combine the advantages of CNN and RNN, Siwei Lai et al. [6] proposed the RCNN model, which first uses the bidirectional cyclic neural network to obtain the context representation, and then
outputs the classification result after convolution and pooling operations. In general, the sentiment classification model based on deep learning has achieved better classification results in most classification tasks than traditional methods based on statistical learning and linguistic ideas. However, most traditional deep learning methods use a single word vector to train the model, which causes the model to rely heavily on the initial value of the input vector.

2.2. Word Vector

Word vector, also known as word embedding, is a distributed representation method based on neural network, which can map words in natural language to dense, real-valued vector space [7]. The most classic and widely used one is word2vec, its training models (Skip-Gram and CBOW) were proposed by Mikolov et al.[8], but the representation method is based only on the distribution hypothesis so that the semantics of the word depends entirely on its context and the semantic information contained in the vocabulary itself is not fully utilized. In response to this problem, in terms of Chinese vocabulary representation, many researchers have done some fine-grained lexical representation research on the morphological characteristics of Chinese characters. Chen X, Xu L et al. proposed a word vector training method (CWE, character-enhanced word embedding model) that uses Chinese characters to enhance the effect of words.[9] Su, T. R., & Lee, H. Y. uses pixel-level information, which exploits character features from font images by convolutional autoencoders, and they proposed Glyph-Enhanced Word Embedding Model (GWE).[10] Yu, J., Jian, X., Xin, H., & Song, Y. proposed a method of Joint Embeddings of Chinese Words, Characters, and Fine-grained Subcharacter Components to learn word embedding (JWE). [11] Shaosheng Cao improved on the basis of Skip-Gram, and trained the stroke n-grams feature information of the words instead of words to further enhance the performance of Chinese word vectors [1]. And this method achieved the result of state-of-the-art at the time.

On the other hand, one problem with word embedding is that a vector (spatial point) represents a word and cannot handle polysemy, that is, the same word is expressed differently in different contexts [12]. In order to solve this problem, the researchers integrated the prior knowledge of the external knowledge base into the model, and made some progress in semantic optimization. Chen et al. use the synset information in WordNet to guide the semantic representation of words [13]. On the basis of the Skip-Gram model, Niu et al. integrated the word sememe of the Chinese knowledge base HowNet into the model, and proposed the SAT model [2]. This work was the first work to utilize sememes in HowNet to improve word representation learning at that time.

3. Method

3.1. Cw2vec Model

cw2vec is a Chinese word embedding method based on stroke level information proposed by Shaosheng Cao et al. at AAAI 2018[1]. The study divided the Chinese strokes into five categories and digitized the stroke information. Each word is represented as a sequence of strokes using the n-gram window swipe method, and each gram and word is represented as a vector to train and calculate the similarity between them.

The experiment uses the inner product sum of all gram vectors and word vectors that make up the target word to represent the similarity between words. Equation (2) is the probability that the context is c given the current word w. Equation (3) is the objective function of the experiment.

\[
sim(w, c) = \sum_{q \in S(w)} \tilde{q} \cdot \tilde{c} \\
p(c|w) = \frac{\exp \left( \sim(w, c) \right)}{\sum_{c' \in V} \exp \left( \sim(s, c') \right)} \\
\mathcal{L} = \sum_{w \in D} \sum_{c \in T(w)} \log \sigma \left( \sim(w, c) \right) + \lambda E_{c \sim P} \left[ \log \sigma \left( -\sim(w, c') \right) \right]
\]
According to the experiments done by Cao et al., compared with the existing Chinese word vector methods, such as word2Vec, Glove [14], CWE [9], GWE [10] and JWE [11], the cw2vec method is in semantic analysis, text classification, Named entity recognition tasks perform better. The article published by Cao et al. verifies that the internal structure of Chinese characters contains rich semantic information. By extracting Chinese character stroke feature information, some semantic information can be effectively extracted, thereby further improving the performance of tasks such as text classification in downstream tasks.

3.2. SAT Model
Yilin Niu et al. integrated the word sememes into the vocabulary representation learning model [2]. The models they propose are based on the Skip-Gram model, and the best performing training model is the Sememe Attention over Target Model, as shown in Figure 3. Compared with the Skip-Gram model, which only considers the context information, the SAT model better understands the word by combining the vocabulary sememes information-assisted model. Specifically, the word sense is disambiguated according to the context word, and the attention mechanism is used to calculate the weight of the context for each sense of the word, and then the word vector is represented by the weighted average of each meaning embedding. The experimental results on the two tasks of word similarity calculation and word analogy reasoning show that the integration of sememes information into lexical representation learning can effectively improve the performance of word vector.

3.3. Double-Channel Convolutional Neural Network Based on Strokes and Sememes

As shown in Figure 4, the Double-Channel Convolutional Neural Network Based on Strokes and Sememes (SS-DCCNN) model is mainly composed of the following six parts.

3.3.1. Input layer: In this paper, different vocabulary representations of the sentences obtained by different word vector mapping methods are used as input to two channels, so that different features in
the sentences are extracted. Specifically, this paper uses the cw2vec model and the SAT model to train
the word vector. The cw2vec model can learn the stroke feature information in the vocabulary, which
can make the SS-DCCNN model of this paper make full use of the unique morphological features of
Chinese characters. The SAT model helps to solve the "word polysemy" problem in the emotional
classification task by extracting and integrating the original information in the HowNet knowledge
base, so that the proposed network model has better experimental results.

3.3.2. Convolutional layer. The convolution kernels with heights of 3, 4, and 5 are used to convolve
the inputs of two different channels, so as to extract the local features of different channel inputs. The
convolution kernel finally learned by the network can be regarded as a feature extraction template.

3.3.3. Pooling layer. The maximum pooling operation is performed on the local features extracted by
different convolution layers of different channels to obtain the most important feature information in
each channel.

3.3.4. Merging layer. A merged layer is used to merge the local features acquired from different
channels to form a feature vector, which combines the different features of the two channels.

3.3.5. Output layer. In the output layer, the softmax classifier is used to output the classification result
of the text to be classified.

4. Experiment

4.1. Experimental Data
This article uses Tan Songbo Hotel Review Corpus Data Set (ChnSentiCorp), collected by Dr. Tan
Songbo from the Chinese Academy of Sciences, and is divided into four subsets. This paper uses the
balanced corpus of ChnSentiCorp-Htl-ba-6000 data to conduct experiments. For example, Table 1 is
shown. Due to the small amount of data in this corpus, the noise and outliers may have a large impact.
On the other hand, in order to further verify the effectiveness of the proposed method, this paper
further crawls a large number of car review data from the Internet to form data sets in different fields
for comparison. The car review data set a total of 70,000, divided into 63,000 training sets and 7000
test sets for experiments, examples are shown in Table 2.

Table 1. Hotel Review Data Example

| positive                                                                 | negative                                                                 |
|-------------------------------------------------------------------------|-------------------------------------------------------------------------|
| 交通方便环境很好，服务态度很好，房间较小。                               | 服务很差！！！服务员的态度很不好！环境也一般。                            |
| (The transportation is convenient, the environment is good, the service attitude is good, and the rooms are small.) | (Service is poor!! The waiter's attitude is very bad! The environment is average.) |
| 地理位置不错，早餐中等偏上，网速一般，在长春还算物有所值。             | 房间比较差，尤其是洗手间，房间隔音和餐饮服务都不好。                    |
| (Good location, medium to high breakfast, average internet speed, good value for money in Changchun.) | (The room is poor, especially the bathroom, the room sound insulation and food service are not good.) |
Table 2. Example of car review data

| positive                                                                 | negative                                                                 |
|-------------------------------------------------------------------------|--------------------------------------------------------------------------|
| 车子外观好看，车内空间大。  
(The car looks good and has a lot of space inside.)                   | 配置太少了，做工太差，噪音有点多。  
(The configuration is too little, the workmanship is too poor, and the noise is a bit more.) |
| 最满意的是空间，性价比超高的车车  
(The most satisfying is the space, cost-effective car)                 | 轮胎太窄、行李架太次、车漆有些薄、没有日间行车灯。  
(The tires are too narrow, the luggage rack is too inferior, the paint is a little thin, and there are no daytime running lights.) |

4.2. Data Preprocessing

This training word vector file uses the Chinese Wikipedia dataset, which contains 2.65 million Chinese Wikipedia articles, 1.53G in size. We use the Skip-Gram and CBOW models of Google's open source Word2vec tool, as well as the GloVe, cw2vec and SAT models to get different word vector files. For uniform comparison, the length of the word vector is all set to 200. The training parameters of these models are the same as those specified in the previous section, and the rest are set with the original paper and refer to the source code published by the author.

For the two text sentiment analysis datasets, this paper first removes the non-Chinese part, and then uses the jieba word segmentation tool to segment the sentences. For the convolution operation of the two channels of the model, the sliding window size of the convolution kernel is set to 3, 4, 5, the number of convolution kernels size is set to 100, the Dropout parameter is set to 0.5, the number of iterations is set to 30 times, and the activation function uses the ReLU function.

4.3. Comparison Experiment

The internal evaluation of word semantics mainly includes word semantic similarity and word analogy tasks. These two tasks are relatively simple and fast, and reflect the semantic relationship of two words through a numerical value. In order to verify the ability of the word vector model to capture lexical semantic closeness and relevance, we evaluated the trained word vector file baseline models (Skip-Gram, CBOW, and glove) in Chinese word similarity and word Analogy tasks. The evaluation files are the wordsim-240 and wordsim-296 and word Analogy data sets [9]. The experimental results are shown in Table 3.

Table 3. The performance of these word vectors on the word similarity and word analogy task

| Model   | Wordsim-240 | Wordsim-297 | Word Analogy |
|---------|-------------|-------------|--------------|
| CBOW    | 57.7        | 61.1        | 64.2         |
| GloVe   | 59.8        | 58.7        | 65.8         |
| Skip-Gram | 58.5      | 63.3        | 73.4         |
| SAT     | 61.2        | 63.3        | 84.5         |
| cw2vec  | 60.4        | 63.6        | 78.1         |

It can be seen from Table 4 that in the internal evaluation task of the word vector, the two word vector models cw2vec and SAT selected in this paper are better in word similarity and word analog task than the three commonly used word vector models. Therefore, they have stronger semantic representation ability.

In order to verify the validity of the word vectors trained by cw2vec and SAT models in text sentiment analysis tasks, we use the following methods for comparative experiments:

1) Rand-CNN: Randomly initialize the word as input, and then use the Text-CNN model for text classification.
2) Skip-Gram-CNN: The word vector is trained using the Skip-Gram model and is used as the initialization input for the Text-CNN model.

3) cw2vec-CNN: The word vector is trained using the cw2vec model and is used as the initialization input for the Text-CNN model.

4) SAT-CNN: The word vector is trained using the SAT model and is used as the initialization input for the Text-CNN model.

5) CNN-multichannel [3]: Using the CNN-multichannel model proposed by Kim, the word vector matrix of tunable parameters and the word vector matrix of non-adjustable parameters are combined as input.

6) SS-DCCNN: Use cw2vec word vector and SAT word vector to map text as input of two different channels, thus blending stroke and sememes features in Chinese word, and then use Convolutional &pooling layer to extract features.

| Methods            | ChnSentiCorp | car reviews |
|--------------------|--------------|-------------|
| Rand-CNN           | 87.83%       | 94.51%      |
| Skip-Gram-CNN      | 89.17%       | 95.41%      |
| cw2vec-CNN         | 89.83%       | 95.47%      |
| SAT-CNN            | 90.50%       | 95.88%      |
| CNN-multichannel   | 90.83%       | 96.13%      |
| SS-DCCNN           | **91.83%**   | **96.34%**  |

As can be seen from Table 4, the two word vector methods used in this paper, SAT-CNN, achieved an accuracy of 90.50% and 95.88%, and cw2vec-CNN reached 89.83% and 95.47%, both of which achieved good results. By comparison, the three models using pre-training word vectors are better than the classification results of random initialization of word vectors, which indicates that the initial word embedding layer can significantly improve the classification accuracy. In these three models, cw2vec-CNN and SAT-CNN are better than Skip-Gram-CNN, which shows incorporating the morphological semantic information of Chinese characters and incorporating HowNet knowledge can achieve better results in the Chinese text sentiment analysis task.

On the other hand, CNN-multichannel and SS-DCCNN are better than models using a single word vector (cw2vec-CNN and SAT-CNN), and the SS-DCCNN model significantly surpassed models using a single word vector (cw2vec-CNN and SAT-CNN) on both data sets, indicating that using the two-channel model is effective in fusing word vectors. Further comparison found that SS-DCCNN is better than CNN-multichannel model when only using the double-channel model, it illustrates the superiority of this model on word vectors, since we use cw2vec and SAT models to train word vectors and then more features has been extracted.

In summary, using the model SS-DCCNN in this paper can get 91.83% (ChnSentiCorp) and 96.34% (car reviews) accuracy. Both data sets are higher than all baseline models because the use of multiple channels to fuse different word vectors can incorporate the prior knowledge contained in these word vectors, so that this article learns more distinguishing features, and let the model have a strong representation ability for Chinese text, and then use the convolution pooling operation and the merged layer to extract and fully fuse the local features obtained from different channels, so as to make better use of these two prior knowledge to improve the effect of the current task.

5. Conclusion
When the traditional convolutional neural network takes the vectorized words as input, on the one hand, it ignores the semantic information contained in the internal structure of the word, and on the other hand, it only relies on a single word vector to cause insufficient representation ability. In response to these two problems, this paper proposes a new methods of text sentiment analysis, SS-DCCNN, based on the characteristics of Chinese commentary corpus. We use the two word vector files obtained as the input of the two-channel model, so that the model integrates the two prior
knowledge of Chinese character morphology and HowNet external knowledge base, so as to improve the classification result. Since the cw2vec word vector incorporates the stroke feature information of Chinese characters, the SAT word vector incorporates the meaning information of the words, and experiments have also proved that the cw2vec-CNN and SAT-CNN based on these two word vector models have achieved good results in sentiment analysis task. Therefore, this paper selects these two models to train the word vector, and merge the different semantic information represented by the two word vectors in the merge layer, so that the model has stronger representation ability to the text. It can be seen from the experimental results that the proposed double-channel convolution neural network model has higher accuracy in text sentiment classification than multiple methods. It is indicated that the integration of Chinese morphological features and external knowledge base information has certain value when performing emotional classification tasks. Future work includes the following two aspects: a) Try some new methods to better extract Chinese character morphology and external knowledge base information features; b) Try to introduce more features as a new channel for the model to further improve the performance of the model.

6. Acknowledgements
This study was supported by National Natural Science Foundation of China (Grant Nos. 61967011), Key Research and Development Program of Jiangxi Province (Grant Nos. 20161BBE50086), Key Science and Technology Program of Jiangxi Provincial Department of Education (Grant Nos. GJJ150299).

7. References
[1] Cao, S., Lu, W., Zhou, J., & Li, X 2018 cw2vec: Learning chinese word embeddings with stroke n-gram information. In Thirty-Second AAAI Conference on Artificial Intelligence.
[2] Niu, Y., Xie, R., Liu, Z., & Sun, M 2017 Improved word representation learning with sememes. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vol. 1, pp 2049-58.
[3] YOON K 2014 Convolutional neural networks for sentence classification. Empirical Methods in Natural Language Processing.
[4] Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom 2014 A convolutional neural network for modeling sentences. arXiv preprint arXiv:1404.2188.
[5] Conneau A, Schwenk H, Barrault L, et al 2016 Very deep convolutional networks for text classification. arXiv preprint arXiv:1606.01781.
[6] Lai, S., Xu, L., Liu, K., & Zhao, J 2015 Recurrent convolutional neural networks for text classification. In Twenty-ninth AAAI conference on artificial intelligence.
[7] Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C 2003 A neural probabilistic language model. Journal of machine learning research, 3(Feb), 1137-55.
[8] Mikolov T, Chen K, Corrado G, et al 2013 Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
[9] Chen, X., Xu, L., Liu, Z., Sun, M., & Luan, H 2015 Joint learning of character and word embeddings. In Twenty-Fourth International Joint Conference on Artificial Intelligence.
[10] Su, T. R., & Lee, H. Y. 2017 Learning chinese word representations from glyphs of characters. arXiv preprint arXiv:1708.04755.
[11] Yu, J., Jian, X., Xin, H., & Song, Y. 2017 Joint embeddings of chinese words, characters, and fine-grained subcharacter components. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 286-91.
[12] Li Fenglin 2019 Research Progress of Word Vector Semantic Representation. Information Science, V37 (5): 155-65.
[13] Chen X, Liu Z, Sun M 2014 A unified model for word sense representation and disambiguation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP): 1025-35.
[14] Pennington J, Socher R, Manning C 2014 GloVe: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP): 1532-43.