Research on English Vocabulary Classification Based on Computer Deep Learning

Meng Yin

1Huangshan University, Anhui, China, 245041

*Corresponding author e-mail: yinmeng@hsu.edu.cn

Abstract. With the popularity of the Internet, the analysis and classification of text data has become a research hotspot in recent years, and text classification is one of them. In this paper, cbow algorithm is used to train the word vector based on deep learning, and fast text is used for text classification experiment.

Keywords: Vocabulary, Computer learning, Deep learning, Mentallexicon, Meaning Construction

1. Introduction
With the advent of the information age, the computer and people's life is constantly compact, playing a more and more important role. For Chinese users, Chinese information processing technology has occupied a position in computer science in China. Chinese information processing mainly involves Chinese conversion, transmission, storage, analysis of and other processing. It is one of the biggest technical difficulties faced by human trans-cognitive intelligence. Human needs to teach computer to understand human natural language, so that machine can think and reason. Artificial intelligence, information retrieval, machine translation and automatic abstract areas such as the key to breakthrough, has long been plagued by many experts and scholars in the field of the study, researchers at present our country has made significant achievements in natural language processing, natural language processing each year related meeting contributors, half are Chinese scholars, and it also shows China's huge advantage in Chinese natural language processing, but still need to have a lot of breakthroughs, such as segmentation technique [1].

2. History and mainstream usage of text classification
With the deepening of Chinese information processing research, Chinese text word segmentation has attracted considerable attention and become a frontier problem in Chinese information processing. In this regard, many text word segmentation methods have been designed. The classical classification algorithms are Naive Bayes NB, K-nearest Neighbor KNN, Decision Tree D Tree, Arithmetical Average Centroid AAC and Support Vector Machine SVM [2].

The text classification tool we adopted is Fast Text, a word vector and text classification tool open source from Facebook, which was opened in 2016. This includes the use of word bags and N-gram representational statements, and the sharing of information between categories through implicit representational learning [3]. Fast text is mainly used for text classification. With the breakthrough of
technology, the learning speed of Fast Text has become very fast and the effect has been getting better and better. Fast text is especially suitable for cases where the classification category is very large and the data set is large enough. However, when the classification category is relatively small or the data set is small, it will be easy to overfit [4].

At the same time, based on the text data, a version of word vector is trained. Word vector is to express words in the form of vector. Word vectors mainly include one-hot, distributed expression of words, etc. One-hot is a very simple word vector, but it has a lot of problems. For example, it is well known that vocabularies are usually large, and if they reach the millionth level, each word will be represented by a million dimensional vector, which can lead to a memory disaster. Another is that any two words are separated from each other and cannot represent the relevant information between words at the semantic level [5].

In the past, one-hot denoted a word by removing its semantic meaning and symbolizing it. So how do you incorporate semantics into word representation? Harris's "distributed hypothesis", which addresses this problem, provides the basis for this hypothesis: if the words in the context are almost similar, then they can be assumed to have similar meanings. The distributed hypothesis was further elaborated and clarified by Firth in 1957: the meaning of a word is determined by its context. CBOW, for example, if a sentence "the children play computer games in their house", at the time of training will be "the children play computer games in their" as the output, then predict the last word is "house". The great advantage of distributed representation is that it has a very powerful representational capability. For example, if n dimensional vectors have k values per dimension, they can be represented as k to the n concepts. So, when we want to complete a task, we can break it down into two steps: Describe its context in some way;(2) Build a model to analyze the relationship between the target word and its context. At present, the commonly used word vector representation in the industry is a distributed representation, which is based on neural network training. Its core is still to describe its context in a certain way and to build a model to describe the relationship between contexts. On top of that, we learned about the drawbacks of one-hot coding with too many dimensions. Therefore, the following improvements were made :(1) the elements in the vector were changed from plastic to floating point and adjusted to be represented by the whole range of real numbers;(2) Compress the previous huge dimension into a smaller dimension space. The essence of word vector is the hidden layer parameter or matrix value of training neural network [6].

3. Technical principle
CBOw and SKIP-gram models are mirror images of each other, and the algorithm flow is shown in Figure 1. CBOw algorithm uses the surrounding words to predict the center words, while Skip garm algorithm uses the center words to predict the surrounding words [7].

![Algorithm flow of CBOw and skip.](image_url)
Word segmentation is the process of rearranging successive word sequences with certain rules, and then recombining them into word sequences. Chinese word segmentation is to slice a sequence of Chinese characters to get a single word. On the surface, word segmentation is actually such, but whether the effect of word segmentation is good or not has a great impact on information retrieval and experimental results, and at the same time, there are also a variety of algorithms behind word segmentation. Then according to the characteristics of these algorithms, they can be divided into three categories, namely rule-based word segmentation, statistics-based word segmentation and semantic-based word segmentation. The statistical word segmentation method is our main research object. The definition of this word segmentation method is to regard each word as consisting of the smallest unit of the word and each word. If the linked words appear in different texts very frequently, then these linked words are likely to be one word. In this way, the frequency of their occurrence can be used to calculate the reliability of the word. The frequency of the combination of adjacent co-occurring words in the statistical materials can be used. When the combination frequency is higher than a certain critical value, we can think that the word group may constitute a word. Among them, there are the following statistical models: N-gram, Hidden Markov, maximum entropy, conditional random field, etc. As shown in Table 1.

**Table 1.** Table of three categories.

| Three categories                     |
|-------------------------------------|
| rule-based word segmentation        |
| statistics-based word segmentation  |
| semantic-based word segmentation    |

4. **Train the principle of word vector**

As the name implies, a word vector is used to represent a word in the form of a vector. In this way, the relationship between words can be measured quantitatively, so as to get the relationship between words. One-hot is the first method used to represent word vectors. In this method, each word is represented as a real vector with the length of the dictionary. Each dimension corresponds to each word in the dictionary, except that the corresponding dimension of this word is 1, all other elements are 0. However, one-hot also has two disadvantages: (1) if the number and type of words in a sentence increase continuously, the dimension of the vector will also increase continuously; (2) The biggest disadvantage is that any two words are isolated from each other, and the relevant information between words cannot be expressed at the semantic level. Due to the shortcoming of One-hot, we also introduced the method of neural network training word vector. Neural network training algorithms are mainly divided into two types: CBOW and SkP-gram algorithms. The training of the CBOW model is to input the word vector of a feature and the word vector of the context related word, so that the output is the word vector of the feature. (3) Text Classification Fast Text is an open source word vector and text classification tool of Facebook. It mainly provides a simple and efficient text classification method and representation learning method, and its performance can be compared with deep learning and the speed becomes faster. Fast text combines some of the most successful ideas in natural language processing with machine learning. These include N-gram representation statements and word bags, and then learn to share information through hidden representations.

5. **Conclusion**

In the past period of time, information processing technology has developed very fast and gradually met people's basic needs for information processing. However, from a long-term perspective, its automatic word segmentation technology is still the most important part, so we still need to constantly improve the previous technology for its future development. In order for English culture to spread around the world, we must solve the problem of English participles so that more people can use English more easily. Therefore, in the future, we can turn our research direction to more convenient operation and word segmentation, so as to optimize the current operation mode and improve the accuracy of future word segmentation, so as to enable computers to freely process English texts,
enhance the vitality of English, and make it adapt to the needs of the development of The Times.

Acknowledgements
Key Plan Project of Supporting Outstanding University Young Talents, Anhui, China (Grant No. GXYQZD2016296); 2. Key Research Project of Education Department of Anhui Province (Grant No. SK2015A529).

References
[1] Tao Kai, Tao Huang. A text classification model based on deep learning [J]. Journal of taiyuan normal university (natural science edition), 20(19):45-51.
[2] Tang vanadium, zhao ying, guo qianling, teng jian. Network Traffic classification method based on double-branch Network Deep learning [A]. Network Application Branch of China Computer Users Association. Proceedings of the 24th Annual Meeting of Network New Technology and Application in 2020, Network Application Branch of China Computer User Association [C].
[3] Wu ling, wang hao, zhang xiaochun, zhou jian, duan aihua. Design and implementation of garbage classification system based on deep migration learning [J]. Journal of shenyang university (natural science edition), 20(32):496-502.
[4] Zhu yuhua, lu jianfeng. Deep learning based difficult classification sample pedestrian detection method [J]. Computer and digital engineering, 2020, 48(10):2420-2424.
[5] Shao xuhui. Training Word Vector and Text Classification based on Deep Learning [J]. Electronic Production, 2020(20):80-82+100.
[6] Liu Audi, DU Xuehui, WANG Na, QIAO Rui. ABAC Access control Strategy Automatic generation Technology based on deep learning [J]. Journal of communication: 1-14 [2020-12-25]
[7] Depth based on reinforcement learning business function chain cross-domain mapping algorithm [J]. Computer application research: 1-5 [2020-12-25]