Computer Life
ISSN: 1819-4818 | Vol. 11, No. 2, 2023

Discovery of Weibo New Words Based on Rules and SVM

Yuanfang Xu

Inner Mongolia Normal University, Hohhot, China

Abstract: This article combines the proposed word features with SVM for Weibo new word recognition and extraction. Firstly, by modifying the segmentation dictionary to simulate Weibo new words, the training and testing corpus are segmented using the segmentation dictionary, and various proposed word features are counted. Then, the positive and negative samples extracted from the training corpus are vectorized using word features, and different kernel functions are selected to obtain Weibo new word classification support vectors through SVM training. By adding relaxation variables to improve the accuracy of classification, the Weibo new word classification support vector obtained from the training corpus and the Weibo new word candidate vector obtained from the test corpus are combined for SVM testing to obtain the calculated value of each candidate Weibo new word. The final Weibo new word recognition result is obtained by comparing the calculated value and threshold. Experiments have shown that the combination of word features and SVM can be used for the recognition and extraction of Weibo new words, and relatively good results have been achieved. This method can be extended to the application field of Weibo new word recognition.

Keywords: Weibo data; Discovering new words; Constraints; SVM.

1. Introduction

Although new vocabulary is constantly emerging on social media platforms such as Weibo [1], providing us with rich language choices for daily and online communication, it also poses new challenges to Chinese word segmentation technology. Currently, Weibo new word recognition technology mainly includes three categories: rule-based, statistical, and a combination of these two methods. Among them, rule-based methods rely on rule sets manually selected, annotated, and constructed by linguistic experts for identifying new Weibo words [2]. Based on statistical methods, it is generally used to first screen out potential Weibo new words that may match, and then use professional Chinese language knowledge and other filtering methods to remove non-Weibo new word strings to obtain the final statistical screening result. Both rule-based and statistical methods have certain drawbacks. The main drawback of rule-based methods is that they are not easily organized into a rule base with sufficient coverage, and the resulting rule base often only has a high recognition accuracy for a specific industry field. The main drawback of statistical methods is that it is difficult to find a complete, suitable, and effective statistical system for Weibo new word recognition [3]. Nowadays, most Weibo new word recognition methods use a combination of rules and statistics. This article uses statistical and partial rule-based methods to identify Weibo neologisms.

2. Discovery of Weibo New Words Based on Rules and SVM

2.1. SVM Theory

SVM is a method dedicated to minimizing the structural risk of statistical learning objectives [4], namely:

$$R(w) \leq Remp(w) + \Phi(n/h) \quad (1)$$

In the formula, R(w) is the real risk, and Remp(w) is the empirical risk, Φ(n/h) is confidence risk. SVM has the advantage of training small samples and testing large samples. Therefore, when solving problems, SVM is independent of the dimensionality of the samples, even if the samples are tens of thousands of dimensions. This makes SVM very suitable for solving classification problems. This article combines SVM and word features to solve the problem of Weibo new word recognition. Of course, this ability is also due to the introduction of kernel functions.

2.2. Kernel function selection

This article selects three different kernel functions for experimental analysis:

Polynomial kernel function:

$$K(x, x\_i) = [(x \cdot x\_i) + 1]\text{^}q \quad (2)$$

Radial basis kernel function:

$$K(x, x\_i) = \exp\left(\left(-||x - x\_i||\text{^}2\right) / \sigma\text{^}2 \right) \quad (3)$$

Sigmoid function:

$$K(x, x\_i) = \tanh\left(\nu(x \cdot x\_i) + c \right) \quad (4)$$

2.3. Multivariate classification

In many cases, actual classification tasks often involve multiple factors. However, if these complex multivariate classification problems can be simplified into simple binary classification, it is possible to adopt two strategies to solve them.

N different problems can be categorized into N sets of two categories of problems. Specifically, each problem distinguishes points belonging to the Cn category from points not belonging to the Cn category by using a linear discriminant function.

By using k (k-1)/2 linear discriminant functions, the dataset can be divided into k different categories. These linear discriminant functions only classify two specific categories.

The process of Weibo new word recognition can be seen as a binary classification task, which distinguishes Weibo new words from non Weibo new words. Weibo neologisms and non Weibo neologisms can be considered as two independent categories. With the help of linear classification technology, Weibo new words can be extracted from scattered data, with only a clear boundary line, which is the support vector for Weibo new word recognition.

2.4. feature selection

In the process of training text, we first conduct statistical analysis on the training corpus to construct a context relationship model and determine the Chinese morpheme features that need to be considered when generating sample feature vectors. These features mainly include the MI[5], IWP[6], MP[7], F_F, Context mentioned earlier, as well as the frequency of a Chinese character morpheme appearing in words, the frequency of its individual occurrence, the probability of word formation, and the necessary attributes of some Weibo new words.

2.5. SVM Additional Constraint Classification

To address the issue of mutexes, we introduced constraint 1. Assuming that after using Support Vector Machine for classification, the candidate words with the highest confidence among the mutexes will be considered as the final output, that is, those that deviate farthest from the threshold of 0 will be considered as the final output.

According to our research, if the difference in the frequency of two or more words appearing in an article does not exceed a specific value, we can consider these words as a unified entity. In this example, we set this specific value to 5.

Firstly, for potential Weibo new vocabulary that meets the constraint conditions, we will process it under constraint condition 2. If these potential new words meet constraint 2, we will select the longest one as the only Weibo new word candidate and convert it into a vector. Then, SVM is used to recognize Weibo new words to obtain the final result. However, if these potential new words do not meet the constraint conditions, we will perform SVM recognition on all candidate words and obtain their corresponding results R. According to constraint 1, we will select the candidate vocabulary with the maximum R and the maximum deviation threshold as the final result, as shown in Fig.1:

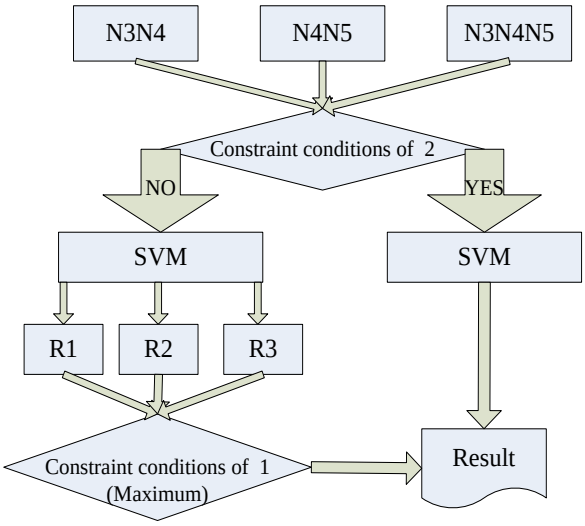

Fig.1 Flow Chart of Combining Constraints with SVM

3. Experimental results and analysis

3.1. 3.1 Experimental training corpus

Firstly, modify the dictionary in the word segmentation software, randomly select 100 words and delete them from the dictionary. Use the program to load the modified dictionary and segment the text. At the same time, remove the part of speech annotation that is not significant for Weibo new word recognition in this article. This is because the word segmentation dictionary has been modified, so the deleted words in the dictionary are simulated as Weibo new words. Through word segmentation and removing part of speech annotations, many scattered strings appear in the segmentation results of the training corpus. These scattered strings are separated due to the simulation of Weibo new words. These scattered strings are extracted through the program and imported into the Weibo new word candidate document as the focus of the next step of processing. Import Weibo new word candidate documents and use the SVM program in this article to extract positive and negative samples from the training data: positive samples refer to vocabulary that has appeared in the training data, while negative samples are negative samples composed of a series of non-vocabulary strings. And the number of scattered words is between 2 and 4, which means that the default Weibo new words are 2 words, 3 words, or 4 words.

Import word feature attribute documents, independent word formation probability documents, and word formation probability documents obtained from Chinese language analysis programs, and vectorize the positive and negative samples obtained from the processed Weibo new word candidate documents to form positive and negative sample feature vectors.

Perform SVM classification training on the vectorized positive and negative sample sets to obtain Weibo new word recognition support vectors, which are used as input for candidate Weibo new word vectors in the recognition test corpus.

3.2. Experimental testing corpus

Similarly, by loading the modified word segmentation software dictionary through the program in this article, the text is segmented and part of speech annotated. The part of speech annotated part is removed, and the scattered strings are defined as candidate Weibo new word strings.

Using the word feature attribute document, independent word formation probability document, and single word formation probability document obtained from the Chinese language analysis program, vectorize the scattered candidate Weibo new word strings to obtain Weibo new word candidate feature vectors for the test corpus.

Taking Weibo new word candidate support vectors and candidate Weibo new word feature vectors as inputs, the input results are tested using the SVM classification program in this article to obtain the initial Weibo new word recognition results.

3.3. Experimental Results and Evaluation

This article selects a 6-month corpus from Sina Weibo. After sorting and integrating, the corpus is 30M in size and contains approximately 500000 words. After extracting the experimental results, a relevant number of Weibo new words are obtained through different features, kernel functions, and other parameters. These data are obtained from the test corpus in this experiment. In the testing of other corpora, the number of Weibo new words obtained varies, this experimental analysis takes this result as an example to analyze the experimental results. The experimental results are shown in Table 1:

Table 1 Statistical analysis of Weibo new word recognition using radial basis functions and different word features

Number	Word Features and Kernel Functions	P (%)	R (%)	F (%)	Change(%)
1	B+RBF	51.23	49.03	52.56	—
2	B+ Context +RBF	61.83	76.27	68.27	15.71
3	B+MI +RBF	63.89	77.07	69.74	17.18
4	B + Context + MI+RBF	67.43	78.37	71.75	19.19
5	B+FF+RBF	61.83	79.79	69.78	17.22
6	B + Context + FF +RBF	71.87	83.46	76.58	24.02
7	B + MI + FF +RBF	72.23	82.32	77.97	25.41
8	B + Context + MI + FF +RBF	75.93	86.17	79.88	27.32
9	B + Context + MI + FF + WWF +RBF	77.89	89.79	83.39	30.83

The experimental results show that when considering more word features, the impact on Weibo new word recognition system is positive. When selecting radial basis function for the same word feature, the experimental effect is optimal. Therefore, this article ultimately uses radial basis function for subsequent word feature increase comparison. The experiment shows that when selecting B + Context + MI + F_F + WWF + RBF model, including word formation probability, morpheme productivity, frequency features When features such as contextual information and mutual information are considered, their recall and accuracy reach their optimal states of 77.89% and 89.79%, respectively. Therefore, in the upcoming experiment, we plan to comprehensively introduce all possible word features for further research.

4. Summary

The main purpose of this study is to use SVM combined with vocabulary feature information to identify new Weibo words. After experimental verification, the method proposed in this article has successfully improved the recall rate of Chinese Weibo new words, indicating that it has played a positive role in improving the recognition ability of Weibo new words. In addition, the corpus we selected has broad representativeness, as SVM exhibits good classification performance in large samples. The experimental results also showed that an increase in the number of vocabularies features significantly improves the accuracy of Weibo new word recognition.

Acknowledgements

Fund projects: Research Project of Inner Mongolia Higher Education Institutions (NJZY21549)

References

[1] Han Xiulong. Research on Weibo New Word Discovery Based on SVM and Feature Correlation [J], Computer Knowledge and Technology, 2018,14,66-69.

[2] Fu Lina, Xiao He, Ji Donghong. New Emotional Word Recognition Based on OC-SVM [J], Computer Application Research, 2015,71946-1048.

[3] Feng Yong, Li Hua. Based on Adaptive Chinese word segmentation and approximation of SVM text classification algorithm [J], computer science, volume thirty-seventh, 2010, first, 251-254, 293.

[4] Qian Qiuyin, Zhang Zhenglan. A method based on multiple SVM classification method of relevance feedback image retrieval [J], computer technology and development, 2009, volume nineteenth, issue eighth, 66-69.

[5] Huang Xiuli, Wang Yu.SVM in unbalanced data set [J], computer technology and development, 2009, volume nineteenth, issue sixth, 190-193.

[6] Li Chengcheng,Xu Yuanfang, Based on support vector and word features new word discovery research, proceedings of 2012 IEEE International Conference on Computer Science and Automation Engineering ,2012,166-168.

[7] Jian-Yun Nie, Unknown Word Detection and Segmentation of Chinese using Statistical and heuristic Knowledge. Communications of COLIPS,2008,5(I&2),47-57.