Research on New Words Discovery of Weibo Based on SVM and Word Features

Yuanfang Xu

Inner Mongolia Normal University, Hohhot 010022, China;

axuyuanfang86@126.com

Keywords: Weibo neologisms; SVM; word features.

Abstract: In order to effectively identify new words in Weibo corpus, a new word discovery method based on SVM and word features is proposed for the unique text characteristics of Weibo corpus. With the help of the good classification of SVM, firstly, positive and negative samples are extracted from the micro blog corpus and the part of speech tagged training corpus, and then vectorized by combining the characteristics of various words calculated from the training corpus, and then the micro blog new words classification support vector is obtained through the training of SVM. In this paper, the word segmentation and part of speech tagging are performed on the test corpus containing simulated new words, and the candidate new words are selected by combining the proposed constraints and relaxation variables. After vectoring with the characteristics of the words themselves, the candidate new words are used as input and the trained SVM classifier is used for calculation. The results are compared with the threshold values. When the results are less than the threshold values, it is determined as a new micro blog The most suitable kernel function of SVM is selected by comparing the experimental results.

1. Introduction

With the continuous development of language, the progress and promotion of the network, and the emergence of new words in the network, people's expressions are more vivid and rich. At the same time, as a key technology of word segmentation in Chinese information processing, micro blog new word discovery technology has an important impact on the accuracy and recall rate of Chinese information processing. With the development of the Internet, Weibo, as a new social network media and information exchange and sharing platform in the Internet, has gradually become a new force of the network based on the diversity of its user level, the immediacy of information in the publishing area, and the freedom of publishing content. New events and new things are emerging on Weibo, of course, there are also a variety of new words on the network The emergence represents the language trend of the Internet, and the effective processing and analysis of micro blog information has become an important issue to be solved.

Although the related technology of micro blog neologism discovery is mature, the research on micro blog neologism discovery based on micro blog information is still less. This paper proposes a method combining SVM and word features to identify the neologisms in micro blog corpus, which can improve the accuracy of micro blog neologism discovery based on some existing algorithms.

2. Weibo neologism discovery based on SVM and constraints

selection of candidate words

The traditional new word discovery algorithms usually use the related Chinese word segmentation tools to segment the corpus first, then analyze the scattered strings after segmentation, and extract new words from them. However, due to the novelty and irregularity of micro blog word formation, some new words are the corresponding combination of the existing words, such as "Wang Zhe pesticide". When using the word segmentation tool to segment words, it will be divided into two unrelated words, "Wang Zhe" and "pesticide", but this word is actually popular in micro blog. In

order to avoid this problem, this paper does not use the existing word segmentation tools to segment the corpus.

2.2 corpus preprocessing

Crawl through 300000 popular Weibo of sina in October 2018. These Weibo cover different topics, including society, science and technology, education, etc. Compared with the traditional Chinese text corpus, the expected text content of Weibo is extremely disordered. In addition to the normal Chinese content, there are a lot of irrelevant interference items, such as emoticons. After obtaining candidate words, we need to label these words to determine new words and non new words. Then, the results of annotation are divided into training set and test set.

2.3 feature selection and calculation of candidate words

In this method, feature vectors are formed by combining the features of words with Weibo data and test data. The features of words selected in this paper include mutual information (MI), word frequency (TF), morpheme productivity (MP), frequency feature (FF), context information (context).

2.4 relaxation variables and penalty factors

Figure dots represents the positive class, while the square dots represent the negative, can be found and a square point appears in the H1 right into the high dimension space in the map, but this point should be negative, this should appear on the left side of H2, this is because the points that can be in linear classification the sample into a linear unclassifiable, also called approximate classification, the new word recognition, word originally in the article number or corpus is less, when this point if direct give up on this point or the point classification may be the correct rate of recall or have a greater impact, so we added a slack variable method to solve the problem mentioned above, we sample point distance requirements:

$$\zeta_i = y_i[(w \cdot x_i) + b] \geq 1 \quad (i = 1,2,\cdots,l) \tag{1}$$

Where w and b are the classifier parameters, l is the total number of samples up string, w as the weight vector, b offset, the dimension of w and space, such as in two-dimensional space g (x) of the w is a 2 dimensional vector. In n-dimensional space, w is an n-dimensional vector. From the sample point recently classification function interval should be greater than 1. If you want to introduce fault tolerance, will give 1 of the hard threshold and a slack variable, which allows

$$\mathrm{y_i}[(\mathrm{w} \cdot \mathrm{x_i}) + \mathrm{b}] \geq 1 - \zeta_{\mathrm{i}} \quad (\mathrm{i} = 1,2,\cdots,\mathrm{l}) \tag{2}$$

Because of the slack variables be nonnegative, therefore the final result is the requirement of interval can be less than 1. But when some points of this interval is less than 1 when the situation (which is also called outlier), mean that we give up the precise classification of these points, and this is a loss of our classifier. But to give up these points also brought a benefit that is to make the surface without moving to the point, so they could be a geometric interval greater (it seems, in a low dimensional space classification boundaries are much smoother). Obviously we must weigh the losses and benefits. The benefits are obvious, classification intervals we get bigger, more benefits.

The loss was added to the objective function, we need a penalty factor C to measure, the original optimization problem into the following:

$$\begin{aligned} &\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} \zeta_i \\ &y_i[(wx_i) + b] - 1 \geq 0 (i = 1,2,...l) \\ &\zeta_i \geq 0 \end{aligned} \tag{3}$$

In Formula 3, not all candidate new words sample points have a relaxation variable corresponding to them. In fact, only "outliers" exist, or it can be assumed that all the point relaxation variables without outliers are equal to 0, that is, for negative classes, outliers deviate from those negative sample points on the right side of H2, and for positive classes, they deviate from those positive sample points on the left side of H1.

For the actual problem, new word recognition problem as the method, only consider the above situation is not enough, because for the classification of Figure 1 can be understood as the number of positive and negative samples of the difference is not great, but for new words than words in terms of number of difference is big, because most of the words in the corpus there is a dictionary of words rather than words, so the formula 4 slack variables and the penalty factor optimization:

$$\begin{aligned} &\min \frac{1}{2} \| w \|^2 + C_+ \sum_{i=1}^{p} \zeta_i + C_- \sum_{j=p+1}^{p+q} \zeta_i \\ &y_i[(wx_i) + b] - 1 \geq 0 (i = 1,2,...l) \\ &\zeta_i \geq 0 \end{aligned} \tag{4}$$

Where i represent the non words part of bulk string, and j represents is identified as part of the powder on new words. For the determination of and take a method to estimate the sample proportion is set in the training corpus, this study simulated the 100 new words, and the training corpus contains approximately words (including the new words) for 3000000 a number, not new words: new words = 30000:1, so this paper set: C_+=0.0002, C_-=0.5, so for the new word is more fair, minimize the misclassification neologisms recognition, so the new word recognition core formula eventually adopted for:

$$\begin{aligned} &\min \frac{1}{2} \| w \|^2 + C_+ \sum_{i=1}^{p} \zeta_i + C_- \sum_{j=p+1}^{p+q} \zeta_i \\ &y_i[(wx_i) + b] - 1 \geq 0 (i = 1,2,...l) \\ &\zeta_i \geq 0 \\ &C_+ = 0.0002 \\ &C_- = 0.5 \end{aligned} \tag{5}$$

3. Experimental results and analysis

Firstly, we test the influence of various word features on the discovery of new words in Weibo. The support vector machine kernel function selects the radial basis function (RBF). At the same time, we select the word feature MP, the word formation probability and the word formation probability as the baseline SVM (B). Then we add context, MI, context + MI, FF, context + FF, MI + FF, context + MI + FF and context + MI + FF + TF to the same test corpus The experimental results are shown in Table 1:

It is found that the more word features are selected, the more positive impact on the accuracy and recall rate of micro blog neologism discovery system. When SVM (B + context + MI + FF + TF) is selected, that is, when the probability of word formation, word frequency, morpheme productivity, frequency features, context information, mutual information and other features are taken into account, the optimal recall rate and correct rate are 71% and 65.66%, respectively In the following experiment, we will introduce all the word features to experiment.

Other conditions are the same. Select all word features and different kernel functions for experiments. The results are shown in Table 2:

Table 1 Experimental results table statistic

Class features	Identify new words	Correct new words	P(%)	R(%)）
SVM(B)	201	78	38	38.80
SVM(B+ Context)	215	86	46	40.00
SVM(B+MI)	213	85	45	39.90
SVM(B+Context+MI)	226	91	52	40.27
SVM(B+FF)	210	83	53	39.52
SVM(B+Context+FF)	199	90	50	45.23
SVM(B+MI+FF)	193	96	56	49.74
SVM(B+Context+MI+FF)	196	123	63	62.76
SVM(B+Context+MI+FF+TF)	198	130	71	65.66

Table 2 Experimental results table statistic

kernel function	penalty Factor (Cost)	Identify new words	Correct new words	P（%）	R（%）
The radial basis kernel function (RBF)	C_+=0.0002 C_-=0.5	196	138	76	70.41
Polynomial kernel function	C_+=0.0002 C_-=0.5	212	118	61	55.66
The Sigmoid kernel function	C_+=0.0002 C_-=0.5	239	126	66	52.72

Through experiments, it is found that the recall rate and accuracy rate of Weibo neologism system are the best when radial basis function (RBF) is used.

4. Summary

This paper proposes a new word discovery method based on SVM and word features. Through different comparative experiments, the method based on feature correlation can improve the recall rate and accuracy rate of new words discovery in Weibo to a certain extent. The next step is to study the effect of the method proposed in this paper on a large-scale corpus, and select the appropriate kernel function and optimize the kernel algorithm to further improve the accuracy rate and recall rate.

References

[1] Qian Qiuyin, Zhang Zhenglan. A Relevant Feedback Image Retrieval Method Based on Multi-Classification SVM [J], Computer Technology and Development, 2009, Volume 19, Number 8, 66-69.

[2] Huang Xiuli, Wang Wei. Application of SVM in non-equilibrium data set [J], Computer Technology and Development, 2009, Vol. 19, No. 6, 190-193.

[3] Lu Hongliang, Chinese New Word Recognition Based on Large-scale Corpus [D], Dalian University of Technology, 2008.

[4] Xu Liang, Research on Chinese new word recognition [D], Dalian University of Technology, 2008.

[5] Lu Hongliang, New Chinese Word Recognition Based on Large-scale Corpus [D], Dalian University of Technology, 2008.

[6] Cui Shiqi, Chinese New Word Detection and Analysis [D], Institute of Computing Technology, Chinese Academy of Sciences, 2006.

[7] Qin Haowei, A Study of Chinese New Word Recognition [J], Computer Engineering, 2004, Volume 30, 369-370.