Research on Automatic Recognition of New Words on Weibo

Yuanfang Xu

Inner Mongolia Normal University, Hohhot, China

xuyuanfang86@126.com

Abstract. To effectively capture emerging vocabulary on Weibo, this article proposes a new Weibo new word recognition strategy that combines Weibo data and support vector machine. Firstly, select positive and negative example sentences from Weibo corpus and trained corpus with part of speech tagging. Then, the lexical features in these sentences are transformed into vectors, and then trained using support vector machines to obtain classification support vectors for Weibo new words. Finally, input the vectorized features into the already trained support vector machine classifier to identify new Weibo words. Based on the experimental results, the system found the optimal feature combination.

Keywords: Weibo new word recognition; SVM; recognition system; Intelligent recognition.

1. Introduction

With the rapid dissemination and evolution of Weibo, various new Weibo words have emerged one after another. The emergence of Weibo neologisms reflects the development trend of online language, therefore, how to more accurately identify and utilize these Weibo neologism resources has become an urgent problem to be solved[1]. Some Weibo new words may only be reordered from known vocabulary[2], while this study uses a support vector machine model as the basic framework, combining the proposed inter word pattern features and various internal word features to train the training samples to obtain Weibo new word classification support vectors. The smoothed test samples are tested and rule filtered to obtain recognition results.

2. Automatic Recognition of Weibo New Words Based on SVM

2.1 Corpus preprocessing

We used web crawlers to capture 150000 hot topics on Sina Weibo in May 2021, covering various topics such as society, science and technology, education, and more. By using a Chinese language analysis program to statistically analyze the corpus, we obtain the probability of word formation (IWP) [3], morpheme productivity (MP) [4], frequency feature (F_F) [5], mutual information (MI) [6], as well as the probability of single word formation and independent word formation extracted through this program.

Firstly, modify the dictionary in the word segmentation software, randomly select 500 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: positive samples refer to the words appearing in the training corpus, negative samples refer to continuous non word strings as negative samples, and the number of

scattered strings is between 2 and 4, which means that the default Weibo new words are 2 word, 3 word, or 4 word. 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.

2.2 Selection and calculation of candidate word features

This method needs to combine the characteristics of the words themselves with Weibo corpus and test corpus to form feature vectors. The selected features of the words themselves in this article include mutual information (MI), word formation probability (IWP), morpheme productivity (MP), frequency feature (F_F), and contextual information (Context) [7]. For the strings S1, S, S2 after word segmentation, definition 1:

$$WWF(S_1,S,S_2) = F(S_1) \bullet T(S_{new}) \bullet F(S_2) \quad (1)$$

S represents a Weibo new word string, T represents the pattern of left and right words in Weibo new words, and the adjacency category AV of string S is defined as the smaller value of the left and right adjacency categories of the string:

$$AV(S) = \min\{L_{AV}(S), R_{AV}(S)\} \quad (2)$$

Among them, $L_{AV}(S)$ and $R_{AV}(S)$ are the left and right adjacency categories of the string S, defined as the types of words or words that appear on the left and right sides of the string respectively.

2.3 Data processing and related work

Firstly, the test dataset was preprocessed, including word segmentation operations. Then, candidate phrases that may become new vocabulary on Weibo were extracted from it. Next, the absolute discount method was used to smooth these phrases, and the candidate phrases that are most likely to become new vocabulary on Weibo were selected by considering intra word features and WWF vectorization technology. Finally, a new feature vector was obtained based on the feature vectors of these candidate phrases, which can be used for further analysis and research., Construct a matrix for Weibo new word recognition support vector and candidate Weibo new word vectors for SVM testing to obtain the final results. The system flowchart is shown in Fig. 1:

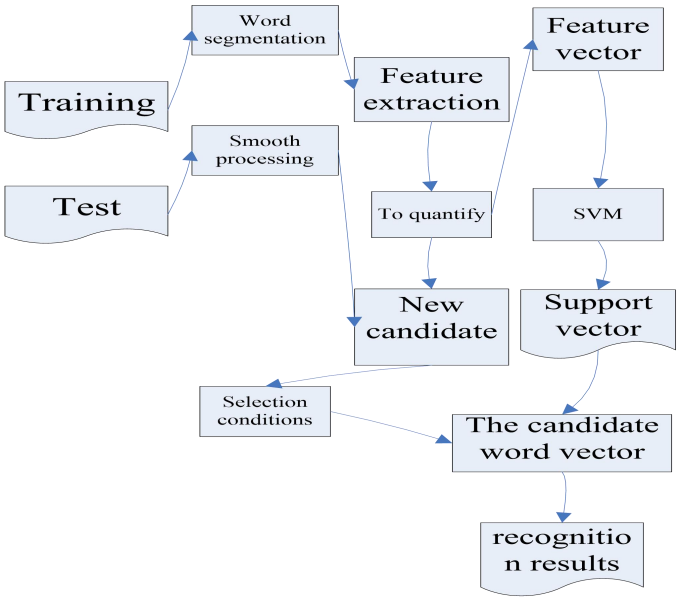

Fig. 1 Structure diagram of Weibo new word automatic recognition system

Due to the fact that the training samples may not include all cases, we smooth out some events that have not been estimated. In this paper, we use the absolute discount method and the estimation formula is:

$$P_r=\begin{cases} m-a/N, 0<m\le M \\ a\cdot K-n_0/N\cdot n_0, m=0 \end{cases}, \quad K=\sum_{m=0}^{M} n_m \text{ , b=2} \tag{3}$$

3. Experimental results and analysis

3.1 Testing Weibo corpus processing

Load the modified word segmentation software dictionary through the program in this article to segment and label the text, remove the part of speech annotation part, and define the scattered strings 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.

3.2 Methods and standards

This method uses accuracy (P), recall (R), and F-measure, which is a comprehensive evaluation based on R and P.

$$\text{F-measure}=(\beta^2+1)\times P\times R/(\beta\times P)+R \quad (\beta=1) \tag{4}$$

3.3 Experimental results and analysis

The experiment evaluated the role of different word features in Weibo new word recognition, and selected SVM's radial basis kernel function as the core algorithm. A basic model F(B) was constructed using three-word features: MP, IWP, and word formation probability. Classify F(B) and obtain the results, as shown in Fig. 2:

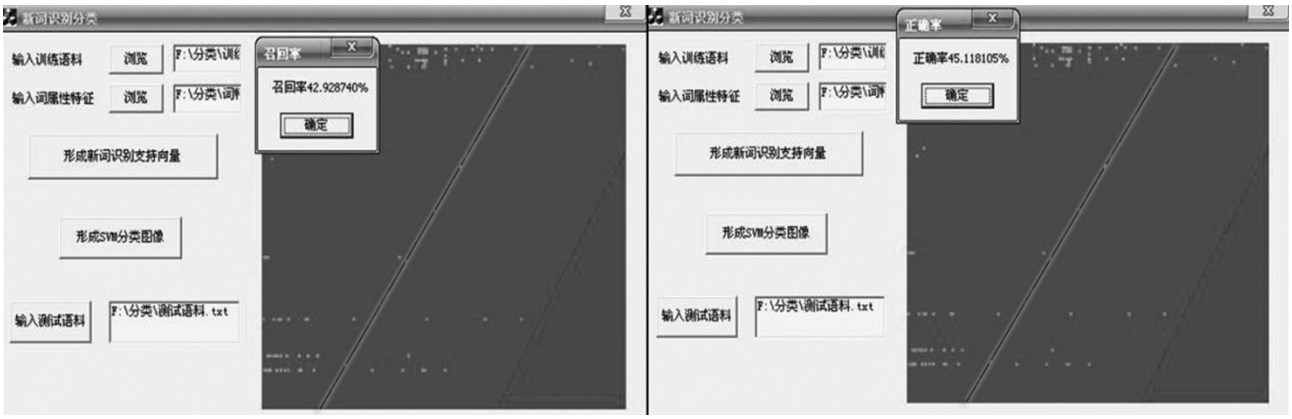

Fig. 2 Operation diagram of Weibo new word recognition system

Next, different word feature combinations are fused into the model, and experiments are conducted on the same test dataset for comparison. The experimental comparison results are shown

in Table 1, After experimental verification, the number of word features in the Weibo new word recognition system has a significant impact on accuracy and retrieval rate. When all possible word features were considered, including word formation probability, morpheme productivity, frequency features, contextual information, and mutual information, the optimal accuracy rate observed in the experiment was 71.78%. Therefore, future research will comprehensively use these word features for further experiments.

Table 1. Statistical Comparison Table of Experimental Results

Numble	Word element	P（%）	R（%）	F（%）	Change(%)
1	F(B)	45.12	42.92	47.85	—
2	F(B+ Context)	55.72	70.16	62.16	14.31
3	F(B+MI)	57.78	70.96	63.63	15.78
4	F(B + Context + MI)	61.32	72.26	65.64	17.79
5	F(B+FF)	55.72	73.68	63.67	15.82
6	F(B + Context + FF)	65.76	77.35	70.45	22.60
7	F(B + MI + FF)	66.12	76.21	71.86	24.01
8	F(B + Context + MI + FF)	69.82	80.06	73.77	25.92
9	F(B + Context + MI + FF + WWF)	71.78	83.68	77.28	29.43

4. Summary

This study uses Weibo data as the foundation to propose a Weibo new word recognition method that combines SVM and word features, and designs and implements a Weibo new word automatic recognition system. After a series of comparative experiments, it has been proven that the system can effectively recognize Weibo new words and improve the accuracy of Weibo new word recognition. The next research focus is to further explore the practical application effects of this method in larger corpora.

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.