Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 5, No. 2, 2023

Research on Weibo New Word Recognition based on Weibo Data and Statistical Information

Yuanfang Xu

Inner Mongolia Normal University, Hohhot, China

Abstract: One of the key challenges in the field of Chinese information processing is the recognition of Weibo new words, which has a profound impact on machine translation and text classification. As Weibo has become the most used social platform for internet users, mining new vocabulary from Weibo data not only helps to deeply understand the data itself, but also provides personalized recommendation services for users. Although a large amount of research has focused on the recognition of Weibo new words, specialized research in this field is still scarce. In this article, we propose a Weibo new word recognition strategy that combines Weibo content features and statistical information. Firstly, extract repetitive vocabulary from Weibo topic names, and then use various methods such as absolute frequency, relative frequency, mutual information, and information entropy to filter for incorrect vocabulary. The experimental results show that by setting appropriate thresholds, incorrect vocabulary can be effectively filtered out, thereby improving recognition performance.

Keywords: Weibo Data; Statistical Information; Mutual Information; Information Entropy.

1. Introduction

With the evolution of language and the progress of the Internet, new vocabulary on Weibo continues to emerge [1], providing people with more vivid and rich ways of expression. At the same time, Weibo new word recognition, as one of the key technologies in Chinese information processing, has a direct impact on the efficiency of Chinese information processing. Weibo, as an emerging social network media, has become an important component of the Internet due to its diverse user base, real-time information dissemination, and infinite content. On this platform, new stories and things are constantly emerging, and Weibo neologisms represent the language development trend of the Internet. Therefore, how to effectively process and analyze Weibo information has become an urgent problem to be solved [2].

The information processing of Chinese faces a unique challenge, which is the lack of symbols that clearly distinguish words like English [3]. Therefore, before segmenting Chinese, it is necessary to first perform Chinese segmentation. However, the new vocabulary on Weibo may lead to many fragments in the segmentation results, thereby affecting the accuracy of segmentation. A study has found that Weibo neologisms may cause half of the segmentation errors [4]. If this new vocabulary can be effectively recognized and processed, it will significantly improve the accuracy of Chinese segmentation and the efficiency of Chinese information processing. In addition, Weibo is the main publishing platform and development center for many online events. By identifying new Weibo words on Weibo data, people can use Weibo content for natural language processing research such as topic tracking, personalized recommendations, and public opinion analysis.

2. Research on Weibo New Word Recognition based on Weibo Data and Statistical Information

2.1. Candidate String Extraction

This article identifies new Weibo words from Weibo data. If all Weibo data is processed, due to the large scale of Weibo corpus and the colloquial nature of Weibo content itself, it will introduce a lot of garbage strings and the processing time will be long, affecting recognition performance and efficiency. To ensure the enhanced applicability of this method, this article will conduct experiments on a small and complete dataset as much as possible.

After in-depth research on Sina Weibo data, we found significant differences between Weibo and other data sources such as forums, blogs, and news corpora: many Weibo not only contain the main text, but also come with a theme tag created by users. This phenomenon is more common in content published by authenticated users and institutions, as well as in long and semantically valuable Weibo accounts. These theme tags are considered part of Weibo and are usually labeled in the form of "#" or "[]". Our statistical data shows that Weibo with self-defined topic tags accounts for 40%, while among Weibo without such tags, 80% of Weibo content is less than 10 words in length. This indicates that the semantic value of Weibo partially depends on the topic tags it contains, and the use of topic tags is a common practice in Weibo with semantic value.

By analyzing the theme tags selected by users, we can quickly gain insight into the core topics or themes involved in the Weibo. Usually, these topic tags contain the key vocabulary of Weibo articles, and the new vocabulary of Weibo is often also a part of the keywords. Therefore, we found that if we only process topic tags instead of full-text processing all corpora, the recognition effect of new vocabulary on Weibo will be better, while also improving efficiency and reducing the probability of meaningless strings, thereby reducing the burden of subsequent screening work.

2.2. String Extraction

This article adopts the most used method for extracting candidate strings, namely the meta incremental model [5]. Considering that topic names are generally not too long, the maximum value of N is 6, which means extracting consecutive strings of length not exceeding from topic names as candidate Weibo new word strings. Additionally, it should

be noted that the topic name may not be a phrase, but rather a sentence containing spaces and punctuation. In this case, the topic name needs to be segmented based on these punctuation marks, and candidate strings need to be extracted from the obtained substrings.

2.3. Candidate String Filtering

Using the N-ary incremental model may generate a large number of strings, but most of these strings are not new vocabulary for Weibo, so they can be considered useless strings and effective filtering strategies need to be implemented. In this article, statistical indicators such as absolute word frequency, relative word frequency, mutual information, and information entropy were used for screening.

2.4. Word Frequency Based Filtering

This article believes that an important feature of Weibo neologisms, especially those used in daily life, is their frequent occurrence. For data such as Weibo, which is large in scale and has content concentration and guidance, this characteristic is more prominent, that is, the probability of Weibo new words as low-frequency words is very small. Therefore, this article considers first using the word frequency of candidate strings for preliminary filtering, retaining candidate strings with word frequency greater than the threshold, and not considering those with fewer occurrences.

2.5. Filtering based on Mutual Information

Mutual Information (MI) [6]: Assuming A is a text string of length n, i.e. S=C1...Cn, M and N are substrings of length n-1, i.e. M=C1...Cn, N=C2...Cn, then:

$$MI(A)=\frac{F(A)}{F(M+N)-F(A)} \quad (1)$$

That is the mutual information between M and N, where F(A) represents the number of times A text string appears in the corpus, and F(M+N) represents the total number of times M and N substrings appear. It can be seen that the larger MI(A), the greater the probability of A becoming a Weibo new word.

2.6. Word Frequency Based Filtering

Mutual information examines the tight internal integration of words, and in actual language environments, it is necessary to simultaneously examine the flexibility of Weibo neologisms as a whole [7], which can be reflected by the variability of their contextual environment. This article uses adjacency context information 0 to describe whether candidate terms themselves are stably combined in different contexts.

This article uses adjacency context information entropy to examine the stability of candidate strings appearing together in different contexts. The higher the adjacency information entropy of a string, the more variable its contextual environment is. The string is more likely to appear in different contexts [8], and the probability of word formation is higher. The filtering method is as follows: first, the corpus is segmented using a word segmentation tool. For each string in the remaining word string set in the previous step, the corpus segmentation result is processed to become an independent unit, and fragments are integrated or extracted from the parent string. For each string S, the adjacency information is calculated by counting the number of occurrences of each adjacency string and S in the adjacent string set, as well as the frequency of S itself, Retain strings with adjacent information greater than the value as the final Weibo new word recognition result.

3. Experimental Results and Analysis

3.1. Experimental Corpus

This study utilizes four months of Sina Weibo data from May to August 2020 for analysis. Firstly, the data from May was manually filtered to identify new Weibo vocabulary and set thresholds for statistical features. Then, use the remaining three months of data to evaluate the recognition ability of these new words.

3.2. Determination of Statistical Feature Thresholds

In order to screen out possible new vocabulary, it is necessary to determine thresholds for mutual information and adjacency information. For the training data in May, 215 new Weibo vocabulary were manually identified. Then, randomly select 60 subsets from these vocabulary as the training set to estimate the range of the threshold. The remaining vocabulary will be used as a validation set to find the optimal threshold.

The artificially identified Weibo new words are extracted using an N-ary incremental model for string extraction and re string filtering. In the filtered training subset, for the case where both the string S and its parent string are included, the information values of all S are calculated. Finally, the threshold range of mutual information is determined by calculating the maximum and minimum mutual information values on this training subset. Then, within this threshold range, Identify and count the accuracy, recall, and F-value of Weibo new words in the validated corpus, and select the best information entropy value as the information threshold for this experiment. After the experiment, the mutual information threshold was determined to be 5.8, at which point the F-value was 78.16%.

Add all manually recognized Weibo new words to the segmentation tool dictionary to assist in word segmentation. Then, on the segmentation results, calculate the contextual adjacency string set of Weibo new words in the training subset, calculate the adjacency information entropy of all Weibo new words in the training subset, and determine the threshold range. Within this range, conduct Weibo new word recognition experiments on the validation set and calculate the accuracy, recall, and F-value, Select the information value that can obtain the best F as the adjacency string information threshold for this experiment. After the experiment, it was determined that the adjacency information entropy threshold was 0.7, and at this point, the F-value was 76.18%.

3.3. Experimental Result

Due to the complexity of manually identifying Weibo new words from a large amount of Weibo data, this article mainly examines the filtering performance of non-Weibo new words in candidate strings. Only one month of test data is used to manually identify Weibo new words and calculate accuracy, recall, and F-value.

In this study, we analyzed the recognition of new vocabulary on Sina Weibo in July and presented the results in Table 1. We have calculated in detail the accuracy, recall rate, and F-value of various indicators after various string filtering mechanisms.

Table 1. New Word Recognition Performance on Weibo Data

Numble	Filter Method	accuracy	recall	F value
1	Not filtered	12.68%	81.87%	25.68%
2	After overlapping string filtering	67.56%	79.88%	73.87%
3	Mutual information	78.90%	76.56%	76.86%
4	Adjacency information entropy	83.67%	75.16%	79.18%

With each addition of a filtering method, the performance of Weibo new word recognition is improved. After statistics, it was found that words with three or more characters accounted for 89.18% of the identified Weibo new words, indicating that this method has good recognition performance for Weibo new words, especially those with longer word lengths. The filtering performance of non-Weibo new word strings in May, June, and August is shown in Table 2. The filtering percentage of non-Weibo new word strings after each filtering method is calculated in sequence.

Table 2. Non-Weibo New Word Filtering Statistics

Numble	Experimental items	Training Set 1	Training Set 2	Training Set 3
1	Number of topics	465	6	18.76
2	Number of non-new word strings	3356	3058	3126
3	After overlapping string filtering	1365	1256	1298
4	After mutual information filtering	952	768	826
5	After filtering adjacency information	812	615	732
6	String filtering percentage	76.87%	79.65%	76.61%

By observing Table 2, it can be found that the research method in this article has good filtering ability for selecting candidate strings of non-Weibo new words. This method can quickly reduce the set of candidate strings with lower experimental costs without losing recognition accuracy.

4. Summary

This article proposes a Weibo new word recognition method that utilizes the characteristics of Weibo content on Weibo datasets to extract candidate strings and filter strings using statistical features. The filtering method considers the degree of internal cohesion of words and the variability of their context. The overlapping string filtering mechanism, mutual information, and adjacency information entropy are effectively used for Weibo new word recognition. After experimental verification, good results have been achieved in the accuracy, recall, and F-value of Weibo new word recognition.

Acknowledgments

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

References

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

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

[3] 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.

[4] 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.

[5] 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.

[6] Su Ning. Based on word features and search engine for Chinese new word identification [J], Journal of Wuhan University, 2010, volume fifty-sixth, issue sixth, 704-710.

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

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