Extraction New Sentiment Words in Weibo Based on Relative Branch Entropy

Liang Yang, Dongyu Zhang, Shaowu Zhang, Peng Fei,
and Hongfei Lin(✉)

School of Computer Science and Technology, Dalian University of Technology,
Dalian, China
{liang, zhangsw, hflin}@dlut.edu.cn

Abstract. There are a lot of new sentiment words appear in Weibo platform every day in the web2.0. As the unpredictable polarity of massive new words are detrimental to sentiment analysis for Weibo, hence how to extract new sentiment words and expand sentiment lexicon is of great importance. Therefore, we propose a relative branch entropy based method, which combines word frequency and adjacent words information to extract new sentiment word in Weibo. After integrated context and other factors, this method improves the accuracy of new sentiment word recognition. Three experiments are implemented on COAE 2014 Weibo corpus to compare the performance of different statistics with the proposed method. Experiment results show that the proposed method has a high accuracy, which demonstrates the effectiveness of this method and verify the promoted effect of new sentiment word extraction on the performance of Weibo sentiment classification.

Keywords: Sentiment word · Relative branch entropy · Sentiment analysis

1 Introduction

Weibo, namely Micro-Blog, is a social networking platform, on which users can share brief real-time information by setting up individual communities or following mechanism through WEB, WAP and other clients. Due to its short, real-time, high efficiency and other characteristics, it has become the dominant communication platform for the expression of ideas and views nowadays. By June 2015, the number of China's Weibo users was 204 million and the use rate of Net-surfers was 30.6% [1]. However, massive new words, especially new sentiment words, occur in Weibo frequently. These new sentiment words have brought great challenge to sentiment analysis, for example, the new word "狗带" as the homophony of "go die" means "go die", that if we don't know its negative emotion, which will have impacts to sentence-level sentiment analysis, such as some sentences similar to "今天的考试,我选择狗带(Today's exam, I choose to go die)".

As sentiment lexicon is the basis of sentiment analysis [2], some existing research firstly identify all the new words, then screen out new sentiment words, while there are also some other researchers extract new sentiment words directly. Among them, the statistics-based method are the most widely used because this approach does not rely

S. Zhang et al. (Eds.): CCIR 2018, LNCS 11168, pp. 158–169, 2018.
https://doi.org/10.1007/978-3-030-01012-6_13

on annotation. But After observed the statistical information, we found that large number of “new words” emerge with very flexible contexts frequently, such as “看事(see thing)”, “我来(I come)” and so on. Unfortunately, it is hard to find a suitable threshold to filter them, which affect the extraction accuracy of new sentiment words definitely. Hence, we propose a novel extraction method for new sentiment word in Weibo based on relative branch entropy. This method takes the context of words into account, which is able to filter out the incorrect neologisms with high frequency.

The main contributions of this paper are that we propose a novel statistics, namely relative branch entropy, to extract new sentiment words, and then set up detailed experiments to compare the performance with different statistics and the proposed method. After that, we further analyze different methods to determine the polarity of new sentiment word, experiment results provide an effective reference for related research.

The rest of the paper is structured as follows: we will introduce related word in the next section. We will emphatically describe the proposed method in Sect. 3. In Sect. 4, we will resent the experiments. Finally, the work is summarized in last Section.

2 Related Work

There are two main thoughts of extraction of new sentiment words at present. The first is to identify new sentiment words directly, for example, Hatzivassiloglou et al. [3] applied “and” or “but” and other conjunctions to obtain the polarity of connected adjectives through POS known words. Qiu et al. [4] employed syntactic analysis to conjecture and extend emotional word set. Huang et al. [5] started from very few seed words and used lexicon patterns to extend sentiment words and patterns iteratively in order to achieve final sentiment words. However, these methods not only requires construction of grammatical rules artificially, but also are ineffective to identify the noun in new sentiment words.

Another line is to identify all the new words firstly, and then screen out new sentiment words. For identifying the new words, there are a large number of researchers studying this issue from different point of view and methods, the methods are summarized as follows:

Rule-Based Methods. Such methods require manual summary and construct rules. Justeson et al. [6] used the regular expression to extract technical terminologies from documents. Isozaki et al. [7] proposed a simple-rule-based generator and decision tree learning method. Chen et al. [8] extracted Chinese new word employing morphological and statistical rules. Zheng et al. [9] established a rule base according to Chinese word formation for new word mining. However, these rule-based methods don’t scale well, hard to maintain, and cannot exhaust all the linguistic phenomena.

Machine Learning Methods. Such methods threat new word identification as a machine learning binary classification problem, for example, Li et al. [10] proposed using an Independent Word Probability(IWP), term frequency in the document and so on as the features of SVM to classify candidate new words. Fu et al. [11] find new word boundary according to train the conditional probability model by using the POS context feature, joint model between words and word formation. Goh et al. [12] trained SVM model by using the Hidden Markov Model to annotate words to obtain character-based labels, and then detect new words by the character sequence. Li et al. [13] proposed applying neologisms pattern as the feature of SVM classifier, combined with rules to get new words. Xu et al. [14] also used SVM with related constraints and slack variables. However, machine learning models require not only heavy engineering of linguistic features, but also ex-pensive annotation of training data.

Statistics-Based Methods. Such methods focus on seeking the statistics of describing new word characteristics, like Pointwise Mutual Information [15], Probability Into Words [16], Rigidity [17] and so on. They select new word by setting the threshold. The first statistical model, Pointwise Mutual Information (PMI), was raised by Church et al. [15] in the 1990 in order to measure the degree of integration between words, since which statistics-based methods developed rapidly. Zhang et al. [18] proposed Enhanced Mutual Information (EMI), based on PMI, to measure the cohesion of co-occurrence word. Huang et al. [19] proposed Branch Entropy (BE) to measure the adjacent character uncertainty of candidate new words. Bu et al. [20] proposed Multi-word Expression Distance (MED) based on the information distance theory.

Then it is required a judgment of polarity for the identified new word. This problem can be considered as a classification problem which uses of existing seed lexicon or the polarity of document consist of candidate word, and the most of them utilized the term or synonym sets in Wordnet[1] [21, 22]. Other common approaches, based on the hypothesis that the similarity between the same polarity sentiment words, aim to calculate the similarity between candidate new words and known sentiment words, such as PMI based [23, 24], word embedding based [25], context similarity based [26] and so on. Among them, the word embedding based method used in this paper.

3 Methodology

In order to detect new words, it is need to use tools to segment Chinese word, so this paper uses Jieba[2] system, which deals with unknown words by HMM [27] and can find out some new words. However, it is inevitable to make some wrong segmentations for the massive emerging new words in the Internet.

To solve this problem, we preprocess Weibo corpus to obtain candidate word set, and use the relative branch entropy proposed to obtain the candidates. Then we train word embedding and compute the similarity between candidates and external sentiment lexicon [2], the larger the similarity is, the larger the probability of a candidate to be a sentiment word is. Finally, we discuss the results of new sentiment word extraction on the Weibo sentiment classification.

[1] http://wordnet.princeton.edu/.

[2] https://github.com/fxsjy/jieba.

3.1 Extraction of New Word Based on Relative Branch Entropy

The new word referred in this paper is the word not in the used lexicon, which contains 77,455 ‘old’ sentiment words, which was provided by the Task 3 of COAE 2014. The new sentiment word is the new word, which contain emotions. We used the lexicon [2] to check the data, and obtained a number of 1,117,299 new words as candidate set *S*.

3.1.1 Rules

The set *S* contains some correct new words, such as “奥迪(Audi)”, “给力(excellent)”, but also contains a large number of incorrect ones, such as “宇大”, “之末” and so on. This may be caused by some characters, namely stop-characters, which should appear alone rather than as a part of a word or phrase. Words contained stop characters are usually incorrect and can be filtered out. Therefore, we used Chinese stop-words list to extract out 264 stop-characters and constructed two types of stop-characters manually as shown in Table 1.

Table 1. Two types of stop-characters

Type	Amount	Examples
词首停用字 (Head-stop-characters)	165	已, 之, 此, 只, 虽
词尾停用字 (Tail-stop-characters)	171	哟, 吗, 了, 么, 呢

3.1.2 Improved Statistics Based on Branch Entropy

In order to find out the correct new words in *S*, we proposed an improved statistic method, named Relative Branch Entropy, which is based on Term Frequency and Branch Entropy.

Term Frequency (TF) is the number of the word appeared in corpus.

Branch Entropy (BE) is an important statistic, proposed by Huang et al. [19], for measuring the uncertainty of the adjacent characters of the candidate new word. The higher *BE* is, the higher uncertainty stands for. The *BE* of the new word is divided into **Left Branch Entropy (*LBE*)** and **Right Branch Entropy (*RBE*)**, and they are computed as follows:

$$\begin{aligned} LBE(w) &= -\sum\nolimits_{l} p(l|w)\log p(l|w) \\ RBE(w) &= -\sum\nolimits_{r} p(r|w)\log p(r|w) \end{aligned} \tag{1}$$

Where *l* is the left adjacent character of *w*, *r* is the right adjacent character of *w*, $p(l|w)$ is the co-occurrence probability of *l* and *w*, then $p(r|w)$ is the co-occurrence probability of *r* and *w*.

Relative Branch Entropy

Only utilizing ***TF*** and ***BE*** is not enough to detect new word, for using ***TF*** can easily filter out some words whose ***TF*** equals 1 like “更会严(more strict)”, but it cannot handle the high ***TF*** and incorrect words like “所行(doings)(*TF* = 22)”, “看事(see thing) (*TF* = 16)” and so on.

Meanwhile, if we take ***BE*** into account, we can calculate out that ***BE*** of “所行(doings)” is 2.1529, ***BE*** of “看事(see thing)” is 2.2200, so when the threshold of ***BE*** is set 2, the incorrect words could be filtered out. However, there are still a lot of new sentiment words whose ***TF*** are high and ***BE*** are low, such as “宅心仁厚(kind-hearted)”(*TF* = 10, *BE* = 0.3250), “淡然处之(take it lightly)” (*TF* = 21, *BE* = 0.1914). To solve the above mentioned problem, we proposed our method, named **Relative Branch Entropy** (***RBE***), and it computed as follows:

$$RBE(w) = \frac{TF(w)}{\min\{LBE(w), RBE(w)\}} \tag{2}$$

3.1.3 Algorithm

In a word, the algorithm of new word extraction is shown as follows:

```
Algorithm : New word extraction algorithm
Input:
  S : a set of candidate word
  HS: head-stop-characters
  TS: tail-stop-characters
  $t_0$ : the threshold of TF
  $t_1$ : the threshold of BE
  $t_2$ : the threshold of RBE
Output:
  A list of new words, NewWords

for word in $S$:
  if the head-character of word in HS:
    continue
  if the tail-character of word in TS:
    continue
  if $TF(word) > t_0 \wedge BE(word) > t_1 \wedge RBE(word) > t_2$ :
    add word into NewWords
return NewWords
```

3.2 Extraction of New Sentiment Word Based on Word2Vec

Word2Vec[3] is a word embedding tool proposed by Mikolov et al. [28], which used of the context information to embed words to vectors. The similarity between words in semantics can be represented by calculating the distance of each word in the vector space. Based on this, we utilized the Skip-Gram model in Word2Vec to train the preprocessed data, and obtain vectors for all the words set V.

In order to identify new sentiment words in new words, we utilize the Affective Lexicon Ontology [2] as the seed word set, then the center sentiment vector $V_{sentiment}$ is computed as follows:

$$V_{sentiment} = \frac{1}{n}\sum_{i=1}^{n} v_i \tag{3}$$

Where v_i is the vector of seed word set.

After that, we can compute the similarity between new words and the center sentiment vector as follows to obtain the final new sentiment words.

$$Similarity(w|w \in NewWords) = \cos(V_w, V_{sentiment}) \tag{4}$$

Where V_w is the vector of w.

4 Experiment

4.1 Data Preparation and Evaluation Metric

In this paper, we will conduct three experiments, and detailed information of corpus is shown in Table 2. And we adapt P@N to evaluate the performance of extraction results for new sentiment word. P@N represents the accuracy of new sentiment word in top N words

Table 2. Information of corpus

Experiment	Corpus	Size	Number of Weibo	Average length	Data format
Experiment.1	COAE 2014 Task.3	1.55 GB	9,999,626	59.24 characters	<Doc_ID> 内容(content) (文字, #话题#, @用户名, 标点) </Doc_ID>
Experiment.2 Experiment.3	COAE 2014 Task.4	5.4 MB	40,000	54.18 characters	<Doc_ID> 内容(content) (文字, #话题#, @用户名, 标点) </Doc_ID>

[3] https://code.google.com/p/word2vec/.

4.2 Parameter Tuning

In order to obtain the thresholds of Algorithm, we firstly set $t_1 = 0$ and results are shown in Table 3, the performance of P@50 across different t_0 and t_2 settings.

Table 3. P@50 results for different t_0 and t_2

t_0	t_2							
	16	20	22	24	25	26	27	28
1	0.64	0.66	0.68	0.74	0.74	0.76	0.76	0.78
2	0.64	0.66	0.68	0.76	0.76	0.78	0.78	0.78
3	0.66	0.68	0.68	0.76	0.76	0.78	0.78	0.78
4	0.66	0.68	0.70	0.76	0.76	0.78	0.78	0.78
5	0.68	0.68	0.70	0.76	0.76	0.78	0.78	0.80

From the algorithm, we can know that the greater the threshold sets, the fewer candidate words get. Therefore, we choose the optimal setting as $t_0 = 2$, $t_2 = 26$.

Then we need to decide the optimal threshold oft_1. After setting $t_0 = 2$ and $t_2 = 26$, we annotate the top 100 words extracted, if the word is new sentiment word, the label is 1, else label is 0. The relevance of new sentiment word and ***BE*** shown in Fig. 2, where the vertical axis is the labels and the horizontal axis is the ***BE*** of each words.

From Fig. 1, we find there is no obvious relevance between whether a word is a new sentiment word and the number of *BE*, so it seems to be difficult to find an optimal threshold, and use it to separate new sentiment word between other words. However, when $t_1 = 0$, experiments achieve good results, hence t_1 set as 0 in this paper.

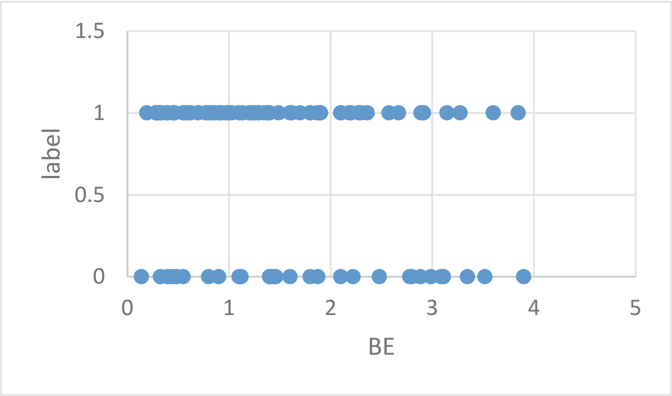

Fig. 1. The distribution of BE

4.3 Experiment Setting

In this section, we will conduct the following experiments:

Experiment. 1
We will compare our method with several baseline methods: PMI [15], EMI [18], MED [20], BE [19] and BA [29].

Experiment. 2
We will predict the polarity of new sentiment words. Firstly, three annotators labeled new sentiment words extracted by our proposed method, then the other two methods were adapted to predict the polarity of these words at the same time.

The first compared method is **Majority Vote** (***MV***), and the polarity is judged according to this rule: if $MV(w) > 0$, the word is positive; if $MV(w) < 0$, the word is negative; otherwise is neutral. ***MV*** is formulated as below:

$$MV(w) = \sum_{w_p \in P} \frac{count(w, w_p)}{|P|} - \sum_{w_n \in N} \frac{count(w, w_n)}{|N|} \tag{5}$$

Where P and N are the positive and negative set of the Affective Lexicon Ontology respectively, and $count(x, y)$ is co-occurrence times of input word x and y.

The second one is **Similarity of Vectors** (***SOV***), the judgement rule is the same as ***MV***: if $P(w) > 0$, the word w is positive; if $P(w) < 0$, the word w is negative; otherwise is neutral. And ***SOV*** is computed as follows:

$$P(w) = \sum_{w_p \in P} \frac{\cos(V_w, V_{w_p})}{|P|} - \sum_{w_n \in N} \frac{\cos(V_w, V_{w_n})}{|N|} \tag{6}$$

Where V_x is the vector of word x.

Experiment. 3
In order to further justify whether expansion of new sentiment word would benefit sentiment classification on Weibo. We use the Weibo corpus in COAE 2014 (Task.4), the corpus consists of 40,000 Weibo posts, and there are 5,000 Weibo posts are sentimental. Firstly, we randomly sampled Weibo posts which contain at least one new sentiment word in the posts. Then two methods are applied for sentiment classification work.

The first method is a **lexicon-based model** (**Lexicon**), and its rule is that counts the number of positive and negative words respectively in each Weibo post, and then classifies the post to be positive or negative. Another method applies the **support vector machine** (**SVM**), where sentiment words are used as features, and 10-fold cross validation is conducted. The **Affective Lexicon Ontology** (**ALO**) [2] and ALO expanded with new sentiment words (denoted by **Expansion**) are utilized as sentiment lexicon resource respectively.

4.4 Results and Analysis

Experiment. 1

The results of different methods in experiment.1 are shown in Fig. 2. We can find that methods ***BE*** and ***RBE*** outperform the three baselines (PMI, EMI and MED) remarkably and consistently. From statistics perspectives, ***BE*** is better than other statistic-based methods. Meanwhile, for our ***RBE*** method integrate ***TF***, ***BE*** and make full use of contextual statistic information of the candidate word, then the accuracy of ***RBE*** is higher than ***BE***. For ***BA***, which use the combination of adjacent strings as candidate words, while it ignores the correct new word before combination, which results in that ***RBE*** outperform ***BA***.

As shown in Fig. 2, the P@N accuracy will decrease as *N* increase, mainly because the number of incorrect words (such as "不见泰山 (be shortsighted)", "欲加之罪(-condemn someone arbitrarily)"), are segmented from longer sentiment words as *N* increase. Due to these words are judged as non-sentiment words, which cause a worse accuracy. In addition, themes of the experiment corpus are complex, the polarities of some words are different in different domains, and this also affects the results.

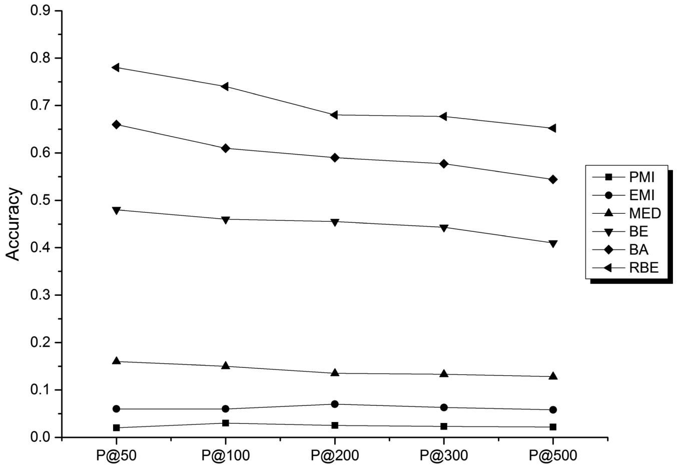

Fig. 2. Results of experiment. 1

Experiment. 2

In this experiment, we extracted top 200 new sentiment words from the similarity ranking list. Since there are three themes (jadeite, mobile phone and insurance) in the corpus, the annotators are requested to consider the domain dependent problem, for example, the word "起胶(light reflection continuously)" is not a sentiment word itself, but from jadeite perspectives, it is used to describe the jade is very valuable, so it should be labeled as positive. Similarly, the word "卡机(not running smoothly)" is also not a sentiment word itself, but it indicates a phone is not smooth, so it expresses a negative feeling in cell phone domain. Based on the above consideration, the label results shown as follows: (Table 4)

Table 4. Label results

Polarity	Amount	Example
Positive	89	起胶 (light reflection continuously), 佳品(treasure)
Negative	82	卡机 (not running smoothly), 骗保(Fraud)
Neutral	29	寒暄(exchange of conventional greetings), 缘分(fate)

Then, we apply ***MV*** and ***SOV*** for new sentiment word tow-classes and three-classes polarity classification respectively, the results is shown in Table 5. As we can see, the performance of ***SOV*** is much better than that of ***MV***, because **SOV** is full use of contextual information by train vectors in neural network. However, the scale of corpus for training is small, so the improvement is not good enough. Another observation in three-classes polarity classification work is more difficult than two-classes polarity classification, for many extracted new sentiment words are nouns, and they are more hard to judge their polarities without domain knowledge.

Table 5. The results of two/three-class polarity classification

Number of classes	Methods	Accuracy
Two-classes	MV	0.845
	SOV	0.865
Three-classes	MV	0.520
	SOV	0.550

Table 6. The results of polarity classification of Weibo posts

Sentiment lexicon	Methods	Accuracy
ALO	Lexicon	0.657
	SVM	0.680
Expansion	Lexicon	0.705
	SVM	0.726

Experiment. 3

In this experiment, we randomly sampled 2,000 Weibo posts (Including 1,123 positive and 877 negative posts) that contain expanded sentiment word from the 5,000 Weibo posts that are official labeled. Results in Table 6 show that **SVM** model outperform **Lexicon** model generally, and expansion of new sentiment words improves the performance to a large degree, both models obtain 6–7% gains. It is an obvious proof for the effectiveness of extended sentiment words.

5 Conclusion

The sentiment lexicon forms the basis of sentiment analysis, due to the manually label work will take a lot of labor; hence it is impossible to cover the increasingly new words in Weibo all the time. Therefore, how to automatically extract new sentiment words from large-scale corpus and use it as an expansion of sentiment resources are great of importance to sentiment analysis researches.

In this paper, we propose a method to extract new sentiment word base on relative branch entropy (*RBE*). This method is full use of term frequency and adjacent information, and almost free of linguistic resources. Comparative experiments show that our proposed method outperforms than other baselines, which verify the effectiveness of the proposed method. Meanwhile, the experiments also demonstrate that expansion of sentiment words will benefit sentence-level sentiment classification.

However, there are still some problems unsolved, such as the wrong segmentations caused by longer words and the evaluation works. We will explore the solutions in our further researches.

Acknowledgements. This research is supported by the National Natural Science Foundation of China (No: 61702080, 61632011), the Fundamental Research Funds for the Central Universities (No. DUT17RC(3)016).

References

1. CNNIC: The 36th China Internet Network Development State Statistical Report. China Internet Network Information Center (2015)
2. Xu, L.H., Lin, H.F., Pan, Y., et al.: Constructing the affective lexicon ontology. J. China Soc. Sci. Tech. Inf. **27**(2), 180–185 (2008)
3. Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In: Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 174–181 (1997)
4. Qiu, G., Liu, B., Bu, J., et al.: Opinion word expansion and target extraction through double propagation. Comput. Linguist. **37**(1), 9–27 (2011)
5. Huang, M., Ye, B., Wang, Y., et al.: New word detection for sentiment analysis. In: Proceedings of the 52th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp. 531–541 (2014)
6. Justeson, J.S., Katz, S.M.: Technical terminology: some linguistic properties and an algorithm for identification in text. Nat. Lang. Eng. **1**(01), 9–27 (1995)
7. Isozaki, H.: Japanese named entity recognition based on a simple rule generator and decision tree learning. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pp. 314–321. Association for Computational Linguistics (2001)
8. Chen, K.J., Ma, W.Y.: Unknown word extraction for Chinese documents. In: Proceedings of the 19th International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, pp. 1–7 (2002)
9. Zheng, J.H., Li, W.H.: A study on automatic identification for Internet new words according to word-building rule. J. Shanxi Univ. Nat. Sci. Edit **25**(2), 115–119 (2002)

10. Li, H., Huang, C.N., Gao, J., et al.: The use of SVM for Chinese new word identification. In: Conference First International Joint Conference on Natural Language Processing, pp. 723–732 (2004)
11. Fu, G., Luke, K.K.: Chinese unknown word identification using class-based LM. Lect. Notes Artif. Intell. **3248**, 704–713 (2005)
12. Goh, C.L., Asahara, M., Matsumoto, Y.: Machine learning-based methods to Chinese unknown word detection and POS tag guessing. J. Chin. Lang. Comput. **16**(4), 185–206 (2006)
13. Li, C., Xu, Y.: Based on support vector and word features new word discovery research. In: IEEE International Conference on Computer Science and Automation Engineering, pp. 287–294. IEEE (2012)
14. Yuanfang, X., Hui, G.: New word recognition based on support vector machines and constraints. In: 2nd International Conference on Information Science and Control Engineering (ICISCE), pp. 341–344. IEEE (2015)
15. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Comput. Linguist. **16**(1), 22–29 (1990)
16. Chen, A.: Chinese word segmentation using minimal linguistic knowledge. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing-Volume 17. Association for Computational Linguistics, pp. 148–151 (2003)
17. Wang, M.C., Huang, C.R., Chen, K.J.: The identification and classification of unknown words in Chinese: a N-grams-based approach. Logico-Linguist. Soc. Jpn. 113–123 (1995)
18. Zhang, W., Yoshida, T., Tang, X., et al.: Improving effectiveness of mutual information for substantial multiword expression extraction. Expert Syst. Appl. **36**(8), 10919–10930 (2009)
19. Huang, J.H., Powers, D.: Chinese word segmentation based on contextual entropy. In: Proceedings of the 17th Asian Pacific Conference on Language, Information and Computation, pp. 152–158 (2003)
20. Bu, F., Zhu, X., Li, M.: Measuring the non-compositionality of multiword expressions. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 116–124. Association for Computational Linguistics (2010)
21. Kim, S.M., Hovy, E.: Determining the sentiment of opinions. In: Proceedings of the 20th International Conference on Computational Linguistics, p. 1367 (2004)
22. Esuli, A., Sebastiani, F.: Sentiwordnet: a publicly available lexical resource for opinion mining. In: Proceedings of LREC, vol. 6, pp. 417–422 (2006)
23. Volkova, S., Wilson, T., Yarowsky, D.: Exploring sentiment in social media: bootstrapping subjectivity clues from multilingual Twitter streams. In: Proceedings of the 51th Annual Meeting on Association for Computational Linguistics, pp. 505–510. Association for Computational Linguistics (2013)
24. Turney, P.D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424. Association for Computational Linguistics (2002)
25. Yang, Y., Liu, L.F., Wei, X.H., et al.: New methods for extracting emotional words based on distributed representations of words. J. Shandong Univ. Nat. Sci. **49**(11), 51–58 (2014)
26. Yu, H., Deng, Z.H., Li, S.: Identifying sentiment words using an optimization-based model without seed words. In: Proceedings of the 51th Annual Meeting on Association for Computational Linguistics, pp. 855–859. Association for Computational Linguistics (2013)
27. Huang, C., Zhao, H.: Chinese word segmentation: a decade review. J. Chin. Inf. Process. **21**(3), 8–20 (2007)
28. Mikolov, T., Chen, K., Corrado, G., et al.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
29. Chen, X., Wang, S.G., Liao, J.: Automatic identification of new sentiment word about microblog based on word association. J. Comput. Appl. **36**(2), 424–427 (2016)