AN ANALYSIS OF THE EMOTIONAL TENDENCY OF NEW WORDS IN CHINESE TEXT BASED ON WORD2VEC

Jiang Quan and Rao Wenbi

Wuhan University of Technology, Wuhan 430070, China

ABSTRACT

At present, there are many new words expressing emotions on the Internet, but the expressions of these new words have rich meanings but lack of accurate definitions, so it is difficult to analyze their emotional tendentiousness, This thesis studies the feasibility and framework design of word2vec based analysis method of emotional neologisms' tendency, and conducts experiments on Weibo corpus. The results show that new words can analyze their emotional tendency from their similar words.

KEYWORDS

Word vector; New word discovery; Emotional word; Tendentiousness analysis; Word2Vec

1. INTRODUCTION

The generation of new words is accompanied by the expression of a variety of emotions, which makes a new word may contain a variety of emotional tendencies, and the importance of each tendency is different. Therefore, this paper proposes a new word emotional tendency analysis method based on word vector, which can analyze the multiple emotional tendencies that new words may contain by quantitative way. At the same time, by training word vector, We can cluster the emotion words with similar emotion tendency and find the synonyms.

1.1. Related work

At present, there are a lot of related work in foreign countries on the discovery of new words and the research on emotional classification of new words. There are also many research results on text emotional analysis. However, domestic research on short text, such as emotional phrases or emotional words, can be divided into two types: one is emotional analysis of short text based on emotional dictionaries and rules, the other is emotional analysis of short text based on machine learning.

Emotion analysis based on emotion dictionary includes Xiao Jiang[1] and others who use similarity method to build emotion dictionary in related fields, Jo's[2] emotion classification method based on "topic sentence" relationship, marking two kinds of labels on words at the same time, and Yang Liyue's[3] Micro blog emotion dictionary, It includes open source emotion dictionary, network emotion dictionary with era characteristics and mood emotion dictionary with obvious emotion tendency. Emotion classification method based on machine learning includes emotion analysis combining co training cooperative training algorithm and SVM proposed by Liu[4] et al. In 2016, dey[5] et al. Used Bayes algorithm for emotion analysis.

Dhinaharan Nagamalai et al. (Eds): NATP, ACSTY, ADCOM, SOFE, CCCIOT, MLSC, SVC - 2020
pp. 79-87, 2020. © CS & IT-CSCP 2020 DOI: 10.5121/csit.2020.100407

However, the discovery of emotional neologisms is not limited to the words found around emotional dictionaries. Many neologisms are full of emotional labels. For example, "Gui Mi" is a noun, but it can express one's intimate feelings towards another. In addition, the emotion of an emotional neologism is not a single classification, but has multiple emotions, Therefore, this paper proposes a word vector based analysis of Chinese neologisms' emotional tendency.

2. DISCOVERY OF NEW WORDS

Chinese word segmentation is an important direction in the field of natural language processing. Unlike western languages, the process of Chinese language processing has a natural boundary between words, so the problem of word segmentation can greatly affect many next steps, such as relationship extraction, automatic summarization, etc. now most of the word segmentation methods are based on the word bank, So, the problem of unregistered words is more important[6,7]. Chinese writing does not have capital letters or proper names, so it is difficult for computers to recognize proper nouns such as people's names and place names. In addition to special names, the emergence and evolution of network terms, brand organization names, abbreviations, abbreviations and other words seem to be completely irregular. With the increasing importance of language processing, The research of Chinese word segmentation is focused on overcoming this problem. Automatic discovery of new words has become a key link.

The traditional method of new words mining is to segment the text first, assuming that the remaining segments of the text that fail to match successfully are new words, and then extract these segments[8]. However, the accuracy of segmentation results depends on the integrity of the segmentation lexicon. If there are no new words in the segmentation lexicon, the segmentation results may lead to the difficulty of mining "new words" into words.

Therefore, new words mining needs to find a new way. A mature idea is that, instead of relying on any established thesaurus, only according to the characteristics of the word itself, all the text fragments that may be words are extracted in a large-scale corpus, no matter whether the word is a new word or an old word. Then, the words in the extracted thesaurus are compared with the existing thesaurus, You can find new words[9-11].

2.1. Cohesion

The primary criterion for judging whether a word can be a word is its congruence[12] .For example, in the training corpus of about 59 million words, there are two fields with a frequency of more than 150, and the frequency of "de she ji" is higher than that of "she ji gan", but "she ji gan" is a word in people's cognition, which shows that the combination of "she ji" and "gan" is closer, But the combination of "she ji" and "de" has not reached the close degree in people's cognition.

The following calculation will prove that the internal solidification degree of the word "she ji gan" is higher than that of "de she ji". If the word "she ji" and the word "gan" appear in the text independently and randomly, then the probability of these three words being put together is calculated. In the 59 million words corpus, the word "she ji" appears 8491 times, The probability of occurrence of the word is about 0.000 0143. The word "gan" appears 59448 times in the corpus, and the probability of occurrence is about 0.000 9321. If the two fields are random and independent of each other, the probability of the word "she ji gan" should be 0.000 0143 × 0.000 9321, about 1.33 × 10^(-7). But in fact, the word "she ji gan" appears 185 times in

the corpus, about 3.13 × 10^(-6), about 46 times of the predicted value. And so on, According to the statistical data, we can find that the occurrence probability of "de" is about 0.0343, so the probability of "de" and "she ji" combining randomly should be 0.0343 × 0.000 0143, about 4.904 × 10^(-6), and the frequency of occurrence should be 290, which is close to the frequency of "de" appearing in the data. The frequency of occurrence of this field is 1816, It is 6.26 times of the predicted value. From the above calculation, it can be seen that the combination of "she ji gan" is closer, and the field is more likely to be a meaningful match, while the appearance of "de she ji" is more likely to be the combination of "de" and "she ji".

However, a notable problem is that there is no prior knowledge in the calculation process. In other words, the word "she ji gan" may be a combination of "she ji" and "gan", or a combination of "she ji" and "gan". Therefore, it is necessary to enumerate multiple combinations of a field during the calculation, and then take the most probable combination.

2.2. Information Entropy

In addition to its internal congruence, there is also a criterion for the word's external freedom[13]. For example, in addition to the usage of "yi bei zi", "zhe bei zi", "xia bei zi ", "shang bei zi", there are not many choices for adding words in front of "bei zi". The words that can appear on the left side of "bei zi" field are relatively limited, Therefore, in the calculation of congruence, "bei zi" is not a word alone, but a whole of "yi bei zi" and "zhe bei zi".

Therefore, the concept of "information entropy" needs to be added. Information entropy reflects how much information will be brought after the result of an event is obtained. If the probability of a result of an event is p, when the result appears, the amount of information will be defined as log (P). The smaller the value of P is, the greater the amount of information will be.

Adjacency entropy is an important statistic put forward by Huang[14] et al. Adjacency entropy statistic uses information entropy to measure the uncertainty of left-hand and right-hand characters of candidate new word T. The higher the uncertainty is, the more chaotic and unstable the strings before and after candidate new word t are, the more likely it is to be a word. For example, in the experimental corpus, the word "bei zi" appeared 1080 times in total, and the word "bei zi" appeared 4030 times in total. The information entropy of their right neighbor word sets was 4.7374 and 6.1655 respectively, It is close in value, but the use cases of "bei zi" are very rich. For example, there are dozens of different uses, such as "jia bei zi", "yong bei zi", "na bei zi", "xin bei zi", "jiu bei zi", "shuai bei zi", "shou bei zi". It is calculated that the information entropy of "bei zi" is 4.9745. However, the left neighbor of "bei zi" is relatively few, Of the 4030 "bei zi" in the corpus, 3240 are "bei zi", 414 are "zhe bei zi ", 261 are "xia bei zi", 78 are "shang bei zi". In addition, there are 15 rare usages, such as " n nei zi " and " liang bei zi ". The information entropy of all the left neighbor words is only 1.3679.

In addition to the left neighbor words, although there are many usages of the left neighbor words in some text fragments, the use cases of the right neighbor words are very poor, such as "guo qin", "tuo er", "e luan", etc., which are not consistent with the common sense.

Therefore, the degree of free use of a phrase or word can be defined as the smaller of the left-hand and right-hand information entropy of the phrase.

3. WORD2VEC : A WORD VECTOR ANALYSIS TOOL

Word vectors have been used to represent words for a long time. They are usually called 1-of-N representation, or unique heat representation, etc., but the dimension used in this method is the size of the whole vocabulary. For each word in the vocabulary, set 0 in the corresponding position of the word to 1. For example, a vocabulary with five words, the vector representation of the second word answer is (0, 1, 0, 0, 0), The vector representation of the fifth word hungry is (0, 0, 0, 0, 1). Therefore, it can be seen that the vocabulary of a vocabulary is generally very large, so the expression of this vocabulary vector is very sparse, and the expression efficiency is not high.

The way to solve this problem is driven representation. Through training, each word is represented as a shorter vector, and each dimension of the vector expresses a semantic information. However, the interpretability of the specific meaning expressed by each dimension of the vector is not good. Before word2vec appeared, Generally, neural networks are used to train word vectors to process words, which can be divided into two models[15-17]: CBOW (continuous bag of words and skip gram) and skip-gram.The input of CBOW model is the word vector of the context word corresponding to a word in the text, and the output is the word vector of the word.For example, sentence fragment " distributed representations which encode therelevant grammatical relations…"the context size is 6 and the output word is encode, the input should be the word vector of the first three words and the last three words of encode. It should be noted that these six words have no order, and the word bag model is used. On the contrary, skip gram model and CBOW model use the word vector of a word as the input and the word vector of the word context as the output. In the above example, the input of skip gram model is the word vector of "encode", while the output is three words of "encode" context, Word2vec uses the data structure of Huffman number instead of neural network model, which is also divided into CBOW and skip gram models.

First, analyze the CBOW model. The first step is to define the dimension size of the word vector as m, and the context size of the field as 2C. So for each word in the training sample, the first C words and the second C words are the input of the CBOW model, and the output is the word vector of all words. The algorithm steps are as follows: algorithm 1.

Algorithm 1. CBOW model algorithm

(1) The Huffman tree was established with training data as samples;
(2) All model parameters θ and word vectors are initialized randomly;
(3) For each sample (context (w), w) in the training set, do the following processing:
(3.1)e=0, calculation $x_w = \frac{1}{2c}\sum_{i=1}^{2c} x_i$

$$f = \sigma(x_w^T \theta_{i-1}^w)$$
$$g = (1 - d_i^w - f)$$
$$e = e + g\theta_{i-1}^w$$
$$\theta_{i-1}^w = \theta_{i-1}^w + g x_m$$

(4) Update every word vector in context(w) (2c in total): $x_i = x_i + e$;
(5) If the gradient converges, the gradient iteration is ended, otherwise, step (3) is returned.

For skip-gram model, the input and output of the model are the same as CBOW model, and the training algorithm is like algorithm 2.

Algorithm 2. Skip-gram model algorithm

(1) The Huffman tree was established with training data as samples;
(2) All model parameters θ and word vectors are initialized randomly;
(3) For each sample (w, context (w)) in the training set, do the following processing:
For i=1 to 2c
E=0
For j=2 to l_w,calculation:

$$f = \sigma(x_i^T \theta_{i-1}^w)$$
$$g = (1 - d_i^w - f)$$
$$e = e + g\theta_{i-1}^w$$
$$\theta_{i-1}^w = \theta_{i-1}^w + gx_i$$
$$x_i = x_i + e$$

(4) If the gradient converges, the gradient iteration is ended, otherwise, step (3) is returned.

In word2vec, in addition to the training model based on Huffman tree, there are also methods based on negative sampling[18]. Because if a word is too remote, the search level of Huffman tree will be more. When using the negative sampling method, the model can be trained only by sampling n different central words as negative examples each time.

4. NEW WORDS DISCOVERY AND ANALYSIS OF EMOTIONAL TENDENCY

Based on word vector, the analysis of new words' affective tendency is to find new words through threshold setting by using the calculation amount of cohesion, information entropy and word frequency mentioned above. Then, by learning the training corpus, word2vec is used to generate the word direction of all words in the word list, Then find out the words with the highest similarity in all the new words list. The structure of the new words discovery and sentiment analysis method is shown in Figure 1.

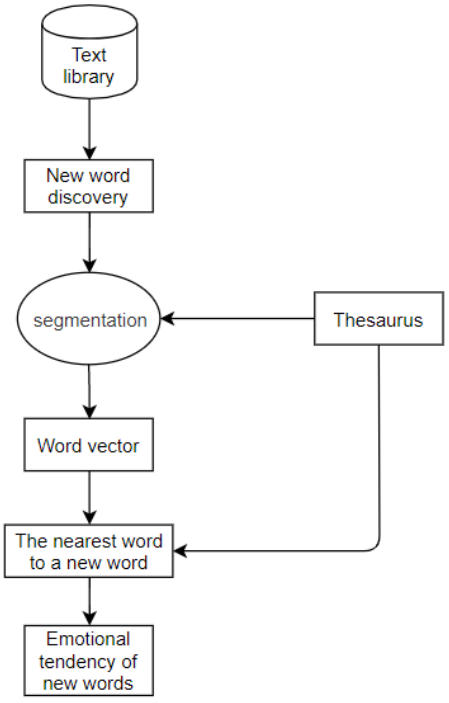

Figure 1 design framework of neologism discovery and sentiment analysis

In this method, all words in the corpus are first divided into fields with length less than 5, and then the cohesion and information entropy of the fields are calculated. After the calculation, the words within a certain threshold are selected as new words. After adding new words into the word segmentation list, the corpus is segmented, and then the word vectors of all words are trained. Then, the new words are calculated, The first n words closest to each word. Finally, the emotional tendency analysis is carried out. The specific construction process is as follows:

Step 1. Extract the text content from the web page by analyzing the XML language or HTML language from the text library to be retrieved

Step 2. Divide the text into several fields whose length is less than 5. Calculate the conglomeration degree and information entropy of each field. According to the best threshold obtained by the experiment, the new word table is selected

Step 3: the text is segmented. Generally, the stop words should be removed after segmentation, but the learning of word vector should be based on the context, and the stop words will also have an impact on the words, so the stop words should not be removed at this step

Step 4. Train the segmented file with word2vec, and adjust the parameters continuously to get satisfactory results. At the end of the training, get the word vectors of all words

Step 5. Find out the first n words which are closest to each word in the new words list, and analyze the emotional tendency of new words through the emotional tendency of these words.

The flow chart of this method is shown in Figure 2.

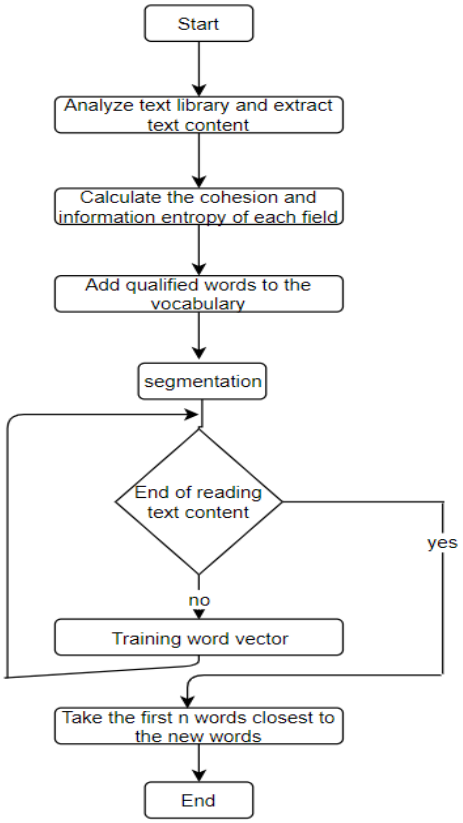

Figure 2 flow chart of new words discovery and emotional tendency analysis method

5. EXPERIMENT AND RESULT ANALYSIS

In order to test the proposed method, this paper grabs 12 million pieces of data from Sina Weibo, mainly analyzing the text content of each Weibo as a corpus.

In the neologism discovery step, after many experiments, this paper sets the threshold value of cohesion for neologism acquisition to 0.35, and the information entropy to 0.5 to 1.5. The example of the added neologism is shown in Table 1.

Table 1 examples of new words discovery

	frequency	Coagulation degree	Information entropy
“zhuakuang”	647	0.808	0.79
“diaosi”	145	0.425	1.08
"miaosha"	109	0.332	0.91
"guimi"	98	0.685	0.88
"kengdie"	75	0.479	0.58

Through the adjustment of parameters during model training, except for the word vector dimension, other parameters are selected, including: the maximum distance between current word and prediction word in a sentence is 3, using CBOW algorithm. The distance between words is calculated by cosine value of two vectors:

$$similarity(A,B) = \cos(A,B) = \frac{A \cdot B}{||A|| \cdot ||B||} \quad (1)$$

The input of the program is five new words in the above examples, and the output is four words with the largest similarity with each word in the training model. When adjusting the word vector space dimension of the model to 100, the output result is shown in Table 2. Each new word corresponds to four words with the closest distance, and the four words and their distance are listed for each new word respectively.

Table 2 analysis results of emotional tendency

Word Spacing	Word1	Word2	Word3	Word4
"zhuakuang"	"lei"	"kelian"	"shuai"	"yun"
	0.7356	0.7038	0.7027	0.6736
“diaosi”	“meinv”	“shuaige”	“jipin”	“jilao”
	0.5975	0.5681	0.5239	0.5068
“miaosha”	“dazhe”	“jingpai”	“huo”	“dao”
	0.5105	0.4987	0.1702	0.4611
“guimi”	“pengyou”	“laoyou”	“qinren”	“jiemei”
	0.6877	0.6133	0.6098	0.5939
“kengdie”	“maodun”	“shuai”	“kelian”	“leiliumanmia”
	0.6242	0.6143	0.6071	0.5985

From the results in Table 2, we can see that the new words trained by word vector can analyze their emotional tendency from the words close to them. For example, the emotion expressed by madness is close to tears, pitifulity and decline. At the same time, we can see the distance between the two words.

6. CONCLUSION AND PROSPECT

In order to more specifically express the emotional meaning and tendentiousness of social new words, this paper proposes a new word emotional tendentiousness analysis method based on word vector. Based on the analysis of the new word discovery method and word vector training tool word2vec, this paper studies the feasibility and architecture design of the new word emotional tendentiousness analysis method based on word2vec, and conducts experiments for microblog corpus. It can be seen from the experimental results that this method has good feasibility and convincing results, but it has not been carried out in the specific classification of new words' emotional tendency, so there are many details to be improved.

The purpose of new words discovery and tendentiousness analysis of emotional words is to better understand the feelings expressed by users through texts, and to provide an exploration method for the mining of unlisted words in Chinese segmentation. Although there are still many difficulties in the research of this direction, it will achieve satisfactory results in the process of continuous in-depth innovation.

REFERENCES

[1] Xiao J,Ding X,He R.[Chinese Micro Blog Sentiment Analysis Based On Domain Sentiment Dictionary]Dian Zi She Ji Gong Cheng. 2015, 23(12): 18–21.

[2] Jo Y, Oh AH. Aspect and sentiment unification model foronline review analysis. ACM International Conference on Web Search and Data Mining. Hong Kong, China. 2011.815–824.

[3] Yang L,Wang Y.[Research On The Construction and Analysis Method of Emotion Dictionary of Micro Blog Emotion Analysis]Ji Suan Ji Ji Shu Yu Fa Zhan, 2019, (2): 1–6.

[4] Liu SH, Li FX, Li FT, et al. Adaptive co-training SVM for sentiment classification on tweets. Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. San Francisco, CA, USA. 2013. 2079–2088.

[5] Dey L, Chakraborty S, Biswas A, et al. Sentiment analysis of review datasets using naive Bayes and K-NN classifier.arXiv: 1610.09982, 2016. Guo S,Xing D.[Sentence Similarity Calculation Based On Word Vector and Its Application]Xian Dai Dian Zi Ji Shu, 2016, 39(13): 99–102, 107.

[6] Guo S,Xing D.[Sentence Similarity Calculation Based On Word Vector and Its Application]Xian Dai Dian Zi Ji Shu, 2016, 39(13): 99–102, 107.

[7] Tang M,Zhu L,Zou X.[A document vector representation based on word2vec]Ji Suan Ji Ke Xue, 2016, 43(6): 214–217, 269.

[8] Feng C,Shi G,Guo Y,etc.[Micro Blog Entity Link Method Based On Word Vector Semantic Classification]Zi Dong Hua Xue Bao, 2016, 42(6): 915–922.

[9] Du L,Li X,Yu G,etc.[New Word Discovery Based On Improved Mutual Information Algorithm and Improvement of Chinese Word Segmentation System]Beijing Da Xue Xue Bao(Zi Ran Ke Xue Ban) , 2016,52(1): 35–40.

[10] Zhang J,Xi Y,Wang B,etc.[An Event Tracking Method of Microblog Based On Word Vector]Ji Suan Ji Gong Cheng Yu Ying Yong, 2016, 52(17): 73–78, 117.

[11] Wang X,Wang Y,Wang L.[Hot Spot Ranking of Network News Based On Neologism Discovery]Tu Shu Qing Bao Gong Zuo, 2015, 59(6): 68–74.

[12] Zhang J,Qu D,Li Z.[Cyclic Neural Network Language Model Based On Word Vector Feature]Mo Shi Shi Bie Yu Ren Gong Zhi Neng, , 2015, 28(4): 299–305.

[13] Li W,Zhang Y,Chen R.[New Words Discovery Based On Internal Combination Degree and Boundary Freedom Degree of Words]Ji Suan Ji Ying Yong Yan Jiu, 2015, 32(8): 2302–2304,2342.

[14] Chen F,Liu Y,Wei C,etc.[New Words Discovery In Open Domain Based On Conditional Random Field Method]Ruan Jian Xue Bao, 2013, 24(5): 1051–1060.

[15] Huang JH, Powers D. Chinese word segmentation based oncontextual entropy. Proceedings of the 17th Asian PacificConference on Language. Sentosa, Singapore. 2003.152–158.

[16] Ghosh M, Sanyal G. Performance assessment of multiple classifiers based on ensemble feature selection scheme for sentiment analysis. Applied Computational Intelligence and Soft Computing, 2018, 2018: 8909357.

[17] Mondal A, Cambria E, Das D, et al. Relation extraction of medical concepts using categorization and sentiment analysis. Cognitive Computation, 2018, 10(4): 670–685.

[18] Zhu YJ, Yan EJ, Wang F. Semantic relatedness andsimilarity of biomedical terms: examining the effects ofrecency, size, and section of biomedical publications on theperformance of Word2Vec. BMC Medical Informatics andDecision Making, 2017, 17: 95.

AUTHORS

Jiang Quan，Mr. of Wuhan Unversity of Technology, China, 430070