Construction of a Chinese Emotion Lexicon from Ren-CECps

Lijuan Wang[1], Changqin Quan[1], Yanwei Bao[1], and Fuji Ren[1,2]

[1] AnHui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine,
School of Computer and Information, HeFei University of Technology,
Hefei, 230009, China
[2] Faculty of Engineering, University of Tokushima,
2-1 Minami-Josanjima, Tokushima 770-8506, Japan
wlj1034900987@163.com, quanchqin@gmail.com,
baoyanwei007@sina.com, ren@is.tokushima-u.ac.jp

Abstract. This paper presents an automatic method to build a Chinese emotion lexicon based on the emotion corpus Ren-CECps. The method includes word extraction and emotion classification. Firstly, sentences are parsed to extract candidate emotional words. By making use of the words co-occurrence in the corpus, we get the similarity between words. And then Support Vector Machine (SVM) is adopted to classify the candidate emotional words. Experiment on the manual labeled words has shown that our classification method achieved high precision. Finally we apply our method on unlabeled corpus to get emotional words.

Keywords: emotion lexicon, corpus, sentence parse, SVM.

1 Introduction

Currently, the number of blogs, network comments and other texts on Internet is increasing every day, and these texts contain huge amount of information. How to quickly recognize the emotional information on characters, events, products, etc., becomes a hot research topic in the field of natural language processing and affective computing [1, 2, 3, 4].

Emotion lexicon plays a very important role in textual emotion classification and recognition. The General Inquirer was developed at Harvard and is a lexical database that uses tags to carry out its tasks. It contains about 3500 entries and each entry consists of a term and a number of tags. The two tags Positive and Negative express valence [5]. MPQA Subjectivity Cues Lexicon is created from the Multi-perspective Question Answering (MPQA) Opinion Corpus, and the lexicon contains 2718 positive words and 4912 negative words [6]. SentiWordNet is developed from WordNet. It is formed by conducting emotion classification and labeling the positive weight and negative weight of the terms in WordNet [7].

D.-S. Huang et al. (Eds.): ICIC 2014, LNAI 8589, pp. 625–634, 2014.

As to Chinese emotion lexicons, HowNet has published a set of emotion words, such as the degree words, positive emotion words and proposal words[1]. But there are no relatively complete Chinese emotion lexicons with precise emotion categories. The manual assignment of emotion categories and intensities needs a great amount of labor and time cost. In order to support the task of emotion recognition in text, this paper proposes an automatic approach to build a Chinese emotion lexicon with precise emotion category. We present a new algorithm of building a Chinese emotion lexicon; the process of our method includes word extraction and emotion classification by making use of the words co-occurrence in the corpus.

The rest of this paper is organized as follows: Section 2 presents a review of methods for building emotion lexicon. Section 3 introduces the emotion corpus Ren-CECps 1.0. Section 4 describes the main algorithm for building emotion lexicon. Section 5 is the experiments. And in Section 6, we come to a conclusion and present some future work.

2 Related Work

There are mainly two methods to build emotion lexicon: lexicon-based method and corpus-based method. Lexicon-based method is to extend the set of emotional labeled words by synonyms and antonyms relations in dictionaries. Corpus-based method is to calculate the similarity between words according to words' co-occurrence and use words semantic similarity computation tendencies.

Li J and Ren F proposed a method to automatically build Chinese emotion lexicon with emotional intensity marked by the use of Tongyici Cilin and HowNet [8]. The advantage of this method is that the initial mass of the existing dictionary words, thus avoiding dependence on the basis of the word.

One approach to creating Chinese emotion lexicon is to use a semantic lexicon, such as WeiPing Liu et al. [9]. They use Chinese emotion words to create a basic emotion dictionary for the different areas. Based on the Chinese word similarity calculation method, proposes a method of calculation the emotional weight of a Chinese word.

Xu Ge et al. [10] described an algorithm based on graph theory and multiple resources to build emotion lexicon. The advantage of the method is used four methods when constructing the similarity matrix; the final result is a weighted sum of the four methods, which improves the accuracy of the similarity matrix. The disadvantage is dependencies of the reference corpus.

Rohwer R et al. [11] proposed an automatic approach of creating lexicon by analyzing statistical data obtained from the corpus. According to the obtained statistical data, words are clustered. Calculating mutual information between the given words and words in the corpus, the most common words in the corpus is clustered by using information theory joint cluster method.

[1] `http://www.keenage.com/html/c_index.html`

The above methods are to construct dictionary for given words, the method proposed in this paper is automatically extracting words from the corpus, and to determine the emotional category of this extracted words.

3 The Emotion Corpus Ren-CECps

The corpus we used in this experience is Ren-CECps1.0. It is a Chinese blog emotion corpus with manual annotations for linguistic expressions of emotion. The frame of emotion annotation includes four levels (document, paragraph, sentence and word). For each level, Eight basic emotion classes (surprise, sorrow, love, joy, hate, expect, anxiety, and anger) are used in the emotional expression space model. Emotion of each level is represented by an 8-dimensional vector: where is a basic emotion class contained in the level, the values of range from 0.0 to 1.0 (discrete) [12]. The annotated files are organized into XML documents. And we can extract emotional keywords from them. The corpus now contains 1487 articles.

4 The Method

In this section, we present our method of automatically extracting words from Ren-CECps1.0 corpus, and classifying the emotional category of the extracted words. The method has two basic principles; one is that the emotion words mostly exist in some specific syntactic dependencies, the other is that the emotion words co-occurred in one sentence always has the same emotion category if the sentence doesn't contain any negative modifiers [13].

Based on these principles, we first divide the corpus into two parts, the first 1000 articles are used to extract the seed emotion words, the remaining articles are used to extract the candidate emotion words, and we select feature words from the seed words. Then we get the co-occurrence of the candidate words and the seed words from the corpus and compute the similarity based on the co-occurrence. All seed words and candidate words have characteristic value of each feature.

4.1 Extracting Candidate Emotional Words

To extract candidate emotional words, Stanford Parser is used to parse sentences. After parsing sentences, we observe the dependencies of words in sentences and find that the emotional words exist mostly in some specific syntactic dependencies. The dependencies include adverbial modifier (advmod), adjectival modifier (amod) and relative clause modifier (rcmod).

Stanford Parser is a statistical parser that works out the grammatical structure of sentences. The parser provides Stanford Dependencies output as well as phrase structure trees. Typed dependencies are otherwise known grammatical relations. The grammatical relations outputted by Stanford parser are arranged in a hierarchy, rooted with the most generic relation, dependencies. The hierarchy contains 48 grammatical relations and includes grammatical relations for NPs (amod – adjective modifier, rcmod - relative clause modifier, det - determiner, partmod - participial modifier, infmod - infinitival modifier, prep - prepositional modifier)[14].

After parsing, we choose the selected grammatical relations and extract the words that may be emotion words as the candidate emotional words in our method. For example, the Chinese segmentation sentence: "我 很 开心 ， 你 喜欢 这 份 礼物 。 (I am very happy that you like this gift.)" The dependencies parsed of this sentence are that "nsubj(开心-3, 我-1), advmod(开心-3, 很-2), conj(喜欢-5, 开心-3), dobj(开心-3,你-4), dep(开心-3, 喜欢-5), det(礼物-8, 这-6), clf(这-6, 份-7), dobj(喜欢-5, 礼物-8)". We extra "开心 (happy)" from advmod dependence and "礼物 (gift)" from det dependence as candidate emotional words.

4.2 Calculating the Similarity between Words

Our method of building an emotion lexicon is based on corpus. And the approach of computing two words' similarity is relied on the co-occurrence number of words.In the computing co-occurrence number procedure, we set each sentence as an independent window and find if there is co-occurrence of words in the window [15]. Finally we get the number of word pairs and the number of words appeared alone in the whole corpus.

The similarity calculation algorithm used in our method is Jaccard. The Jaccard similarity is a common index for binary variables and is computed as following formula.

$$\text{Jacsim}=\text{num}(A*B)/(\text{num}(A)+\text{num}(B)-\text{num}(A*B)) \tag{1}$$

In the formula, and is respectively the presence number of word A and B in the corpus, is the co-occurrence number of word A and word B in the corpus [16].

5 Experiments

In this section, we mainly introduce experiment preparation, evaluate our algorithm using the manual labeled words from the bottom 487 articles and summarize the experimental results.

5.1 Preprocessing

In order to be consistent with the emotional corpus' category, the lexicon we constructed includes eight basic emotion categories. The seed words are extracted from the top 1000 articles of the corpus. To select the feature words, we rank the seed words according to the words' number of occurrences and keep the top 15 words of each category as feature words. To extract candidate words, we parse the sentences in the segmentation articles using Stanford Parser. If the sentence doesn't contain the negative dependency, we extract the words in advmod, amod and rcmod dependencies as candidate words.

The process to obtain high quality emotional seed words is as follow:

Step1: We extract the emotional key words and their emotion vector.

Step2: We compute the average emotion vector of each word and get 6752 emotion words with emotion vector.

Step3: We select a threshold β. If vector of a word has a value more than β, we put the word to the category that the value more than β indicates and this category is the domain category of the word.

In order to select an appropriate threshold, we calculate the number of words that have one, two or three domain categories on different threshold. Experimental data is the 6752 words obtained on the previous step. The results are as follows:

Table 1. The number of emotion words on different threshold

β	N_1	N_2	N_3
0.4	4260	268	8
0.5	3153	91	0
0.6	1697	24	0
0.7	671	6	0
0.8	127	2	0

In the table, we find 0.6 and 0.7 can be selected as an appropriate threshold. Observing the emotion words and their vectors, we find some words have emotion intensity value just big than 0.6, and emotion category of the word is obvious. Such as "美满(happy) (0,0,0,0,0.67,0.31,0,0)", "贼(thief) (0,0,0,0.6,0,0,0,0)". So we choose 0.6 as the threshold, this can provide more data for subsequent experiments.

Step4: According to the intensity value, we rank words in each category and save about the preceding 50 words in each rank list as seed emotion words.

5.2 Experimental Settings

Our method uses SVM model [17] to predict the emotion category of candidate words. We calculate the similarity between seed words and feature words as feature values and constitute the training file. At the same time, we calculate the similarity between candidate words and feature words as feature values and form the testing file.

In order to examine the effectiveness of SVM classification method, we classify tagged words in the last 487 articles of the corpus and obtain a high precision. After examining, we predict the candidate words' category using our method. Then we delete the words that don't have co-occurrence words that the words' category is the same with the classification result.

5.3 Experimental Results

The Seed Emotion Words and the Feature Words

The seed emotion words and the feature words respectively contain 493 words and 120 words. Table 2 lists the seed emotion words; table 3 lists the feature words.

Table 2. The seed emotion words

Anger	气愤(indignant) 生气(angry) 争吵(quarrel) … (60 words)
Surprise	大吃一惊(surprise)难以置信(incredible)惊讶(confound)…(55 words)
Sorrow	惨剧(disaster)悲痛(grieved)悲伤(sad)绝望(despair)…(67 words)
Love	尽善尽美(perfect)热爱(love) 珍爱(cherish)… (66 words)
Joy	欢天喜地(joy)欢呼(cheer)高高兴兴(happy)…(67 words)
Hate	深恶痛绝(hate) 谩骂(diatribe) 厌恶(disgust) … (58 words)
Expect	期待(expect)希望(hope) 祝愿(wish) 追寻(pursue)… (57 words)
Anxiety	恐惧(fear)忐忑不安(uneasy) 烦躁(irritable)… (63 words)

Table 3. The feature words

Anger	争吵(quarrel)批评(criticism) 抱怨(complain)…(15 words)
Surprise	莫名其妙(surprise) 不可思议(incredible) … (15 words)
Sorrow	忧伤(sad)伤感(sorrow)绝望(despair)…(15 words)
Love	爱人(lovers)珍贵(precious)偏爱(preference)… (15 words)
Joy	快乐(happy) 高兴(joy) 幸运(lucky) … (15 words)
Hate	恨(hate) 贪婪(greedy) 无耻(shameless) … (15 words)
Expect	希望(hope) 期待(expect) 渴望(desire) … (15 words)
Anxiety	恐惧(fear) 担心(worry) 惊慌(panic) … (15 words)

The candidate words that we parsed from the sentences of the rest 487 segmentation articles using Stanford Parser. We extract the nouns, verbs and adjectives words in advmod, amod and rcmod dependencies from the sentences that don't contain the negative dependency as candidate words. After deleting the stop words and the words that the seed words list contained, we get 5,131 words as candidate emotion words.

Using the Manual Labeled Words from the Bottom 487 Articles to Test Our Classification Method

Table 4. The experimental results of labeled words in the last 487 articles

category	number of manual labeled words	number of correctly classified words	Accuracy
Anger	6	2	33.33%
Surprise	0	0	/
Sorrow	34	16	47.06%
Love	168	105	62.50%
Joy	25	16	64.00%
Hate	34	13	38.24%
Expect	11	4	36.36%
Anxiety	103	53	51.46%
total	381	209	54.86%

In this section, we use the manual labeled last 487 articles of the Ren-CECps corpus. The number of labeled words that have emotional intensity more than 0.6 in these articles is 664. We use the seed words' characteristic value file as the test file, and we get the similarity of the manual labeled words and the feature words from the remaining articles, organized the testing file. We find 381 words that have co-occurrence with the seed words. After classifying, we find 209 words have been classified correctly. Table 4 lists the number of each category, the number of correctly classified words and the accuracy rate of each category.

Results of Extracting the Candidate Emotional Words and Classification

Our method extracts candidate words from the unlabeled last 487 articles of the Ren-CECps corpus. Using the method described in 4.1, we get 5131 candidate emotional words. After calculating the words co-occurrence with feature words and their feature value, we finally get 762 words that have co-occurrence with feature words and classify the 762 words.

For the emotion words co-occurred in one sentence always have the same emotional category if the sentence doesn't contain any negative modifiers, we delete some candidate emotional words from the 762 words. If a candidate word's co-occurred words don't contain any word that the emotional category of the word is the same with the word's classification results. Finally we get 229 emotional words. Table 5 lists the emotional words got from the unlabeled last 487 articles of the Ren-CECps corpus.

Table 5. The emotional words got from the unlabeled last 487 articles

Category	words
Anger	错误(error) 生活(life) 折磨(torture) … (18 words)
Surprise	感动(moving) 大叫(shouted) 眼睛(eye) … (11 words)
Sorrow	错过(miss) 悲哀(sorrow) 痛(pain) … (36 words)
Love	宝贵(precious) 微笑(smiling) 旅途(journey) … (23 words)
Joy	享受(enjoy)浪漫(romantic)烦(bother) 自然(nature) … (48 words)
Hate	/
Expect	重生(rebirth)花朵(flower)梦想(dream)伤害(hurt)…(64 words)
Anxiety	后遗症(sequela) 灾难(disaster) 急(emergency) … (29 words)

5.4 Analysis of Experimental Results

In this section, we analyze the accuracy of results in chapter 5.3.3 and cause of words misclassification.

The Accuracy Analysis

We manual pick out the right words for each category. The accuracy is listed in table 6.

Table 6 shows the accuracy for each category. Correctly classified words are the words that all annotators mark the words are correctly classified. The accuracies vary from 36.36% to 50.00%, and the average value is more than 40%, this value is considered as good performance for an eight classification experiments. According to table 4, we find that the classification method can get a better performance than the result showed in table 6. The reason maybe we extract some neutral words or some words only have emotions in some context.

Table 6. The accuracy of our method

category	number of classified words	number of correctly classified words	Accuracy
Anger	18	9	50.00%
Surprise	11	4	36.36%
Sorrow	36	16	44.44%
Love	23	9	39.13%
Joy	48	20	41.67%
Hate	0	0	/
Expect	64	24	37.50%
Anxiety	29	11	37.93%
total	229	93	40.61%

Problem Analysis

The main reason is that we extract some neutral words as candidate words and these words have been classified, such as "生活 (life)", "眼睛 (eye)" and "自然 (nature)". This kind of words always simultaneously occurs with some modifying words and is extracted according to the grammatical relation.

Another important reason for words misclassification is that emotion category of some words is determined by the context. For example, "花朵 (flower)" is a neutral word. But in sentence "青少年是祖国的花朵 (Teenagers are flowers of the motherland) 。" The word "花朵 (flower)" is an emotion word that means expectation. In another sentence "花园里盛开了五颜六色的花朵 (Colorful flowers are in full bloom in the garden) 。" The word "花朵 (flower)" is an emotion word that express someone's love of flowers.

Besides, some emotion words are misclassified. For example, "伤害 (hurt)" is an emotion word in sorrow category, but we classified it into expect category; "烦 (bother)" is an emotion word in anger category, but we classified it into joy category.

6 Conclusion and Future Work

In this paper, we present a method of building a Chinese emotion lexicon; the process of our method includes word extraction and emotion classification by making use of the words co-occurrence in the corpus. The lexicon we built contains words and their emotional category. A major disadvantage of our method is that error classified word cannot be automatically removed. In future, we will do our efforts on this respect. In addition, we will continue to explore more information such as emotional intensity and one word with multiple categories to improve our lexicon.

Acknowledgements. This research has been partially supported by National Natural Science Foundation of China under Grant No.61203312 and the National High-Tech Research & Development Program of China 863 Program under Grant No. 2012AA011103 and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, and Key Science and Technology Program of Anhui Province, under Grant No. 1206c0805039.

References

1. Ren, F.J., Quan, C.Q., Matsumoto, K.: Enriching Mental Engineering. International Journal of Innovative Computing, Information and Control 9(8), 1–12 (2013)
2. Quan, C.Q., Ren, F.J.: Unsupervised Product Feature Extraction for Feature-oriented Opinion Determination. Information Sciences (2014), doi: http://dx.doi.org/10.1016/j.ins.2014.02.063
3. Ren, F.J., Quan, C.Q.: Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: an application of affective computing. Information Technology and Management 13(4), 321–332 (2012)
4. Quan, C.Q., Ren, F.J., He, T.T.: Recognition of Word Emotion State in Sentences. IEE J. Transactions on Electrical and Electronic Engineering 6(1), 35–41 (2011)
5. Stone, P.J., Hunt, E.B.: A computer approach to content analysis: studies using the general inquirer system. In: Proceedings of the Spring Joint Computer Conference, May 21-23, pp. 241–256. ACM (1963)
6. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 347–354. Association for Computational Linguistics (2005)
7. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: LREC 2010, vol. 10, pp. 2200–2204 (2010)
8. Li, J., Ren, F.: Creating a Chinese emotion lexicon based on corpus Ren-CECps. In: 2011 IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), pp. 80–84. IEEE (2011)
9. Liu, W., Zhu, Y., Li, C.: Research on building Chinese Basic Semantic Lexicon. Journal of Computer Applications 29(11), 2882–2884 (2009)
10. Ge, X., Meng, X.F., Wang, H.F.: Build Chinese emotion lexicons using a graph-based algorithm and multiple resources. In: Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics (2010)
11. Rohwer, R., Freitag, D.: Towards full automation of lexicon construction. In: Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics, pp. 9–16. Association for Computational Linguistics (2004)
12. Quan, C., Ren, F.: A blog emotion corpus for emotional expression analysis in Chinese. Computer Speech & Language 24(4), 726–749 (2010)

13. Zhu, Y.L., Min, J., Zhou, Y.: Semantic orientation computing based on HowNet. Journal of Chinese Information Processing 20(1), 14–20 (2006)
14. De Marneffe, M.C., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: Proceedings of LREC, vol. 6, pp. 449–454 (2006)
15. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, & Computers 28(2), 203–208 (1996)
16. Cheng, Y., Huang, H., Qiu, L.: Similarity-Based Dynamic Multi-Dimension Concept Mapping Algorithm. Minimicro Systems 27(6), 975 (2006)
17. Drucker, H., Wu, D., Vapnik, V.N.: Support vector machines for spam categorization. IEEE Transactions on Neural Networks 10(5), 1048–1054 (1999)