Available online at www.sciencedirect.com

SciVerse ScienceDirect

Physics Procedia 22 (2011) 597 – 603

Physics Procedia

2011 International Conference on Physics Science and Technology (ICPST 2011)

Chinese Subjective Sentence Extraction Based on Dictionary and Combination Classifiers

Wei Chen[a], Yanquan Zhou[a], Xin Wang[a]

*a * School of Computer Science and Technology, Beijing University of Posts and Telecommunications, Beijing , China*

Abstract

For extracting of Chinese subjective sentence, this paper proposes a new dictionary-based extraction method and a novel classifier combination strategy. For the first method, we use the training data to score the subjective dictionary, which was composed of indicative verb, indicative adverbs, sentiment words, interjection and punctuation. Then we use the dictionary to score the test data, and filter the sentences by setting a reasonable threshold. New classifier combination strategies base on the maximum error correction capability. To enhance the accuracy, the method improves the traditional single error correction and achieves the dual error correction both in positive and negative classes. Experimental results show that the two methods are effective .And the final results show that the combination of two ways achieves a satisfactory subjective sentence extraction performance.

Keywords: Subjective Sentence Extraction; Subjective Dictionary; Sentiment analysis; Combination Classifiers； Support Vector Machine; Maximum Entropy;

1. Introduction

In recent years, with the increasing amount of network information, huge redundancy information makes it difficult to quickly and accurately obtain valuable information in a short time. However, subjective sentence extraction intends to make a distinction between subjective comment sentences and objective description sentences. Its importance is mainly reflected in two aspects. On the one hand, it needs to distinguish the subjectivity and objectivity text. It can not only help users quickly to search for evaluation information of product, but also facilitate the product market survey for the manufactures in a timely manner. On the other hand, subjective sentence extraction is the prerequisite and basis for the sentiment polarity classification of text, opinion holder extraction and other research. Such as the sentiment polarity analysis of text, the first work is to extract subjective sentences.

Rest of the article is organized as follows: the Section II introduces the overall relevant work of subjective sentence extraction area. Section III describes the building process of subjective dictionary and the method of sentence scoring and threshold selection. Section IV describes a new strategy of multiple classifiers combination based on maximum error correction. Section V presents the experiments and results analysis. Section VI makes a summary and prospects for future research.

* Corresponding author. Tel.:+86-010-62283459; fax:86-010-62283459.
E-mail address: cw3954826@bupt.edu.cn.

 doi:10.1016/j.phpro.2011.11.092

2. Related Work

Chinese subjective sentence extraction as a relatively new field of study, still in the exploratory stage, the current classification is still relatively simple algorithm. In the field of Chinese studies, Tianfang Yao and Siwei Peng (2007)[1] proposed some words and punctuation as features, like personal pronoun, interjection, inaccurate numbers and dates, opinionated verb, non-standard punctuation and punctuation with emotional color, to compare classification performance of Conjunctive Rule, Naïve Bayes, SMO, Id3 algorithms based on experimental data in specific areas. Qiang Ye et al(2007)[2] proposed a method to automatically determine the subjectivity strength of Chinese sentences using the combination patterns of continuous two parts of speech.

Most of the above methods do not consider the use of combination classifiers and dictionary-based sentence extraction strategy. This paper presents a new dictionary-based extraction method and a novel classifier combination strategy, which used different classifiers to achieve complementarily between classifiers. And the method has been used to achieve the extraction of subjective sentences.

3. Building subjective dictionary and scoring sentences algorithm

In this paper, by artificial way, we have summarized the four categories of words with subjective tendencies, including: indicative verb, indicative adverbs, sentiment words, interjections and punctuation of 9670 words. The subjectivity dictionary is established Table1.

Table 1. Type information of words in subjectivity dictionary:

word category	Definition	sample
Indicative verb	That express subjective views	think (认为)
Indicative adverb	one for adverbs of degree, or the other for adverbs pointing with mood or attitude	very(非常), precisely(恰恰)
Sentiment word	adjectives with some kind of emotional tendencies	Beautiful(美丽的), ugly(丑陋的)
Interjection and punctuation	That express particular feelings	Ah(啊), '？ ', '!'

Table 2. Weights of word category

word category	Weight
Indicative verb	2
Indicative adverb	0.5
Sentiment word	1
Interjection and punctuation	1

Although these words have a tendency to subjectivity, but tend to differ on the degree. For example, the indicative verb "think(认为)" than indicative adverb "very(很)" to better performance of the sentence subjectivity. Therefore, in order to quantify the subjective nature of these words, we set these words different weights. Weights are set as Table 2.

And considering the words for the same class maybe have different degrees in orientation, we have these words for statistical score based on the training corpus. Scoring formula is set as follows:

$$t_i = (subnum_i - objnum_i) / num_i \quad (1)$$

In the formula, t_i is the score of i-th word in dictionary. $subnum_i$ is the occurrence number of this word in the subjective corpus. $objnum_i$ is the occurrence number of this word in the objective corpus. $num_i = subnum_i + objnum_i$, num_i is the total occurrence number of this word in the entire train corpus. Statistics show that less than 5% of the scores are negative. In order to establish the subjective dictionary, so the words which have negative score will be removed. Thus the construction of subjectivity dictionary has been completed.

Then, the subjectivity dictionary is used to score the test sentences. Scoring formula:

$$V_k = \sum_{i=1}^{m} \omega_i \frac{subnum_i - objnum_i}{num_i} - P \quad (2)$$

In the formula, V_k is the score of k-th sentence in test corpus. If $V_k > 0$, this sentence is judged as subjective sentence. If $V_k \le 0$, the sentence will not be marked. $i = 1...m$ represents each word of the sentence. ω_i is the weight of word category

for i-th word. P is the manually set threshold. We need to select a reasonable threshold, to ensure high precision and recall rates.

The selection of threshold needs to use the subjective dictionary to score the training data, and set different thresholds for selecting the best threshold. The best threshold of training data is classified as the final threshold of test corpus. Specific experiments in section 5.2.

4. A new strategy of multiple classifiers combination based on Maximum error correction

4.1. Feature Selection

Reference the subjective and objective sentences; we think both of them imply their own unique characteristics on semantic and grammatical levels. So, this article refers to Bo zhang [3], who proposed the objective and subjective feature selection methods. For the classifiers, we select the following night candidate features in Table 3.

Table 3. Candidate features

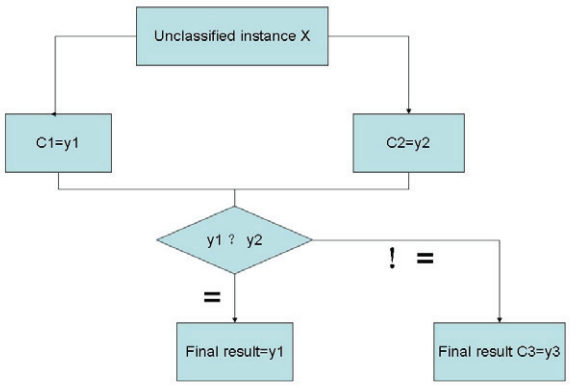
Candidate Feature	F1	F2	F3	F4	F5	F6	F7	F8	F9
Information	Sentiment word	Indicative verb	Indicative adverb	Interjection& Punctuation	1-POS	2-POS	3-POS	1-Word	2-Word

N-POS means a combination of sequence of N continuous parts of speech. N-Word means a combination of sequence of N consecutive words. Ultimately, we identified 745-dimensional features as the final standard features.

4.2. Classifier combination strategy based on the single Maximum error correction

Different types of classifiers can provide complementary information of processed objects from different perspective. Therefore, integrating the output of the classifiers in some way for "complementary" may make classification accuracy improved. So the combination system will have better performance than single classifier. The traditional combination methods such as voting method can not use the complementary information between classifications well, so results are generally. This paper draws a classifier combination strategy based on the maximum error correction principle [4]. We improve this strategy and make a new combination method. Finally, using this method, we extract Chinese opinion sentence. In this article, we treat the subjective sentence as positive class instance and the objective sentence as negative class instance.

Maximum error correction principle means classifier C2 can best correct the error of classifier C1. Here, it represents C2 can best classify the case which is misclassified by C1. This classifier C2 is identified as the classification which has the strongest error-correcting capability to C1. This makes the C1 has a strong complementary to C2. This strategy can greatly enhance the effect of C1. So the overall classification accuracy will be improved. When the classification results of C1 and C2 are inconsistent, we need to choose an arbitration classifier [8] to determine the final result. The role of arbiter is to determine which category the case should belong to. Therefore, the requirement for the arbiter is also the maximum correction principle. There is a traditional combination strategy which we called single maximum error correction (SMEC). It is shown in Fig 1.



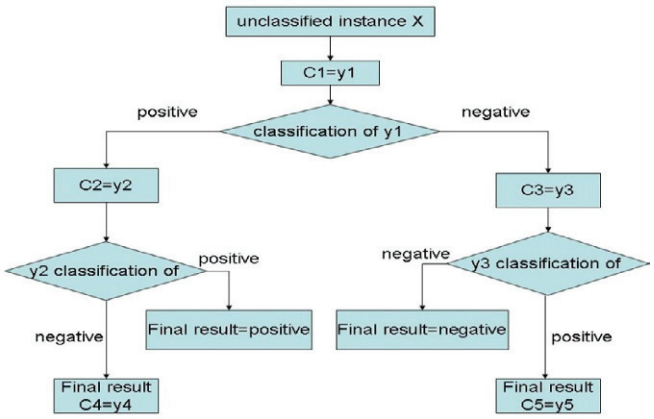


Fig.1. Combination with single Maximum error correction（SMEC） Fig.2. Combination with dual Maximum error correction（DMEC）

The combination based on the single maximum error correction is good in some cases. However, in the classification problem for subjective and objective sentences, the two types of instances often have different distribution characteristics and distribution curve. This leads to the classification performance of each classification for both types is uneven. There are some classifications whose classification performance for one type instances is far better than their effect for the other. This has resulted in poor overall effect. So, to correct the positive and negative results of C1, we choose classifiers respectively in accordance with the maximum error correction principle. Then, the effect will be better. We call the new classifier combination strategy as the dual maximum error correction (DMEC).

4.3. Classifier combination strategy based on the dual Maximum error correction

First, we select the classifier, which has the best performance for both positive and negative instances in all classifiers, as the first classifier C_1. After that, for the negative instances which were misclassified as positive class by C1, we select the classifier with maximum error correction as C2. And, for the positive instances which were misclassified as negative class by C1, we select the classifier with maximum error correction as C3. Note that, c1 and c2 may be the same classifier.

When the classification results of C1, C2 or C1, C3 appear inconsistent, we need to use the arbiter to make the final decision. Still using the maximum error correction principle, we select two classifiers C4 and C5 as the arbiter for C1, C2 and C1, C3. Note that, c3 and c4 may be the same classifier. C2 and c5 may be the same classifier. New classifiers combination based on this algorithm is in Fig 2.

5. Experiment and Analysis

5.1. Data Set

The corpus data in this experiment is from the Singapore's ‘Lianhe Zaobao’. It covers many topics such as music, economic, sports, movies and so on. For each topic, we collect some subjective sentences and some objective sentences. The standard for determining subjective sentence is: If the sentence expresses a certain view, expectation, projection, evaluation, or emotion tendency, whether it is published by first person or third person, we will judge the sentence to be a subjective sentence. After manual annotation, we select 2600 subjective sentences and 2600 objective sentences, some of which are randomly selected to form the test data with 500 subjective ones and 500 objective ones. In the end, the training data contains 4200 sentences and the test data T contains 1000 sentences.

5.2. Dictionary-based subjective Sentences Extraction Experiment

First, the subjective dictionary is established from training data by the method in section3. Second, we score each word in the dictionary with the formula (1). Then, we use the dictionary to score the training data with the formula (2). For choosing the best threshold, we set different thresholds and record the results as Table 4. Because we hope to get the precision as high as possible. So, we want the precision is higher than 90%. As Table 4 indicates, the threshold 3.0 achieves the best result. Therefore, we select this threshold as the ultimate threshold for test data T.

Table 4. Performance of different thresholds Table 5. Performance of the method in test data

Threshold	Precision	Recall	F-measure
0.0	53.08%	99.42%	69.21%
1.0	73.29%	91.09%	81.23%
2.0	87.62%	71.14%	78.52%
3.0	***93.59%***	***50.09%***	***65.26%***
4.0	96.58%	33.66%	49.92%
5.0	98.29%	21.95%	35.88%
6.0	99.66%	14.33%	25.06%

Threshold	Precision	Recall	F-measure
3.0	98.5%	53.2%	69.1%

Now, we use the subjective dictionary and the ultimate threshold 3.0 to score the test data T with the formula (2). The result is as Table 5.

It can be seen that the performance in test data is of high confidence. The result is as same as the statistic data in Table 4. This method achieves 98.5% precision with 53.2% recall. It can be seen that the method really is a high-precision approach and still effective in the test data.

Through the above experiments, we find that this extraction method based on subjective dictionary can get a high accuracy. But the recall rate is low. This method can not achieve complete division for corpus. Next, we will use the classifiers combination method described in Section 4 to classify the sentences whose score is below the threshold.

5.3. Multiple Classifiers Combination Experiments

We utilize the ICTCLAS [5] system, which was developed by Chinese Academy of Sciences, to segment and make POS tagging on all experiment data at first. Here, we also select the features mentioned in section 4.1 as the ultimate features. In addition, we selected the following five representative classification algorithms as the candidate classifiers for our experiment：Support Vector Machine（SVM）； Maximum Entropy（ME）； Naive Bayes（NB）;k-nearest neighbor classification（KNN）； Decision Tree Algorithm(C4.5). Maximum Entropy classification was achieved by the toolkit from Dr. Zhang Yue【6】. Other classification algorithms were achieved by corresponding modules of weka【7】.

According to the classifier combination algorithm based on DMEC, 1200 sentences are randomly selected from 4200 training data to form the test set S. The remaining 3000 sentences form the training set D. Table 6 shows the results of each classifier.

Table 6. Performance of each classifier

Table 7. Performance of each method in test data T

Classifier	Sub_F	Obj_F	Ave_F
SVM	**66.3%**	**65.8%**	**66.1%**
ME	66.1%	62.0%	64.1%
NB	53.2%	66.7%	59.9%
KNN	50.5%	62.8%	56.6%
C4.5	60.9%	59.8%	60.3%

method	Sub_F	Obj_F	Ave_F
SVM	82.7%	83.1%	82.9%
ME	82.1%	80.3%	81.2%
NB	75.1%	79.3%	77.2%
KNN	63.1%	71.1%	67.1%
C4.5	79.4%	80.2%	79.8%
Voting	85.3%	86.2%	85.7%
SMEC	87.9%	87.6%	87.7%
DMEC	**89.9%**	**89.8%**	**89.8%**

The 'Sub_F' means the F-measure for subjective sentences. The 'Obj_F' means the F-measure for objective sentences. From the performance, we select SVM as the optimum classifier C_1. Then, According to the algorithm in section 4.3, we get KNN as classifier C_2, ME as classifier C_3, ME as classifier C_4, C4.5 as classifier C_5. So, we get the combination system.

To compare the effect，we also make the combination system based on SMEC in section 4.2. So, we select SVM as classifier C_1, ME as classifier C_2, C4.5 as classifier C_3.

We compare the above two combination methods, voting method with each classifier methods. We do the experiments on the 1000 test sentences T. Table 7 shows the results of each method.

We can observe that the combination system based on DMEC achieves wonderful performance. This method uses the complementary information of different classifiers better. So, it has better balance and stability.

5.4. Combination of Dictionary-based Sentences Extraction and DMEC System

In section 5.2, we find the dictionary-based sentences extraction method (**DSE**) obtains high precision. So we consider combining it with combination system based on DMEC. At first, we use the dictionary-based method to extract sentence whose score is higher than threshold 3.0 as subjective sentence and then put the remaining test data into the classifiers system based on DMEC. Performance of combination approach is showed in Table 8.

Table 8. Performance of combination approach

Combination	Precision	Recall	F-measure
DSE	93.59%	50.09%	65.26%
DMEC	89.2%	90.8%	89.9%
DSE + DMEC	**89.1%**	**93.4%**	**91.2%**

As expected, the combination of DSE&SVM achieves the best performance which precision reaches 89.1%, recall is up to 93.4%. Fig 3 gives a clearer show of results mentioned above.

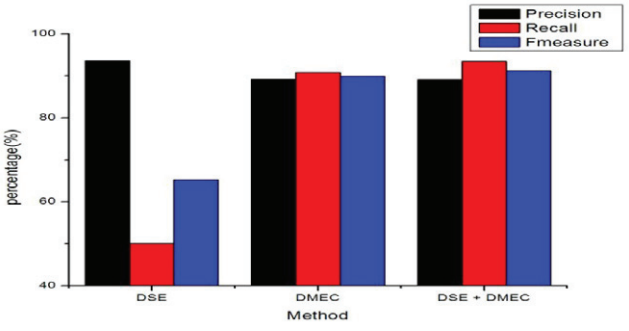


Fig.3. Performance of subjective sentence extraction

6. Conclusion and Outlook

This paper proposes a new dictionary-based sentence extraction method and a new classifier combination method based on the dual Maximum error correction. Through detailed experiments, we verify the availability of two methods. Eventually, the combined method of DSE and DMEC performed well, achieving 91.2% F-measure.

For future work, as the semantic information and syntactic structure have deeper meaning for Chinese information processing. We plan to propose a combined method of syntactic structure, semantic information and the classifiers to extract subjective sentences. To make the current system has a better performance.

References

[1]Tianfang Yao, Siwei Peng, "A Study of the Classification Approach for Chinese Subjective and Objective Texts", Third Conference on National Information Retrieval and Content Security Academic, pp.117-123, 2007.

[2]Qiang Ye, Ziqiong Zhang, Rob Law, "Automatically Measuring Subjectivity of Chinese Sentences of Sentiment Analysis to Reviews on the Internet", China Journal of Information Systems, 1(1)：pp.79-91, 2007.

[3]Bo Zhang，Yanquan Zhou，Yu Mao，"Extracting opinion sentence by combination of SVM and syntactic templates"， Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010）.

[4]CAI Xi, GUO Gong-de, HUANG Tian-qiang. Multiple classifiers fusion method based on semi-supervised learning. Computer Engineering and Applications, 2009, 45 (25) :218-221.

[5]http://ictclas.org/

[6]http://ir.hit.edu.cn/~taozi/ME.htm

[7]http://www.cs.waikato.ac.nz/ml/weka/