PROCEEDINGS OF SCIENCE

Explore the Application of S-HAL in the Identification of the Sentiment Polarity of Chinese Words

Tao Xu[12]
Hangzhou Dianzi University
Hangzhou, 310018, China
E-mail: txu@mail.xjtu.edu.cn

The task of identifying the sentiment polarity of terms is to decide whether an individual term is subjective or objective, and to classify a subjective term as positive or negative. Accurate identification of sentiment polarity of terms is of vital importance in sentiment analysis, as well as a challenging task. As a specific semantic subspace, S-HAL (Sentiment Hyperspace Analogue to Language) shows the advantages in modelling and distinguishing semantic orientation characteristics of terms. In this article, we present a new method to identify the sentiment polarity of Chinese words, which is based on S-HAL model and standard supervised learner. A series of empirical evaluation results demonstrate that S-HAL-based identification method could outperform the known method and the way of combining multiple classifiers can balance the identification performance between subjective and objective. Similar to the way for building SentiWordNet from WordNet, the method presented in this article provides a solid technical basis for construction of Chinese SentiHowNet based on HowNet.

ISCC 2017
16-17 December 2017
Guangzhou, China

[1]Speaker
[2]This work is supported by the National Natural Science Foundation of China (Grant Numbers 61402142)

© *Copyright owned by the author(s) under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).*

PoS(ISCC 2017)027

1.Introduction

Since being proposed by Dave et al in WWW'2003, sentiment analysis has received growing attention in the past decade [1]. A fundamental task in sentiment analysis is sentiment detection and polarity identification for terms. It involves identifying whether a given term is subjective or objective, and classifyingthe sentiment polarity of a subjective term as positive or negative [2-3]. To simplify the issue, this task can be treated as a ternary categorization problem under categories Positive, Negative and Objective [3]. Sentiment detection and polarity identification for terms is beneficial for finishing some other key tasks in sentiment analysis, such as identification of the sentiment polarity of a paragraph, computation of the strength of the sentiment polarity of a paragraph and extraction of the opinion of a whole text. The existing researches into sentiment detection and polarity identification of terms can be divided into two categories: constructing for sentiment lexicon and developing for automatic identification. The former refers to development on lexicons manually annotated by experts [4-5]. The lexicons possess high accuracy but limited coverage. Automatic identification of sentiment polarity of terms generally adopts learning methods to generate the sentiment polarity for a given term [3, 6-9]. Automatic identification methods are able to cover a wider vocabulary, and therefore can meet more applications. However, many automatic identification approaches are limited by low computation speed and relatively low accuracy rate of identification.

The most effective automatic identification method under the three categories (Positive, Negative and Objective) is proposed by Esuli and Sebastiani (referred to below as the "E-S method") [3]. It trains a ternary classifier to detect the sentiment polarity of a given term by using machine learning methods and online semantic materials. The E-S method possess considerable coverage and relatively high accuracy (nearly 70%), thus it was utilized as a technical basis for building the SentiWordNet [6] that is the most famous lexical resource for sentiment analysis research and application in English language.

Publicly available online glossary on internet is crucial for generating the semantic representation of any given term in E-S method, and therefore E-S method's effectiveness in Non-English language (such as Chinese, Japanese, Spanish and etc.) may be affected because there exist relatively less electronic glossary resources. In this study, we aim at developing an accurate method to automatically and rapidly identify sentiment polarity of Chinese words without the online support via internet. According to the task requirement, a pipeline method is proposed based on the previous work [10]. With the proposed method, we first built a Chinese S-HAL that is a sentiment polarity representation model. In S-HAL, the sentiment polarity feature of Chinese words is captured via a carefully designed numeric vector space, and the sentiment polarity feature vector of any given word can be acquired by querying S-HAL. Next, from the obtained sentiment polarity feature vector, several classifiers with different strategies and a finally assembed classifier are trained to determine the sentiment polarity of Chinese words.

The next parts of this article can be summarized as follows. The definition and construction procedure on S-HAL model are introduced in section 2, which is the fundation of the present work. Section 3 presents a pipeline method for automatic sentiment polarity identification on the basis of the S-HAL model. Section 4 contains the experiment design and result analysis, and Section 5 concludes this paper.

2. Related Work

This section briefly describes the definition and construction procedure of S-HAL, which

PoS(ISCC 2017)027

is an important foundation for the proposed work. S-HAL is indicated as a specific semantic space model used for modeling sentiment polarity feature in the previous work [10]. The construction procedure of S-HAL can be described as follows.

Consider a set of words with definite sentiment polarity, denoted by S. A sliding window of length $2K$-1 is moved across a massive text corpus at one-word increments. All words $\underline{w_{i-K+1}, w_{i-K+2}, \cdots w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, \cdots w_{i+K-1}}$ in sliding window are considered as co-occurring with the target word w_i. For any word pair $<w_i, w_{i+j}>$, $j \in \{-K+1, -K+2, .., -2, -1, 1, 2, .., K-1\}$, if $w_{i+j} \in S$, the co-occurrence weight between w_i and w_{i+j}, denoted by $n(w_i, w_{i+j})$, is calculated by using equation 2.1.

$$n(w_i, w_{i+j}) = K - |j| \tag{2.1}$$

After moving the window over the whole corpus, an accumulated co-occurrence weight matrix is produced, which is S-HAL space. The resulted S-HAL space is an $N \times |S|$ matrix, as shown in equation 2.2, where $|S|$ denotes the size of S and N is the target vocabulary size. The sentiment polarity of word t_i can be represented by row vector in S-HAL.

$$S\text{-}HAL = \begin{vmatrix} & s_1 & s_2 & \cdots & s_{|S|} \\ t_1 & w_{t_1,s_1} & w_{t_1,s_2} & \cdots & w_{t_1,s_{|S|}} \\ t_2 & w_{t_2,s_1} & \ddots & \iddots & \vdots \\ \vdots & \vdots & \iddots & \ddots & \vdots \\ t_N & w_{t_N,s_1} & \cdots & \cdots & w_{t_N,s_{|S|}} \end{vmatrix}_{N\times|S|} \tag{2.2}$$

3. Proposed Method

This section proposes a pipeline approach for identifying the sentiment polarity of Chinese words, which is based on classfying sentiment polarity vectors derived from S-HAL. The overall framework of method is shown in Figure 1.

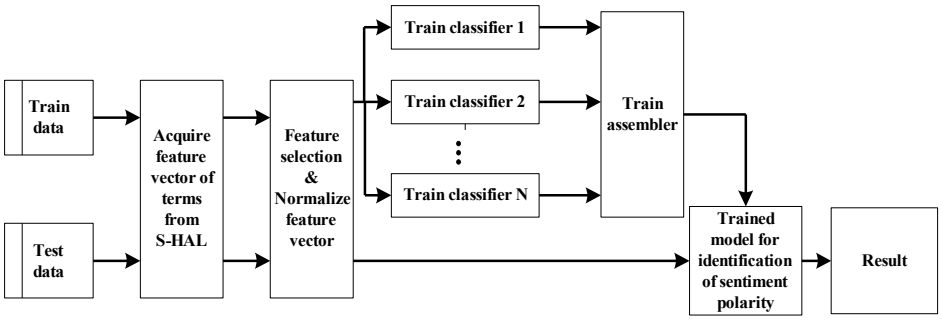

Figure 1: The framework of a pipeline method for sentiment polarity identification

3.1 Step 1: Acquiring Sentiment Polarity Feature Vector

After the construction of S-HAL is finished, this step follows. According to section 2, sentiment polarity feature information of each targeted word is captured by a numeric vector characterized by S. Therefore, the sentiment polarity feature vector of word t_i in training and testing dataset could be acquired from S-HAL, which is denoted by $[w_{t_i a_1}, w_{t_i a_2}, \ldots w_{t_i a_{|S|}}]$, where $a_j \in S$.

3.2 Step 2: Performing Feature Selection and Normalization

After acquiring the sentiment polarity feature vectors from S-HAL, feature selection can be used to eliminate the effects caused by noise. Feature selection algorithm based on

PoS(ISCC 2017)027

information gain (IG) is used to filter discriminating features from original sentiment polarity feature set during the process. As an effctive and efficient method of feature selection, IG-based feature selection can be defined as follows:

$$\hat{S} = \{a_i \mid IG(a_i) > \delta, a_i \quad S\} \tag{3.1}$$

$$\begin{aligned} IG(a_i) = &-\sum_{k=1}^{|C|} P(c_k)\log P(c_k) + P(a_i)\sum_{k=1}^{|C|} P(c_k \mid a_i)\log P(c_k \mid a_i) \\ &+ P(\overline{a}_i)\sum_{k=1}^{|C|} P(c_k \mid \overline{a}_i)\log P(c_k \mid \overline{a}_i) \end{aligned} \tag{3.2}$$

where $\hat{S}$ denotes the resulting feature subset, $\delta > 0$ denotes the selection threshold, C = {*Negative*, *Non-polarity*, *Positive*} is the categories set, $P(c_k)$ is the probability which category c_k appears, $P(a_i)$ is the probability which feature a_i appears, and $P(\overline{a}_i)$ is the probability which feature a_i does not appear.

After the feature selection, a more efficient and refined feature subset representing sentiment polarity is obtained. In order to make following classification algorithm work better, it is necessary to reweight and normalize the sentiment polarity feature vector. In this research work, equation 3.3 and 3.4 are adopted to weight and normalize feature vector, separately.

$$w_{t_i a_j} = w_{t_i a_j} * \log \frac{N_{vector}}{vf(a_j)} \tag{3.3}$$

$$w_{t_i a_j} = \frac{w_{t_i a_j}}{(\sum_{l=1}^{\hat{N}} (w_{t_i a_l})^2)^{\frac{1}{2}}} \tag{3.4}$$

where $a_j \in \hat{S}$, N_{vector} is the total number of vectors, $vf(a_j)$ is the number of vectors containing feature a_j, and $\hat{N}$ is the capacity of resulting feature subset $\hat{S}$.

3.3 Step 3: Training A Group of Classifiers for Sentiment Polarity Identification

After Step 1 and Step 2, each word of training set will be encoded by a normalized numeric vector. Next, encoded vectors could be sent to some standard classification learning algorithms to generate four ternary classifiers under the categories *Positive*, *Negative* and *Non-polarity* in this step. The first classifier is obtained by directly training a ternary under the categories *Positive*, *Negative* and *Non-polarity*. The second classifier is trained by combining two binary classifiers including the classifier (under the categories *Polarity* and *Non-polarity*) and the classifier (under the categories *Positive* and *Negative*). The third classifier is trained by combining two binary classifiers including the classifier (under the categories *Negative* and *Non-negative*) and the classifier (under the *Positive* and *Non-polarity*). The fourth classifier is trained by combining two binary classifiers including the classifier (under the categories *Positive* and *Non-positive*) and the classifier (under the categories *Negative* and *Non-polarity*).

3.4 Step 4: Combining All Classifiers to Generate A Final Identification Model

After the completion of Step 3, a total of four ternary classifiers are yielded at last. In this step, these ternary classifiers are combined to produce a final identification model that outputs a score for estimating the strength of sentiment polarity of terms. Specifically, the classification results of ternary classifiers is first encoded according to 1 (denotes *Positive* category), 0 (denotes *Non-polarity* category) and -1 (denotes *Negative* category), and then a neural network is trained to estimate the strength of sentiment polarity of terms as an assembler. In the design

PoS(ISCC 2017)027

approach of this paper, the output of neural network, denoted by $sentiScore(\ )$, is a continuous real number, therefore we define a categorization threshold λ, $0<\lambda<1$, to identify the sentiment polarity of terms finally. Given a term t_i, we have

If $sentiScore(t_i)\geqslant\lambda$, then t_i is *Positive*;

If $sentiScore(t_i)\leqslant-\lambda$, then t_i is *Negative*;

If $-\lambda < sentiScore(t_i) < \lambda$, then t_i is *Non-polarity.*

4. Experiments

In this section, two different experiments are concluded to evaluate the proposed method. In the first experiment, we compared the proposed method with other well-known methods under the same experimental conditions. The second experiment analyzed the effect of feature selection on final Chinese words' sentiment polarity identification.

4.1 Data sets and Evaluations

4.1.1 Corpora and lexicons

To construct the Chinese S-HAL model, the Chinese Sogou CS corpus were employed (the Sogou CS corpus is available at http://www.sogou.com/labs/). The SogouCS corpus contains a total of 2,820,059 pages and approximately 530 million words after removing all stop words and numerical symbols. After further removing all infrequent words that occurred less than 40 times in the corpus, the SogouCS corpus contains a total of 116,233 distinct words, denoted by *V*. In experiment, these,233 distinct words were used as the target vocabulary for S-HAL model construction.

For training and testing supervised learning classifier, we used two lexicons that were hand-labeled with a sentiment polarity (***Positive*** or ***Negative***): *H lexicon* and *T lexicon*. The *H lexicon* is a list of words with definite sentiment polarity, released by Hownet (the *H* lexicon is available at www.keenage.com/html/c_index.html) [11]; the *T lexicon* is a list of words with definite sentiment polarity created by Li (the *T lexicon* is available at http://nlp.csai.tsinghua.edu.cn/site/index.php?page=resources) [12]. There are 2,074 words that overlap between the *H lexicon* and the *T lexicon*. So, a total of 10,482 distinct words after *H lexicon* and *T lexicon* were merged into a larger lexicon. In the experiments, we define the larger lexicon as the $H+T$ *lexicon*, and use it as the initial base-space of S-HAL model. Correspondingly, we define the list of words belong *V* but not belong $H+T$ as $\overline{H+T}$ *lexicon*.

4.1.2 Experimental Dataset Construction

Based on the lexicons described above, we built the experimental dataset below and used to train and test the supervised classifier for identification of sentiment polarity of Chinese words. Experimental dataset consists of the *H+T lexicon* and a set of 20,000 words were randomly selected from $\overline{H+T}$ *lexicon*. A five-fold cross-validation were carried out on all 30,482 words in the experimental dataset.

4.1.3 Evaluation Measures

To evaluate and analyze the performance of present method, the standard *Accuracy* and *F-measure* were used and we aggregated the *F-measure* scores over three categories by using the Macro- and Micro-averages of the F-measure scores. The details of the evaluation measures can be refered to [10].

PoS(ISCC 2017)027

4.2 Experimental Results

4.2.1 Experiment 1: Comparison of Proposed Method Against E-S Method

Experiments 1 is conducted to analyze and assess the validity of the proposed method for identifying Chinese words' sentiment polarity. For the purpose of comparison, the E-S method presented in [3] was implemented to classify the same data sets.

While implementing the proposed method, we adopted the baseline parameter configuration. Specifically, the *SogouCS* corpus was adopted for training S-HAL; the target vocabulary of S-HAL was the set V defined in Section 4.1; the list of 10,482 words contained in the *H+T lexicon* was used as the initial base-space of S-HAL; the sliding window for training S-HAL had a length of 10 words; the feature selection step for sentiment polarity feature vector was skipped. For the reasons of overcoming impact of random factors in process of randomly selecting a set of 20,000 words from $\overline{H+T}$ *lexicon*, the experiment was repeated 10 times, and the average results were computed.

Positive		**Negative**	
grateful	splendid	distorted	wrong
energetic	brilliant	corrupted	agonizing
mellow	virtuous	psychopathic	poor
remarkable	beautiful	morose	negative
excellent	honest	dishonest	nasty
positive	fortunate	hidebound	tragic
harmonious	polite	rude	horrible
comfortable	nice	unfortunate	stupid
lenient	peaceful	inferior	unhealthy
correct	superior	silly	shameful

Table 1: Seeds set used in the E-S method

The E-S method was implemented with the seeds set in Table 1 which is based on the work in [13], and 5-iterations antonym and synonym expansion were conducted. The final classifier training used the SVM classification algorithm in libsvm package.

Method	**Accuracy**		**Macro F-measure**		**Micro F-measure**	
	Avg (10 times)	**Best**	**Avg (10 times)**	**Best**	**Avg (10 times)**	**Best**
Proposed Method	0.754	0.766	0.731	0.745	0.770	0.781
E-S Method [3]	0.629	0.685	0.605	0.644	0.641	0.699

Table 2: Accuracy, Macro F-measure, and Micro F-measure of the proposed method vs. E-S method

Table 2 shows the results for our method and E-S method across the same experimental dataset. Comparison on the results indicates that identification effectiveness of the proposed method significantly outperforms the E-S method. Similar to the E-S method, our approach trains a group of supervised classifiers to identify sentiment polarity of terms based on the feature vector representation of words. However, themethod in this paper achieves significant performance improvement due to the use of co-occurrence information in the large corpus. Therefore, co-occurrence information with definite sentiment words has a stronger semantic functionality.

4.2.2 Experiment 2: Effect of the Number of Dimensions of the Sentiment Polarity Feature Vector

In this section, we explore the effect of varying the size of definite sentiment words set

PoS(ISCC 2017)027

used for S-HAL base space. A toal of 13 different sentiment words set sizes were evaluated by carefully manually setting IG thresholds. The other experimental setup is the baseline configuration introduced in Section 4.2.1. To overcome the impact of random factors as same as experiment 1, the same experiment was conducted 10 times for each selected feature subset, and the average results as final results were delivered.

Figure 2 plots the accuracy, Macro-F1 and Micro-F1 curves with varying the size of selected feature subset. The figure shows that feature selection in S-HAL construction procedure can effectively improve the accuracy of final identification of sentiment polarity. Specifically, if the $H+T$ *lexicon* containing 10,482 features is used as the initial complete attribute set, a subset about 6,000 most discriminate features extracted by IG-based feature selection can lead to the highest identification accuracy for Chinese words' sentiment polarity. Macro-F1 and Micro-F1 measures also reveal the similar nature with the accuracy measure. We believe that this is because of the enormous redundancy and noise existing in $H+T$ *lexicon*. So, feature selection from S-HAL base space is an essential part for improving the effectiveness of the proposed method.

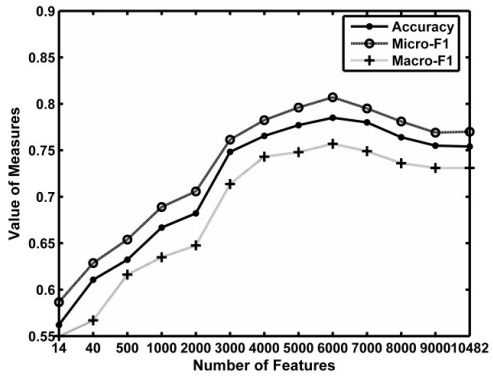

Figure 2: Effect of varying the number of dimensions of the sentiment polarity feature vector

5. Conclusions

In this paper, we present a novel idetification method for sentiment polarity contained in Chinese words. On the same Chinese language test bed, our method outperformed the famous E-S method that was published in [3]. E-S method is proposed for English language and is difficult to be perfectly transplanted into Chinese language without effectiveness loss. In addition, the E-S method needs the online support of internet when constructing semantic feature vectors of words, so the identification speed is limited and it is difficult to use big data processing. The proposed method works better than the E-S method in Chinese language and is faster without need for online support of internet, therefore this method will greatly facilitate the use of various applications of emotional analysis as a foundational tool. Moreover, in English corpora resource, SentiWordNet greatly facilitates the research on sentiment analysis and effective computing. Nevertheless, there is still a lack of a similar resource in Chinese field. Our work presented in this paper shall provide a solid technical basis for construction of Chinese SentiWordNet.

References

[1] K. Dave, S. Lawrence and D. M. Pennock. *Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews [C]*. In Proceedings of the 12th International Conference on World Wide Web, Budapest. 519-528(2003).

PoS(ISCC 2017)027

[2] A. Esuli and F. Sebastiani. *Determining the Semantic Orientation of Terms through Gloss Classification [C]*. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management. 617-624(2005).

[3] A. Esuli and F. Sebastiani. *Determining Term Subjectivity and Term Orientation for Opinion Mining* [*C*]. In Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics. 193-200(2006).

[4] H. D. Lasswell and J. Z. Namenwirth. *The Lasswell Value Dictionary [M]*. Yale University Press(1969).

[5] A. Neviarouskaya, H. Prendinger and M. Ishizuka. *SentiFul: A Lexicon for Sentiment Analysis [J]*. IEEE Transactions on Affective Computing. 2: 1-15(2011).

[6] A. Esuli and F. Sebastiani. *SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining [C]*. In Proceedings of the 5th Conference on Language Resources and Evaluation. 417-422(2006).

[7] T. Wilson, J. Wiebe and P. Hoffmann. *Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis [J]*. Computational linguistics. 35(3): 399-433(2009).

[8] M. Thelwall, K. Buckley and G. Paltoglou. *Sentiment strength detection for the social web [J]*. Journal of the Association for Information Science and Technology. 63(1): 163-173(2012).

[9] F. H. Khan, U. Qamar and S. Bashir. *SentiMI: Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection [J]*. Applied Soft Computing. 39: 140-153(2016).

[10] T. Xu, Q. Peng and Y. Cheng. *Identifying the semantic orientation of terms using S-HAL for sentiment analysis [J]*. Knowledge-Based Systems. 35: 279-289(2012).

[11] Z. Dong and Q. Dong. *HowNet and the Computation of Meaning [M]*. World Scientific Publishing Co. Inc. (2006).

[12] J. Li and M. Sun. *Experimental Study on Sentiment Classification of Chinese Review using Machine Learning Techniques*. In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering. 393-400(2007).

[13] Y. Zhu, J. Min, Y. Zhou and et al. *Semantic Orientation Computing Based on HowNet [J]*. Journal of Chinese Information Processing. 20: 14-20(2006)(In Chinese).

PoS(ISCC 2017)027