A New Method for Sentiment Classification in Text Retrieval

Yi Hu[1], Jianyong Duan[1], Xiaoming Chen[1,2], Bingzhen Pei[1,2], and Ruzhan Lu[1]

[1] Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai, China, 200030
[2] School of Computer Science and Engineering,
Guizhou University, Guiyang, China, 550025
{huyi, duan_jy, chen-xm, peibz, rz-lu}@cs.sjtu.edu.cn

Abstract. Traditional text categorization is usually a topic-based task, but a subtle demand on information retrieval is to distinguish between positive and negative view on text topic. In this paper, a new method is explored to solve this problem. Firstly, a batch of Concerned Concepts in the researched domain is predefined. Secondly, the special knowledge representing the positive or negative context of these concepts within sentences is built up. At last, an evaluating function based on the knowledge is defined for sentiment classification of free text. We introduce some linguistic knowledge in these procedures to make our method effective. As a result, the new method proves better compared with SVM when experimenting on Chinese texts about a certain topic.

1 Introduction

Classical technology in text categorization pays much attention to determining whether a text is related to a given topic [1], such as sports and finance. However, as research goes on, a subtle problem focuses on how to classify the semantic orientation of the text. For instance, texts can be for or against "racism", and not all the texts are bad. There exist two possible semantic orientations: positive and negative (the neutral view is not considered in this paper). Labeling texts by their semantic orientation would provide readers succinct summaries and be great useful in intelligent retrieval of information system.

Traditional text categorization algorithms, including Naïve Bayes, ANN, SVM, etc, depend on a feature vector representing a text. They usually utilize words or n-grams as features and construct the weightiness according to their presence/absence or frequencies. It is a convenient way to formalize the text for calculation. On the other hand, employing one vector may be unsuitable for sentiment classification. See the following simple sentence in English:

— Seen from the history, the great segregation is a pioneering work.

Here, "segregation" is very helpful to determine that the text is about the topic of racism, but the terms "great" and "pioneering work" may just be the important hints for semantic orientation (support the racism). These two terms probably contribute

less to sentiment classification if they are dispersed into the text vector because the relations between them and "segregation" are lost. Intuitively, these terms can provide more contribution if they are considered as a whole within the sentence. We explore a new idea for sentiment classification by focusing on sentences rather than entire text.

"Segregation" is called as Concerned Concept in our work. These Concerned Concepts are always the sensitive nouns or noun phrases in the researched domain such as "race riot", "color line" and "government". If the sentiment classifying knowledge about how to comment on these concepts can be acquired, it will be helpful for sentiment classification when meeting these concepts in free texts again. In other words, the task of sentiment classification of entire text has changed into recognizing the semantic orientation of the context of all Concerned Concepts.

We attempt to build up this kind of knowledge to describe different sentiment context by integrating extended part of speech (EPOS), modified triggered bi-grams and position information within sentences. At last, we experiment on Chinese texts about "racism" and draw some conclusions.

2   Previous Work

A lot of past work has been done about text categorization besides topic-based classification. Biber [2] concentrated on sorting texts in terms of their source or source style with stylistic variation such as author, publisher, and native-language background.

Some other related work focused on classifying the semantic orientation of individual words or phrases by employing linguistic heuristics [3][4]. Hatzivassiloglou et al worked on predicting the semantic orientation of adjectives rather than phrases containing adjectives and they noted that there are linguistic constraints on these orientations of adjectives in conjunctions.

Past work on sentiment-based categorization of entire texts often involved using cognitive linguistics [5][11] or manually constructing discriminated lexicons [7][12]. All these work enlightened us on the research on Concerned Concepts in given domain.

Turney's work [9] applied an unsupervised learning algorithm based on the mutual information between phrases and the both words "excellent" and "poor". The mutual information was computed using statistics gathered by a search engine and simple to be dealt with, which encourage further work with sentiment classification.

Pang et al [10] utilized several prior-knowledge-free supervised machine learning methods in the sentiment classification task in the domain of movie review, and they also analyzed the problem to understand better how difficult it is. They experimented with three standard algorithms: Naïve Bayes, Maximum Entropy and Support Vector Machines, then compared the results. Their work showed that, generally, these algorithms were not able to achieve accuracies on the sentiment classification problem comparable to those reported for standard topic-based categorization.

3   Our Work

3.1   Basic Idea

As mentioned above, terms in a text vector are usually separated from the Concerned Concepts (**CC** for short), which means no relations between these terms and CCs. To avoid the coarse granularity of text vector to sentiment classification, the context of each CC is researched on. We attempt to determine the semantic orientation of a free text by evaluating context of CCs contained in sentences. Our work is based on the two following hypothesizes:

- ♦ H_1. A sentence holds its own sentiment context and it is the processing unit for sentiment classification.
- ♦ H_2. A sentence with obvious semantic orientation contains at least one Concerned Concept.

H_1 allows us to research the classification task within sentences and H_2 means that a sentence with the value of being learnt or evaluated should contain at least one described CC. A sentence can be formed as:

$$word_{-m}word_{-(m-1)}...word_{-1}CC_iword_1...word_{(n-1)}word_n \quad . \tag{1}$$

CC_i (given as an example in this paper) is a noun or noun phrase occupying the position 0 in sentence that is automatically tagged with extended part of speech (**EPOS** for short)(see section 3.2). A word and its tagged EPOS combine to make a 2-tuple, and all these 2-tuples on both sides of CC_i can form a sequence as follows:

$$\begin{bmatrix} word_{-m} \\ epos_{-m} \end{bmatrix} \begin{bmatrix} word_{-(m-1)} \\ epos_{-(m-1)} \end{bmatrix} ... \begin{bmatrix} word_{-1} \\ epos_{-1} \end{bmatrix} CC_i \begin{bmatrix} word_1 \\ epos_1 \end{bmatrix} ... \begin{bmatrix} word_{(n-1)} \\ epos_{(n-1)} \end{bmatrix} \begin{bmatrix} word_n \\ epos_n \end{bmatrix}. \tag{2}$$

All the words and corresponding EPOSes are divided into two parts: m 2-tuples on the left side of CC_i (from –m to -1) and n 2-tuples on the right (from 1 to n). These 2-tuples construct the context of the Concerned Concept CC_i.

The sentiment classifying knowledge (see sections 3.3 and 3.4) is the contribution of all the 2-tuples to sentiment classification. That is to say, if a 2-tuple often co-occurs with CC_i in training corpus with positive view, it contributes more to positive orientation than negative one. On the other hand, if the 2-tuple often co-occurs with CC_i in training corpus with negative view, it contributes more to negative orientation. This kind of knowledge can be acquired by statistic technology from corpus.

When judging a free text, the context of CC_i met in a sentence is respectively compared with the positive and negative sentiment classifying knowledge of the same CC_i trained from corpus. Thus, an evaluating function **E** (see section 3.5) is defined to evaluate the semantic orientation of the free text.

3.2   Extended Part of Speech

Usual part of speech (POS) carries less sentiment information, so it cannot distinguish the semantic orientation between positive and negative. For example, "hearty" and "felonious" are both tagged as "adjective", but for the sentiment classification, only

the tag "adjective" cannot classify their sentiment. This means different adjective has different effect on sentiment classification. So we try to extend words' POS (EPOS) according to its semantic orientation.

Generally speaking, empty words only have structural function without sentiment meaning. Therefore, we just consider substantives in context, which mainly include nouns/noun phrases, verbs, adjectives and adverbs. We give a subtler manner to define EPOS of substantives. Their EPOSes are classified to be positive orientation (PosO) or negative orientation (NegO). Thus, "hearty" is labeled with "pos-adj", which means PosO of adjective; "felonious" is labeled with "neg-adje", which means NegO of adjective. Similarly, nouns, verbs and adverbs tagged with their EPOS construct a new word list. In our work, 12,743 Chinese entries in machine readable dictionary are extended by the following principles:

- ♦ To nouns, their PosO or NegO is labeled according to their semantic orientation to the entities or events they denote (pos-n or neg-n).
- ♦ To adjectives, their common syntax structure is {Adj.+Noun*}. If adjectives are favor of or oppose to their headwords (Noun*), they will be defined as PosO or NegO (pos-adj or neg-adj).
- ♦ To adverbs, their common syntax structure is {Adv.+Verb*/Adj*.}, and Verb*/Adj*. is headword. Their PosO or NegO are analyzed in the same way of adjective (pos-adv or neg-adv).
- ♦ To transitive verb, their common syntax structure is {TVerb+Object*}, and Object* is headword. Their PosO or NegO are analyzed in the same way of adjective (pos-tv or neg-tv).
- ♦ To intransitive verb, their common syntax structure is {Subject*+InTVerb}, and Subject* is headword. Their PosO or NegO are analyzed in the same way of adjective (pos-iv or neg-iv).

3.3 Sentiment Classifying Knowledge Framework

Sentiment classifying knowledge is defined as the importance of all 2-tuples <*word, epos*> that compose the context of CC_i (given as an example) to sentiment classification and **every** Concerned Concept like CC_i has its own positive and negative sentiment classifying knowledge that can be formalized as a 3-tuple K:

$$K := (CC, S^{pos}, S^{neg}) \ . \tag{3}$$

To CC_i, its S_i^{pos} has concrete form that is described as a set of 5-tuples:

$$S_i^{pos} := \left\{ (< word_\xi, epos_\xi >, wordval_\xi, eposval_\xi, \alpha_\xi^{left}, \alpha_\xi^{right}) \right\} \ . \tag{4}$$

Where S_i^{pos} represents the positive sentiment classifying knowledge of CC_i, and it is a data set about all 2-tuples <*word, epos*> appearing in the sentences containing CC_i in training texts with positive view. In contrast, S_i^{neg} is acquired from the training texts with negative view. In other words, S_i^{pos} and S_i^{neg} respectively reserve the features for positive and negative classification to CC_i in corpus.

In terms of S_i^{pos}, the importance of $< word_\xi, epos_\xi >$ is divided into $wordval_\xi$ and $eposval_\xi$ (see section 4.1) which is estimated by modified triggered bi-grams to fit the

long distance dependence. If $<word_\xi, epos_\xi>$ appears on the left side of CC_i, the "side" adjusting factor is α_i^{left}; if it appears on the right, the "side" adjusting factor is α_i^{right}. We also define another factor β (see section 4.3) that denotes dynamic "positional" adjusting information during processing a sentence in free text.

3.4   Contribution of *<word, epos>*

If a *<word, epos>* often co-occurs with CC_i in sentences in training corpus with positive view, which may means it contribute more to positive orientation than negative one, and if it often co-occurs with CC_i in negative corpus, it may contribute more to negative orientation.

We modify the classical bi-grams language model to introduce long distance triggered mechanism of $CC_i \rightarrow <word, epos>$. Generally to describe, the contribution *c* of each 2-tuple in a positive or negative context (denoted by ***Pos_Neg***) is calculated by (5). This is an analyzing measure of using multi-feature resources.

$$c(<word, epos> | CC_i, Pos\_Neg) := \alpha\beta \exp\big(\Pr(<word, epos> | CC_i, Pos\_Neg)\big) \quad \alpha, \beta > 0 \quad (5)$$

The value represents the contribution of *<word, epos>* to sentiment classification in the sentence containing CC_i. Obviously, when α and β are fixed, the bigger $\Pr(<word, epos> | CC_i, Pos\_Neg>)$ is, the bigger contribution *c* of the 2-tuple *<word, epos>* to the semantic orientation Pos_Neg (one of {positive, negative} view) is.

It has been mentioned that α and β are adjusting factor to the sentiment contribution of pair *<word, epos>*. α rectifies the effect of the 2-tuple according to its appearance on which side of CC_i, and β rectifies the effect of the 2-tuple according to its distance from CC_i. They embody the effect of "side" and "position". Thus, it can be inferred that even the same *<word, epos>* will contribute differently because of its side and position.

3.5   Evaluation Function *E*

We propose a function *E* (equation (6)) to evaluate a free text by comparing the context of every appearing *CC* with the two sorts of sentiment context of the same *CC* trained from corpus respectively.

$$E = (1/N)\sum_{i=1}^{N}\big(Sim(S_i^{'}, S_i^{pos}) - Sim(S_i^{'}, S_i^{neg})\big) \cdot \quad (6)$$

N is the number of total Concerned Concepts in the free text, and *i* denotes certain CC_i. *E* is the semantic orientation of the whole text. Obviously, if $E \geq 0$, the text is to be regarded as positive, otherwise, negative.

To clearly explain the function *E*, we just give the similarity between the context of $CC_i (S_i^{'})$ in free text and the **positive** sentiment context of the same CC_i trained from corpus. The function *Sim* is defined as follows:

$$Sim(S_i^{'}, S_i^{pos}) = \left(\prod_{\xi=-1}^{-m} \alpha_{\xi}^{left} \beta_{\xi}^{left} \right) \exp \left(\sum_{\xi=-1}^{-m} \Pr(<word_{\xi}, epos_{\xi} > | CC_i, positive) \right) . \tag{7}$$

$$+ \left(\prod_{\gamma=1}^{n} \alpha_{\gamma}^{right} \beta_{\gamma}^{right} \right) \exp \left(\sum_{\gamma=1}^{n} \Pr(<word_{\gamma}, epos_{\gamma} > | CC_i, positive) \right)$$

$\left(\prod_{\xi=-1}^{-m} \alpha_{\xi}^{left} \beta_{\xi}^{left} \right) \exp \left(\sum_{\xi=-1}^{-m} \Pr(<word_{\xi}, epos_{\xi} > | CC_i, positive) \right)$ is the positive orientation of the left

context of CC_i, and $\left(\prod_{\gamma=1}^{n} \alpha_{\gamma}^{right} \beta_{\gamma}^{right} \right) \exp \left(\sum_{\gamma=1}^{n} \Pr(<word_{\gamma}, epos_{\gamma} > | CC_i, positive) \right)$ is the right one.

Equation (7) means that the sentiment contribution *c* of **each** *<word, epos>* calculated by (5) in the context of CC_i within a sentence in free text, which is $S_i^{'}$, construct the overall semantic orientation of the sentence together. On the other hand, $Sim(S_i^{'}, S_i^{neg})$ can be thought about in the same way.

4 Parameter Estimation

4.1 Estimating Wordval and Eposval

In terms of CC_i, its sentiment classifying knowledge is depicted by (3) and (4), and the parameters *wordval* and *eposval* need to be leant from corpus. Every calculation of $\Pr(<word, epos> | CC_i, Pos\_Neg)$ is divided into two parts like (8) according to statistic theory:

$$\Pr(<word_{\xi}, epos_{\xi} > | CC_i, Pos\_Neg) = \Pr(epos_{\xi} | CC_i, Pos\_Neg) \times \Pr(word_{\xi} | CC_i, Pos\_Neg, epos_{\xi}) \cdot (8)$$

$eposval := \Pr(epos_{\xi} | CC_i, Pos\_Neg)$ and $wordval := \Pr(word_{\xi} | CC_i, Pos\_Neg, epos_{\xi}) \cdot$

The "*eposval*" is the probability of $epos_{\xi}$ appearing on both sides of the CC_i and is estimated by Maximum Likelihood Estimation (MLE). Thus,

$$\Pr(epos_{\xi} | CC_i, Pos\_Neg) = \frac{\#(epos_{\xi}, CC_i) + 1}{\sum_{epos} \#(epos, CC_i) + |EPOS|} . \tag{9}$$

The numerator in (9) is the co-occurring frequency between $epos_{\xi}$ and CC_i within sentence in training texts with Pos_Neg (certain one of {positive, negative}) view and the denominator is the frequency of co-occurrence between all EPOSes appearing in CC_i's context with Pos_Neg view.

The "*wordval*"is the conditional probability of $word_{\xi}$ given CC_i and $epos_{\xi}$ which can also be estimated by MLE:

$$\Pr(word_{\xi} | CC_i, Pos\_Neg, epos_{\xi}) = \frac{\#(word_{\xi}, epos_{\xi}, CC_i) + 1}{\sum_{word} \#(word, epos_{\xi}, CC_i) + \sum_{word} 1} . \tag{10}$$

The numerator in (10) is the frequency of co-occurrence between $< word_\xi , epos_\xi >$ and CC_i, and the denominator is the frequency of co-occurrence between all possible words corresponding to $epos_\xi$ appearing in CC_i's context with Pos_Neg view.

For smoothing, we adopt add–one method in (9) and (10).

4.2 Estimating α

The α_ξ is the adjusting factor representing the different effect of the $< word_\xi , epos_\xi >$ to CC_i in texts with Pos_Neg view according to the side it appears, which means different side has different contribution. So, it includes α_ξ^{left} and α_ξ^{right} :

$$\alpha_\xi^{left} = \frac{\# \text{ of } < word_\xi , epos_\xi > \text{ appearing on the left side of } CC_i}{\# \text{ of } < word_\xi , epos_\xi > \text{ appearing on both sides of } CC_i}, \qquad (11)$$

$$\alpha_\xi^{right} = \frac{\# \text{ of } < word_\xi , epos_\xi > \text{ appearing on the right side of } CC_i}{\# \text{ of } < word_\xi , epos_\xi > \text{ appearing on both sides of } CC_i}. \qquad (12)$$

4.3 Calculating β

β is positional adjusting factor, which means different position to some CC will be assigned different weight. This is based on the linguistic hypothesis that the further a word get away from a researched word, the looser their relation is. That is to say, β ought to satisfy an **inverse proportion** relationship with position.

Unlike *wordval, eposval* and α which are all private knowledge to some CC, β is a dynamic positional factor which is independent of semantic orientation of training texts and it is only depend on the position from CC. To the example CC_i, β of $< word_\mu , epos_\mu >$ occupying the μth position on its left side is β_μ^{left}, which can be defined as:

$$\beta_\mu^{left} = (1/2)^{|\mu|-1}(2-(1/2)^{m-1})^{-1} \quad \mu = -1 \sim -m . \qquad (13)$$

β of $< word_\nu , epos_\nu >$ occupying the νth position on the right side of CC_i is β_ν^{right}, which can be defined as:

$$\beta_\nu^{right} = (1/2)^{\nu-1}(2-(1/2)^{n-1})^{-1} \quad \nu = 1 \sim n . \qquad (14)$$

5 Test and Conclusions

Our research topic is about "Racism" in Chinese texts. The training corpus is built up from Chinese web pages and emails. As mentioned above, all these extracted texts in corpus have obvious semantic orientations to racism: be favor of or oppose to. There are 1137 texts with positive view and 1085 texts with negative view. All the Chinese texts are segmented and tagged with defined EPOS in advance. They are also marked posi-

tive/negative for supervised learning. The two sorts of texts with different view are respectively divided into 10 folds. 9 of them are trained and the left one is used for test.

For the special domain, there is no relative result that can be consulted. So, we compare the new method with a traditional classification algorithm, i.e. the popular SVM that uses bi-grams as features. Our experiment includes two parts: a part experiments on the relatively "long" texts that contain more than 15 sentences and the other part experiments on the "short" texts that contain less than 15 sentences. We choose "15" as the threshold to distinguish long or short texts because it is the mathematic expectation of "length" variable of text in our testing corpus. The recall, precision and F1-score are listed in the following Experiment Result Table.

Table. Experiment Result

	Texts with Positive View (more than 15 sentences)		Texts with Negative View (more than 15 sentences)	
	SVM	Our Method	SVM	Our Method
Recall(%)	80.6	73.2	68.4	76.1
Precision(%)	74.1	75.3	75.6	73.8
F1-score(%)	77.2	74.2	71.82	74.9
	Texts with Positive View (less than 15 sentences)		Texts with Negative View (less than 15 sentences)	
	SVM	Our Method	SVM	Our Method
Recall(%)	62.1	63.0	62.1	69.5
Precision(%)	65.1	70.1	59.0	62.3
F1-score(%)	63.6	66.4	60.5	65.7

The experiment shows that our method is useful for sentiment classification, especially for short texts. Seen from the table, when evaluating texts that have more than 15 sentences, for enough features, SVM has better result, while ours is averagely close to it. However, when evaluating the texts containing less than 15 sentences, our method is obviously superior to SVM in either positive or negative view. That means our method has more potential value to sentiment classification of short texts, such as emails, short news, etc.

The better result owes to the fine description within sentences and introducing linguistic knowledge to sentiment classification (such as EPOS, α and β), which proved the two hypothesizes may be reasonable. We use modified triggered bi-grams to describe the importance among features ({<word, epos>}) and Concerned Concepts, then construct sentiment classifying knowledge rather than depend on statistic algorithm only.

To sum up, we draw the following conclusions from our work:

♦ Introducing more linguistic knowledge is helpful for improving statistic sentiment classification.

♦ Sentiment classification is a hard task, and it needs subtly describing capability of language model. Maybe the intensional logic of words will be helpful in this field in future.
♦ Chinese is a language of concept combination and the usage of words is more flexible than Indo-European language, which makes it more difficult to acquire statistic information than English [10].
♦ We assume an independent condition among sentences yet. We should introduce a suitable mathematic model to group the close sentences.

Our experiment also shows that the algorithm will become weak when no *CC* appears in sentences, but this method is still deserved to explore further. In future, we will integrate more linguistic knowledge and expand our method to a suitable sentence group to improve its performance. Constructing a larger sentiment area may balance the capability of our method between long and short text sentiment classification.

References

1. Hearst, M.A.: Direction-based text interpretation as an information access refinement. In P. Jacobs (Ed.), *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval.* Mahwah, NJ: Lawrence Erlbaum Associates (1992)
2. Douglas Biber: Variation across Speech and Writing. Cambridge University Press (1988)
3. Vasileios Hatzivassiloglou and Kathleen McKeown: Predicting the semantic orientation of adjectives. In Proc. of the 35th ACL/8th EACL (1997) 174-181
4. Peter D. Turney and Michael L. Littman: Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report EGB-1094, National Research Council Canada (2002)
5. Marti Hearst: Direction-based text interpretation as an information access refinement. In Paul Jacobs, editor, Text-Based Intelligent Systems. Lawrence Erlbaum Associates (1992)
6. Bo Pang and Lillian Lee: A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Proceedings of the 42nd ACL (2004) 271--278
7. Sanjiv Das and Mike Chen: Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proc. of the 8th Asia Pacific Finance Association Annual Conference (2001)
8. Vasileios Hatzivassiloglou, Janyce Wiebe: Effects of Adjective Orientation and Gradability on Sentence Subjectivity. COLING (2000) 299-305
9. Peter Turney: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classication of reviews. In Proc. of the ACL (2002)
10. Bo Pang, Lillian Lee and Shivakumar Vaithyanathan: Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proc. Conf. on EMNLP (2002)
11. Warren Sack: On the computation of point of view. In Proc. of the Twelfth AAAI, page 1488. Student abstract (1994)
12. Richard M. Tong: An operational system for detecting and tracking opinions in on-line discussion. Workshop note, SIGIR Workshop on Operational Text Classification (2001)