Construction and application of a Web text-oriented integrated sentiment feature library mined by a big corpus

Meijuan Liu [1*] , Shicai Yang [2]

[1] *School of Foreign Languages, Zhejiang Ocean University, Zhoushan, 316022, China*

[2] *Ningbo University of Technology, Ningbo, 315016, China*

Received 18 October 2014, www.cmnt.lv

Abstract

This paper describes the implementation of a system containing a web text-oriented integrated sentiment feature library (hereinafter referred to as WTISFL) and its application in sentiment analysis. Sentiment library plays an important part in sentiment analysis. A quicker and more complete method of constructing new sentiment library is presented in the paper. Firstly, the structure of WTISFL which sentiment analysis needed is proposed. Besides, WTISFL is mined from the big corpus database. Moreover, our sentiment word set is extended on the basis of existing sentiment resources, semantic similarity calculation of HowNet and computation of Chinese Synonym Thesaurus. Finally, the WTISFL is checked manually. Based on the above WTISFL, Web texts are studied from the perspective of sentiment analysis with the method of maximum entropy classifier. The experiment shows that WTISFL in this paper is extremely effective in sentiment analysis, which can evidently improve the performance of web texts sentiment classification.

Keywords: Integrated Sentiment Feature Library; Sentiment Classification; Maximum Entropy; Web Texts

1 Introduction

In recent years, as a newly soon-to-launch information platform, web texts including network comments, blogs and microblogs, are playing an increasingly important role in daily life, and gaining more and more people's attention. In the meantime, it is found that the phenomenon of group polarization is far more apparently on the Internet than it in reality. Therefore, it is of vital significance to analyze reviews from web texts for getting to know word-of-mouth evaluation of a product, public opinion on hot social issues, and so on. Analyzing the sentiment tendency of the web texts has a great value either in business world or political territory. Consequently, sentiment analysis on web texts has become a novel hotter research area in natural language text processing[1-2].

Traditional methods of sentiment analysis mainly contain the analytical approach based on dictionary[3] and machine learning method[1][4-5], between of which machine learning method is divided into the supervised[1] and the unsupervised[4-5].

An analytical approach based on dictionary is to construct a sentiment dictionary containing positive and negative sentiment words, by which to define the sentiment tendency of a text. Reference[3] studied how to build a universal and domain-related sentiment dictionary.

Supervised machine learning approach covers Naive Bayes (NB), maximum entropy (ME), support vector machines(SVM), and so on. Reference [1] achieved firstly sentiment classification on texts with the method of machine learning. Reference[4] classified subjective and objective sentences by adopting the method based on graph. Reference[1] compared respectively the results of sentiment classification based on supervised machine learning in various classification algorithms, all kinds of features and selection strategy of character weight parameters.

Unsupervised machine learning approach classifies unsupervisedly the labeled seed word set with the unlabeled sample modeling. Reference[5] selected "excellent" and "poor" as positive and negative sentiment benchmark words. After obtaining the spot mutual information between each word and the benchmark word, the sentiment tendency of each word is calculated by the difference between of them.

However, just considering the use of sentiment words in sentiment analysis is far not enough. The following questions have been ignored in most researches.

While we are analyzing ambiguous determiners in the corpus of reviews [6], it is found that there are some words with indefinite sentiment tendency, which are called ambiguous sentiment words in our study, such as "大/big", "小/small", "高/high", "低/low", "圆滑/tactful", etc.

For example, the sentiment tendency of the word

*Corresponing author's E-mail: azalea1212@126.com

"高/high" is not definite, which is closely related to the context. That is to say, when the word "高/high" is used in different contexts, it shows different sentiment tendency. For instance, when "高/high" modifies "质量/quality" in the phrase "高质量/high quality", the sentiment tendency of the phrase is positive. On the contrary, when "高/high" modifies "油耗/ oil consumption" in the phrase "油耗高/oil consumption is high", it is negative.

In the corpus of reviews, there exist a large number of sentences including many ambiguous words, such as "big", "small", "high", and "low", etc. For example, among 200,000 sentences of car reviews, there are about 10,000 reviews containing "big", and about 8,000 reviews containing "small".

Therefore, with the development of sentiment classification technology and the increased demand for its application, polarity disambiguation related to ambiguous sentiment words has become a research hotspot[6-7][11].Researches[1][3][6] show that supervised machine learning method needs to utilize abundant of corpus, which costs a large number of human power. As for unsupervised machine learning, it has a low accuracy. What the method based on dictionary emphasized is to construct a sentiment dictionary, which makes a direct effect on the result of experiment.

As there are a lot of pending labeled sentiment words on the internet, direct labeling takes a large amount of time and effort. Given that the features of web texts like timeliness, short text, irregular expression, and a large amount of information, we implement a system including construction and application of WTISFL mined by a Big Corpus.

The system is divided into the following three models: Model of constructing library will conduct the tasks including mining sentiment words, extending sentiment polar words in library, and manual checking finally. Classifier training model works with data downloading, feature mining and model training, etc. Application model of sentiment analysis focuses on feature extraction from web texts, sentiment classification, and results returning. The system structure is shown in Figure 1.

This structure of the paper is as follows: In section two, it introduces preparatory work and related foundations applied in the paper. A WTISFL is designed in section three. In section four, a way of constructing WTISFL is presented by adopting multiple methods of natural language processing (NLP) and a big corpus. Based on the above WTISFL, section five analyzes the sentiment tendency of web texts with the maximum entropy (ME) classifier.

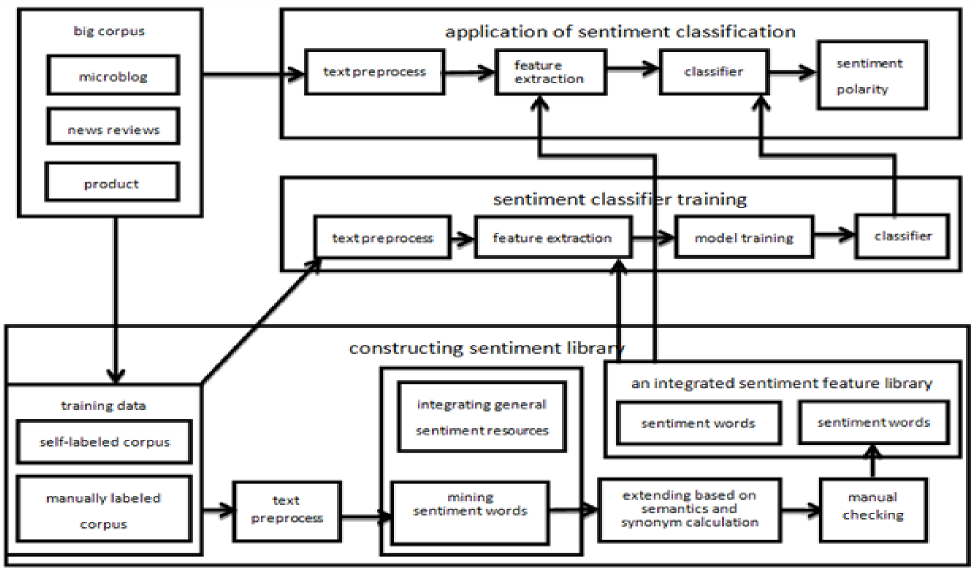

FIGURE 1. Figure1 System Structure Diagram

2 Basic work: constructing and preprocessing a corpus

Common text classification method usually contains the following steps: preparation of training corpus, text preprocessing, feature mining, selection of classification algorithm and sort application, etc.

2.1 CONSTRUCTING A BIG CORPUS

Data accumulation is the cornerstone of sentiment classification because feature mining and model classification mainly use corpus as materials. Therefore, the quality and the quantity of training corpus are of crucial importance to the result of classification.

Corpus itself can be divided into the labeled and the unlabeled. As for the labeled corpus, such as merchants' comments and reviews on products, sentiment classification can be determined by star ratings.

With regard to the unlabeled corpus like comments on the current news, a classification model or a great deal of manual annotation should be finished before the corpus is used. Moreover, because of the subjectivity of manual annotation, it is necessary to have an in-depth communication with the annotators to reach the reliability and the availability of corpus.

Up to now, we have selected roughly three millions of corpus, which cover more than ten categories like e-business, news, Film and Television, music, etc.

2.2 TEXT PREPROCESSING

There is a method of full word list in mining the sentiment polarity of words in sentiment classification, namely all the words are regarded as polar words, the advantage of which is that words are entirely reserved, but will increase characteristic dimension and make calculation more complicated.

What we adopt is “a method of polar word list” to mine some words or phrases which can represent the positive or negative polarity in documents. For instance, in the following positive example:

#淡淡晒奖品#盼望已久的奖品收到了感谢 / I Received the long-awaited prizes. Thanks! @奥e-tron 多方拍照得瑟下 /Take some photos and show them off everywhere. In the above two texts, what positive polar words we can extract are “奖品/prizes”, “感谢/Thanks”, “得瑟/show off”.

As there are too many noises in the corpus, it is necessary to preprocess the text before mining sentiment polarity of words. Text preprocessing involves in some related techniques, such as parsing, denoising and seeking a best match, etc.

Tools used for word segmentation and part-of-speech tagging in the paper is researched by ICTCLAS[12]. There are several major functions containing Chinese word segmentation, part-of-speech tagging, recognition of named entity and new words, and support for user dictionary. Word segmentation accuracy by ICTCLAS word reaches 98. 45%.

As the corpus data contains review data like microblog data and non-microblog data, which requires special treatment in preprocessing the special data. The process of preprocessing mainly includes the following:

1) For all the microblog data, it is necessary to filter out noise information in it, such as user names, hashtags, etc.

2) For all the training data, it needs to take away the invalid links, like "http:// ", "www." , "url.” and the link beginning with “WWW.”.

3)Substitute some special characters in the text, for example, replace "<” with “<", ""” with ““““,">” with “>“, etc.

A best match is to make sure that the suggested feature can accurately reflect positive or negative tendency. For instance, in the phrase “ 威信扫地/ a loss of prestige”， if the word “威信/ prestige” is extracted individually, it will convey positive sentiment information. The word “扫地/sweep” is neutral in sentiment. However, the phrase “威信扫地/ a loss of prestige “ itself is negative. Therefore, a longest match method is employed to analyze a word or a phrase like the example “威信扫地/ a loss of prestige “.

3 Introduction of WTISFL

Different from the general sentiment library like NTUSD[16], our library takes into account ambiguous sentiment words and their factors which could make influence on sentiment tendency. Our WTISFL is made up of general sentiment words, sentiment ambiguities, sentiment impact factors, among of which general sentiment words refer to positive and negative words unrelated to context.

3.1 GENERAL SENTIMENT WORDS POSITIVE AND NEGATIVE WORDS UNRELATED TO CONTEXT

General sentiment words here refer to positive or negative sentiment words which keep expressing the same sentiment tendency towards different evaluation objects.

Positive words are commendatory terms, known as “good words”, which always express the emotional sense of appreciation, praise, incentives, love, respect and happiness in any contexts. For instance, positive words like “真诚的/sincere”,”勤奋/diligence”,”欢快的/cheerful” remain expressing commendatory sense in any thematic articles. However, general negative words refer to derogatory terms, known as “bad words”, which always express the emotional sense of disparagement, negation and hatred in any language environments. Words like “虚伪/Hypocrisy”,”懒惰的/ lazy”,”狂暴/fury”,”鲁莽的/ reckless”, are always showing derogatory sense in any contexts. In a word, we define the above discussing sentiment words, whose sentiment tendency is not affected by contexts, as general sentiment words.

In network texts, there are some sentiment words which do not exist in a formal dictionary. Especially those network buzzwords including some English words expressing affection, some abbreviations of Chinese Pinyin, and some emotion icons, have a strong subjective sense of sentiment tendency. For example, buzzwords like “苦逼/tormented” and “TMD/fuck” are evidently showing derogatory sense, yet the emotion icon “ ╰_╯ ” denotes “anger” in any thematic articles, which is irrelevant to its close words or the context itself.

3.2 SENTIMENT AMBIGUITIES

Sentiment ambiguities refer to words which show inconsistent sentiment polar tendency towards different evaluation objects. In other words, the sentiment tendency of these words is not unique. These words appearing alone have no commendatory or derogatory sense, but show obviously positive or negative meaning while combining with other words. For instance, "利润很高/Profits are very high", and "利润很低 /Profits are very low" show a strong sense of sentiment tendency, yet the individual words "利润/profit", "高/high", and "低/low" themselves do not show any sentiment tendency at all.

3.3 SENTIMENT IMPACT FACTORS

Sentiment impact factors refer to words that have an effect on the tendency or the degree of strength of a sentiment word, like the use of some degree adverbs, some conjunctions, or some negatives, etc.

3.3.1 Negative words

Negative words can turn over the sentiment polarity of words. By extracting from corpus and adding manually some common negative words, a negative word list with the size of 68 words could be acquired.

For example, words like " hardly/几乎不" ; "rarely/很少" ; "seldom/很少" ; "scarcely/极少" could play a reversely transformational role in the context. The sentence "Some villagers had received scarcely any education. /有些村民几乎没有接受过任何教育." is totally opposite to "Some villagers had received some education./有些村民接受过一些教育."

3.3.2 Degree words

A degree word list is drawn from "degree level words (Chinese)" by HowNet in a total of 140 words. A degree word is a word to define or modify an adjective or adverb in some extent, like "too/太", "very/非常","much/很","almost/几乎","nearly/几乎","enough/充分", "略微/a little" . The use of a degree word can intensify or weaken the original sentiment tendency in a certain extent. For example, in the sentence "He drives very carefully./他驾驶非常小心." The degree word"非常/very" "intensifies the degree of his driving. However, in sentence" "He can speak a little English./他能说一点英语.", "a little/一点" here weakens the ability he speak English.

3.3.3 Conjunctions

There are two common semantic relations between sentences: coordinating relation and master-slave relation. Master-slave relation involves the relationships like concession, adversative, assumption, purpose and condition. The usage of conjunctions will affect the expressing focus of sentiment in the sentence. For example, adversative conjunctions are the words which can turn over the meaning of adjacent texts. What affection an adversative clause expresses is the most important in the surrounding sentences. In the sentence "它曾经是涡轮增压器的先锋，但它的技术如今已不再引人注目了。/It was once a pioneer of turbocharging, but its technology no longer stands out", the conjunction "但是/but" denotes an adversative relationship in the context.

4 Constructing a Web text-oriented integrated sentiment feature library

General sentiment library like NTUSD [16],a Chinese sentiment dictionary compiled by National Taiwan University, whose simplified Chinese edition contains 2812 positive words and 8276 negative words, has such drawbacks as the insufficiency of corpus and the imperfection of structure.

We present a method of constructing sentiment library implemented by the following steps.

Step1 Mining the Sentiment Polarity of Words Based on the Big Corpus

It will save a great deal of workload of manual work to mine sentiment words based on the big corpus as there are hundreds of thousands of frequent words in Chinese. And there are a great many new unlabeled words, especially words on web texts. The implementation of detailed process on how to mine WTISFL will be introduced in section 4.1.

Step2 Integrating the Existing Sentiment Resources

We integrate general sentiment words from traditional sentiment resources including the Dictionary of Positive Words, the Dictionary of Negative Words, HowNet and other network resources like Terms of Adverse Drug Reaction, Codes of Diseases and Symptoms, etc.

Step3 Extending the Library Based on HowNet

Based on a calculating semantic similarity formula and an approach of HowNet lexical semantic similarity computation described in section 4.2, we can calculate the distance between new words and benchmark words in Chinese, and identify the sentiment polarity of a new word.

Step4 Extending the Library Based on Chinese Synonym Thesaurus

With the smallest unit of word group, Chinese Synonym Thesaurus searches word groups where sentiment words are located. We extend the words with the same meaning into the sentiment dictionary and check manually.

A word set containing 30,000 sentiment polar words is obtained finally.

4.1 MINING THE SENTIMENT POLARITY OF WORDS BASED ON THE BIG CORPUS

After text preprocessing, what should we do is to select some words from the innumerable words as polar words for training models.

There are several approaches commonly used for feature selection, such as TF-IDF, chi-square, mutual information, information gain, X2 statistics, cross entropy, Fisher discriminant, etc. TF-IDF will be focused in the following.

The main idea of TF-IDF is that if a certain word or a phrase appears with a high frequency in a paper and is rarely used in others, this word or the phrase is concluded as having a good sort distinguishing capacity and is suited for classification.

TF-IDF can be briefly understood as below:

TF: word frequency, which denotes the number of times that feature t, appears in document D. For example, a paper on Ma Yun, words like "Alibaba", "Taobao" can be predicted to have a high TF value.

DF: the number of documents including feature t. A higher DF, a lower distinguishing function to different documents that feature X has. For instance, words like "我/I" , "的/of" usually have a highest DF.

IDF: defined as IDF =log(|D|/DF), |D| is the overall number of documents. Inversely proportional to DF, a higher IDF value, a more important significance to distinguish documents. Final definition:

$$TF\text{-}IDF=TF*IDF \qquad (1)$$

As the texts we trained and processed are a little short, DF has roughly the same value as TF. Consequently, the value of TF is sufficient alone. In addition, we also calculate the frequency of polar words occurring in counter-examples. For example, for the positive polar word "赞/praise", the value of TF in positive polar data must be greater than it in negative one. If the difference is greater than a certain domain value, the feature of which will be integrated into the candidate set of polar words.

4.2 AN APPROACH OF HOWNET LEXICAL SEMANTIC SIMILARITY COMPUTATION

HowNet[13] is a commonsense knowledge base that describes the concepts of Chinese and English words as objects and shows the relationship between concepts and the property of the concepts as its basic content. Two basic notions in HowNet are: "concepts" and "sememes". "Concepts" refer to the description of lexical semantics in which a word can be expressed as several concepts. A knowledge-based language is used to describe "concepts", and "words" in the language are called "sememes". "Sememes" are regarded as the smallest semantic units describing "concepts". Each concept is characterized as a set of sememes in the HowNet, which are organized as a tree hierarchy system of sememes by hyponymy relations. Additionally, other relationships do exist among sememes, such as synonymy, antonymy, opposition, property-host, part-whole, material-product, event-role relations, etc. Hence, it is not only a single tree structure, but also a complicated net organization of sememes. However, hyponymy relation is the most important one among all sememe relations. From the perspective of hyponymy sememe relation, all the fundamental sememes constitue a level system of sememes, which is the baseline of semantic similarity computation .

According to the approach of HowNet lexical semantic similarity computation proposed by Liu Qun[10], the similarity of two independent words can be simplified as the similarity between two concepts. Supposing two Chinese words W1 and W2, if W1 has n concepts: S11, S12, …, S1n ; and W2 has m concepts: S21, S22, …, S2m, the similarity of W1 and W2 is the maximum among each concepts, that is,

$$Sim(W1 , W2) = \max_{i, j} (Sim(S1i , S2j)) \qquad (2)$$

Concepts are expressed via sememes and the semantic similarity computation is the base of computing concept similarity. Sememe similarity is calculated by the semantic distance, of which sememes are in the tree hierarchy system.

$$Sim(p1, p2) =alpha/(dis+alpha) \qquad (3)$$

In the formula above, p1 and p2 are two sememes, dis is a positive integer, the overall path length of p1 and p2 in the semantic level system, and alpha is an adjustable parameter.

5 Sentiment analysis on Web text based on maximum entropy method

There are some sorting algorithms like statistical-based Bayesian algorithm, KNN algorithm, maximum entropy model, support vector machine approach, rule-based decision-making tree method, a more complicated technique of neural network. This paper mainly verifies the effect of our constructed sentiment library and analyzes microblog sentiment tendency by adopting maximum entropy model.

5.1 MAXIMUM ENTROPY (ME) CLASSIFIER

Maximum entropy model was firstly employed to deal with natural language in reference[8]. The model is extensively applied in NLP areas like machine translation, word segmentation, syntactic analysis, part-of-speech tagging, word sense disambiguation, and so on. Kamal's [9] study indicated that maximum entropy model performed better than Naive Bayes in NLP.

On the basis of experimental data and given probability, maximum entropy provides a method of machine learning to verify the classification of a sentence. Maximum entropy statistic modeling is a technique of model selection, which is to select maximum entropy distribution as the best among the proper distributions, when some constraint factors can't decide the uniqueness of a systematic distribution, maximum entropy distribution is regarded as the most suitable.

Determining a web text is positive or negative in this paper involves various factors. Supposing X is the vector quantity composed of these factors, the value of variable quantity y is the sentiment type of commendatory or derogatory sense. p(y/X) is the probability of sentiment tendency of a certain web text in the system.

Maximum entropy model requires p(y/X) to meet some certain constrained conditions and make the following defined entropy obtain a maximum value：

$$H(p) = -\sum_{X,\ y} p(y \mid X)\log p(y \mid X) \quad . \qquad (4)$$

In fact, the constrained conditions here refer to all the known facts, which can be stated as the following formula:

$$f_i(X,y) = \begin{cases} 1, & if(X,y)\ \ satisfies\ \ certain\ \ condition \\ 0, & else \end{cases} \qquad (5)$$

$$i = 1,2,3,\cdots,n$$

$f_i(X,y)$ is the feature of maximum entropy model; n is the sum of all features.

What can be seen is that all the features describe the relationship between vector quantity X and semantic role y.

As with other machine learning methods, it is necessary for a maximum entropy classifier to train data sets and get a prediction about new data sets except for training. There are some frequently-used preprocessing periods for training a classifier, such as characteristic extraction, word segmentation and part-of-speech tagging to generate the feature data from the original as a training data set.

5.2 EXPERIMENT

Our raw experimental data comes from riders' car comments on Sina auto forum, each of which shows commentator's commendatory or derogatory sense. In these reviews, commentators are requested to make some comments on auto merits or demerits. Accordingly, we have collected abundant of manual labeled data and chose 20,000 comments in it. In the paper, 80% linguistic materials are randomly selected for training set and 20% are for testing set.

Preprocessing including word segmentation and part-of-speech tagging can extract feature data like characteristic of a word, part-of-speech to train and test maximum entropy classifier.

Based on WTISFL established in the paper, maximum entropy classifier can make use of the features like general sentiment words, ambiguous sentiment words from preprocessing Chinese comments. To train and test the maximum entropy classifier, it is helpful to make full use of features like word feature, part of speech, and so on. On the basis of extracted features from WTISFL, maximum entropy classifier is trained to test the data set, and its result is compared with experiment results based on NTUSD.

Compared to feature selection based on NTUSD, experimental results show that, by adopting Maximum Entropy classifier, our selection based on WTISFL can extract more subtle features and reach a higher accuracy, from 76.7% to 86.5%.

With the method of maximum entropy classifier, our approach based on WTISFL could extract more delicate features, which has been proved in the experiment. Compared with feature selection based on NTUSD, it achieves a higher accuracy, from 76.3% to 85.6 % .

Sentiment tendency analysis on network texts in NTUSD dictionary achieves poor result. The main problem remains that the sentiment dictionary is constructed imperfectly. Particularly, because of the insufficiency of sentiment words, it is difficult to judge the sentiment tendency of a word.

Example1, no sentiment words in NTUSD like "离谱/ outrageous"

"见过有史以来最离谱的平板车. /It is the most outrageous dray ever recorded ."

.Example2, Ambiguous sentiment words, the following ambiguous cases, like

"速腾价格太高 Sagitar's price is too high", "速腾油耗高/ Sagitar has a heavy fuel use", "帝豪性价比高 /Emgrand is cost-effective"

As a classification method, the maximum entropy model on the basis of characteristic selection of our WTISFL has a better performance in accuracy and feasibility, which has been demonstrated in the experiments.

There are still some further work we should undertake in the future, due to the following limitations like an objective evaluation criterion, the insufficiency of sentiment knowledge base, and so on.

5.3 FUTURE WORK

There are still some shortcomings in our experiment, such as the inadequacy of sentiment knowledge base, a lack of objective standard for sentiment knowledge because of various understandings, etc. Based on the current work, we are going to take part in Aspect Based Sentiment Analysis task on International Workshop on Semantic Evaluation (SemEval-2015)[15].

Acknowledgements

This work was supported by grants from Zhejiang Provincial Federation of Social Sciences (Grant No: 2010Z46) and Zhejiang Ocean University (Grant No: X10M24).

References

[1] Pang B, Lee L, Vaithyanathan S 2002 Thumbs up: Sentiment classification using machine learning techniques *Proceedings of the Association for Computational Linguistics* 79-86

[2] Seki Yohei, Ku Lun-Wei, Sun Le, Chen Hsin-His, Kando Noriko 2010 Overview of Multilingual Opinion Analysis Task at NTCIR-8: A Step Toward Cross Lingual Opinion Analysis *Proceedings of the Eighth NTCIR Workshop*

[3] Du Weifu, Tan Songbo et al 2009 A New Method to Compute Semantic Orientation *Journal of Computer Research and Development* (10) 1713-20

[4] Pang B, Lee L 2007 A study of the classification approach for Chinese subjective and objective texts *Proceedings of the 3th National Conference on Information Retrieval and Content Security* 117-23

[5] Sun Yan, Zhou Xueguang, Fu Wei 2013 Unsupervised Topic and Sentiment Unification Model for Sentiment Analysis *Acta Scientiarum Naturalium Universitatis Pekinensis* **49**(1)

[6] Yang Shicai, Liu Meijuan 2010 YSC-DSAA: An Approach to Disambiguate Sentiment Ambiguous Adjectives Based On SAAOL *Proceedings of the 5th International Workshop on Semantic Evaluation* 440–3

[7] Pak A and Paroubek P 2010 Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives *Proceedings of the 5th International Workshop on Semantic Evaluation* 436-9

[8] Berget Adam L, Della Pietra S A, Della Pietra V J 1996 A Maximum Entropy Approach Natural Language Processing. *Computational Linguistics* **22**(1) 39-71

[9] Nigam K, McCallum A, Thrun S, Mitchell T M 2000 Text Classification from Labeled and Unlabeled Documents using EM *Machine Learning* **39**(2) 103-34

[10] Liu Qun and Li Jian Su 2002 Word Similarity Computing Based on How-net *Computational Linguistics and Chinese Language Processing* **7**(2) 59-76

[11] Liu Meijuan, Yang Shicai, Chen Qiaofen 2012 Sentiment Classification on Chinese Reviews Based on AMBIGUOUS SENTIMENT CONFINED LIBRARY *Proceedings of IEEE CCIS2012*

[12] ICTCLAS http://ictclas.org/

[13] HowNet http://www.keenage.com

[14] SemEval-2015 Task 12: Aspect Based Sentiment Analysis http://alt.qcri.org/semeval2015/task12/

[15] Chinese Synonym Thesaur *http://www.datatang.com/data/42306/*

[16] NTUSD *http://nlg18.csie.ntu.edu.tw:8080/opinion/pub1.htm*

Authors

Meijuan Liu, born in 1976, Zhejiang, China

Current position, grades: Master of Applied Linguistics, lecturer
Scientific interest: Applied Linguistics Analysis; Computational Linguistics
Experience: teaching experience of 15 years, has completed two scientific research projects.

Shicai Yang, born in 1969, Zhejiang, China

Scientific interest: Computational Linguistics, Machine Learning
Publications： more than ten papers and 1 book