Advanced Engineering Forum *Online: 2011-09-09*
ISSN: 2234-991X, Vol. 1, pp 338-342
doi:10.4028/www.scientific.net/AEF.1.338

A Smart Sentiment Analysis System in Word, Sentence and Text Level

Hui He[1, a], Bo Chen[2,3,b]

[1]School of Control and Computer Engineering, North China Electric Power University, Beijing, 102206, P.R. China
[2]Beijing University of Posts and Telecommunications, Beijing 100876, P.R. China
[3]Postdoctoral Working Station, China United Network Communications Group Company Limited, Beijing 100033, P.R. China

[a]hh1012@gmail.com, [b]chb615@gmail.com

Keywords: Sentiment Analysis, Maximum Entropy, LMR Template, Indri, Domain Lexicon

Abstract. Recently, sentiment analysis of text is becoming a hotspot in the study of natural language processing, which has drawn interesting attention due to its research value and extensive applications. This paper introduces a smart sentiment analysis system, which is to satisfy three aspects of sentiment analysis requirement. These are Chinese sentiment word recognition and analysis, sentiment related element extraction and text orientation analysis. Promising results and analysis are presented at the end of this paper.

Introduction

With the rapid development of Web2.0, more and more Internet users generate their online comments, and opinions in some popular web applications, such as micro-blog, BBS. Therefore, text sentiment analysis is becoming a novel research topic, which has drawn interesting attention due to its research value and extensive applications. Text sentiment analysis, which is also called opinion mining, is to recognize the orientation of online reviews. There are three important tasks of sentiment analysis in the state-of-the-art research. They are sentiment extraction, sentiment classification, sentiment retrieval and summarization.[1] They could be taken as word level, sentence level and text level sentiment analysis. [2]

In this paper, a smart sentiment analysis system (SSAS) is described, which satisfy the three main tasks of sentiment analysis. SSAS contains three parts. The first is Chinese sentiment word recognition. A LMR-template is introduced to recognize the word sentiment orientation. The second part is sentiment related elements extraction. We adopt knowledge engineering method to organize a domain lexicon, which is used to judge the sentiment related elements. In the third part, an improved Maximum Entropy algorithm is proposed to classify sentences. Then a polarity model gives the text orientation by using of the sentence sentiment results. At last, text sentiment retrieval is presented by combining the third part results and Indri.

The remainder of this paper is structured as follows. The related work of text sentiment analysis is introduced in section 2. Section 3 describes the smart sentiment analysis system in detail in word level, sentence level and text level respectively. Section 4 gives the promising results in Chinese Opinion Analysis Evaluation in 2008. Analysis and conclusions are presented in Section 5.

Related Work

Sentiment analysis involves several challenging research tasks. It includes three main tasks: sentiment extraction, sentiment classification, sentiment retrieval and summarization. These three tasks are related with each other.

The first task is called sentiment extraction. It ranges over several kind of information extraction for sentiment analysis, for example, sentiment holder extraction, sentiment word recognition, sentiment related elements extraction, sentiment unit identification and so on. Methods of sentiment extraction can be divided to into two categories: methods based on corpus [4] and methods based on lexicons [3][5].

The second task is sentiment classification. It includes two steps. First, a word or a sentence or a text is to be judged as subjective or objective. If it is subjective, the second sub-task is to recognize whether the sentiment holder is positive negative to the sentiment object. Methods [6-10], such as Naïve Bayes, Maximum Entropy, SVM, are used in sentiment classification in each level.

The last main task is sentiment retrieval and summarization. Because of the large amount of online reviews, sentiment retrieval and summarization is helpful and necessary for users to obtain useful information. Some evaluations, like Blog TREC[11], NTCIR[12], COAE[13], are involved.

Smart Sentiment Analysis System

The smart sentiment analysis system satisfies three aspects of sentiment analysis requirement. Fig. 1 shows the system overview. After text preprocessing, word segmentation and POS tagging, different models are designed for sentiment analysis in three levels. LMR-template and Maximum Entropy are combined to recognize the sentiment words. A domain lexicon constructed by knowledge engineering method is used for sentiment related element extraction. An improved Maximum Entropy with priors is applied in sentence sentiment classification. The results of text sentiment are calculated by the sentence orientation. At last, text sentiment retrieval is presented by combining the third part results and Indri.

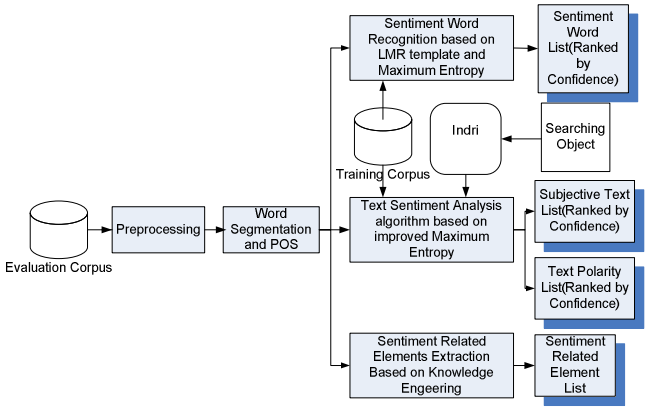

Fig. 1 System Overview

Sentiment Word Recognition

The LMR sentiment word template follows the hypothesis that Chinese text sentiment expression owns its internal mechanisms, which can be expressed through the word order. This template simulates the order arrangement, which obtains sentiment expressions.

Letters in the "LMR" have different meanings. "M" is the word which is required to be judged. "L" is the word on the left of the M word. And, "R" means the word on the right side of the M word. Thus, the word sequence which contains $2n+1$ words can be devoted as: $L_n L_{n-1} \ldots L_1 M R_1 \ldots R_{n-1} R_n$. Some information that gets from the L word and R word are helpful to judge the polarity of the M word.

In the process of extracting the sentiment words, all output values make up of the finite set Y which is the result of the word's polarity. Y is influenced and constrained by the contextual information X. The goal is to construct a stochastic model that accurately represents the behavior of the random process. With the given contextual information $x \in X$, the output is the conditional probability $p(y \mid x)$, which is denoted by $y \in Y$.

In the Maximum Entropy model based on LMR template, x is the feature information in the LRM template, such as the word, its position, its POS. y is the result of the polarity of the M word.

For example, a feature function can be designed as follows:

$$f(x,y) = \begin{cases} 1 & if \quad y = negand x = v \\ 0 & otherwise \end{cases} \tag{1}$$

Given a sequence X and sentiment labels set Y, the probability $y \in Y$ estimated by Maximum Entropy model for M word is :

$$p_\lambda(y \mid x) = \frac{1}{Z_\lambda(x)} \exp(\sum_i \lambda_i f_i(x, y)) \tag{2}$$

The parameter λ_i is introduced for the features. It represents the weight of feature f_i and indicates how important the feature is. $Z_\lambda(x)$ is a normalizing constant with every x satisfying $\sum_y p_y(y \mid x) = 1$. $Z_\lambda(x)$ is defined as:

$$Z_\lambda(x) = \sum_y \exp(\sum_i \lambda_i f_i(x, y)) \tag{3}$$

Sentiment related element extraction

Because of the lack of useful corpus, knowledge engineering method is adopted to organize a domain lexicon. We download web pages from pconline.com, dangdang.com, and extract structural information like product attributes to construct a domain lexicon. Then, a sentiment word list in Hownet is used for sentiment unit finding.

Sentiment classification

An improved Maximum Entropy with priors is proposed to classify sentences. The sentiment or polarity is decided according to the proportion of the sentences' sentiment or polarity.

First, twenty thousand sentences are used for training samples. Unigram and bigram features are used in training process. The term frequency in training samples is used as its weight. Parameters of the improved Maximum Entropy are adjusted optimally by training samples.

Then, after testing corpus preprocessing, segmentation and part-of-speech tagging, each text is partitioned in accordance with its punctuations, such as comma, period, and semicolon. The improved Maximum Entropy is applied to classify the sentiment or polarity of these sentences. Hence, sentiment information for each text is as follow:

- The total number of sentence in text, denoted as *SenNum*.
- The number of sentence with sentiment, denoted as *SenSubNum*;
- The number of sentence with Positive sentiment, denoted as *SenPosNum*;
- The number of sentence with Negative sentiment, denoted as *SenNegNum*;
- The sum of all sentence confidence in the text determined by the improved Maximum Entropy, denoted as *SenSubSum*;
- The maximum confidence in the text, denoted as *SenSubMax*;
- The sum of positive sentence confidence, denoted as *SenPosSum*;

The sentiment text is ranked by its text confidence. The text confidence is denoted as *TextSub* and defined as followed:

$$TextSub = (SenSubMax + SenSubSum / SenSum) / \max TextSub \tag{4}$$

maxTextSub is the maximum confidence in the testing corpus. It is a normalizer that makes the scope of *Textsub* to be [0,1].

Similarly, the polarity is decided by:

$$PosScale = SenPosNum / SenSubNum \tag{5}$$

PosScale is sentiment index, and it is used in:

$$TextPol = \begin{cases} -1 & if \quad PosScale < 0.4 \\ 1 & if \quad PosScale > 0.6 \\ 0 & otherwise \end{cases} \tag{6}$$

TextPol is the sentiment tag. "-1" is for negative. "1" is for positive. And, "0" is neutral.

Sentiment retrieval

Indri is applied as for ad-hoc text retrieval. Text sentiment classification is used for retrieval texts. And, the results are ranked by its correlation confidence and sentiment confidence. The correlation confidence with a query term by Indri is denoted as *corScore*.

Because the corpus retrieval by Indri contains some texts, which are related with the query term but are not sentiment texts, (6) is revised as:

$$TextOrientation = \begin{cases} 0 & if \quad PosScale = \#DIV/0! \\ 2 & if \quad PosScale < 0.4 \\ 4 & if \quad PosScale > 0.6 \\ 3 & otherwise \end{cases} \tag{7}$$

TextOrientation is the sentiment tag set for retrieval texts. "0" is related but without sentiment. "2" is negative. "3" is neutral. "4" is positive.

The retrieval texts are ranked by :

$$retrievalScore = corSore + TextSub \tag{8}$$

retrievalScore is index for both the query correlation confidence and sentiment confidence. *TextSub* is calculated by (4). Since *TextSub* ranges in [0,1], which is much less than the scope of *corScore*, the results is not influenced by *TextSub*.

Results and Analysis

The smart sentiment analysis system was tested in COAE2008 [13]. The evaluation contains six tasks. Task1 and task2 are Chinese sentiment word recognition. Task3 is sentiment related element extraction. Task4 and task5 are text sentiment classification. Task6 is text sentiment retrieval.

Task1 and Task2

The results of Task1 and task2 is shown in Table 1 and Table 2. run1 takes word and its POS as features. run2 uses word, word position and its POS as its features.

Table 1 Results of Task1

run	P@100	P@1000	Right_by_Lexicon
Best	1	0.984	3097
SSAS-task1-run1	0.91	0.691	2581
SSAS-task1-run2	**0.97**	**0.956**	2354
Median	0.925	0.9335	2602

Table 2 Results of Task2

run	P@100	P@1000	POS_Right_by_Lexicon	NEG_Right_by_Lexicon
Best	0.89	0.925	2966	3095
SSAS-task1-run1	0.7	0.578	1842	2522
SSAS-task1-run2	0.51	0.52	1755	2529
Median	0.795	0.738	2627	3006

It can be seen from Table1 that the most of the results in task1 is close to the median. Compared to run1, run2 combined with the word position is better than the median in P@100 and P@1000. That means the word position is very important in sentiment analysis. The results of Task2 are worse than median. We speculate that the features are not enough in LMR template. In the future, more effective features will be discussed in our work.

Task3

Table 3 Results of Task3

		Results of attribute extraction			Results of attribute extraction		
runid	DataSet	Strict			Lenient		
		Precision	Recall	F-measure	Precision	Recall	F-measure
SSAS	Car	0.2857	0.03767	0.06657	0.5844	0.09247	0.1597
	Camera	0.2653	0.09403	0.1388	0.5529	0.206	0.3002
	Phone	0.2549	0.04563	0.07741	0.576	0.1207	0.1995
	NoteBook	0.3275	0.09461	0.1468	**0.6268**	0.2096	0.3141
	All	0.2771	0.06181	0.1011	0.5769	0.1451	0.2319
Average of all results		0.36325	0.25708	0.28047	0.61034	0.44948	0.49103
Best of all results		0.5966	0.4577	0.4419	0.7968	0.6971	0.6786

The results are worse than the median. The testing corpus contains large amount of oral Chinese, but the Hownet and the domain lexicon are written Chinese. The difference makes the results inaccuracy. We will explore statistic methods in the future in sentiment related element extraction.

Task4 and task5

Results of task4 and task5 are shown in Table 4 and Table 5.

Table 4 Results of task4

run	Raccuracy	Acc10	Acc1000	accuracy_by_lexicon
Best	0.363	1	0.698	0.95075
SSAS-task4-run1	0.1677	**0.7**	0.273	0.7585
Median	0.2488	0.4	0.387	0.84

Table 5 Results of Task5

run	Raccuracy	Acc10	Acc1000	accuracy_by_lexicon
Best	0.1981	0.8	0.397	0.905
SSAS-task5-run1	**0.1622**	**0.7**	**0.358**	0.5725
Median	0.16165	0.4	0.323	0.6055

Results of task4 are worse than the median. However, the results of task5 is better than the median. We infer that the error is caused by those training samples, which has no sentiment.

Task6

Table 6 Results of Task6

run	MAP	R-Prec	bPref	P@10
Best	0.444	0.4999	0.4817	0.8
SSAS-task6-run1	**0.4178**	**0.4759**	**0.4455**	0.8
Median	0.3686	0.4477	0.4069	0.69

Results of task6 are shown in Table 6. Most of the results are better than the median, even the best. That means the SSAS we designed has high performance in opinion retrieval or text sentiment retrieval.

Conclusions and future work

This paper introduces a smart sentiment analysis system in three levels: word, sentence, text, respectively. SSAS satisfies three aspects of sentiment analysis requirement. The promising results in Chinese Opinion Analysis Evaluation in 2008 are given.

Our work is still worth further studying. In the future, we will explore more effective features in our system and try different methods in these sentiment analysis tasks.

Supported by the Fundamental Research Funds for the Central Universities

References

[1] Zhao YY, Qin B, Liu T, “A Survey of Sentiment Analysis”, Journal of Software, Vol.21,No. 8, 1834-1848, 2010.
[2] Huang XJ, Zhao J. Sentiment Analysis for Chinese Text. Communications of CCF, 4(2), 2008.
[3] Rao D, Ravichandran D. Semi-supervised Polarity Lexicon Induction. In: Proceedings of EACL-2009, p675-682, 2009.
[4] Wiebe J. Learning Subjective Adjectives from Corpora. In: Proceedings of AAAI. 2000.
[5] Kim SM, Hovy E. Automatic Detection of Opinion Bearing Words and Sentences. In: Proceedings of IJCNLP-2005, p61-66, 2005.
[6] Yao TF, Peng SW. A Study of the Classification approach for Chinese Subjective and Objective Texts. In: Proceedings of the NCIRCS-2007, p117-123, 2007.(in Chinese with English abstract)
[7] Pang B, Lee L. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In: Proceedings of ACL-2004, p271-278, 2004.
[8] Hu MQ, Liu B. Mining and Summarizing Customer Reviews. In: Proceedings of KDD-2004, p168-177, 2004.
[9] Turney P. Thumbs up Or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of ACL-2002, p417-424, 2002.
[10] Pang B , Lee L, Vaithyanathan S. Thumbs Up? Sentiment Classification Using Machine Learning Techniques. In: Proceedings of EMNLP-2002, p79-86, 2002.
[11] Information on http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG/
[12] Information on http://research.nii.ac.jp/ntcir/
[13] Zhao J, Xu HB, Huang XJ. Overview of Chinese Opinion Analysis Evaluation 2008. COAE2008, 1-20, 2008.