Chinese Sentiment Analysis Using Maximum Entropy

Huey Yee Lee
ePulze Sdn Bhd, C-41-2, Block C,
Jaya One, No 72a, Jalan Universiti,
46200, Petaling Jaya, Selangor Darul
Ehsan, Malaysia.
leehueyyee@hotmail.com

Hemnaath Renganathan
ePulze Sdn Bhd, C-41-2, Block C,
Jaya One, No 72a, Jalan Universiti,
46200, Petaling Jaya, Selangor Darul
Ehsan, Malaysia.
hemnaath@live.com

Abstract

This paper presents the use of Maximum Entropy technique for Chinese sentiment analysis. Berger, Vincent and Stephen (1996) prove that Maximum Entropy is a technique that is effective in a number of natural language processing applications. In this paper, Maximum Entropy classification is used to estimating the polarity of given comments of from electronic product. These messages are classified into either positive or negative. Apart from presenting the results obtained via Maximum Entropy technique, we also analyze the feature selection and pre-processing of the comments for training and testing purpose.

1 Introduction

Nowadays, there are lots of comments about products, movies, hotels, or restaurants available in on-line documents such as blogs, Facebook, Twitter and Amazon. As part of the effort to better classify information for users, researchers have actively investigated sentiment analysis. Sentiment analysis, attempts to gather the overall opinion towards the comments – for example, whether a product feedback is positive or negative.

This system is useful for consumers who want to research the sentiment of products before making a purchase, and companies that want to monitor the public opinion of their products. Labeling these comments correspondently, their sentiment would provide succinct summaries to both readers and organizations.

Research in sentiment analysis has been done mainly using the English language. In this paper however, we examine the effectiveness of applying machine learning techniques which is Maximum Entropy classification to the Chinese sentiment analysis. Maximum Entropy is a technique that uses probability distribution estimation and widely used for a variety of natural language tasks. The challenging aspect that is different from ordinary sentiment analysis is that the comments to be analysis are in Chinese language.

Section 2 presents the related work and Section 3 describes the idea of using Maximum Entropy classification. Meanwhile Section 4 discusses the pre-processing methods and feature selection techniques for constructing Maximum Entropy model which include segmentation, conjunction rules, stop words and punctuation elimination, negation, and lastly is keyword-based. The experimental result will be present on Section 5 and finally the conclusion and future work in Section 6.

2 Related Work

There has been a wide research effort on analyzing sentiment in languages other than English by applying bilingual resources and machine translation techniques to employ the sentiment analysis approach existing for English. For an example, Yao *et al.* (2006) had proposed a method of determining sentiment orientation of Chinese words using a bilingual lexicon and achieve precision and recall of 92%.

So far, many researchers have conducted on sentiment classification. These researches have fallen into two categories which is machine learning techniques and semantic orientation. Machine learning technique attempt to train a sentiment classifier based on occurrence frequencies of the various words in the documents. There are several Machine learning

Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011, pages 89–93,
Chiang Mai, Thailand, November 13, 2011.

methods, such as Naïve Bayes (NB), Maximum Entropy (ME), Support Vector Machine (SVM), unsupervised learning and etc. Hemnaath and Low (2010) propose sentiment analysis using Maximum Entropy and Support Vector Machine.

Meanwhile, semantic orientation is to classify words into two classes, such as 'positive' or 'negative', and then count in the overall score of the text. Yuen *et al.* (2004) presents a method for inferring the semantic orientation of a Chinese word from their association with strongly-polarized Chinese morphemes.

Among several machine learning algorithms, Maximum Entropy is the convenient for natural language processing, since it allows the unrestricted use of contextual features, and combines them in a principled way. Besides, Wang and Acero (2007) also mentioned that the Maximum Entropy model has a convex objective function and consequently they converge to a global optimum with respect to a training set. Because of these advantages, Maximum Entropy Classification is selected to develop Chinese sentiment analysis.

3 Maximum Entropy Classification

Maximum Entropy is a machine learning method based on empirical data and provides the probabilities for which sentence belongs to a particular class. Kamal *et al.* (1999) found that Maximum Entropy works better than Naïve Bayes for their experiment. The fundamental principle of Maximum Entropy is that the distribution should be uniform. Besides, constraints for the model that characterize the class-specific expectations for the distribution are derived from labeled training data.

When using maximum Entropy, the first step is to identify a set of feature functions which define a category. For an example, in case of documents features could be the words that belong to the documents in that category; for each feature, measure its expected value over the training data and treat this as constraint for the model distribution. Maximum Entropy models are feature-based models. In a two-class scenario, it is the same as using logistic regression which corresponds to the Maximum Entropy classifier for independent observations.

Like any learning technique, the outputs generated from the process are relied on the given dataset of input. The dataset is analyzed, and from it, a model is generated, encapsulating all the rules about the process that could be inferred from the dataset. This model is then used to predict the output of the process, when supplied with sets of input that is not found in the sample dataset.

Each of these rows of dataset represents a training event. Each training event has an outcome which consists of the predicates and lead to the outcome of the event. Each time it runs, the model is built from the training dataset. In order to train a classifier, it's usually requires several stage of pre-processing to hand-label the training data.

4 Pre-processing

4.1 Stage 1: Segmentation

Unlike western languages, normally sentences in Chinese text are represented by strings of Chinese characters without spaces between words. Therefore, Chinese sentences are ongoing problems in information retrieval (IR) and computational linguistics. Each Chinese character represents a meaning, while two or more characters combined to form a word that has different meanings. Therefore, segmentation needed to retrieve the meaning of the sentence. For an example,

今天的天气很好 (Today's weather is very nice)

If the sentence is separated by characters, each character has their own meaning.

今	This
天	Day
的	The
天	Sky
气	Gas
很	Very
好	Nice

Meanwhile, by using segmentation, it can identify which character should be combined to form the word and carry the actual meaning of the sentences.

After segmentation: 今天 的 天气 很 好

今天	Today
的	The
天气	Weather
很	Very
好	Nice

Thus, to process any word-based or token-based linguistic processing on Chinese,

segmentation plays an important role in determining word boundaries. Wang and Christopher (2007) previously mentioned indexing of Chinese document is impossible without a proper segmentation algorithm. Before either task can take place, the sentence must be broken into tokens; it must be segmented and it is the necessary stage in pre-processing of Chinese sentiment analysis.

4.2 Stage 2: Conjunction Rules

The main purpose of applying conjunction rules is to extract the accurate meaning or expression from a given sentence using grammar rules. Generally, a sentence only expresses one opinion orientation unless there is some certain conjunction such as BUT, ALTHOUGH, HOWEVER, WHILE and etc word which changes the direction of the sentence. Conjunction rules explanations are shown as below.

1. Although (虽然，尽管，虽，虽说)

   Although (Phrase A), (Phrase B).

   E.g. 虽然这相机很好，可惜电池寿命很短。
   (Although this camera is nice, too bad has short battery life.)

   In this case, phrase A will be cut off, and phrase B will be remain as sentence sentiment.

2. But (但，但是，而，不过，却，可是，然而，只是，可是，可，只，然)

   (Phrase A), but (Phrase B).

   E.g. 这相机的外观不美，但很耐用。
   (The camera appearance is not beautiful, but very durable.)

   In this case, phrase A also will be cut off, and phrase B will be remain as sentence sentiment.

3. Although…, but.... (虽然…但是，虽然…可是，尽管…却，……)

   Although (Phrase A), (Phrase B),but (Phrase C).

   E.g. 虽然这相机很好，可惜电池寿命很短，但我还是喜欢用它。
   (Although this camera is nice, too bad has short battery life, but I still like it.)

   For this example, phrase A and phrase B will be cut off, while phrase C is remain as new sentence for sentiment.

By applying conjunction rules, the sentences become more understandable and straightforward; it is because having two polarities in a sentence which can affect the result of sentiment analysis. Hemnaath and Low (2010) proved that with conjunction rules, accuracy of sentiment analysis can be increased by approximately 5%.

4.3 Stage 3: Stop words and Punctuation Elimination

The next important stage in pre-processing of sentiment analysis is to simplify the text. Zou *et al.* (2006) claimed that in modern information retrieval system, effective indexing can be achieved by removal of stop words. Stop words are very common words that appear in the text that carry little meaning; they serve as a syntactic function but do not indicate subject matter.

For an example, words “and”, “of”, and “the” are appearing frequently in the document. They can affect the retrieval effectiveness because they have a very high frequency and tend to diminish the impact of frequency differences among less common words, thus affecting the training process in sentiment. Also, these stop words may result in a large amount of unproductive processing. The removal of the stop words and punctuation also changes the document length, subsequently affect the learning algorithm.

Those stop words and punctuation that having minor help in determining polarity of text can be removed. Ibrahim (2006) previously also showed that identifying a stop words list or a stoplist and eliminate them from text processing is essential to an information retrieval system.

4.4 Stage 4: Negation

One issue of accurate sentiment analysis identified in recent of research is negation detection. The treatment is very relevant for all NLP applications that involve deep text understanding. Li *et al*. (2010) showed that the negation word feature is an important feature for sentiment analysis. Negation needed to discriminate between factual and non-factual

information in information extraction for sentiment analysis which process the actual meaning of the texts.

This is the process by which a negation word, such as 'not' inverts the evaluative value of as affective word. For an example, 'not good' is similar to saying 'bad'. By adapting a technique proposed by Das and Chen (2001), a tag 'NOT_' was added to every word between a negation word and the first punctuation mark following the negation word. Applying this, a new corpus variation was obtained.

我 不 喜欢 这 电影 (I do not like this movie)
我 不_喜欢 这 电影 (I do NOT_like this movie)

In unigrams, the value of 'like' is positive, but there is a negation word 'not', therefore a 'NOT_' is replaced and joint with the consequent word. As a result, 'NOT_like' can indirectly affect the value of the word; subsequently affect the polarity of entire sentence. In Chinese, instead of 'NOT_' , '不_' was applied, and the list of supported negation includes '不', '不是', '没', '没有', '无', '别' and etc.

4.5 Stage 5: Keyword-based

Comparison of keywords is an extra feature for sentiment analysis. Kaji and Kitsuregawa (2007) mentioned recognizing polarity requires a list of polar words and phrases such as 'good', 'bad' and 'high performance' etc. At first, lists of positive and negative polarities keywords are obtained by using the NTU Sentiment Dictionary. Consequently, the numbers of positive keywords and negative keywords that appear in the input sentence are counted. The polarity with the higher count returns as an extra feature for sentiment analysis.

5 Experiment

5.1 The analysis data

Our test-corpus is derived from product reviews harvested from the website IT168, which can be downloaded from http://product.it168.com. All the reviews have been tagged by their authors as either positive or negative. The corpus consists of 10 sub-corpora containing a total of 7818 reviews, distributed between 10 product types which are monitor, mobile phone, digital camera, MP3 player, computer part, video camera, networking, office equipment, printer and computer peripheral.

From these reviews, 2909 of both positive and negative comments are used as training data, while 1000 comments for both polarities are used for testing purpose. In addition, the Entropy model is manually set to discriminate >=0.5 as positive and <0.5 as negative.

5.2 Experiment 1

Segmentation is an initial and compulsory stage in Chinese sentiment analysis, without applying any other pre-processing stage for training and testing data, the overall accuracy for sentiment analysis is 81.65%. Table 1 shows that by apply each pre-processing, the overall accuracy is increased compared to the one without any pre-processing. It proves that the pre-processing stages discussed are functional in sentiment analysis.

Pre-processing	Accuracy		
	positive	negative	overall
Segmentation	77.2%	86.1%	81.65%
Conjunction rules	73.0%	90.8%	81.90%
Stopwords & punctuation elimination	80.9%	85.5%	83.20%
Negation	81.4%	86.1%	83.75%
Keyword - based	78.9%	87.6%	83.25%

Table 1: Result of sentiment analysis by applying pre-processing steps separately.

5.3 Experiment 2

In this experiment, each review for both training and testing datasets is going through 5 stages of pre-processing according the sequences as shown in Table 2. Table 2 shows the result that the overall accuracy of sentiment analysis is increased stage by stage which is increased from 81.65% to 87.05%.

Pre-processing	Accuracy		
	positive	negative	overall
Segmentation	77.2%	86.1%	81.65%
Conjunction rules	73.0%	90.8%	81.90%

Stopwords & punctuation elimination	80.2%	86.3%	83.25%
Negation	85.1%	86.8%	85.95%
Keyword - based	87.3%	86.8%	87.05%

Table 2: Result of sentiment analysis by applying pre-processing steps in stages.

6 Conclusion and Future Work

In this paper, we explore the steps of pre-processing that can reduce the features and extract the polarity of the Chinese texts. It is proved Maximum Entropy classification can achieve high accuracy for classifying sentiment when using these steps. Empirical results from the experiments demonstrate the feasibility of our approach. Besides classifying positive and negative sentences, entropy model results that lie between 0.48 – 0.52 can be pronounced as neutral sentences.

In our future work, we plan to future improve the feature selection techniques for constructing Maximum Entropy model, which is Word Sense Disambiguation. We believe that identify the actual meaning of the words can help to increase the accuracy of Chinese sentiment analysis. In addition, our next work will draw on more heavily Chinese Natural Language Processing technique, such as Chinese parsing and semantic annotation. We look forward to addressing these challenges in future work.

References

Berger, A. L., Vincent, J. D., and Stephen, A. D. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics, 22(1)*: 39-71.

Das, S., and Chen, M. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. *8th Asia Pacific Finance Association Annual Conference (APFA).*

Hemnaath, R., and Low, B. W. 2010. Sentiment Analysis Using Maximum Entropy and Support Vector Machine. *Semantic Technology and Knowledge Engineering in 2010.* Kuching,Sarawak.

Ibrahim, A. E.-K. 2006. Effect of Stop Wprds Eliminationfor Arabic Information Retrievel: A Comparative Study. *International Journal of Computing & Information Sciences*, 119-133.

Kaji, N., and Kitsuregawa, M. 2007. Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational*, 1075–1083. Prague.

Kamal, N., John, L., and Andrew, M. 1999. Using Maximum Entropy for Text Classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering.*

Li, S., Zhang, H., Xu, W., Chen, G., and Guo, J. 2010. Exploiting Combined Multi-level Model for Document Sentiment Analysis. *2010 International Conference on Pattern Recognition*, 4141-4144.

Wang, F. L., and Christopher, C. Y. 2007. Mining Web Data for Chinese Segmentation. *Journal of The American Society For Information Science and Technology,* 58(12): 1820–1837.

Wang, Y. Y., and Acero, A. 2007. Maximum Entropy Model Parameterization with TF*IDF Weighted Vector Space Model. *IEEE Automatic Speech Recognition and Understanding Workshop*, 213-218. Kyoto, Japan: Institute of Electrical and Electronics Engineers, Inc.

Yao, J., Wu, G., Liu, J., and Zheng, Y. 2006. Using bilingual lexicon to judge sentiment orientation of Chinese words. *Computer and Information Technology, 2006. CIT '06. The Sixth IEEE International Conference.*

Yuen, R. W., Chan, T. Y., Lai, T. B., Kwong, O., and T'sou, B. K. 2004. Morpheme-based Derivation of Bipolar Semantic Orientation of Chinese Words. *Proceedings of the 20th international conference on Computational Linguistics,* 1008-1014. USA.