Active Learning on Sentiment Classification by Selecting Both Words and Documents

Shengfeng Ju and Shoushan Li

Natural Language Processing Lab, Soochow University,1 Shizi Street, Suzhou, China 215006
{shengfeng.ju,shoushan.li}@gmail.com

Abstract. Currently, sentiment analysis has become a hot research topic in the natural language processing (NLP) field as it is highly valuable for many real applications.. One basic task in sentiment analysis is sentiment classification which aims to predict the sentiment orientation (positive or negative) of a document. Current approaches to this problem are mainly based on supervised machine learning technologies. The main drawback of such approaches lies in their needs of large amounts of labeled data. How to reduce the annotation cost has become an important issue in sentiment classification. In this study, we propose a novel active learning approach to select both "informative" word and document samples for annotation. Experimental results show that our approach apparently outperforms random selection or uncertainty sampling on documents.

Keywords: Chinese information processing, sentiment analysis, active learning, dual supervision.

1 Introduction

With the rapid development of Internet, the information on the Internet is more and more abundant. The various comments are valuable information to both customers and producers, which can be used for learning the satisfaction degree of the product or service. In order to acquire and analyze this kind of subjective information automatically, text sentiment analysis has got a rapid development which has aroused close attention from both academic and business research groups [1]. Sentiment Classification is a basic task of sentiment analysis, which is focused on the classification of semantic orientation, in other words, to classify the sentiment orientation as sentimental categories, such as positive, neutral and negative.

The research on sentiment classification has been carried out for many years, and currently the main methods for this task are generally based on supervised learning [1-2]. However, a significant disadvantage of supervised learning is that it requires a large amount of labeled samples during its training process while obtaining large amount of labeled samples is a very time-consuming and laborious work. Therefore, how to reduce the sample scale to be labeled and maintain a desired classification performance is an important issue which is really worth deep research. To achieve this, active learning is a method which can reduce the scale of

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 49–57, 2013.

labeling samples by choosing some “high-quality” samples actively for manual annotation, and it results in using the minimum number of labeled samples while keeps the classification performance at a high level.

The traditional active learning approaches focus on how to select the documents which could contribute most to the classification work, and select the most uncertain documents for the classifier usually before. Meanwhile the latest research shows that adding information of the words during active learning process helps improve the quality of the final classification [3], that’s to say the words with additional information (usually emotional words) make great contribution to the classification results. For instance, the word “comfortable” can be labeled as a positive word while “bad” can be labeled as a negative word in the hotel fields, therefore we can improve the classification performance with annotation of such words which contain additional information. In other words, we can have better classification performance when have both "informative" word and document samples for annotation. However, the cost of word and document annotation is different: the complexity of document makes the annotation of it much more costly of time and labor than that of word. In order to save the annotation cost, it is possibly preferable to choose the accurate sentiment words for effective annotation. In this study, based on the unit annotation scale, we calculate the information of every word and document respectively when the same annotation scale is given, then select the most helpful documents and words for manual annotation.

The remaining part of this paper is organized as follows: In the second section, we will introduce the previous research of active learning in sentiment classification; In the third section, we will introduce the classification approaches based on the coordination of word and document; The fourth section describes the active learning approaches based on collaborative selection of word and document; The fifth section i the results and analysis of the experiments and the last section is the conclusion of this paper.

2 Related Work

Recently, sentiment classification has gradually become a hot research topic in natural language processing. [3] employ the machine learning approaches based on supervised learning to conduct sentiment classification for the first time. The following studies aim at improving the performance of the supervised learning with various methods, such as extracting the subjective sentences [4], looking for the higher level classification character [5] and taking advantage of the theme related information [6]. Generally, sentiment classification research has been carried out on different text particle sizes, for example, document level [7-11] and word level [12-13]. This paper mainly focuses on document level, but also uses sentimental information for help in word level.

Active learning is an important research branch in machine learning for a long history[14-15]. The active learning algorithms can be roughly divided into three categories: The first category is to select the text which can reduce the inaccuracy

of the classifier mostly for manual annotation, like error reduction sampling approach [16]; The second category is to select the most uncertain texts of the classification results from the classifier for manual annotation, like uncertainty sampling approach [17]; The third category is based on the differences of predict result from multiple classifiers, like Query By Committee approach (QBC) [18].

In the research of sentiment classification, the relative achievements of active learning on both word and document are rare so far, compared to that of the traditional active learning on only document. Melville and Sindhwani (2009)[19-21] compare the performance of annotating both word and document with document only, and found out that the former one had better classification performance. However, they do not fully consider the different annotation cost of word and document during the selection procedure of word and document. Furthermore, the word polarity categories are not true label but a kind of simulation approach by feature selection method where the polarity categories of some words are not correct. Relatively speaking, we have fully considered the annotation cost of word and document in our approach and used the true polarity category labels in the experiments.

3 Sentiment Classification Method of Learning from Both Words and Documents

Most of the existing classification approaches are performed by training annotation of document. To incorporate the classification knowledge in labeled words, we adopt the classification method proposed by [19] that focus on the combination of document and dictionary when annotate both word and document. This method is based on Bayes classification method [22] and it degenerates to the ordinary simple Bayes when the dictionary is empty.

Simple Bayes classifier is a classification approach based on Bayes' principle and it has simple model and high operating speed as one of the most popular machine learning approaches. It has an assumption as prerequisite that the document features are independent of each other in a given document. It calculates maximum likelihood estimates to get the category of the document. The calculation formula is as below:

$$\arg\max_{c_i} P(c_i)\prod_k P(w_k \mid c_i) \tag{1}$$

Where $P_e(w_k \mid c_i)$ means conditional probability of the words w_k from training corpus to compute the document belongs to the sentiment category, $P_f(w_k \mid c_i)$ is the posterior probability of computing document belongs to the sentiment category by sentiment dictionary. α is weight ratio for both sides. We set α as 0.5 and $P(c_i)$ (prior probability of positive and negative category) as 0.5.

When learning the weight of $P_e(w_k \mid c_i)$, we calculate the estimate of the Laplasse of conditional probability word w_k in category c_i of document by ordinary simple Bayes approach. While the detailed derivation process of $P_f(w_k \mid c_i)$ has been completed by Melville etc. (2009)[12], so we just describe the calculation method of $P_f(w_k \mid c_i)$, which requires the parameters as below:

V : Collection of all words
P : Collection of positive words
N : Collection of negative words
U : Collection of neutral words , that's to say , $(V-(P+N))$
m : Number of words in collection V, that's to say , $|V|$
p : Number of words in collection P, that's to say , $|P|$
n : Number of words in collection N, that's to say , $|N|$

All the words can be divided into three categories: positive, negative and neutral. The neutral words contain two kinds of words: one are those included in collection V but haven't been annotated and the other are those have been annotated annually but can't be determined whether positive or negative. We denote the probability of positive word w_+ in positive document as $P_f(w_+ \mid +)$ and the probability of negative word w_- in negative document as $P_f(w_+ \mid +)$. Similarly, we denote the probability of neutral words w_u in positive document and negative document as $P_f(w_u \mid +)$ and $P_f(w_u \mid -)$ respectively. The detailed calculation formulas are listed as below:

$$P_f(w_+ \mid +) = P_f(w_- \mid -) = \frac{1}{p+n} \tag{2}$$

$$P_f(w_+ \mid -) = P_f(w_- \mid +) = \frac{1}{p+n} \times \frac{1}{r} \tag{3}$$

$$P_f(w_u \mid +) = \frac{n(1-1/r)}{(p+n)(m-p-n)} \tag{4}$$

$$P_f(w_u \mid -) = \frac{p(1-1/r)}{(p+n)(m-p-n)} \tag{5}$$

The parameter r means the proportion of the probability of positive words in positive documents and negative documents. In this study, we set the value to 100, i.e., r=100, as suggested by [19].

4 Active Learning with Collaborative Selection on Both Words and Documents

4.1 Annotation Costs

Previous active learning methods mainly focus on the selection of “informative" document, while the active learning approach in this paper annotate words and documents simultaneously. However, the annotation cost of word and document is different. Therefore, it is necessary to obtain the detailed annotation cost of each word and document.

In order to compare the cost of time between word and document, we propose the concept of unit annotation time and unit annotation scale:

Unit annotation time: the average annotation time of a document.

Unit annotation scale: the number of annotation words in unit annotation time.

So, we can annotate multiple words during the unit annotation time. We random sorted all the documents relating to package and hotel field together with all the words relating to these two fields, and assigned the sorted documents to two students (A and B) with Bachelor degree for manual annotation, then recorded the number of annotation words or documents in 15 minutes. Table 1 shows the final results as below:

Table 1. Annotation ratio

Annotation student	A	B
Number of document annotated	116	102
Number of word annotated	1848	1748

According to the table above, we can get the annotation overhead ratio of word and document: (1848+1748) / (116+102) =16.5. So we set the number of words can be annotated in unit annotation time as 16.5.

4.2 Selection of Sentiment Words

Sentiment words usually play a key role in the classification process of sentiment classification, so we should try to select sentiment words for annotation. Sentiment words can be divided into two categories: positive words and negative words. However, there are a large number of words are neutral words which cost much more annotation effort compared with sentiment words. Therefore, the effective selection of sentiment words for manual annotation can help save annotation cost very well.

Generally speaking, whether the word is a sentiment word is closely related to the part of speech. For instance, the probability of the adjective is sentiment word is apparently higher than that of the noun. Therefore, more attention should be paid to the adjective rather than noun while searching for sentiment words.

In order to calculate the information of the part of speech to approve that the adjective has the biggest chance to be a sentiment word, we had an experiment that

manually annotated 200 words randomly from all categories of part of speech respectively, and recorded the probability of the word is sentiment as weight. Table 2 shows the final results as below:

In addition, the frequency of a word in a document is an important basis of the importance of the word, in other words, the more a word appears in the document, the more it influences the document and the more important the word is.

$$V\left(POS\right)=\frac{\text{number of the words which are both emotional words and POS}}{\text{number of the all words which part of speech are POS}} \tag{6}$$

$$Weight(w)=V\left(POS(w)\right)\times\log(F(w)) \tag{7}$$

Note: $POS(w)$ is part of speech, $V(POS(w))$ is the weight of the part of speech which can be referred from Table 2. For instance, if the word is adjective (adj.), then $V(adj.)=0.77$. $F(w)$ means the total number of documents where the word appears.

Table 2. Statistical information of part of speech

part of speech	adjective	verb	noun	others
Weight of sentiment	0.77	0.36	0.11	0.08

4.3 Sort of Word and Document Based on Weight

We can obtain the weight of every word according to formula (7) when selecting word and document for manual annotation. In addition, we can calculate the weight of every word in each document, and it will help calculate the weight of each document.

$$Weight(d)=\frac{\sum_{w\in d}Weight(w)}{k\log\left(L(d)\right)} \tag{6}$$

$Weight(d)$ is the weight of document, $\sum_{w\in d}Weight(w)$ is the sum weight of the words in document, k is a constant which means the annotation scale of words in unit annotation time, and we set 16.5 in this paper. L is the length the document. We prefer the document with shorter length while with the same weight, because the longer the document is, the higher probability of containing words without classification information it has. These words which contain no classification information will cause noise, so we set the weight of document divided by the length of the log.

We can sort the words and documents by calculation of weight of them with formula (7) and formula (8), then select the words and documents with maximum weight to annotate.

In general, our procedure is as follows. 1) Input large number of unlabeled documents, use segmentation and part of speech annotation tool to divide words and annotate them. 2) Obtain the part of speech and occurrence frequency of each word.

3) Calculate the weight of word and document with formula (7&8). 4) Select the words and documents with maximum weight for manual annotation. 5) Train the model using the labeled word and document, then classify the test samples by Pooling Multinomials classifier.

5 Experiments

We collected the Chinese sentiment text from package and hotel field, and the corpus was from the comments on amazon website. Each field contained 2400 samples, in which 1200 positive and 1200 negative. We selected 400 positive and 400 negative samples as test samples, while all the rest as training samples. We used software called ICTCLAS by Chinese Academy of Sciences Institute for segmentation and part of speech annotation firstly. The evaluating indicator is standard accuracy in the experiment.

The comparison of different approaches in this paper:

Random Sampling (RAND): select the document randomly for manual annotation;

Uncertainty Sampling (UNCE): select the document with most uncertainty for manual annotation, it's based on the result by Bayes classifier;

Document-word co-selecting (DW): select the document and word with highest weight for manual annotation, it's based on the result of formula (8).

Since the first approach of selection is random, we performed five times and took the average results to make the final statistics stable.

Fig 1 and Fig 2 shows the performance of the three different selection categories in package and hotel field, and we can find that the performance of our DW approach is apparently better than the other two approaches when the number of text samples is small. For instance, our active learning approach based on co-selecting on words and documents performed much better than the other two approaches when only 20 or 40 unit annotation time cost. This improvement was six percentages in package field while even reached to more than 10 percentages in hotel field.

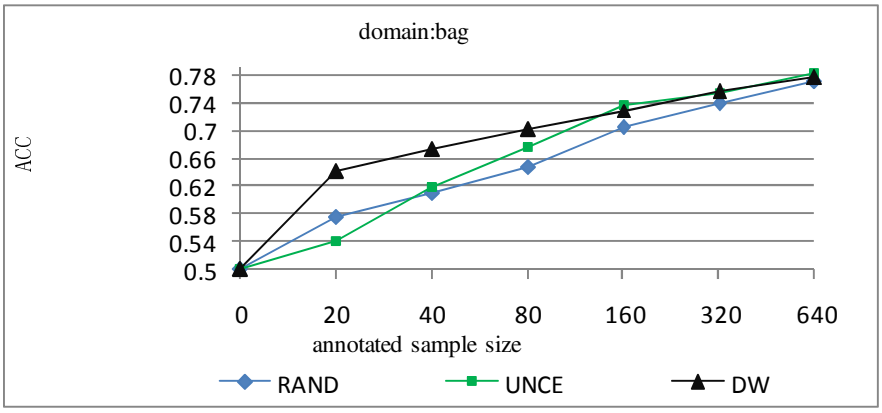

Fig. 1. Classification performance of the three approaches in each field respectively

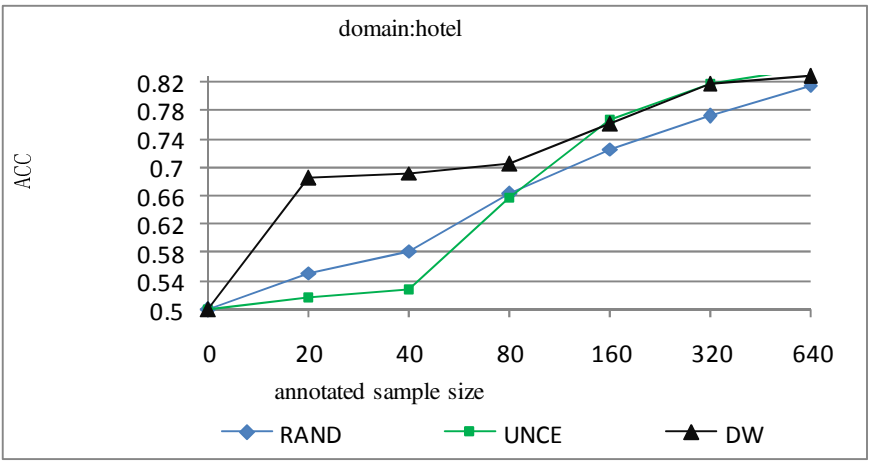

Fig. 2. Classification performance of the three approaches in each field respectively

After investigation of selected samples, we can find that the samples contained large amount of sentiment words at the beginning stage. This result shows the importance of annotating some sentiment words as classification resource at the beginning stage. As the sentiment scale increased to some extent, our approach performed similarly with UNCE, but still had obvious advantage on RAND.

6 Conclusion

This paper proposes a novel active learning approach for sentiment classification, where both "informative" words and documents are actively selected for training the classifier. To enable selecting words and documents simultaneously, we evaluate their annotation costs and informativeness values. The experimental results demonstrate that the proposed active learning approach greatly reduces the annotation cost and significantly outperforms random sampling and uncertainty sampling.

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of EMNLP 2002, pp. 79–86 (2002)
2. Li, S., Zong, C.: Multi-domain Sentiment Classification (short paper). In: Proceedings of ACL 2008, pp. 257–260 (2008)
3. Melville, P., Gryc, W., Lawrence, R.: Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification. In: Proceedings of KDD 2009, pp. 1275–1284 (2009)
4. Pang, B., Lee, L.: A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts. In: Proceedings of ACL 2004, pp. 271–278 (2004)
5. Riloff, E., Patwardhan, S., Wiebe, J.: Feature Subsumption for Opinion Analysis. In: Proceedings of EMNLP 2006, pp. 440–448 (2006)

6. McDonald, R., Hannan, K., Neylon, T., Wells, M., Reynar, J.: Structured Models for Fine-to-coarse Sentiment Analysis. In: Proceedings of ACL 2007, pp. 432–439 (2007)
7. Cui, H., Mittal, V., Datar, M.: Comparative Experiments on Sentiment Classification for Online Product Reviews. In: Proceedings of AAAI 2006, pp. 1265–1270 (2006)
8. Li, S., Huang, C., Zong, C.: Multi-domain Sentiment Classification with Classifier Combination. Journal of Computer Science and Technology (JCST) 26(1), 25–33 (2011)
9. Li, S., Lee, S., Chen, Y., Huang, C., Zhou, G.: Sentiment Classification and Polarity Shifting. In: Proceeding of COLING 2010, pp. 635–643 (2010b)
10. Li, S., Huang, C., Zhou, G., Lee, S.: Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification. In: Proceedings of ACL 2010, pp. 414–423 (2010a)
11. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis: Foundations and Trends. Information Retrieval 2(12), 1–135 (2008)
12. Hatzivassiloglou, V., McKeown, K.: Predicting the Semantic Orientation of Adjectives. In: Proceedings of ACL 1997, pp. 174–181 (1997)
13. Wiebe, J.: Learning Subjective Adjectives from Corpora. In: Proceedings of AAAI 2000 (2000)
14. McCallum, A., Nigam, K.: Employing EM in pool-based active learning for text classification. In: Proceedings of ICML 1998, pp. 350–358 (1998)
15. Long, J., Yin, J., Zhu, E., Zhao, W.: Active learning research. Research and Development of Computer 45, 300–304 (2008)
16. Roy, N., McCallum, A.: Toward Optimal Active Learning through Sampling Estimation of Error Reduction. In: Proceedings of ICML 2001, pp. 441–448 (2001)
17. Lewis, D., Gale, W.: Training Text Classifiers by Uncertainty Sampling. In: Proceedings of SIGIR 1994, pp. 3–12 (1994)
18. Argamon-Engleson, S., Dagan, I.: Committee-Based Sample Selection For Probabilistic Classifiers. Journal of Artificial Intelligence Research, 335–360 (1999)
19. Melville, P., Sindhwani, V.: Active Dual Supervision: Reducing the Cost of Annotating Examples and Features. In: Proceedings of NAACL 2009, pp. 49–57 (2009)
20. Sindhwani, V., Melville, P.: Document-Word Co-Regularization for Semi-supervised Sentiment Analysis. In: Proceedings of ICDM 2008, pp. 1025–1030 (2008)
21. Sindhwani, V., Hu, J., Mojsilovic, A.: Regularized co-clustering with dual supervision. In: NIPS, pp. 1505–1512 (2008)
22. Zong, C.: Statistical natural language processing. Tsinghua University Publishing (2008)