Classifying Chinese Texts in Two Steps

Xinghua Fan[1, 2, 3], Maosong Sun[1], Key-sun Choi[3], and Qin Zhang[2]

[1] State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China
fanxh@tsinghua.org.cn, sms@tsinghua.edu.cn
[2] State Intellectual Property Office of P.R. China, Beijing, 100088, China
zhangqin@sipo.gov.cn
[3] Computer Science Division, Korterm, KAIST, 373-1 Guseong-dong Yuseong-gu, Daejeon 305-701, Korea
kschoi@cs.kaist.ac.kr

Abstract. This paper proposes a two-step method for Chinese text categorization (TC). In the first step, a Naïve Bayesian classifier is used to fix the fuzzy area between two categories, and, in the second step, the classifier with more subtle and powerful features is used to deal with documents in the fuzzy area, which are thought of being unreliable in the first step. The preliminary experiment validated the soundness of this method. Then, the method is extended from two-class TC to multi-class TC. In this two-step framework, we try to further improve the classifier by taking the dependences among features into consideration in the second step, resulting in a Causality Naïve Bayesian Classifier.

1 Introduction

Text categorization (TC) is a task of assigning one or multiple predefined category labels to natural language texts. To deal with this sophisticated task, a variety of statistical classification methods and machine learning techniques have been exploited intensively[1], including the Naïve Bayesian (NB) classifier [2], the Vector Space Model (VSM)-based classifier [3], the example-based classifier [4], and the Support Vector Machine [5].

Text filtering is a basic type of text categorization (two-class TC). It can find many real-life applications [6], a typical one is the ill information filtering, such as erotic information and garbage information filtering on the web, in e-mails and in short messages of mobile phone. It is obvious that this sort of information should be carefully controlled. On the other hand, the filtering performance using the existing methodologies is still not satisfactory in general. The reason lies in that there exist a number of documents with high degree of ambiguity, from the TC point of view, in a document collection, that is, there is a fuzzy area across the border of two classes (for the sake of expression, we call the class consisting of the ill information-related texts, or, the negative samples, the category of TARGET, and, the class consisting of the ill information-not-related texts, or, the positive samples, the category of Non-TARGET). Some documents in one category may have great similarities with some other documents in the other category, for example, a lot of words concerning love

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 302 – 313, 2005.

story and sex are likely appear in both negative samples and positive samples if the filtering target is erotic information. We observe that most of the classification errors come from the documents falling into the fuzzy area between two categories.

The idea of this paper is inspired by the fuzzy area between categories. A two-step TC method is thus proposed: in the first step, a classifier is used to fix the fuzzy area between categories; in the second step, a classifier (probably the same as that in the first step) with more subtle and powerful features is used to deal with documents in the fuzzy area which are thought of being unreliable in the first step. Experimental results validate the soundness of this method. Then we extend it from two-class TC to multi-class TC. Furthermore, in this two-step framework, we try to improve the classifier by taking the dependences among features into consideration in the second step, resulting in a Causality Naïve Bayesian Classifier.

This paper is organized as follows: Section 2 describes the two-step method in the context of two-class Chinese TC; Section 3 extends it to multi-class TC; Section 4 introduces the Causality Naïve Bayesian Classifier; and Section 5 is conclusions.

2 Basic Idea: A Two-Step Approach to Text Categorization

2.1 Fix the Fuzzy Area Between Categories by the Naïve Bayesian Classifier

We use the Naïve Bayesian Classifier to fix the fuzzy area in the first step. For a document represented by a binary-valued vector $d = (W_1, W_2, \ldots, W_{|D|})$, the two-class Naïve Bayesian Classifier is given as follows:

$$\begin{aligned} f(d) &= \log \frac{\Pr\{c_1|d\}}{\Pr\{c_2|d\}} \\ &= \log \frac{\Pr\{c_1\}}{\Pr\{c_2\}} + \sum_{k=1}^{|D|} \log \frac{1-p_{k1}}{1-p_{k2}} + \sum_{k=1}^{|D|} W_k \log \frac{p_{k1}}{1-p_{k1}} - \sum_{k=1}^{|D|} W_k \log \frac{p_{k2}}{1-p_{k2}} \end{aligned} \quad (1)$$

where $\Pr\{\bullet\}$ is the probability that event $\{\bullet\}$ occurs, c_i is category i, and $p_{ki}=\Pr\{W_k=1|c_i\}$ (i=1,2). If $f(d) \geq 0$, the document d will be assigned the category label c_1, otherwise, c_2.

Let:

$$Con = \log \frac{\Pr\{c_1\}}{\Pr\{c_2\}} + \sum_{k=1}^{|D|} \log \frac{1-p_{k1}}{1-p_{k2}} \quad (2)$$

$$X = \sum_{k=1}^{|D|} W_k \log \frac{p_{k1}}{1-p_{k1}} \quad (3)$$

$$Y = \sum_{k=1}^{|D|} W_k \log \frac{p_{k2}}{1-p_{k2}} \quad (4)$$

where *Con* is a constant relevant only to the training set, X and Y are the measures that the document *d* belongs to categories c_1 and c_2 respectively.

We rewrite (1) as:

$$f(d) = X - Y + Con \tag{5}$$

Apparently, *f(d)*=0 is the separate line in a two-dimensional space with X and Y being X-coordinate and Y-coordinate. In this space, a given document *d* can be viewed as a point (x, y), in which the values of x and y are calculated according to (3) and (4).

As shown in Fig.1, the distance from the point (x, y) to the separate line will be:

$$Dist = \frac{1}{\sqrt{2}}(x - y + Con) \tag{6}$$

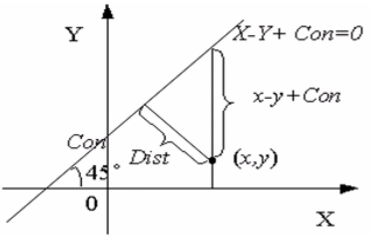

Fig. 1. Distance from point (x, y) to the separate line

Fig. 2 illustrates the distribution of a training set (refer to Section 2.2) regarding *Dist* in the two-dimensional space, with the curve on the left for the negative samples, and the curve on the right for the positive samples. As can be seen in the figure, most of the misclassified documents, which unexpectedly across the separate line, are near the line. The error rate of the classifier is heavily influenced by this area, though the documents falling into this area only constitute a small portion of the training set.

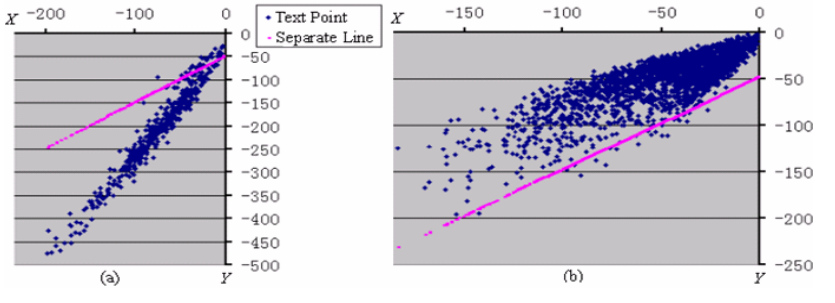

Fig. 2. Distribution of the training set in the two-dimensional space

Thus, the space can be partitioned into reliable area and unreliable area:

$$\begin{cases} Dist_2 \le Dist \le Dist_1, & \text{Decision for } d \text{ is unreliable} \\ Dist > Dist_1, & \text{Assigning the label } c_1 \text{ to } d \text{ is reliable} \\ Dist < Dist_2, & \text{Assigning the label } c_2 \text{ to } d \text{ is reliable} \end{cases} \tag{7}$$

where $Dist_1$ and $Dist_2$ are constants determined by experiments, $Dist_1$ is positive real number and $Dist_2$ is negative real number.

In the second step, more subtle and powerful features will be designed in particular to tackle the unreliable area identified in the first step.

2.2 Experiments on the Two-Class TC

The dataset used here is composed of 12,600 documents with 1,800 negative samples of TARGET and 10,800 positive samples of Non-TARGET. It is split into 4 parts randomly, with three parts as training set and one part as test set. All experiments in this section are performed in 4-fold cross validation.

CSeg&Tag3.0, a Chinese word segmentation and POS tagging system developed by Tsinghua University, is used to perform the morphological analysis for Chinese texts. In the first step, Chinese words with parts-of-speech verb, noun, adjective and adverb are considered as features. The original feature set is further reduced to a much smaller one according to formula (8) or (9). A Naïve Bayesian Classifier is then applied to the test set. In the second step, only the documents that are identified unreliable in terms of (7) in the first step are concerned. This time, bigrams of Chinese words with parts-of-speech verb and noun are used as features, and the Naïve Bayesian Classifier is re-trained and applied again.

$$MI_1(t_k, c) = \sum_{i=1}^{n} \Pr\{t_k, c_i\} \log \frac{\Pr\{t_k, c_i\}}{\Pr\{t_k\} \Pr\{c_i\}} \tag{8}$$

$$MI_2(t_k, c) = \sum_{i=1}^{n} \log \frac{\Pr\{t_k, c_i\}}{\Pr\{t_k\} \Pr\{c_i\}} \tag{9}$$

where t_k stands for the kth feature, which may be a Chinese word or a word bigram, and c_i is the ith predefined category.

We try five methods as follows.

Method-1: Use Chinese words as features, reduce features with (9), and classify documents directly without exploring the two-step strategy.

Method-2: same as Method-1 except feature reduction with (8).

Method-3: same as Method-1 except Chinese word bigrams as features.

Method-4: Use the mixture of Chinese words and Chinese word bigrams as features, reduce features with (8), and classify documents directly.

Method-5: (i.e., the proposed method): Use Chinese words as features in the first step and then use word bigrams as features in the second step, reduce features with (8), and classify the documents in two steps.

Note that the proportion of negative samples and positive samples is 1:6. Thus if all the documents in the test set is arbitrarily set to positive, the precision will reach 85.7%. For this reason, only the experimental results for negative samples are considered in evaluation, as given in Table 1. For each method, the number of features is set by the highest point in the curve of the classifier performance with respect to the number of features (For the limitation of space, we omit all the curves here). The numbers of features set in five methods are 4000, 500, 15000, 800 and 500+3000 (the first step + the second step) respectively.

Table 1. Performance comparisons of the five methods in two-class TC

Performance \ Method used	Method-1	Method-2	Method-3	Method-4	Method-5
Precision	78.04%	93.35%	93.15%	95.86%	97.19%
Recall	88.72%	88.78%	94.17%	91.11%	93.94%
F_1	82.67%	91.00%	93.65%	93.42%	95.54%

Comparing Method-1 and Method-2, we can see that feature reduction formula (8) is superior to (9). Moreover, the number of features determined in the former is less than that in the latter (500 vs. 4000). Comparing Method-2, Method-3 and Method-4, we can see that Chinese word bigrams as features have better discriminating capability meanwhile with more serious data sparseness: the performances of Method-3 and Method-4 are higher than that of Method-2, but the number of features used in Method-3 is more than those used in Method-2 and Method-4 (15000 vs. 500 and 800). Table 1 shows that the proposed method (Methond-5) has the best performance (95.54% F1) and good efficiency. It integrates the merit of words and word bigrams. Using words as features in the first step aims at its better statistical coverage, -- the 500 selected features in the first step can treat a majority of documents, constituting 63.13% of the test set. On the other hand, using word bigrams as features in the second step aims at its better discriminating capability, although the number of features becomes comparatively large (3000). Comparing Method-5 with Method-2, Method-3 and Method-4, we find that the two-step approach is superior to either using only one kind of features (word or word bigram) in the classifier, or using the mixture of two kinds of features in one step.

3 Extending the Two-Step Approach to the Multi-class TC

We extend the two-step method presented in Section 2 to handle the multi-class TC now. The idea is to transfer the multi-class TC to the two-class TC. Similar to two-class TC, the emphasis is still on the misclassified documents given by a classifier, though we use a modified multi-class Naïve Bayesian Classifier here.

3.1 Fix the Fuzzy Area Between Categories by the Multi-class Bayesian Classifier

For a document represented by a binary-valued vector d = (W_1, W_2, …, $W_{|D|}$), the multi-class Naïve Bayesian Classifier can be re-written as:

$$c^* = \arg\max_{c_i \in C}(\log \Pr\{c_i\} + \sum_{k=1}^{|D|} \log\ (1\text{-}p_{ki}) + \sum_{k=1}^{|D|} W_k \log\ \frac{p_{ki}}{1\text{-}p_{ki}}) \tag{10}$$

where $\Pr\{\bullet\}$ is the probability that event $\{\bullet\}$ occurs, $p_{ki}=\Pr\{W_k=1|c_i\}$, (i=1,2, …, |C|), C is the number of predefined categories. Let:

$$MV_i = \log \Pr\{c_i\} + \sum_{k=1}^{|D|} \log\ (1\text{-}p_{ki}) + \sum_{k=1}^{|D|} W_k \log\ \frac{p_{ki}}{1\text{-}p_{ki}} \tag{11}$$

$$MV_{\max\_F} = \underset{c_i \in C}{\text{maximum}}(MV_i) \tag{12}$$

$$MV_{\max\_S} = \underset{c_i \in C}{\text{second\_maximum}}(MV_i) \tag{13}$$

where MV_i stands for the likelihood of assigning a label $c_i \in C$ to the document d, MV_{max_F} and MV_{max_S} are the maximum and the second maximum over all MV_i ($i \in |C|$) respectively. We approximately rewrite (10) as:

$$f(d) = MV_{\max\_F} - MV_{\max\_S} \tag{14}$$

We try to transfer the multi-class TC described by (10) into a two-class TC described by (14). Formula (14) means that the binary-valued multi-class Naïve Bayesian Classifier can be approximately regarded as searching a separate line in a two-dimensional space with MV_{max_F} being the X-coordinate and MV_{max_S} being the Y-coordinate. The distance from a given document, represented as a point (x, y) with the values of x and y calculated according to (12) and (13) respectively, to the separate line in this two-dimensional space will be:

$$Dist = \frac{1}{\sqrt{2}}(x - y) \tag{15}$$

The value of *Dist* directly reflects the degree of confidence of assigning the label c^* to the document d.

The distribution of a training set (refer to Section 3.2) regarding *Dist* in this two-dimensional space, and, consequently, the fuzzy area for the Naïve Bayesian Classifier, are observed and identified, similar to its counterpart in Section 2.2.

3.2 Experiments on the Multi-class TC

We construct a dataset, including 5 categories and the total of 17756 Chinese documents. The document numbers of five categories are 4192, 6968, 2080, 3175 and

1800 respectively, among which the last three categories have the high degree of ambiguity each other. The dataset is split into four parts randomly, one as the test set and the other three as the training set. We again run the five methods described in Section 2.2 on this dataset. The strategy of determining the number of features also follows that used in Section 2.2. The experimentally determined numbers of features regarding the five methods are 8000, 400, 5000, 800 and 400 + 9000 (the first step + the second step) respectively.

The average precision, average recall and average F_1 over the five categories are used to evaluate the experimental results, as shown in Table 2.

Table 2. Performance comparisons of the five methods in multi-class TC

Method / Performance	Method-1	Method-2	Method-3	Method-4	Method-5
Average Precision	92.14%	97.03%	98.36%	97.99%	98.58%
Average Recall	91.13%	97.38%	98.17%	98.03%	98.55%
Average F_1	91.48%	97.20%	98.26%	98.01%	98.56%

We can see from Table 2 that the very similar conclusions as that in the two-class TC in Section 2.2 can be obtained here:

1) Formula (8) is superior to (9) in feature reduction. This comes from the performance comparison between Method-2 and Method-1: the former has higher performance and higher efficiency that the latter (the average F_1, 97.20% vs. 91.48%, and the number of features used, 400 vs. 8000).

2) Word bigrams as features have better discriminating capability than words as features, along with more serious data sparseness. The performances of Method-3 and Method-4, which use Chinese word bigrams and the mixture of words and word bigrams as features respectively, are higher than that of Method-2, which only uses Chinese words as features. But the number of features used in Method-3 is much more than those used in Method-2 and Method-4 (5000 vs. 400 and 800).

3) The proposed method (Methond-5) has the best performances and acceptable efficiency. In term of the average F_1, the performance is improved from the baseline 91.48% (Method-1) to 98.56% (Method-5). In the first step in Method-5, the number of feature set is small (only 400), but a majority of documents can be treated by it. The number of features exploited in Method-5 is the highest among the five methods (9000), but it is still acceptable.

4 Using Dependences Among Features in Two-Step Categorization

In this section, a two-step text categorization method taking the dependences among features into account is presented. We do the same task with the Naïve Bayesian Classifier in the first step, exactly same as what we did in Section 2 and Section 3. In the

second step, each document identified unreliable in the first step are further processed by exploring the dependences among features. This is realized by a model named the Causality Naïve Bayesian Classifier.

4.1 The Causality Naïve Bayesian Classifier (CNB)

The Causality Naïve Bayesian Classifier (CNB) is an improved Naïve Bayesian Classifier. It contains two additional parts, i.e., the k-dependence feature list and the feature causality diagram. The former is used to represent the dependence relation among features, and the latter is used to estimate the probability distribution of a feature dynamically while taking its dependences into account.

K-Dependence Feature List (K-DFL): CNB allows each feature node Y to have a maximum of k features nodes as parents that constitute the k-dependence feature list representing the dependences among features. In other words, $\prod(Y) = \{Y_d, C\}$, where Y_d is the set of at most k features nodes, C is the category node, and $\prod(C) = \Phi$.

Note that we can build a K-DFL for each feature under each class c_t, which represents different dependence relations under different class.

Obviously, there exists a 0-dependence feature list for every feature in the Naïve Bayesian Classifier, from the definition of K-DFL.

The algorithm of constructing K-DFL is as follows: Given the maximum dependence number k, mutual information threshold θ and the class c_t. For each feature Y, repeat the follow steps. 1) Compute class conditional mutual information $MI(Y_i, Y_j| c_t)$, for every pair of features Y_i and Y_j, where $i \neq j$. 2) Construct the set $Si=\{ Y_j \mid MI(Y_i, Y_j| c_t) > \theta\}$. 3) Let $m= \min (k, | S_i|)$, select the top m features as K-DFL from S_i.

Feature Causality Diagram (FCD): CNB allows each feature Y, which occurs in a given document, to have a Feature Causality Diagram (FCD). FCD is a double-layer directed diagram, in which the first layer has only the feature node Y, and the second layer allows to have multiple nodes that include the class node C and the corresponding dependence node set S of Y. Here, $S=S_d \cap S_F$, S_d is the K-DFL node set of Y and $S_F=\{X_i| X_i$ is a feature node that occurs in the given document. There exists a directed arc from every node X_i at the second layer to the node Y at the first layer. The arc is called causality link event L_i which represents the causality intensity between node Y and X_i, and the probability of L_i is $p_i=Pr\{L_i\}=Pr\{Y=1|X_i=1\}$. The relation among all arcs is logical OR. The Feature Causality Diagram can be considered as a sort of simplified causality diagram [9][10].

Suppose feature Y's FCD is G, and it parent node set $S=\{X_1, X_2,\ldots,X_m\}$ ($m \geq 1$) in G, we can estimate the conditional probability as follows while considering the dependences among features:

$$\Pr\{Y=1 \mid X_1=1,\cdots,X_m=1\} \cong \Pr\{Y=1 \mid G\} = \Pr\{\bigcup_{i=1}^{m} L_i\} = p_1 + \sum_{i=2}^{m} p_i \prod_{j=1}^{i-1}(1-p_j) \tag{16}$$

Note that when m=1, $\Pr\{Y=1 \mid X_1=1\} = \Pr\{Y=1 \mid G\} = \Pr\{Y=1 \mid C\}$.

Causality Naïve Bayesian Classifier (CNB): For a document represented by a binary-valued vector d=(X_1 ,X_2 , …,$X_{|d|}$), divide the features into two sets X_1 and X_2, X_1= {X_i| X_i=1} and X_2= {X_j| X_j=0}. The Causality Naïve Bayesian Classifier can be written as:

$$c^* = \arg\max_{c_t \in C}(\log \Pr\{c_t\} + \sum_{i=1}^{|X^1|} \log \Pr\{X_i \mid G_i\} + \sum_{j=1}^{|X^2|} \log(1 - \Pr\{X_j \mid c_t\})) \quad (17)$$

4.2 Experiments on CNB

As mentioned earlier, the first step remains unchanged as that in Section 2 and Section 3. The difference is in the second step: for the documents identified unreliable in the first step, we apply the Causality Naïve Bayesian Classifier to handle them.

We use two datasets in the experiments. one is the two-class dataset described in Section 2.2, called Dataset-I, and the other one is the multi-class dataset described in Section 3.2, called Dataset-I.

To evaluate CNB and compare all methods presented in this paper, we experiment the following methods:

1) Naïve Bayesian Classifier (NB), i.e., the method-2 in Section 2.2;

2) CNB without exploring the two-step strategy;

3) The two-step strategy: NB and CNB in the first and second step (TS-CNB);

4) Limited Dependence Bayesian Classifier (DNB) [11];

5) Method-5 in Section 2.2 and Section 3.2 (denoted TS-DF here).

Experimental results for two-class Dataset-I and multi-class Dataset-II are listed in Table3 and Table 4. The data for NB and TS-DF are derived from the corresponding columns of Table 1 and Table 2. The parameters in CNB and TS-CNB are that the dependence number k=1 and 5, the thresholdθ= 0.0545 and 0.0045 for Dataset-I and Dataset-II respectively. The parameters in DNB are that dependence number k=1and 3, the thresholdθ= 0.0545 and 0.0045 for Dataset-I and Dataset-II respectively.

Table 3. Performance comparisons in two-class Dataset-I

Method / Performance	NB	CNB	TS-CNB	DNB	TS-DF
Precision	93.95%	94.08%	94.08%	93.31%	97.19%
Recall	88.78%	89.00%	89.00%	90.61%	93.94%
F_1	91.00%	91.46%	91.46%	91.93%	95.54%

Table 3 and Table 4 demonstrate that 1) The performance of the Naïve Bayesian Classifier can be improved by taking the dependences among features into account, as evidenced by the fact that CNB, TS-CNB and DNB outperform NB. By tracing the experiment, we find an interesting phenomenon, as expected: for the documents

identified reliable by NB, CNB cannot improve it, but for those identified unreliable by NB, CNB can improve it. The reason should be even though NB and CNB use the same features, but CNB uses the dependences among features additionally. 2) CNB and TS-CNB have the same capability in effectiveness, but TS-CNB has a higher computational efficiency. As stated earlier, TS-CNB uses NB to classify documents in the reliable area and then uses CNB to classify documents in the unreliable area. At the first glance, the efficiency of TS-CNB seems lower than that of using CNB only because the former additionally uses NB in the first step, but in fact, a majority of documents (e.g., 63.13% of the total documents in dataset-I) fall into the reliable area and are then treated by NB successfully (obviously, NB is higher than CNB in efficiency) in the first step, so they will never go to the second step, resulting in a higher computational efficiency of TS-CNB than CNB. 3) The performances of CNB, TS-CNB and DNB are almost identical, among which, the efficiency of TS-CNB is the highest. And, the efficiency of CNB is higher than that of DNB, because CNB uses a simpler network structure than DNB, with the same learning and inference formalism. 4) TS-DF has the highest performance among the all. Meanwhile, the ranking of computational efficiency (in descending order) is NB, TS-DF, TS-CNB, CNB, and DNB.

Table 4. Performance comparisons in multi-class Dataset-II

Method / Performance	NB	CNB	TS-CNB	DNB	TS-DF
Average Precision	97.03%	97.95%	97.95%	98.18%	98.58%
Average Recall	97.38%	98.35%	98.35%	97.91%	98.55%
Average F_1	97.20%	98.15%	98.15%	98.04%	98.56%

5 Related Works

Combining multiple methodologies or representations has been studied in several areas of information retrieval so far, for example, retrieval effectiveness can be improved by using multiple representations [12]. In the area of text categorization in particular, many methods of combining different classifiers have been developed. For example, Yang et al. [13] used simple equal weights for normalized score of each classifier output so as to integrate multiple classifiers linearly in the domain of Topic Detection and Tracking; Hull at al. [14] used linear combination for probabilities or log odds scores of multiple classifier output in the context of document filtering. Larkey et al. [15] used weighted linear combination for system ranks and scores of multiple classifier output in the medical document domain; Li and Jain [16] used voting and classifier selection technique including dynamic classifier selection and adaptive classifier. Lam and Lai [17] automatically selected a classifier for each category based on the category-specific statistical characteristics. Bennett et al. [18] used voting, classifier-selection techniques and a hierarchical combination method with reliability indicators.

6 Conclusions

The issue of how to classify Chinese documents characterized by high degree ambiguity from text categorization's point of view is a challenge. For this issue, this paper presents two solutions in a uniform two-step framework, which makes use of the distributional characteristics of misclassified documents, that is, most of the misclassified documents are near to the separate line between categories. The first solution is a two-step TC approach based on the Naïve Bayesian Classifier. The second solution is to further introduce the dependences among features into the model, resulting in a two-step approach based on the so-called Causality Naïve Bayesian Classifier. Experiments show that the second solution is superior to the Naïve Bayesian Classifier, and is equal to CNB without exploring two-step strategy in performance, but has a higher computational efficiency than the latter. The first solution has the best performance in all the experiments, outperforming all other methods (including the second solution): in the two-class experiments, its F_1 increases from the baseline 82.67% to the final 95.54%, and in the multi-class experiments, its average F_1 increases from the baseline 91.48% to the final 98.56%.

In addition, the other two conclusions can be drawn from the experiments: 1) Using Chinese word bigrams as features has a better discriminating capability than using words as features, but more serious data sparseness will be faced; 2) formula (8) is superior to (9) in feature reduction in both the two-class and multi-class Chinese text categorization.

It is worth point out that we believe the proposed method is in principle language independent, though all the experiments are performed on Chinese datasets.

Acknowledgements

The research is supported in part by the National 863 Project of China under grant number 2001AA114210-03, 2003 Korea-China Young Scientists Exchange Program, the Tsinghua-ALVIS Project co-sponsored by the National Natural Science Foundation of China under grant number 60520130299 and EU FP6, and the National Natural Science Foundation of China under grant number 60321002.

References

1. Sebastiani, F. Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1):1-47, 2002.
2. Lewis, D. Naive Bayes at Forty: The Independence Assumption in Information Retrieval. In Proceedings of ECML-98, 4-15, 1998.
3. Salton, G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading, MA, 1989.
4. Mitchell, T.M. Machine Learning. McCraw Hill, New York, NY, 1996.
5. Yang, Y., and Liu, X. A Re-examination of Text Categorization Methods. In Proceedings of SIGIR-99, 42-49,1999.
6. Xinghua Fan. Causality Reasoning and Text Categorization, Postdoctoral Research Report of Tsinghua University, P.R. China, April 2004. (In Chinese)

7. Dumais, S.T., Platt, J., Hecherman, D., and Sahami, M. Inductive Learning Algorithms and Representation for Text Categorization. In Proceedings of CIKM-98, Bethesda, MD, 148-155, 1998.
8. Sahami, M., Dumais, S., Hecherman, D., and Horvitz, E. A. Bayesian Approach to Filtering Junk E-Mail. In Learning for Text Categorization: Papers from the AAAI Workshop, 55-62, Madison Wisconsin. AAAI Technical Report WS-98-05, 1998.
9. Xinghua Fan. Causality Diagram Theory Research and Applying It to Fault Diagnosis of Complexity System, Ph.D. Dissertation of Chongqing University, P.R. China, April 2002. (In Chinese)
10. Xinghua Fan, Zhang Qin, Sun Maosong, and Huang Xiyue. Reasoning Algorithm in Multi-Valued Causality Diagram, Chinese Journal of Computers, 26(3), 310-322, 2003. (In Chinese)
11. Sahami, M. Learning Limited Dependence Bayesian Classifiers. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, 335-338, 1996.
12. Rajashekar, T. B. and Croft, W. B. Combining Automatic and Manual Index Representations in Probabilistic Retrieval. Journal of the American society for information science, 6(4): 272-283,1995.
13. Yang, Y., Ault, T. and Pierce, T. Combining Multiple Learning Strategies for Effective Cross Validation. In Proceedings of ICML 2000, 1167–1174, 2000.
14. Hull, D. A., Pedersen, J. O. and H. Schutze. Method Combination for Document Filtering. In Proceedings of SIGIR-96, 279–287, 1996.
15. Larkey, L. S. and Croft, W. B. Combining Classifiers in Text Categorization. In Proceedings of SIGIR-96, 289-297, 1996.
16. Li, Y. H., and Jain, A. K. Classification of Text Documents. The Computer Journal, 41(8): 537-546, 1998.
17. Lam, W., and Lai, K.Y. A Meta-learning Approach for Text Categorization. In Proceedings of SIGIR-2001, 303-309, 2001.
18. Bennett, P. N., Dumais, S. T., and Horvitz, E. Probabilistic Combination of Text Classifiers Using Reliability Indicators: Models and Results. In Proceedings of SIGIR-2002, 11-15, 2002.