Master thesis

A Study on Automatic Chinese Text Classification

January 2011

Division of Information Engineering
Graduate School of Engineering
Mie University

Luo Xi

Abstract

Automatic text classification (ATC) is the task to automatically assign one or more appropriate categories for a document according to its content or topic. Traditionally, text classification is carried out by human experts as it requires a certain level of vocabulary recognition and knowledge processing. With the rapid explosion of texts in digital form and growth of online information, text classification has become an important research area owing to the need to automatically handle and organize text collections.

The applications of this technology are manifold, including automatic indexing for information retrieval systems, document organization, text filtering, spam filtering, and even hierarchical categorization of web pages.

Many standard machine learning techniques have been applied to automated text classification problems, and K Nearest Neighbor system (kNN) and Support Vector Machines (SVM) have been reported as the top performing methods for English text classification. Unfortunately, perfect precision cannot be reached in Chinese text classification and the inherent errors caused by word segmentation always remain as a problem.

The purpose of this research is to evaluate the effectiveness of feature extraction, feature transformation and dimension reduction techniques, and to improve the accuracy of Chinese text classification using various techniques.

In this paper, we perform Chinese text classification using N-gram (uni-gram, bi-gram and mixed uni-gram/bi-gram) frequency feature instead of word frequency feature to represent documents and propose the use of mixed uni-gram/bi-gram after feature transformation. We further propose a serial approach based on feature transformation and dimension reduction techniques to improve the performance. Then we compare the results of three different types of SVM kernel functions. Experimental results show that our proposed approach is efficient and effective for improving the performance of Chinese text classification.

Furthermore, we propose a novel feature selection method based on part-of-speech analysis. According to the components of Chinese texts, we utilize the words' part-of-speech (POS) at-

tributes to filter lots of meaningless features. The results show that suitable combination of part-of-speech can lead to better classification performance.

Contents

Chapter 1

Introduction

1.1 Background

Automatic text classification (ATC) is the task to automatically assign one or more appropriate categories for a document according to its content or topic [1]. Traditionally, text classification is carried out by human experts as it requires a certain level of vocabulary recognition and knowledge processing. With the rapid explosion of texts in digital form and growth of online information, text classification has become an important research area owing to the need to automatically handle and organize text collections.

The applications of this technology are manifold, including automatic indexing for information retrieval systems, document organization, text filtering, spam filtering, and even hierarchical categorization of web pages.

Many standard machine learning techniques have been applied to automated text classification problems, and K Nearest Neighbor system (kNN) and Support Vector Machines (SVM) have been reported as the top performing methods for English text classification [2]. Unfortunately, perfect precision cannot be reached in Chinese text classification and the inherent errors caused by word segmentation always remain as a problem.

1.2 Objective of Research

The purpose of this research is to evaluate the effectiveness of feature extraction, feature transformation and dimension reduction techniques, and to improve the accuracy of Chinese text classification using various techniques.

In this paper, we perform Chinese text classification using N-gram frequency feature instead of word frequency feature to represent documents on TanCorpV1.0 [3] which is a new large corpus

special for Chinese text classification. We explain the impact of the different assumptions on N-gram (uni-gram and bi-gram) frequency feature and propose to use the combination of different N-gram (mixed uni-gram/bi-gram) to battle with artificial assumption. We further propose to conduct the combination of uni-gram and bi-gram after feature transformation. Experimental results show that our proposed approach can best represent Chinese documents compared with others.

The limitation of using absolute frequency as the feature vector is dependency on text length which usually leads into lower performance. We experimentally evaluate the effectiveness of proposed approach based on feature transformation techniques including normalizing absolute frequency to relative frequency and power transformation. The results show a significant improvement in performance.

N-gram extraction on a large corpus will yield a large number of possible N-grams. In the experiments, the original dimensionality of bi-gram frequency feature is 212,819. Such a high dimensionality of the feature space may be problematic in terms of computational time and storage resources. Experiments prove that Principal Component Analysis (PCA) is an efficient and effective way to reduce the dimensionality.

Then we compare the results of three different types of SVM kernel functions: Linear-SVM, Poly-SVM and RBF-SVM. The results show that RBF-SVM produces the highest performance.

Furthermore, we propose a novel feature selection method based on part-of-speech analysis. According to the components of Chinese texts, we utilize the words' part-of-speech (POS) attributes to filter lots of meaningless features. The results show that suitable combination of part-of-speech can lead to better classification performance.

The rest of the thesis is organized as follows: In Chapter 2 we describe the proposed method and methodologies used. Chapter 3 describes Part-of-Speech Analysis. The experiments are described and the results and analysis are presented in Chapter 4. Finally, we conclude in Chapter 5 and point out some future directions.

Chapter 2

The Proposed Method

Fig.2.1 shows the general steps for automatic text classification based on the proposed approach. The first step for a Chinese processing system is to represent documents with N-gram frequency feature. Then the feature vector was generated. After that, it is passed on to the feature transformation and dimension reduction phases where various feature vectors were generated and the dimensionality of the data was reduced. These two phases significantly influence the result of the classification. Finally, the classifier evaluates the incoming information and makes a final decision. In the following sections, we introduce our approach in more detail.

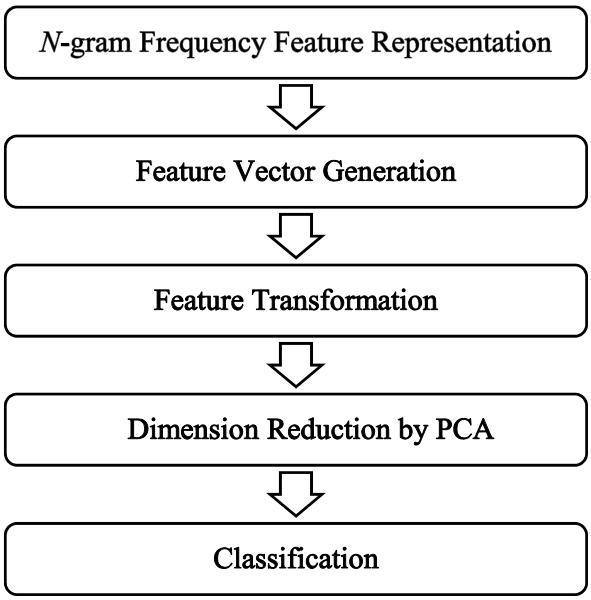

Fig. 2.1: General steps for automatic text classification based on the proposed approach

1. 我是一个中国人

2. 中国是发展中国家

Fig. 2.2: Examples of Chinese documents

1. 我 是 一 个 中 国 人

2. 中 国 是 发 展 中 国 家

Fig. 2.3: Examples of uni-gram sequences

2.1 N-gram Frequency Feature Representation

Unlike English and other western languages, there is no natural delimiter between Chinese words and even no uniform smallest semantic units. This means that the word segmentation is necessary before any other preprocessing and the use of a dictionary is required. The inherent errors caused by word segmentation always remain as a problem.

In this paper, we use a method independent of languages which represents documents with character N-grams [4]. The notion of character N-grams has been in use for many years mainly in the field of speech processing. Fairly recently, this notion has attracted even more interest in other fields of natural language processing, as illustrated by the works of Greffenstette [5] on language identification and that of Damashek [6] on the processing of written text. Amongst other things, these researchers have shown that the use of character N-grams instead of words as the basic unit of information does not lead to information loss.

A character N-gram is a sequence of N consecutive characters. Sequences of one character (N=1) are called uni-gram or mono-gram (1-gram). Sequences of two characters (N=2) are called bi-gram (2-gram). Fig.2.2 shows two examples of Chinese documents. Fig.2.3 and Fig.2.4 show the results of uni-gram and bi-gram sequences.

The use of N-gram frequency feature instead of word frequency feature in text classification tasks offers several advantages. One of them is that by using N-grams, we do not need to perform word segmentation. In addition, no dictionary or language specific techniques are needed and N-grams are also language independent.

1. 我是 是一 一个 个中 中国 国人

2. 中国 国是 是发 发展 展中 中国 国家

Fig. 2.4: Examples of bi-gram sequences

When we establish N-gram frequency feature, we artificially introduce an assumption over the relationship among adjacent words. Uni-gram is based upon the assumption that all words appear in the corpus independently. Bi-gram assumes that only contiguous words correlate with each other. In this sense, single N-gram frequency feature could model the language phenomena with some compromise. To make full use of the power of different N-gram frequency features, we propose to combine uni-gram and bi-gram to represent documents which called mixed uni-gram/bi-gram (1+2-gram).

Obviously, one way is to combine uni-gram and bi-gram frequency feature before any other processing. Since uni-gram and bi-gram are two relatively independent methods, we further propose to conduct the combination of uni-gram and bi-gram after feature transformation (normalization to relative frequency and power transformation). Fig.2.5 and Fig.2.6 show the two different combinations of uni-gram and bi-gram.

In the experiments, the following four methods were used to compare N-gram frequency feature representation.

Method 1 (1-gram): Use uni-gram frequency feature to represent documents.

Method 2 (2-gram): Use bi-gram frequency feature to represent documents.

Method 3 (1+2-gram-before FT): Use mixed uni-gram/bi-gram frequency feature to represent documents and the combination of uni-gram and bi-gram was performed before feature transformation (FT).

Method 4 (1+2-gram-after FT): Same as Method 3 except the combination of uni-gram and bi-gram was performed after feature transformation.

2.2 Feature Vector Generation

In order for a machine learning system to recognize a document there should be a way of representing it. This is usually done by the use of feature vectors. First a lexicon including all different features (N-grams) in training data was generated. Then the feature vector represents the frequency of the N-grams in a document. The form of the feature vector X can be denoted

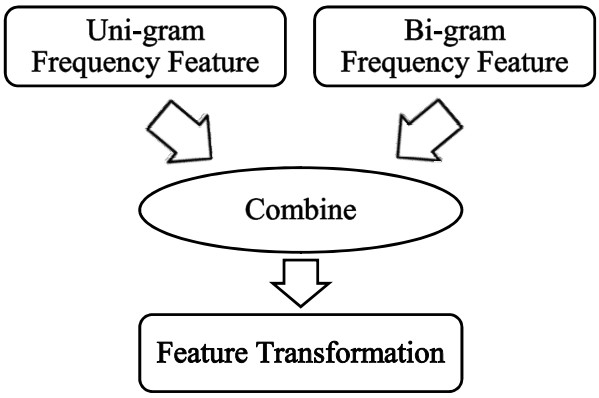

Fig. 2.5: Combination of uni-gram and bi-gram frequency feature before feature transformation

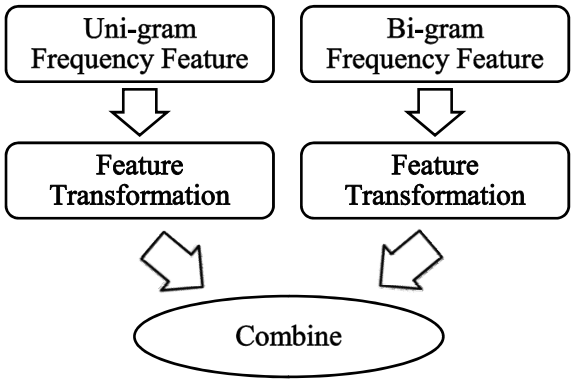

Fig. 2.6: Combination of uni-gram and bi-gram frequency feature after feature transformation

as:

$$X = [x_1 \; x_2 \; ... \; x_n]^T \tag{2.1}$$

where n is the number of N-grams in the lexicon (lexicon size), x_i is the frequency value of i^{th} N-gram and T refers to the transpose of a vector. And the feature vector represents the frequency of specific N-grams in the document. The dimensionality of the feature vector is determined by the lexicon size.

Assume that the two documents of uni-gram sequences in Fig.2.3 represent a text collection. The lexicon (word list) including all different N-grams was generated as shown in Fig.2.7.

Fig.2.8 shows the feature vector (absolute frequency) obtained for each document from the

{ 我 是 一 个 中 国 人 发 展 家 }

Fig. 2.7: Word list

1. $[\ \mathbf{1}\ \mathbf{1}\ \mathbf{1}\ \mathbf{1}\ \mathbf{1}\ \mathbf{1}\ \mathbf{1}\ \mathbf{0}\ \mathbf{0}\ \mathbf{0}\ ]^{T}$
2. $[\ \mathbf{0}\ \mathbf{1}\ \mathbf{0}\ \mathbf{0}\ \mathbf{2}\ \mathbf{2}\ \mathbf{0}\ \mathbf{1}\ \mathbf{1}\ \mathbf{1}\ ]^{T}$

Fig. 2.8: Absolute Frequency (AF)

word list.

2.3 Feature Transformation Techniques

2.3.1 Normalization to Relative Frequency

The feature vector generated by above process is composed of the absolute frequencies (AF). In practice, textual data vary in content and length. The drawback of the AF is dependency on text length which usually leads into lower performance. This is because text length may differ within the same class of documents consequently more complexity of learning. In order to normalize the lengths of documents, absolute frequency is transformation to relative frequency (RF):

$$y_i = \frac{x_i}{\sum_{j=1}^{n} x_j} \tag{2.2}$$

where x_i is the absolute frequency of feature i and n is the lexicon size. Fig.2.9 shows the results after transformed to relative frequency.

2.3.2 Power Transformation

The distribution of absolute/relative frequencies are generally skewed. Therefore in our approach power transformation [7] is applied to improve the symmetry of the distribution:

$$z_i = x_i^v \quad (0 < v < 1) \tag{2.3}$$

This transformation generates Gaussian-like sample distribution. When power transformation

1. $\left[\frac{1}{7}\ \frac{1}{7}\ \frac{1}{7}\ \frac{1}{7}\ \frac{1}{7}\ \frac{1}{7}\ \frac{1}{7}\ 0\ 0\ 0\right]^T$
2. $\left[0\ \frac{1}{8}\ 0\ 0\ \frac{1}{4}\ \frac{1}{4}\ 0\ \frac{1}{8}\ \frac{1}{8}\ \frac{1}{8}\right]^T$

Fig. 2.9: Relative Frequency (RF)

1. $[\,1\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\,]^T$
2. $\left[0\ 1\ 0\ 0\ \sqrt{2}\ \sqrt{2}\ 0\ 1\ 1\ 1\right]^T$

Fig. 2.10: Absolute Frequency with Power Transformation(AFPT)

is applied to the relative frequency with v = 0.5, the length of transformed vector is normalized to 1 which leads to higher classification performance [8]. Therefore in the experiments, v is set to 0.5.

Fig.2.10 shows the results when power transformation was applied to absolute frequency which called absolute frequency with power transformation (AFPT) and Fig.2.11 shows the results when power transformation was applied to relative frequency which called relative frequency with power transformation (RFPT).

2.4 Dimension Reduction

In ATC, high dimensionality problem arises because of the increase in the number of features to be used. The high dimensionality of the feature space may be problematic in terms of computational time and storage resources. In order to solve this problem, the dimensionality is required

1. $\left[\frac{1}{\sqrt{7}}\ \frac{1}{\sqrt{7}}\ \frac{1}{\sqrt{7}}\ \frac{1}{\sqrt{7}}\ \frac{1}{\sqrt{7}}\ \frac{1}{\sqrt{7}}\ \frac{1}{\sqrt{7}}\ 0\ 0\ 0\right]^T$
2. $\left[0\ \frac{1}{\sqrt{8}}\ 0\ 0\ \frac{1}{2}\ \frac{1}{2}\ 0\ \frac{1}{\sqrt{8}}\ \frac{1}{\sqrt{8}}\ \frac{1}{\sqrt{8}}\right]^T$

Fig. 2.11: Relative Frequency with Power Transformation(RFPT)

Table. 2.1: The original and reduced dimensionality for N-gram frequency feature representation

	Original Dimensionality	Reduced Dimensionality
1-gram	4,608	2,154
2-gram	212,819	5,777
1+2-gram	217,190	7,932

to be reduced without deterioration of the performance.

2.4.1 Dimension Reduction by Feature Selection

N-gram extraction on a large corpus will yield a large number of possible N-grams. In the experiments, the original dimensionality of bi-gram frequency feature is 212,819. In fact, only some of them will have significant frequency values in vectors representing the documents and good discriminating power.

Therefore, before generating feature vectors, some characters were removed with reference to a stop list prepared in advance to reduce the features. In this work, a stop list of 243 characters was used to remove useless characters, including numbers, letters, interpunctions and other symbols.

Even when a stop list was used, a lot of features still remained. Yang and Pedersen [9] have shown that it is possible to reduce the dimensionality by a factor of 10 with no loss in effectiveness and a reduction by a factor of 100 bringing about just a small loss. Hence in the experiments, features with frequency value of 25 or less in all training data were removed to reduce the high dimensionality. Table.2.1 summarized the original dimensionality and reduced dimensionality for N-gram frequency feature representation.

2.4.2 Dimension Reduction by PCA

Principal Component Analysis (PCA) was applied to further reduce the high dimensionality. PCA involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

PCA can be done in the following several steps: From the set of training documents $\chi = \{X_1, X_2, \cdots, X_N\}$ the total covariance matrix Σ of the training sample is computed using the

expressions:

$$\Sigma = \frac{1}{N} \sum_{X \in \chi} (X - M)(X - M)^T \tag{2.4}$$

$$M = \frac{1}{N} \sum_{X \in \chi} X \tag{2.5}$$

where M is the mean feature vector of the training sample.

The corresponding eigenvalues λ_i and eigenvectors Φ_i of total cvariance matrix of the training sample were obtained by the definition:

$$\Sigma \Phi_i = \lambda_i \Phi_i, (i = 1, 2, \ldots, n) \tag{2.6}$$

provided that eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

Using eigenvectors corresponding to $m(m \leq n)$ largest eigenvalues, principal components z_i are defined by the linear transformation:

$$z_i = {\Phi_i}^T X, (i = 1, 2, \cdots, m) \tag{2.7}$$

The reduced dimension of feature vectors X was obtained from m principal components selected to compose m-dimension of feature vector.

2.5 Classification

Support Vector Machines (SVM) is a relatively new statistical learning technique first invented by Vladimir Vapnik [10]. It can be seen as a new method for training classifiers based on polynomial functions, radial basis functions, or other functions.

Fig.2.12 illustrates the margin and optimal hyperplane for an SVM trained with samples from two classes. SVM tries to separate a given set of binary labeled training vectors with an optimal hyperplane. The optimum is reached for hyperplane that maximizes the separating margin between the two classes of the training vectors having relatively small number of support vectors.

In the experiments, we used SVM^{light} package [11]. We adopted three different types of SVM kernel functions: Linear Kernel (Linear-SVM), Polynomial Kernel (Poly-SVM) and Radial Basis Function (RBF-SVM).

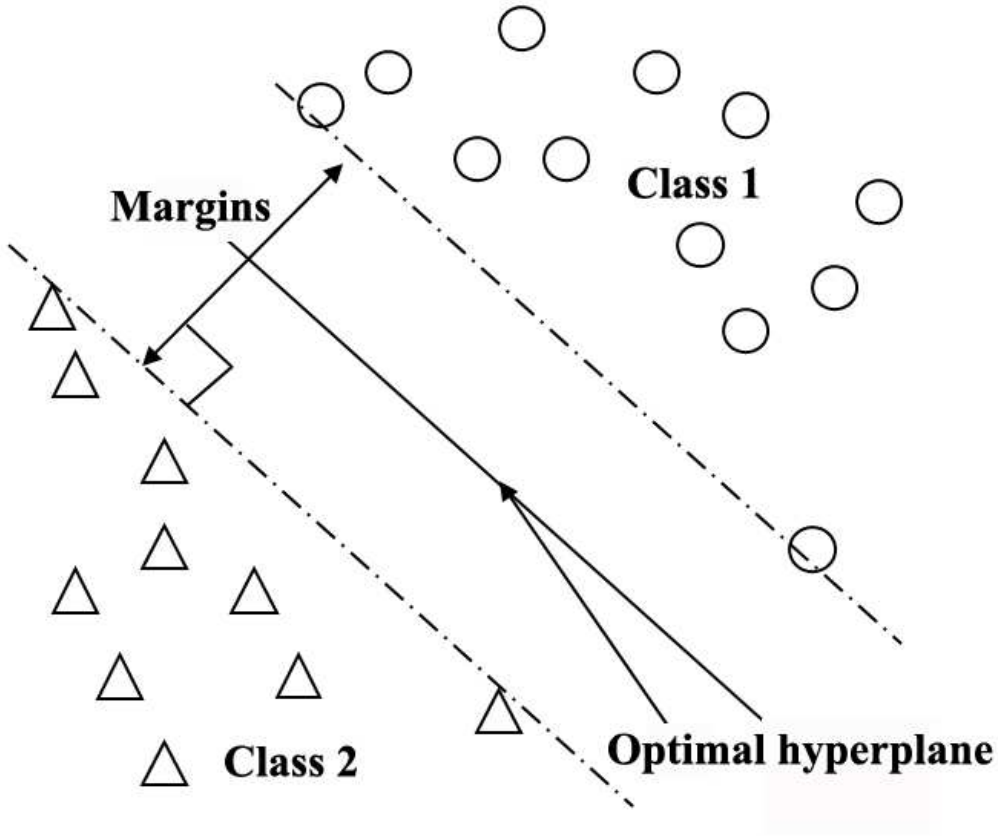

Fig. 2.12: The margin and optimal hyperplane for an SVM trained with samples from two classes

2.6 Evaluation

For evaluation, we adopt the most commonly used performance measures introduced by Van Rijsbergen [12], including recall (R), precision (P), and F-measure (F). These measures are regarded as the standard evaluation methods for classification systems in automatic text classification. The formulae for recall and precision are given below:

$$R = \frac{TP}{TP + FN} \tag{2.8}$$

$$P = \frac{TP}{TP + FP} \tag{2.9}$$

The terms used to express recall and precision for category w_i are given in the following contingency table.

Table. 2.2: Contingency table for category w_i

Category w_i		Expert Judgments	
		Yes	No
Classifier Judgments	Yes	TP	FP
	No	FN	TN

where TP, TN, FN and FP are the number of true positives, true negatives, false negatives and false positives, respectively.

It is a normal practice to combine recall and precision in some way so that classifiers can be compared in terms of a single rating. F-measure is defined as:

$$F = \frac{2RP}{R+P} \tag{2.10}$$

For ease of comparison, we summarize the F-measure over the different categories using the Micro-averaged F-measure and Macro-averaged F-measure. The micro-averaged performance is viewed as a per-document average since it gives equal weight to every document. The macro-averaged performance is considered per-category average because it gives equal weight to every category. Using these averages, we can observe the effect of different kinds of data on a text classification system.

Chapter 3

Part-of-Speech Analysis

In this chapter we proposed the use of linguistic information in the pre-processing phase of Chinese text classification. We present several experiments evaluating the selection of features based on part-of-speech analysis.

According the components of Chinese texts, we utilize the words' part-of-speech (POS) attributes to filter lots of meaningless words.

In this research, we focus on the feature selection process and we aim to explore the effects of different POS on Chinese text classification effectiveness. In feature selection, we explore the use of nouns, verbs, adjectives, adverbs and pronouns to see whether the different POS do actually make a difference in classification effectiveness. We then choose the best POS feature set and employed a feature selection approach.

Therefore, here we make an analysis of different POS and perform a set of experiments to find out suitable combination of POS which can lead to better classification performance.

3.1 Part-of-Speech Tagging

In order to conduct part-of-speech (POS) analysis, lexical analysis system is required. We used a software called 3GWS-Demo (the demo version of the 3rd Generation Word Segmenter) to perform word segmentation and POS tagging. It was developed by Fajava Intelligence Technology (Beijing) Ltd. The main functions are including Chinese word segmentation, POS tagging and user dictionary. The POS tagger incorporated in the software is reported to have more than 94% accuracy for Chinese texts. The POS standard used in the corpus is defined by the Institute of Computing Technology Chinese Academy of Sciences which called ICTPOS (Institute of Computing Technology, part-of-speech set). ICTPOS contains totally 99 POS tags (including 22 big tags, 66 small tags and 11 sub-tags). Fig.3.1 shows an original Chinese document and Fig.3.2

据英国媒体 2 月 21 日报道，只要天气允许，福塞特打算在本周的某个时间，驾驶“维珍大西洋环球飞行者”号（Virgin Atlantic GlobalFlyer）飞机开始此次冒险之旅，希望成为世界上第一个独自驾驶飞机完成无间断环球航行的人。

Fig. 3.1: Example of Chinese document

据/p 英国/ns 媒体/n 2 月/t 21 日/t 报道/v ，/wd 只要/c 天气/n 允许/v ，/wd 福塞/nrf 特/d 打算/v 在/p 本/rz 周/qt 的/ude1 某个/rz 时间/n ，/wd 驾驶/v “/wyz 维珍/nz 大西洋/nsf 环球/n 飞行/vn 者/k ”/wyy 号/n （/wkz Virgin/x Atlantic/x GlobalFlyer/x ）/wky 飞机/n 开始/v 此次/rz 冒险/vi 之/uzhi 旅/ng ，/wd 希望/v 成为/v 世界/n 上/f 第一/m 个/q 独自/d 驾驶/v 飞机/n 完成/v 无/v 间断/vi 环球/n 航行/vi 的/ude1 人/n 。/wj

Fig. 3.2: Example of word segmentation and POS tagging by 3GWS-Demo

shows the output of the word segmentation and POS tagging for this document.

3.2 Combination of Suitable Part-of-Speech

From all possible tags, we just considered five big POS tags: nouns, verbs, adjectives, adverbs and pronouns. We tested the dimensionality of these five POS and all POS. In the experiments, features with frequency value of 5 or less in all training data were removed. Table.3.1 summarized the original dimensionality and reduced dimensionality.

We followed two steps in choosing suitable POS combination. Firstly we found out which POS contribute more in classification performance. In this step we removed one group of POS and then evaluate the classification performance to find out the hierrachy of the POS in describing a category’s content.

For simplicity let us assume that there are two sets of feature vectors: the feature set generated using only nouns and the feature set generated using only verbs.

Table. 3.1: The original and reduced dimensionality

	Original Dimensionality	Reduced Dimensionality
Nouns	19,478	5,111
Verbs	10,203	3,488
Adjectives	2,196	827
Adverbs	980	494
Pronouns	269	167
Five POS	32,094	9,819
All POS	39,842	11,330

We denote nouns with a superscript u and verbs with a superscript v in equations (3.1) to (3.5). Consequently, we can define the feature vectors as:

$$\boldsymbol{x}^{(u)} = \left[x_1^{(u)}\ x_2^{(u)}\ \ldots\ x_{n_1}^{(u)}\right]^T \tag{3.1}$$

for the noun features.

The verb features can be expressed as:

$$\boldsymbol{x}^{(v)} = \left[x_1^{(v)}\ x_2^{(v)}\ \cdots\ x_{n_2}^{(v)}\right]^T \tag{3.2}$$

Let us denote the original feature set as $\boldsymbol{A}$ and the remaining feature vectors after excluding nouns can be defined as:

$$\boldsymbol{R} = \boldsymbol{A}\ \ominus\ \boldsymbol{x}^{(u)} \tag{3.3}$$

Equation (3.3) can be used in excluding other POS such as verbs, adjectives, adverbs and pronouns.

Secondly, the POS which describe the category more, are combined and their effect in classification is observed.

The combination of noun and verb feature vectors can be defined as:

$$\boldsymbol{Q} = \boldsymbol{x}^{(u)}\ \oplus\ \boldsymbol{x}^{(v)} \tag{3.4}$$

$$= \left[x_1^{(u)} \ldots x_{n_1}^{(u)}, x_1^{(v)} \ldots x_{n_2}^{(v)}\right]^T \tag{3.5}$$

Equation (3.5) can be also used to combine other POS such as adjectives, adverbs and pronouns.

Chapter 4

Experiments and Results

4.1 Data for Experiments

Experimental data were obtained from a Chinese corpus called TanCorpV1.0 [3] which is a new large corpus special for Chinese text classification. It is collected and processed by Songbo Tan. The corpus is categorized in two hierarchies. The first hierarchy contains 12 big categories (art, car, career, computer, economy, education, entertainment, medical, property, region, science and sport) and the second hierarchy consists of 60 subclasses. It is totally composed of 14,150 texts. This corpus can serve as three categorization datasets: one hierarchical dataset (TanCorpHier) and two flat dataset (TanCorp-12 and TanCorp-60). In our experiments, we use TanCorpHier.

In the experiments, 150 texts were selected randomly from the corpus for each big category, and totally 1800 texts were used. The ratio of training data to test data is set as 2:1. In order to retain the independence of the data used and to promote the validity of the results, 150 texts of each category were divided into three groups of 50 each. For each time, one of the groups was used as test data, and the remaining two groups were used as training data. And the average values are considered as the final result.

4.2 Performance Evaluation of the Proposed Method

4.2.1 The Effect of Feature Transformation and Principal Component Analysis Techniques

Experiments were performed to compare the results using different feature vectors: absolute frequency (AF), relative frequency (RF), absolute frequency with power transformation (AFPT)

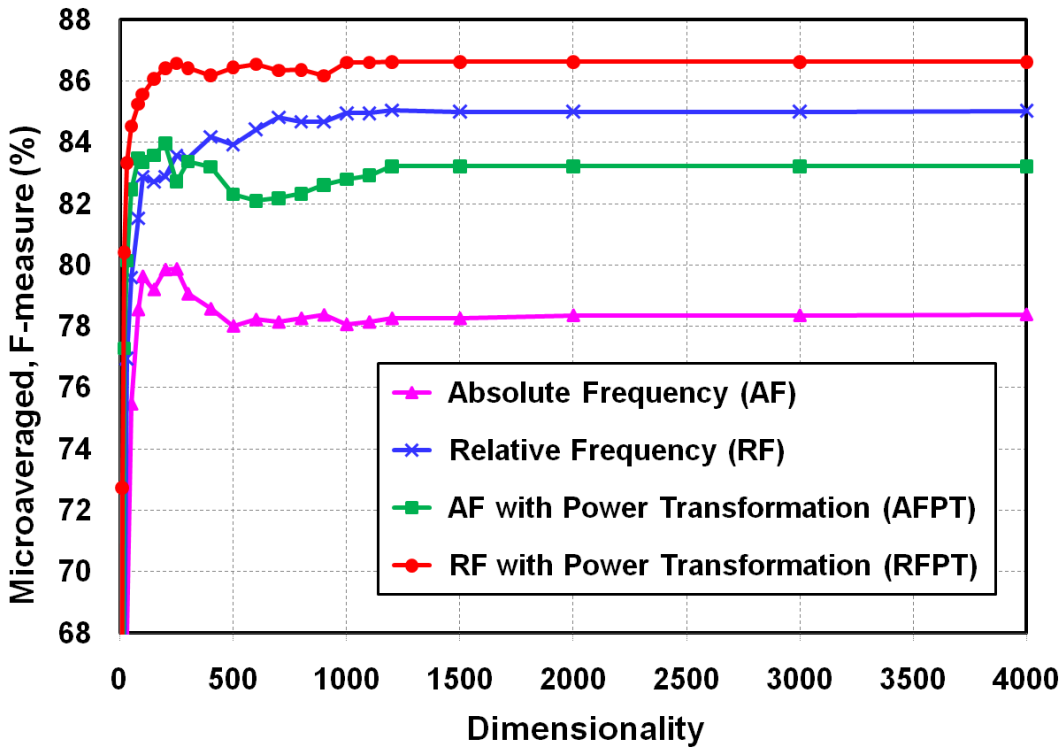

Fig. 4.1: Micro-averaged F-measure vs. Dimensionality for bi-gram with linear kernel

Table. 4.1: The best Micro-averaged F-measure for bi-gram with linear kernel in %

	AF	RF	AFPT	RFPT
Micro-averaged F-measure	79.87	85.05	83.96	86.62

and relative frequency with power transformation (RFPT).

Fig.4.1 shows the relationship between the dimensionality and the Micro-averaged F-measure for bi-gram with linear kernel. Table.4.1 shows the best performance achieved by various feature vectors.

The performance was significantly improved by employing relative frequency instead of absolute frequency. The best Micro-averaged F-measure was improved from 79.87% to 85.05%. Power transformation technique further improved the performance, from 79.87% to 83.96% for absolute frequency and from 85.05% to 86.62% for relative frequency. The relative frequency with power transformation (RFPT) gives the best performances throughout all dimensionality.

As shown in Fig.4.1, for both AF and AFPT, the best performance was achieved at lower dimensionality (200 dimensionality) and after that the performance decreased slightly. For RF and

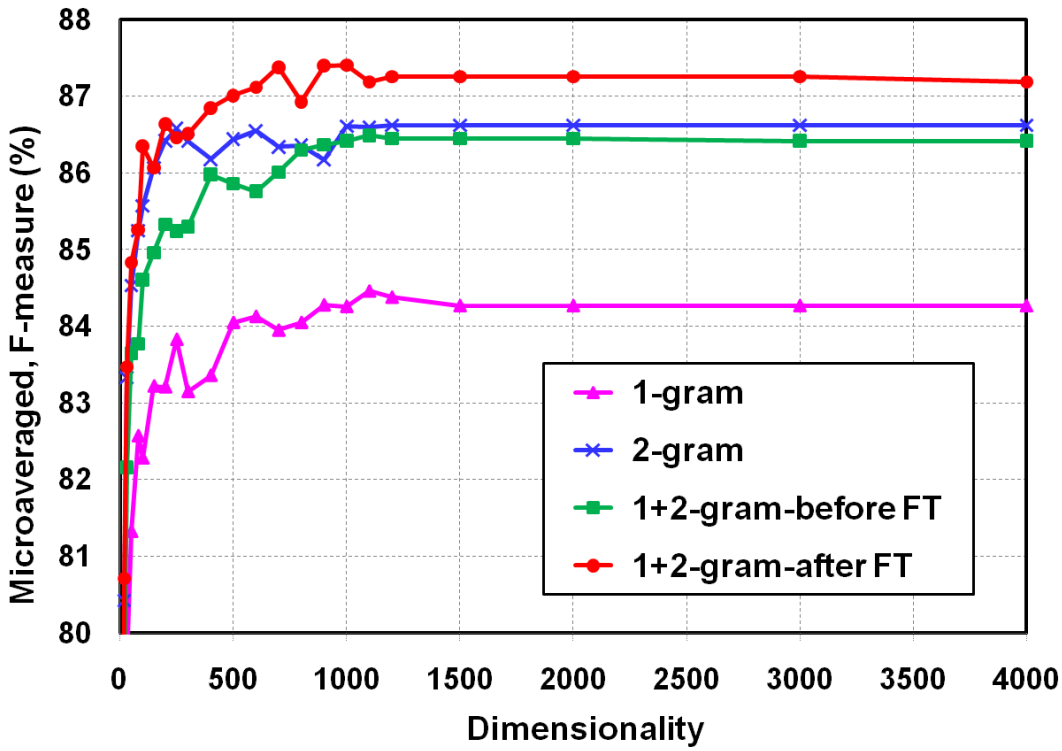

Fig. 4.2: Micro-averaged F-measure vs. Dimensionality for different N-gram frequency feature representation methods of RFPT with linear kernel

RFPT, the performance increased before 1000 dimensionality and then became stable. For all kinds of feature vectors, there is nearly no performance improvement after 1100 dimensionality. Principal Component Analysis (PCA) improves the efficiency of the text classification by reducing the dimensionality.

4.2.2 Comparing N-gram Frequency Feature Representation

Fig.4.2 shows the relationship between the dimensionality and the Micro-averaged F-measure for different N-gram frequency feature representation methods of RFPT with linear kernel. Fig.4.3 and Fig.4.4 show the best performance comparison of RFPT in Micro-averaged and Macro-averaged F-measure respectively. The result shows that throughout all dimensionality, 1-gram has the worst results. It also indicates that both 2-gram and 1+2-gram can well represent Chinese documents, and 1+2-gram-after FT produce the highest effectiveness. The best Micro-averaged F-measure (88.63%) and Macro-averaged F-measure (89.27%) were both achieved by 1+2-gram-after FT.

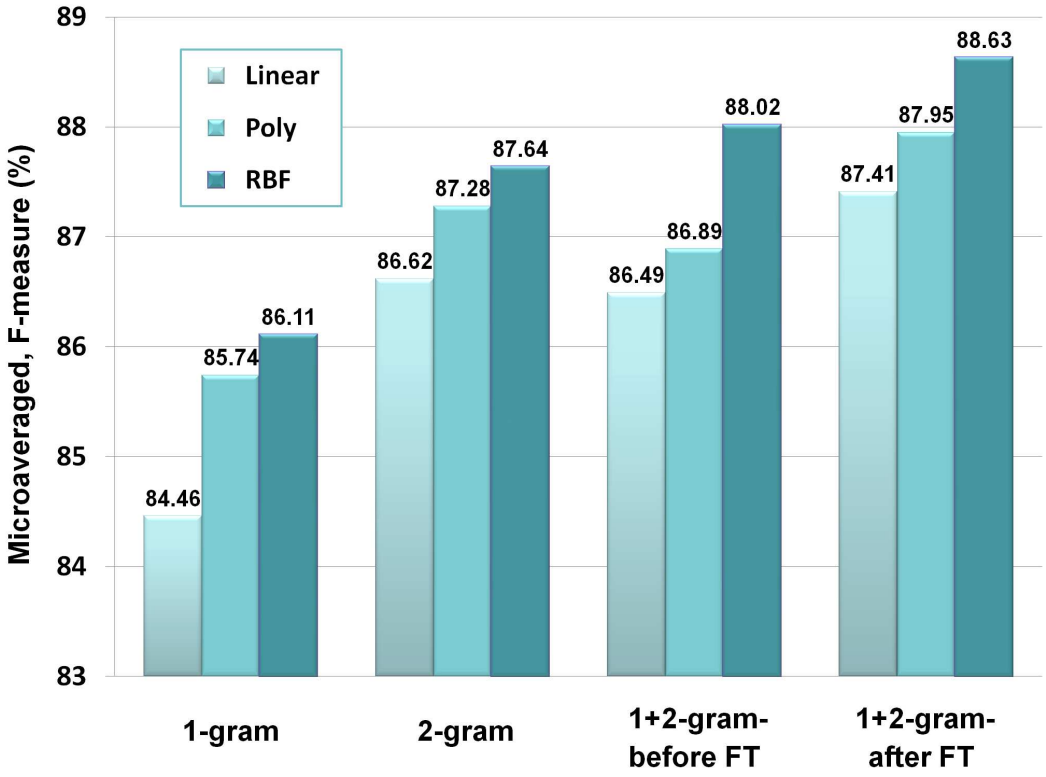

Fig. 4.3: The best Micro-averaged F-measure of RFPT

4.2.3 Comparison of SVM Kernel Functions

Fig.4.3 and Fig.4.4 also show the best performance comparison of three types of SVM kernels in Micro-averaged and Macro-averaged F-measure respectively.

Either for Micro-averaged or Macro-averaged F-measure, the tendency of the results of Linear, Poly, and RBF kernels is similar. RBF kernel exhibits the best F-measure for each N-gram frequency feature representation method. We can obtain the order of performance of different SVM kernels as follows:

RBF-SVM $>$ Poly-SVM $>$ Linear-SVM

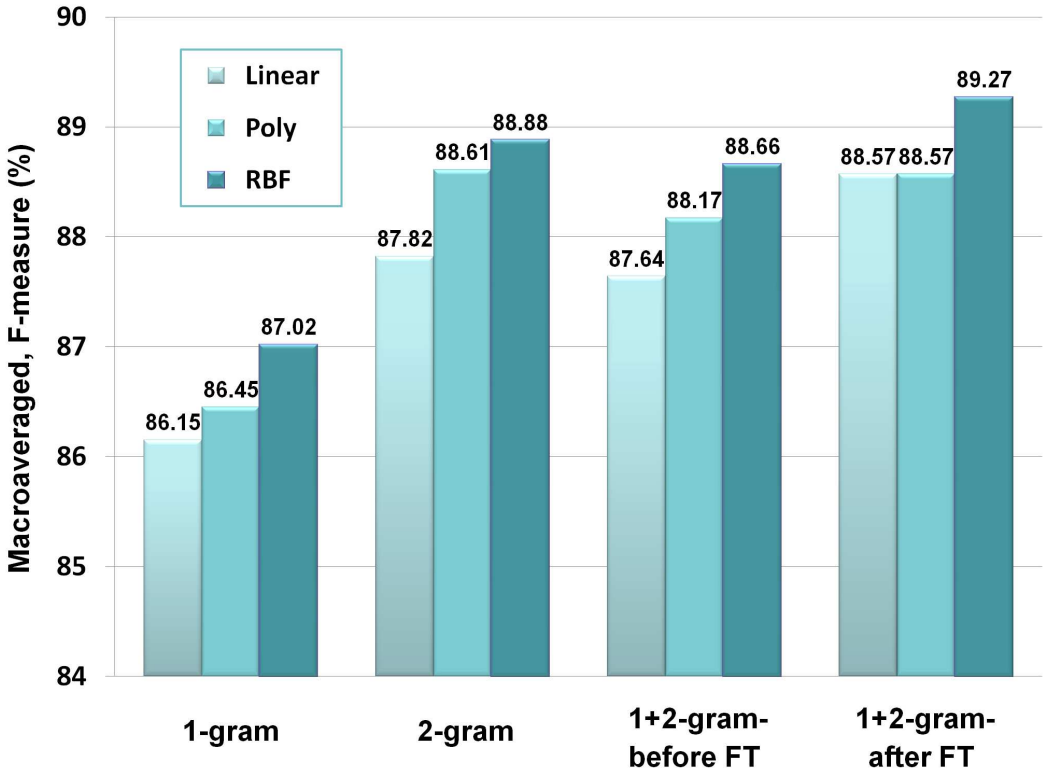

Fig. 4.4: The best Macro-averaged F-measure of RFPT

4.3 The Results of Part-of-Speech Analysis

Two sets of experiments based on part-of-speech (POS) analysis were carried out. The first set of experiment was performed to find out which POS contributes most to higher performance. The second set of experiment was performed to find out suitable combination of POS which can lead to better classification performance.

4.3.1 The Results of Excluding One Part-of-Speech

Fig.4.5 shows the best Micro-averaged F-measure comparison of three types of SVM kernels of RFPT after excluding one POS.

We can see that when nouns were removed, the results were lower than 79% which achieved the lowest classification performance compared with other POS. This means that nouns are the most important features to represent Chinese documents and contribute most to higher performance.

When verbs were removed, the classification performance was only decreased from 87.19%

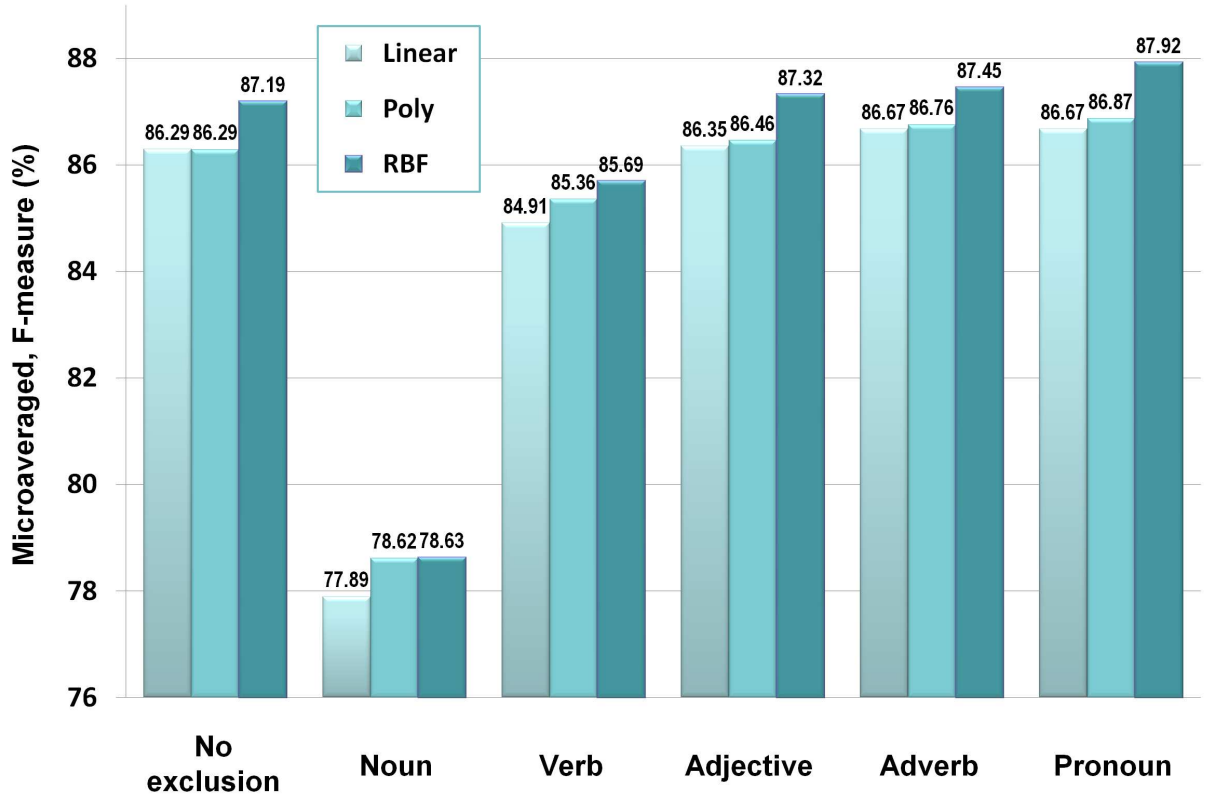

Fig. 4.5: The best Micro-averaged F-measure after excluding one POS

to 85.69% for RBF kernel which means that verbs contribute less in representing Chinese documents. This is because although most of the verbs are relevant to the category, it can also be used in many other categories. This is due to the nature of verbs, where it is used to describe an action.

When adjectives were removed, there is no performance decrease but rather an improvement. In other words, adjectives decrease the ability of features to discriminate one category from another. The same case was observed when adverbs or pronouns were removed in the text classification task. In fact, adjectives are commonly used to modify a noun or pronoun, while adverbs are used to modify a verb, an adjective, another adverb or a whole sentence. These can be found throughout the dataset, across all categories. This explains the increased classification results obtained when the adjectives or adverbs were removed in the text classification task.

4.3.2 The Results of Suitable Combination of Part-of-Speech

Fig.4.6 shows the best Micro-averaged F-measure comparison of three types of SVM kernels of RFPT for different combinations of POS.

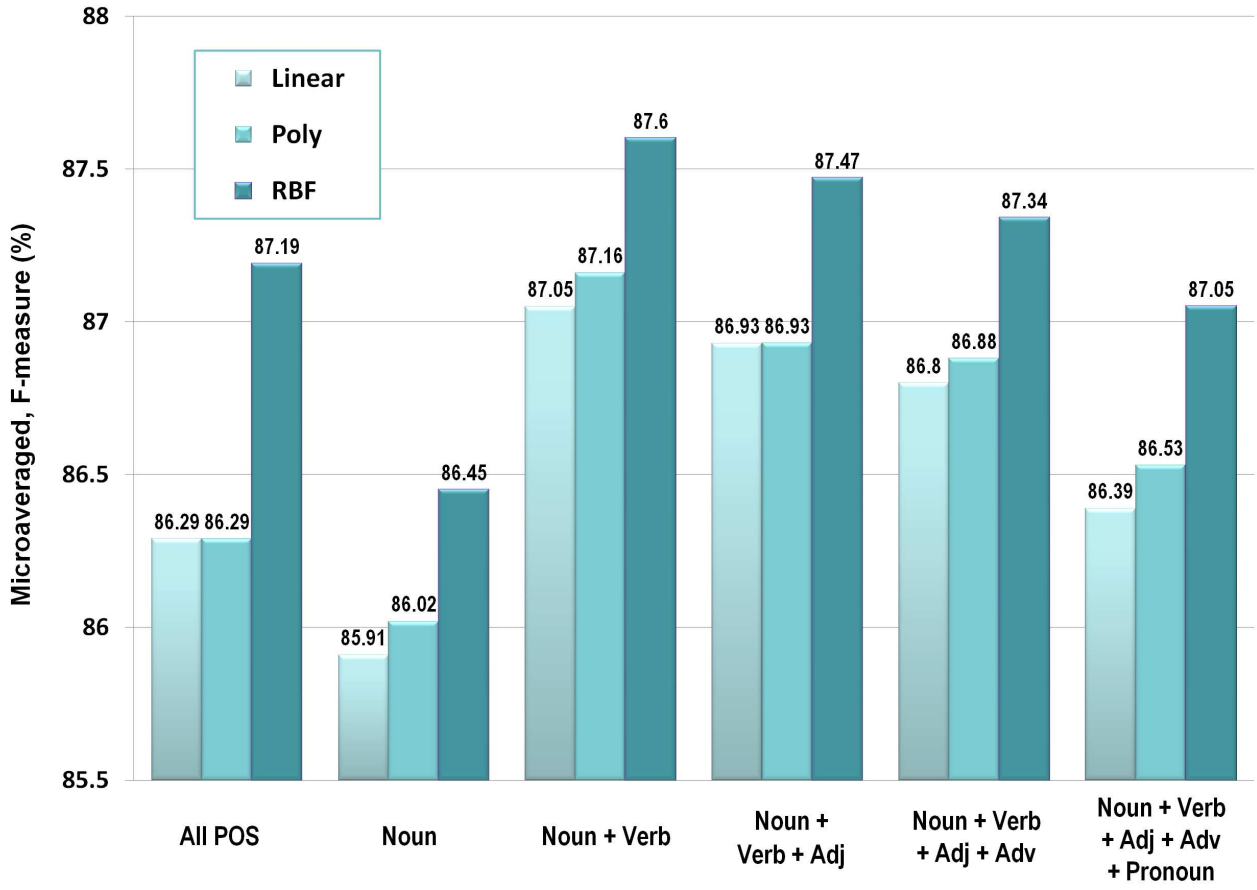

Fig. 4.6: The best Micro-averaged F-measure for different combinations of POS

From the first set of experiments, we found that nouns contribute most to higher performance. Fig.4.6 shows that the performance of using only nouns as features is 86.45% for RBF kernel which is only slightly lower than the use of all POS. One advantage of using only nouns as features is that it has a much smaller dimensionality as compared to using all POS.

Then nouns were combined with verbs. The performance was improved from 86.45% to 87.6% which is higher than that of all POS and also the best performance as compared to all the other combinations of POS.

After combined adjectives with nouns and verbs, the performance was slightly decreased to 87.47%. When adverbs and pronouns were added, the performances were further reduced to 87.34% and 87.05%.

Chapter 5

Conclusion

5.1 Conclusion

In this paper, we perform Chinese text classification using N-gram frequency feature representation. We propose to use the combination of uni-gram and bi-gram after feature transformation (1+2-gram-after FT) which proved to be the most efficient method to represent Chinese documents. We further propose a serial approach based on feature transformation and dimension reduction techniques to improve the performance of Chinese text classification. The experimental results show that normalizing absolute frequency to relative frequency followed by power transformation (RFPT) significantly improved the performance. Principal Component Analysis (PCA) effectively reduced the dimensionality without deterioration of the performance. Then we propose the use of RBF-SVM as classifier which produced the highest performance. Furthermore, we have explored the roles of the different POS (nouns, verbs, adjectives, adverbs and pronouns) in feature selection and we found that nouns best describe a category' contents. And the combination of nouns and verbs as features is able to perform better than all the other combination of POS, with a relatively much smaller feature set.

5.2 Future Work

Future work includes:

1. Conduct the experiment using TFIDF (term frequency by inverse document frequency).
2. Extensive experimental evaluation using more texts on more categories.
3. Evaluate the performance using other classifiers and compare the results.
4. Text classification of Chinese OCR output.

AppendixA

A.1 Programs

Directory: /home/xserve0/users/luoxi/program/

∼**/ngram**	program for generating N-gram features
∼**/makeVec**	program for generating feature vector
∼**/transFeature**	program for feature transformation
∼**/dimension**	program for reducing feature dimension
∼**/pca**	program for PCA
∼**/svm**	program for SVM
∼**/fmeasure**	program for calculating F-measure
∼**/pos**	program for extracting feature set based on part-of-speech analysis

Please refer to README.txt in each folder for details.

A.2 Experimental Data

Directory: /home/xserve0/users/luoxi/data/

∼**/TanCorpV1.0**	Chinese corpus: TanCorpV1.0
∼**/ngram**	experimental data for N-gram
∼**/pos**	experimental data for part-of-speech analysis

Please refer to README.txt in each folder for details.

Acknowledgements

I would like to express my sincerest gratitude to my supervisor, Professor Fumitaka Kimura, Associate Professor Tetsushi Wakabayashi and Dr. Wataru Ohyama, for their great guidance and endless support throughout the course of this thesis work. Their support, encouragement and guidance helped me to gain a solid understanding of this field. I would like to thank them for all their help and advice over the years I have pursued my education and research at Mie University.

I would also like to extend my gratitude to the staff in the department of information engineering. I am grateful to Secretary Miyuki Tanaka, for helping the departments to run smoothly and for assisting me in many different ways.

I am indebted to my many student colleagues in human interface laboratory for providing a stimulating and fun environment in which to learn and grow. I am especially grateful to Akihiro Kokawa who was particularly helpful, patiently teaching me a lot to achieve this work.

I would like to express my sincerest appreciation and love to my husband, Chen Chao, for his unwavering support and love. Without his encouragement and understanding it would have been impossible for me to finish this work. My special gratitude is due to my brother, my sisters and their families for their loving support. I owe my sincere gratitude to Han Xuexian, who gave me the opportunity to study in the Department of Information Engineering at the Mie University of Japan and gave me untiring help during my difficult moments.

Lastly, I wish to send my regards and thanks to my parents. They bore me, raised me, supported me, taught me, and loved me. To them I dedicate this thesis.

Reference

[1] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys, Vol. 34, No. 1, (March 2002), 1-47.

[2] Y. Yang and X. Liu, A Re-examination of text categorization methods. In Proceedings, 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), 42-49, 1999.

[3] Songbo Tan and Yuefen Wang, Chinese text categorization corpus-TanCorpV1.0. http://www.searchforum.org.cn/tansongbo/corpus.htm

[4] D. Jurafsky & J.H. Martin, An Introduction to natural language processing, computational linguistics, and speech recognition, Speech and Language Processing, Prentice Hall, 2000.

[5] Greffenstette, Comparing two language identification schemes, Proceedings of JADT-95, 85-96, 1995.

[6] M. Damashek, Gauging similarity with N-grams: language-independent categorization of text, Science, 267, 843-848, 1995.

[7] K. Fukunaga, Introduction to statistical pattern recognition, Academic Press, Inc, (1990), 76-77.

[8] L. S. P. Busagala, W. Ohyama, T. Wakabayashi, and F. Kimura, Machine learning with transformed features in automatic text classification, In Proceedings of ECML/PKDD-05 Workshop on Sub-symbolic Paradigms for Learning in Structured Domains (Relational Machine Learning), pages 11-20, 2005

[9] Y. Yang and J. O. Pedersen, A Comparative study on feature selection in text categorization, In Proceedings of ICML-97, 14th International Conference on Machine Learning (Nashville, TN, 1997), 412-420.

[10] Corinna Cortes and V. Vapnik, Support-Vector Networks, Machine Learning, 20, 1995.

[11] T. Joachims, Learning to classify text using support vector machines: Methods, Theory and Algorithms, Kluwer Academic Publishers Boston Dordrecht London, 2001.

[12] C. van Rijsbergen, Information Retrieval, Butterworths, London, 1979.