

# The Use of SVM for Chinese New Word Identification

Hongqiao Li<sup>\*</sup>, Chang-Ning Huang<sup>#</sup>, Jianfeng Gao<sup>#</sup>, Xiaozhong Fan<sup>\*</sup>

<sup>#</sup>Microsoft Research Asia, Beijing  
{cnhuang, jfgao}@microsoft.com

<sup>\*</sup>Beijing Institute of Technology, Beijing  
{lhqtxm, fxz}@bit.edu.cn<sup>1</sup>

## Abstract

We present a study of new word identification (NWI) to improve the performance of a Chinese word segmenter. In this paper the distribution and types of new words are discussed empirically. In particular, we focus on the new words of two surface patterns, which account for more than 80% of new words in our data sets: NW11 (two-character new word) and NW21 (a bi-character word followed with a single character). NWI is defined as a problem of binary classification. A statistical learning approach based on a SVM classifier is used. Different features for NWI are explored, including in-word probability of a character (IWP), the analogy between new words and lexicon words, anti-word list, and frequency in documents. The experiments show that these features are useful for NWI. The F-scores of NWI we achieved are 64.4% and 54.7% for NW11 and NW21, respectively. The overall performance of the Chinese word segmenter could be improved by  $R_{\text{ov}}$  24.5% and F-score 6.5% in PK-close test of the 1st SIGHAN bake-off. This achieves the performance of state-of-the-art word segmenters.

## 1 Introduction

New word identification (NWI) is one of the most critical issues in Chinese word segmentation – a fundamental research problem in Chinese natural

language processing (NLP). Recent studies (e.g. Sproat and Emerson, 2003; Chen, 2003) show that more than 60% of word segmentation errors result from new words that are not stored in a dictionary. Chinese NWI is challenging because of the two main reasons. First, new words appear constantly. Statistics show that more than 1000 new Chinese words appear every year (Thesaurus Research Center of Commercial Press, 2003). These words are mostly domain-specific technical terms (e.g. 视窗 ‘Windows’) and time-sensitive political/social/cultural terms (e.g. 三个代表 ‘Three Represents Theory’, 非典 ‘SARS’, 海归 ‘oversea returned students’). Only a small amount of them will be stored as words in the dictionary, while most of them remain as OOV (out of vocabulary). Second, there are no word boundaries in Chinese text, so in most cases, NWI is better performed simultaneously with Chinese word segmentation which itself is a challenging task.

While previous approaches explore the use of one or two most promising linguistically-motivated features and detect new words heuristically, we believe it is better to utilize all available features and to make a decision statistically: whether a character sequence in certain context is a new word or not. In this study, we define NWI as a binary classification problem, and use a statistical learning approach based on a SVM (Support Vector Machine) classifier. We then investigate various linguistic and statistical features that can be used in the classifier to improve the performance of NWI. In addition, other classifiers (e.g. Decision Tree, Naïve Bayes, kNN) are also suitable for NWI, but we don’t attempt to compare these classifiers in this paper. These features include: in-word prob-

---

<sup>1</sup> This work was done while Hongqiao Li was visiting Microsoft Research Asia.

ability of a character, the analogy between new words and lexicon words, anti-word list and the frequency in documents. We evaluate the performance of NWI in terms of F-score (i.e. a balance of precision and recall,  $\beta=1$ ) and  $R_{\text{ooV}}$  (i.e. the recalling ratio of OOV words), using SIGHAN bakeoff corpus. Our models achieve a F-score of 64.4% for NW11 (two-character new word) and 54.7% for NW21 (bi-character word followed with a single character). Enhancing a Chinese word segmenter by using NWI engine as a post procession, we improve the  $R_{\text{ooV}}$  and F-score of the segmenter by 24.5% and 6.5% respectively

The rest of this paper is structured as follows. Section 2 presents previous work. Section 3 defines the new words in this study. Section 4 describes our approach in detail. Section 5 presents experimental results. Section 6 presents error analysis and discussion. Finally, we draw conclusions and propose future work in Section 7.

## 2 Previous Work

Previous approaches focus on the use of one or two linguistically-motivated features to detect new words heuristically. For example, Chen (2003) used only one feature for NWI: the probability of a character being inside a word, referred to as IWP afterwards. He then assumed that two adjacent characters form a new word if the product of their IWP is larger than a pre-set threshold. Chen reports that an improvement of 11% for  $R_{\text{ooV}}$  and 0.08% for F-score after his word segmenter has been enhanced by the NWI engine.

In addition to IWP, Wu (2002) used another feature: the likelihood score that represents, given a word as well as its part-of-speech tag and length, how likely a character appears in certain position within the word. Wu then reports an F-score of 56% for NWI. Wu also integrates it into a parser and it turns out that about 85% of the identified new words are real words. Other features explored previous include mutual information, context dependency, relative frequency and so on (e.g. Gao, 2002; Nie, 1995; Luo, 2003; Chiang, 1992).

We think that all of these features, either proposed previously or to be described below, are valuable for NWI and in many cases can comple-

ment each other. So in this study, we consequently define the NWI as a binary classification problem, and explore ways of combining various feature functions in a statistical classifier. We also notice that to make our method feasible, all features should be easily obtained from corpus or lexicon. Before we present our approach, we first define the scope of new words we will explore in this study.

## 3 Problem Statement

General speaking, a new word is any word that is not stored in a lexicon. But in this paper, we focus on identifying those words that cannot be detected by a certain Chinese word segmenter described in Gao et al. (2003). In Gao's segmenter, Chinese words are defined as one of the following four types: lexicon words (LW), morphologically derived words (MDW), factoids<sup>2</sup>, and named entities (NE)<sup>3</sup>. Though the four types except the lexicon words have been considered as new words in previous research, they are not defined as new words in this study. We focus on new words (NW) that are mostly time-sensitive concepts and can be hardly grouped into any word type.

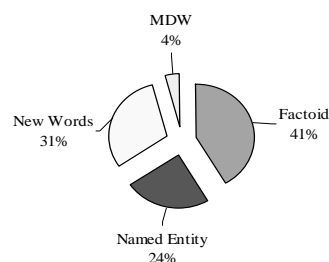


Figure 1. The percentage of each OOV word type We investigate the distribution of new words on PK corpus<sup>4</sup> (4.7MB training set and 89KB test set), which is a subset of the first SIGHAN bakeoff corpora. Figure 1 shows the distributions of words that are not stored in the dictionary (OOV), including words of the type MDW, factoid, NE and NW. We see that NW amount to approximately 31% of

<sup>2</sup> There are ten types of factoid in Gao's segmenter: date, time, percentage, money, number, measure, e-mail, phone number, and URL.

<sup>3</sup> There are three types of named entities: person name, location and organization.

<sup>4</sup> Available at <http://www.sighan.org/bakeoff2003>

OOV, which indicates a substantial improvement space of NWI to the performance of the word segmenter. These new words can be classified using two dimensions. First, from a semantic perspective, NW can be classified into (1) specific-domain concepts, such as 非典 ‘SARS’, 抽射 ‘slap shot’, 洞穿 ‘goal’, 草菇 ‘straw mushroom’, 牡丹花 ‘peony’, and (2) abbreviations, such as 网协 ‘network association’, 抗寒 ‘cold-proof’, 执委 ‘council’, 工价 ‘wage’, 婚检 ‘health care for marriage’. From a surface pattern perspective, they can be classified into: (1) NW11 (two-character<sup>5</sup> new words, ‘1+1’), such as 下岗 ‘out of work’, 旧债 ‘dead horse’, 羊年 ‘year of the goat’. (2) NW21 (a bi-character word followed with a single character, ‘2+1’), such as 杜鹃花 ‘azalea’, 黄金周 ‘golden week’, 世纪坛 ‘century monument’; (3) NW12 (a single character followed with a bi-character word, ‘1+2’), such as 外资金 ‘foreign fund’, 大世界 ‘the big world’; (4) NW22 (two bi-character words, ‘2+2’), such as 卫生设备 ‘sanitary ware’ and (5) others. Figure 2 shows the distribution of NW with different surface patterns and Table 1 shows some examples.

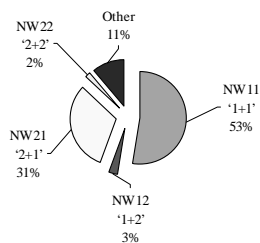


Figure 2. The percentage of each type of NW

Since NW11 and NW21 amount most of new words (i.e. 84% of all new words in PK corpus), they are the focus of this study. We explore them in more detail below. We observe that each character of a NW11 is usually the abbreviated form of another word (or more precisely, lexicon word). For example, the NW11 解读 ‘unscramble’ is composed by 解 ‘jie3’ and 读 ‘du2’, which are respectively the abbreviated forms of 解释 ‘explain’ and 阅读 ‘read’. Other examples include 网

协 ‘network association’ as 网络 ‘network’ and 协会 ‘association’; 执委 ‘executive council’ as 执行 ‘executive’ and 委员会 ‘council’. There are of course a few exceptions such as 羊年 ‘year of the goat’, where both 羊 ‘yang2’ and 年 ‘nian3’ are one-character words. We also observed that NW21s are usually dominated by the last character, which had strong feature to use as a suffix, such as 花 ‘hua1’ of 杜鹃花, 牡丹花; 量 ‘ of 发电量 ‘power generation’, 活动量 ‘quantity of activity’; 业 of 出租业 ‘taxi industry’, 养殖业 ‘plant industry’.

NW types	Samples
NW11	解读, 投射, 养护, 征召, 脱销, 喜迁, 初审, 修学, 达政, 执委, 冻鸡, 早船, 旧俗, 下岗, 年味, 羊年, 海派, 网协
NW21	杜鹃花, 黄金周, 家乡菜, 售票处, 太平鼓, 特色餐, 通行费, 文化村, 住房梦, 专利法, 美食节, 人情贷
NW12	外资金, 大世界, 蹊高跷, 全过程, 总会长, 荡秋千
NW22	卫生设备, 劳动部长, 交通部门, 极端分子
Others	农牧渔业, 县市区, 亚冬会, 党政军群, 十六大, 多云转晴, 不良贷款率, 军事五项

Table 1. Some samples of new words in PK test. There are in general two approaches to NWI. First is to **construct a live lexicon off-line**. That is, we extract new words from large corpus using statistical features such as the frequency of a character string. We then update the existing lexicon using the extracted new words for on-line applications. For example, Chiang(1992) uses statistical knowledge, Nie(1995) uses statistical and heuristic knowledge. The second approach is to **detect new words on-line**. That is, new words in a sentence or a document are identified on the fly. This is the approach we apply in this study.

#### 4 A Binary Classifier for NWI

We define NWI as a binary classification problem. Of the great number of classifiers we experimented, including Perceptron, Naïve Bayes, kNN, SVM and so on, we choose SVM as our basic classifier due to its robustness, efficiency and high performance. Other classifiers maybe are also suitable for NWI, but we don’t attempt to compare these classifiers in this paper. SVMs classify data by mapping it into a high (possibly infinite) dimensional *feature space* and constructing a maximum margin

<sup>5</sup> In this paper a character means a monosyllable morpheme.

*hyperplane* to separate the classes in that space. A detail description is given by Burges(1998). We used SVM<sup>light6</sup> (Joachims, 1999), which is an implementation of the Support Vector Machine described in Vapnik (1995). We now consider the features for NW1.

#### 4.1 IWP(c)

IWP(c) is the probability that a single character *c* is in a word. It is estimated using Equation 1, where \* is a wild-card matching any Chinese character, and C(.) represents the number of occurrence in a corpus given that the variable is a word.

$$IWP(c) = \frac{C(c*) + C(*c*) + C(*c)}{C(c) + C(c*) + C(*c*) + C(*c)} \quad (1)$$

IWP(c) is also used by Chen (2003) and Wu (2002) to detect NW1s: if the product of IWP values of two adjacent characters is larger than a pre-set threshold  $\lambda$ , the two characters form a NW11, as shown in Equation 2. We use this method as a baseline in this study.

$$\begin{aligned} \text{IF } IWP(a)IWP(b) > \lambda \\ \text{THEN } a \text{ and } b \text{ form a NW11} \end{aligned} \quad (2)$$

#### 4.2 IWP(c, pos)

We notice that some characters are more likely to occur in certain positions than in others within a word. For example, 性 ‘xing4’ usually occurs in the last position of a word, while 老 ‘lao3’ in the first position. Therefore, we take into account the position within a word in estimating IWP. The extended feature is IWP(c, pos), where pos is the position of the character *c* in a word, and can be assigned by three values: pos = 1, 2 and 0, indicating the first, middle and last positions, respectively. The values of IWP(c, pos) can be estimated by Equations 3 - 5.

$$IWP(c,1) = \frac{C(c*)}{C(c*) + C(*c*) + C(*c)} \quad (3)$$

$$IWP(c,2) = \frac{C(*c*)}{C(c*) + C(*c*) + C(*c)} \quad (4)$$

$$IWP(c,0) = \frac{C(*c)}{C(c*) + C(*c*) + C(*c)} \quad (5)$$

<sup>6</sup> Available at <http://svmlight.joachims.org/>

For example, using PK corpus as training data, we have following IWP values:

$$\begin{aligned} IWP(\text{解}) &= 0.988 & IWP(\text{解},1) &= 0.624 \\ IWP(\text{解},2) &= 0.017 & IWP(\text{解},0) &= 0.359 \\ IWP(\text{读}) &= 0.734 & IWP(\text{读},1) &= 0.74 \\ IWP(\text{读},2) &= 0.036 & IWP(\text{读},0) &= 0.224 \end{aligned}$$

$IWP(\text{解}) * IWP(\text{读}) = 0.7252$ , it turns out that 解读 is a NW11 according to Equation 2 if  $\lambda = 0.70$ . Table 2 shows some examples.

(解, 1)	解说 3 解答 11 解开 5 解剖学 2 解码 2 解惑 4 解放初 3 解放后 4 解放军报 11 解放鞋 1 解难 5 解放路 2 解渴 1 解除 14 解散 4 解愁 4 ...
(解, 2)	调制解调器 2 不解之缘 2 排忧解难 9 慷慨解囊 2 难解难分 4 浑身解数 1 调解率 1 ...
(解, 0)	谅解 8 和解 12 瓦解 2 理解 101 见解 16 了解 274 分解 11 费解 1 讲解 12 曲解 1 注解 1 ...
(解, -1)*	(难解的) 3 (学者解梦) 1 (忧解过) 1 (以解企业) 2 (民解愁) 1 ...

\* -1 means the character was used as a word alone, the number appended at the sample is its frequency.

Table 2. Some samples of single Chinese character which is in words or as a word

#### 4.3 Analogy to New Words: F<sub>ANA</sub>

We find that some characters can produce words with the same word patterns<sup>7</sup>. For example 上 ‘shang1’ can produce: 上班 ‘go to work’, 上午 ‘morning’, 上面 ‘upside’ et al.; 下 ‘xia4’ can also produce following words by the former patterns: 下班 ‘knock off’, 下午 ‘afternoon’, 下面 ‘downside’ et al. Other examples include 机 ‘ji1’ and 车 ‘che1’ with word patterns: 候\_, 开\_, 客\_, 整\_, et al.; 有 ‘you3’ and 无 ‘wu2’ with: \_害, \_理, 机 et al.. We can learn all these word patterns between different characters in the lexicon. Using them we can make an analogy to new words, such as that the new word 上载 ‘upload’ can be inferred from the lexicon word 下载 ‘download’ by the pattern \_\_载 and the analogy between 上 and 下, by the same token, 飞车 ‘flying car’ from 飞机

<sup>7</sup>In this paper, word pattern is defined that one character in bi-character words is fixed and the remainder are variable. For example, \_\_班 is a word pattern, some characters such as 上, 下, 早, 晚, 白, 夜 can fill the blank to form Chinese word 上班, 下班, 早班, 晚班, 白班, 夜班.

‘plane’, 无方 ‘without means’ from 有方 ‘with means’.

In what follows, we will describe in detail how to value the analogy between new words and lexicon words. First, we give a hypothesis: the two characters can appear more times in the same word patterns, the analogy between them is more reliable. For example, according to our lexicon 下 has the most same word patterns with 上, and that there is a strong preference for them to produce analogous words. Equation 6 shows the valuating principle, where  $a, c, x$ , represent a Chinese character respectively,  $C(\cdot)$  is just the same as 4.1.

$$\text{ANA}(a, x) = \frac{\sum_c \{W(ac)W(xc) + W(ca)W(cx)\}}{\sum_c \{W(ac) + W(ca) + W(xc) + W(cx)\}}$$

$$W(ac) = \begin{cases} 1 \text{ or } C(ac), & \text{if } ac \text{ is in lexicon} \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

A matrix of the analogy between all characters can be obtained. If  $W(ac)$  is set as the indicator function, the matrix can be computed simply according to a lexicon or other segmented corpus, and there is no need to count the frequency of each lexicon word.

If character position is taken into count, there are two values between different characters counted by (7) and (8)

$$\text{ANA}_0(a, x) = \frac{\sum_c W(ca)W(cx)}{\sum_c \{W(ca) + W(cx)\}} \quad (7)$$

$$\text{ANA}_1(a, x) = \frac{\sum_c W(ac)W(xc)}{\sum_c \{W(ac) + W(xc)\}} \quad (8)$$

Second, using the quantified analogy value between different characters we can draw an analogy to new word. Only the character that has the maximum value of the analogy with the character in new word is used. Equation 9 and 10 show how to make the analogy to new word  $ab$  from lexicon word  $xb$  and  $ax$  apart.

$$F_{ANA}^a(ab) = \max_x \{W(xb) \text{ANA}(a, x)\} \quad (9)$$

$$F_{ANA}^b(ab) = \max_x \{W(ax) \text{ANA}(x, b)\} \quad (10)$$

When the character position is taken into count, Equation 11 and 12 are used.

$$F_{ANA'}^a(ab) = \max_x \{W(xb) \text{ANA}_1(a, x)\} \quad (11)$$

$$F_{ANA'}^b(ab) = \max_x \{W(ax) \text{ANA}_0(x, b)\} \quad (12)$$

For example,  $a=\text{下}$ ,  $b=\text{岗}$  ‘gang3’, when  $x=\text{上}$  it gets the maximum value of the analogy and:

$F_{ANA'}^a(ab)=32/(134+91)$ , if  $W(ac)$  is an indicator function in Equation 6;  $F_{ANA'}^a(ab) = 1612/(1678+1603)$ , if  $W(ac) = C(ac)$ . That means there are 32 common word patterns of 上 and 下 in the first position, such as 上 任, 上 游, 上 台, 上 车, 上 面, 上 午, 上 班 et al..

#### 4.4 Anti-word List: $F_A$

As an opposite of IWP, we collect a list of character pairs (called the anti-word list) where the two adjacent characters are most unlikely to form a word. Table 2 shows some examples in the last row. But if all these pairs were recorded, it would need a large amount of memory. In this paper, only when the IWP of one character in the pair is more than 0.5, it will be recorded. We then define a binary feature  $F_A(ab)$ . If  $ab$  are in the anti-word list,  $F_A(ab) = 1$ , otherwise  $F_A(ab)=0$ .

#### 4.5 Frequency: $F_F$

Another important feature of new words is its replicability. A new word usually appears more than once in a document especially the new domain concept, such as 十六大 ‘16<sup>th</sup> NCCPC’. The number of times a new word  $w$  is repeated in the given document are named  $F_F(w)$ , for example  $F_F(\text{十六大})=7$  in PK test set. According to our statistic data, the average appearance times of new words, in PK test data is 1.79.  $F_F(w)$  divided by the total number of word tokens in the text is used as the feature of new word frequency. We also notice that if the processing unit is a sentence this feature will be useless.

All the above features were used for NW11. For NW21, because there was no certain feature of bi-character Chinese words for new word identification, only  $IWP(b)$ ,  $IWP(b,0)$ ,  $F_A$  and  $F_F$  are used as

features. The baseline model for NW21 simply used  $IWP(b)$  only and a threshold.

## 5 Experiments

We test our approach using a subset of the SIGHAN bake-off data: PK corpus (4.7MB training and 89KB test). The training text has been segmented, and contains news articles from People Daily of 1998. We divided the training set into 20 parts. At each step, we use one part as development set and the remainder as training. Because only the lexicon word with the maximum analogy value is used, so the number of features is not more than 8, training time for NWI models is not more than 2 minutes and testing time is not more than 1 second with a standard PC (PIII 800Mhz, 512MB).

### 5.1 Contribution of the Features

We investigate the relative contribution of each feature by generating many versions of the SVM classifier. Precision (P), Recall (R) and F score (a balance of P and R,  $F=2PR/(P+R)$ ) are used for evaluations in these and following experiments. The results are shown in Table 3.

Model		P	R	F
NW11	Baseline( $\lambda=0.675$ )	0.5799	0.5657	0.5728
	IWP	0.5174	0.7056	0.5970
	IWP+pos	0.5048	0.6796	<b>0.5793</b>
	IWP+F <sub>ANA</sub>	0.5154	0.6537	<b>0.5763</b>
	IWP+F <sub>ANA'</sub>	0.5333	0.6926	0.6026
	IWP+F <sub>ANA'</sub> <sup>c</sup>	0.5331	0.6969	0.6041
	IWP+F <sub>F</sub>	0.5271	0.7143	0.6066
	IWP+F <sub>A</sub>	0.5489	0.7532	0.6350
	IWP+F <sub>ANA'</sub> +F <sub>A</sub> +F <sub>F</sub>	0.5635	0.7489	0.6431
	IWP+F <sub>ANA'</sub> <sup>c</sup> +F <sub>A</sub> +F <sub>F</sub>	0.5748	0.7316	<b>0.6438</b>
NW21	Baseline( $\lambda=0.95$ )	0.4066	0.3776	0.3915
	IWP+F <sub>A</sub> +F <sub>F</sub>	0.3861	0.8243	0.5258
	IWP+F <sub>A</sub> +F <sub>F</sub> +pos	0.4094	0.8243	<b>0.5471</b>

+ means the feature is added into SVM, F<sub>ANA</sub> means two features: F<sub>ANA</sub><sup>a</sup> and F<sub>ANA</sub><sup>b</sup>, F<sub>ANA'</sub> means the character position is concerned, F<sub>ANA'</sub><sup>c</sup> means the character position and word frequency are both concerned.

Table 3. the result of several NWI models

From the table, we can see that:

- (1) Using all described features together, the SVM achieves a very good performance of NWI. The best F score is about 7.1% better than that of the baseline model in detecting NW11 and about 15.6% better in detecting NW21.

- (2) The use of  $IWP(c, pos)$  hurts the performance of NW11 identification, but it did work for NW21 identification. The reason is that the two characters of a NW11 do not have fix position property in common lexicon words, but the last character of NW21 doses have the strong feature to use as a suffix. For example, Though 读 is the last character of new word 解读, it is more often as the first character of words than the last, that is  $IWP(解,1) > IWP(解,0)$ .
- (3) F<sub>ANA</sub> has no effect, but F<sub>ANA'</sub> is indeed effective for NWI. The reason may be that the analogy between different characters is related with its position inside words. For example 友 'you3' and 吧 'ba1' have product analogous words only when they are located at the finial character position in a word, such as 酒吧 'saloon', 酒友 'pot companion', 网吧 'internet bar', 网友 'net friend'; whereas in the first position 友 'you3' has the analogy of 喜 'xi3' with three common word patterns: \_\_人, \_\_爱, \_\_好. The effects of F<sub>ANA'</sub> also proved that our hypothesis and valuating approach for the analogy to new words in section 4.3 are right.
- (4) F<sub>ANA'</sub><sup>c</sup> has more effective than F<sub>ANA'</sub>. It means that it is more reliable if word frequency is used to value the analogy to new word. The reason maybe is that the frequency of words has some effect on new word formation.
- (5) F<sub>A</sub> is very useful for NWI. Many interferential new word candidates could be filtered according to this feature.
- (6) F<sub>F</sub> is also useful for NWI. The reason is clear that frequency is an important property of words.

### 5.2 Test in Gao's Segmenter

We perform another two experiments to find how NWI improves the performance of a Chinese segmenter. The first is based on PK-close test in SIGHAN and the other is based on PK open test. In these experiments Gao's segmenter is selected and NWI is used as post procession. But there are the different segmentation standards between Gao's segmenter and PK corpus, such as in PK

corpus surname is apart from the person name, for example 邓小平 ‘Deng Xiaoping’ is segmented as 邓 ‘Deng’ and 小平 ‘Xiaoping’, but in Gao’s segmenter the whole name is a word. So some adjustment is in need to adapt Gao’s segmenter to PK corpus. Table 4 shows the results, where *Adjst* means the adjustment on the output of the segmenter.

PK-close	R	P	F	OOV	R <sub>OOV</sub>	R <sub>IV</sub>
Gao’s +Adjst	.952	.924	.938	.069	.580	.979
Gao’s +Adjst +NWI	.948	.937	<b>.942</b>	.069	<b>.683</b>	.968
PK-open	R	P	F	OOV	R <sub>OOV</sub>	R <sub>IV</sub>
Gao’s +Adjst	.959	.942	.950	.069	.696	.978
Gao’s +Adjst +NWI	.953	.947	<b>.951</b>	.069	<b>.752</b>	.968

Table 4. the result of Gao’s segmenter with NWI in PK corpus

The first line of the result table shows the performance of the segmenter without NWI and the second line is the result after NWI. We could see the R<sub>OOV</sub> is improved 24.5% and the F-score is improved about 6.5% in PK-close test; R<sub>OOV</sub> 13.5% and F-score 2.0% in PK open test. But the R<sub>IV</sub> drops a little for the reason that some two neighboring single character words are incorrectly combined into a new word. If we integrated the NWI model into Gao’s segmenter, it would drop less.

### 5.3 Comparison

We also compare our approach with previous ones that described in Chen (2003), Wu (2003), and Zhang (2003). These segmenters have got excellent performance in 1st SIGHAN Chinese word segmentation bakeoff. Table 5 shows the results.

PK-close	R	P	F	OOV	R <sub>OOV</sub>	R <sub>IV</sub>
Wu’s segmenter	.955	.938	.947	.069	.680	.976
Chen’s segmenter	.955	.938	.946	.069	.647	.977
PK-open	R	P	F	OOV	R <sub>OOV</sub>	R <sub>IV</sub>
Wu’s segmenter	.963	.956	.959	.069	.799	.975
ICTCAS	.963	.943	.953	.069	.743	.980

\* Wu’s segmenter is S10 in the bakeoff, Chen’s segmenter is S09 and ICTCAS is S01 in 1<sup>st</sup> SIGHAN bakeoff.

Table 5. The results of some segmenters in PK corpus

We find that: although there is not so much linguistic knowledge in Gao’s segmenter and NWI, the performance of Gao’s segmenter with NWI has reached the same level as these outstanding segmenters, especially R<sub>OOV</sub> is the best in PK-close test and the second in PK open test. So we con-

clude that the SVM provides a flexible statistical framework to effectively incorporate a wide variety of knowledge for NWI.

## 6 Error Analysis and Discussion

There are two kinds of errors of NWI. The first is incorrectly combined two-character words into a word but it is not a word according the test set. The other is there are some new words that were not identified. The former error accounted about 60% and the latter accounted about 40%. The basic cause for these errors might come from that it is almost impossible to give new words a certain definition. The inconsistency between training set and test set is another reason. Table 6 and Table 7 show some examples of these errors.

Unidentified	Context fraction
上图	上图：随着新春的来临，
问明	问明情况后，…
比武	…组织的军事比武中，
等额	…调整幅度 等额 调整。
还手	更无还手之力。
年味	…郑州“文化”年味浓。
烧烤鸡	烧烤鸡及杂碎、
年画展	天津杨柳青年画展、
鲜花店	…都有鲜花店。

Table 6. Some Unidentified NWs

Incorrect	Context fraction
旅澳	旅澳知名侨领、
明器	…建筑明器馆》
极大	，极大地推动了…
专口	…学生专口和
猫冬	连队“猫冬”时间长
花人	…有很多买花人。
饺子宴	千人饺子宴…
欢乐游	…民俗健身欢乐游”
大家心	…见大家心存疑惑，

Table 7. Some incorrectly identified NWs

We could find that some ‘errors’ are ambiguous, such as 上图 ‘former figure’, it could be seen as a phrase, 旅 ‘lv3’ 澳 ‘ao4’ could be seen as an abbreviation word 旅澳. More than one half of unidentified new words and incorrectly identified new words belonged to this kind. But some were serious error such as 花 ‘hua1’ 人 ‘ren2’ could not be combined in this context by no means.

Besides the indefinite property of new word, there are other reasons why NWI is very difficult:

the very low frequency of new word and the candidates are hundreds of times more than real new words in text. According to our experiments many could be filtered by using anti-word list, but high running memory is needed.

## 7 Conclusion and Future Work

Our work includes several main contributions. First, the distribution and the formal types of new word in real text have been analyzed. NW11 and NW21 were found as the main surface patterns of new words; Second, several features were explored from the statistical and linguistic knowledge of new word, especially the feature of the analogy between new words and lexicon words; Third, our experiments have showed that the SVM based binary classification is useful for NWI.

Now we only concern the uni-gram of new word and NWI is as a post procession of Gao's segmenter. As future work, we would like to integrate NWI into the segmenter. For example, we can define NWI as a new word type, and the whole classifier as a feature function in the log-linear models that are used in Gao's segmenter.

## Acknowledgements

We would like to thank the members of the Natural Language Computing Group at Microsoft Research Asia, and the three anonymous reviewers for their insightful comments and suggestions.

## References

- Aitao Chen. 2003. Chinese Word Segmentation Using Minimal Linguistic Knowledge. In proceedings of *the Second SIGHAN Workshop on Chinese Language Processing*, July 11-12, 2003, Sapporo, Japan.
- Andi Wu. 2003. Chinese Word Segmentation in MSR-NLP. In proceedings of *the Second SIGHAN Workshop on Chinese Language Processing*, July 11-12, 2003, Sapporo, Japan.
- Andi Wu. Zixin Jiang 2000. Statistically-Enhanced New Word Identification in a Rule-Based Chinese System. In proceedings of *the Second Chinese Language Processing Workshop*, Hong Kong, China. pp. 46-51
- C. Burges. A tutorial on support vector machine for pattern recognition. Kluwer Academic Publishers, 1998.
- Geutner, Petra. 1996. Introducing linguistic constraints into statistical language modeling. In *ICSLP96*, Philadelphia, USA. pp. 402-405.
- Huaping Zhang, Hongkui Yu, Deyi Xiong et al. 2003. HMM-based Chinese Lexical Analyzer ICTCLAS. In proceedings of *the Second SIGHAN Workshop on Chinese Language Processing*, July 11-12, 2003, Sapporo, Japan.
- Jianfeng Gao, Mu Li and Chang-Ning Huang. 2003. Improved source-channel models for Chinese word segmentation. In *ACL-2003*. Sapporo, Japan, 7-12, July, 2003
- Jianfeng Gao, Joshua Goodman, Mingjing Li, Kai-Fu Lee. Toward a unified approach to statistical language modeling for Chinese. *ACM Transactions on Asian Language Information Processing*, Vol. 1, No. 1, pp 3-33. 2002
- Jianfeng Gao, Joshua Goodman, Guihong Cao and Hang Li. 2002a. Exploring asymmetric clustering for statistical language modeling. In *ACL2002*, July 6-12. University of Pennsylvania, Philadelphia, PA, USA.
- Jian-Yun Nie, et al. (1995) Unknown Word Detection and Segmentation of Chinese using Statistical and Heuristic Knowledge, *Communications of COLIPS*, vol 5, No. 1 &2, pp.47.
- Richard Sproat and Tom Emerson. 2003. The First International Chinese Word Segmentation Bakeoff. In proceedings of *the Second SIGHAN Workshop on Chinese Language Processing*, July 11-12, 2003, Sapporo, Japan.
- Shengfen Luo, Maosong Sun. Two-Character Chinese Word Extraction Based on Hybrid of Internal and Contextual Measures. In proceedings of *the Second SIGHAN Workshop on Chinese Language Processing*, July 11-12, 2003, Sapporo, Japan.
- Thesaurus Research Center of Commercial Press. 2003. Xinhua Xin Ciyu Cidian. Commercial Press, Beijing, 2003.
- T. H. Chiang, Y. C. Lin and K.Y. Su (1992). Statistical models for word segmentation and unknown word resolution, In proceedings of *the 1992 R. O. C. Computational Linguistics Conference*, pp. 121-146, Taiwan.
- T. Joachims. 2000. Estimating the Generalization Performance of a SVM Efficiently. In proceedings of *the International Conference on Machine Learning*, Morgan Kaufman, 2000.
- Vladimir N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer, 1995.
- Yubin Dai, Christopher S. G. Khoo and Tech Ee Loh. 1999. A new statistical formula for Chinese word segmentation incorporating contextual information. In *SIGIR99*, pp. 82-89.