Based on Support Vector and Word Features New Word Discovery Research

Li Chengcheng and Xu Yuanfang

School of Computer and Information Engineering, Inner Mongolia Normal University, Hohhot, China
nmlcc@sohu.com, xuyuanfang86@126.com

Abstract. Chinese word segmentation is difficult to deal with ambiguity and unknown words recognition, this paper proposes the new word mode features as well as various word internal patterns from the training corpus of positive and negative samples to quantify extraction, and then through the training of support vector machine to get new support vector classification. On the test corpus with absolute discounting method new candidate extraction and selection, and with the training corpus to extract word patterns to quantify the new support vector classification for support vector machine test, through a portion of the rule filter to get the final word recognition results.

Keywords: natural language processing, support vector machine, word recognition, word feature.

1 Introduction

With the rapid development of the economy, Chinese has also been enriched and developed continuously, plenty of Chinese new words appear in people's life. New words appear to bring greater challenges for Chinese word segmentation. New words in the presence of Chinese word segmentation results in too many "loose string", has a great effect on word segmentation accuracy [1]. Therefore, new word detection has become the Chinese automatic word segmentation is one of the difficulties and bottlenecks. How to identify for Chinese *n*neologism has become an important research topic.

In the new word detection methods, mainly based on rules, based on statistics, two methods of hybrid method [2] .The rule-based method main idea is based on the word formation features or design features, establishing rules for professional thesaurus or pattern library, then by matching rules to discover new words. Based on the statistical method, the general is the use of statistical methods to extract candidate string, and then use the language knowledge exclusion is not new word garbage string. Or is the calculation of correlation, find related degree the biggest character and the character combination. The rule of the method mainly is confined to one area, and requires the establishment of rule base. Statistical methods are generally limited to, find a short new words.

Y. Yuan, X. Wu, and Y. Lu (Eds.): ISCTCS 2012, CCIS 320, pp. 287–294, 2013.

This method uses support vector machine (SVM) model as a framework, the training samples for training support vector, using support vector on the test sample to be tested to get initial results, and add rules to get the final result. Support vector machine has high efficiency and good prediction of unknown data, and can effectively integrate a variety of features as well as some Chinese characters morpheme constraint information and other characteristics; the new Chinese words and expressions found the problem also has good applicability.

2 Basic work

2.1 Basic Principles of SVM

Support vector machine is a kind of classification method based on boundary. Its basic principle is that (to 2D data as an example): if the training data according to their classification is assembled in different regions of the planar points. The classification algorithm based on SVM goal is: through training to find these classifications between boundaries, boundary curve is called a non-linear classification, is called linear dividing line. For those data (such as N dimension), they can be viewed as points in an N - dimensional space, while the classification boundary is in an N-dimensional space surface, called the super surface (super than N dimensional space dimension less). Nonlinear classifier using hyper surface type of boundary, while the linear classifier using hyper planes types of boundary [3] .

SVM is based on the linearly separable cases the optimal classification hyper plane proposed [4]. The optimal classification hyper plane is the demands of classification hyper plane can not only two kinds of error separated and to make the two categories in the classification of the largest distance.

2.2 Word Feature Selection

New word identification can be viewed as a linearly no separable classification problems [5], the new words and new words through the hyper plane that separates, so choose how word internal and external features become the key to improve the correct rate.

In this method the identification of new words, word feature selection is a very important link, this article selects the word feature.

- Context information (Context) [6]: in Chinese sentences, words from the words or phrases have a certain relationship between structure, new word is a new word, but it should play a role, so that the sentence structure is more compact, more logical.
- Word probability (IWP) [7]: refers to a Chinese characters morpheme in the corpus with other morphemes to form words by using the probability.

  Characteristic that a Chinese character into words is the Chinese characters morpheme can own characteristics. Defined as:

$$\text{IWP (z)} = {\text{number(in(z))}}\Big/{\text{number(all(z))}}$$

Where number (in (z)) representing the Z in the article with the word probability, number (all (z)) represents the total number of articles appeared in Z.

- Morphological productivity (MP) [8]: where n (z) for the specific structure of different morphological quantity, N (z) for a given structure. The total time. MP more easily derived words. Defined as:

$$\text{MP (z)} = {n(z)}\Big/{N(z)}$$

- Frequency characteristics （F_F)[9]: on large-scale corpus for word recognition, a new words often appear repeatedly, so words with repeatability, defining a new word occurrence number is n, the N is defined as the total number of words:

$$F_F = {n}\Big/{N} \cdot$$

- Mutual information （MI） [10]: assume that A is the length of N text strings, s=$c_1 \dots c_n$, M and N is the length of n-1 substring, Is M,N mutual information, $f(A)$ represents a text string in the corpus the number of occurrences, M and N substring of the total number of times, we can see that MI(A)is larger, the larger the probability of words:

$$\text{MI (A)} = {f(A)}\Big/{f(M+N) - f(A)}$$

2.3 Corpus Processing and Related Work

The training corpus processing, Context, IWP, MP, MI ,F_F statistics, defined as the word internal models feature table, through the process of corpus segmentation, the segmentation results are negative training example, text extraction, text of the dictionary can be correctly recognized words, namely a neologism, marked as 1. Negative text segmentation is a scattered word, labeled as negative 1, according to the word combination pattern; this paper chooses 1+2, 2+1 and 1+3 mode markers. Identification of positive and negative words in text and internal models feature tables are combined to form a training corpus vector attribute matrix, through the SVM test procedure test to get new words recognition support vector.

Table 1. New words common combination mode

Word count	There may be new pattern
Two words	1+1
Three words	1+1+1 1+2 2+1
Four words	1+1+1+1 1+3 3+1 1+2+1 2+1+1

On the test corpus processing, the corpus segmentation, through the above patterns to extract candidate words, through the discount method for screening candidate words, the new word identification support vector and the candidate word vector to construct a matrix through SVM testing to get the final result.

Table 2. Extraction of positive and negative samples

CATEGORY	STRING
POSITIVE CASES	里程碑 转机建制 一国两制 交响曲 星之路 红运当头
NEGATIVE CASES	隐车族 心体谐一 捂地惜建 午动族 养卡人 城铁商圈

We found in the experiments, a lot of new part may be identified into words, and words are not complete, For example: The three words "同名门"，"同名"，"名门"，Through calculation we can see that they are IWP, MP , and F_F difference is not big, the system is likely to name and a cut out, but not with a word recognition, we count them several word appears in the article number, respectively is 28, 27, 31, differ not quite, so when the words in articles appearing in the difference in the number does not exceed a certain value we can think they are the whole existence is a word, the system will cut off the longest string, is "同名门". We define the difference threshold of 5.

Table 3. Ome new characteristics

WORD	IWP	MP	$1/F_f$
同名门	0.893619	0.010866	28
同名	0.902986	0.019826	27
名门	0.927982	0.021067	31

Due to our training sample may not cover all cases; this has not been estimated for some event handling our smoothing, using absolute discounting method, estimation formula:

$$P_r = \left\{ \begin{array}{l} m - a/N, 0 < m \le M \\ a \cdot K - n_0 / N \cdot n_0, m = 0 \end{array} \right\}, \text{K}= \sum_{m=0}^{M} n_m \text{ , take b=2.}$$

Through our observation, we find that a sentence can be extracted from a plurality of candidate words, for example:

“很多企业举万科模式标识的大旗。”

This sentence contains a new word "万科" in the dictionary, we have to have it removed, so here it as a new word.

After processed dictionary word segmentation has been "举万","万科" and "举万科 " three candidate words, after SVM classification results, the three word candidate is likely to have been considered new words. But in fact, there is only one word, "万科" is right. We call these words as mutually exclusive words. The first constraint is to solve the problem of mutually exclusive words. In the above case, we assume that, in the SVM classification, mutually exclusive words with the highest confidence for the final result, namely deviation threshold of 0largest candidate word for the final results.

3 The Experiment Results and Analysis

3.1 Method and Standards

The method adopts the correct rate (P), recall rate (R) [11] and F-measure, the recall rate (R) is a measure of the system to find out new words ability, that all should be identified new word, be system correctly recognized words proportion. Correct rate (P) is a measure of system is not new, i.e. all identified new words new words in correct proportion. F-measure is based on R and P to give a comprehensive evaluation.

$$\text{F-measure}=\left(\beta^2+1\right)\times P\times R \Big/ \left(\beta\times P\right)+R \quad (\beta=1)$$

3.2 The Experimental Results and Analysis

First we with the word patterns MP, IWP and the RBF kernel function as the basic word features for new word detection system is obtained by training the results in order to control foundation After joining Context, MI respectively train again that the training results, in order to draw on new word identification of word internal attributes Classification of image details as well as the system flow chart refer to Fig.1 and Fig. 2.

Table 4. The experimental results of statistical table

WORD ELEMENT	P （%）	R （%）	F （%）	Change
T(B)	40.78	57.86	47.85	—
T(B+ CONTEXT)	45.72	60.36	52.03	4.18
T(B+MI)	47.78	60.86	53.53	5.68
T(B+CONTEXT+MI)	50.13	62.26	55.54	7.69
T(B+ F_F)	46.62	62.96	53.57	5.72
T(B+CONTEX+ F_F)	55.26	67.35	60.71	12.86
T(B+MI+ F_F)	56.12	66.21	60.75	12.90
T(B+CONTEXT+MI+ F_F)	59.82	70.06	64.54	16.69
T(B+CONTEXT+MI+ F_F+RULE)	61.78	73.68	67.20	19.85

Through the experimental results it can be seen, along with the word feature added elements of Chinese new word identification precision and recall rates have increased, this several word internal characteristics of properties on the new word identification has contribution, but also can be observed when joining MI, relative to the base results in a 15.20% increase, there are other MI tests increase rate of more than 15.20%, which indicates that the MI on this experiment has an important role for Chinese new word identification. Additional rules, new words recognition also has been increasing in a certain.

Considering the different kernel functions for new word recognition results, choose three kinds of different kernel functions and using all words characteristic experiment, results were obtained as follows:

Table 5. Different kernel function the experimental results of statistical table

kernel	Penalty factor	P	R
RBF kernel function	C+=0.0001 C-=0.3	61.78	72.68
Polynomial kernel function	C+=0.0001 C-=0.3	43	41.93
Sigmoid kernel function	C+=0.0001 C-=0.3	37	32.15

The experiment found that the use of RBF kernel function when new words recognition recalling rate and correct rate.

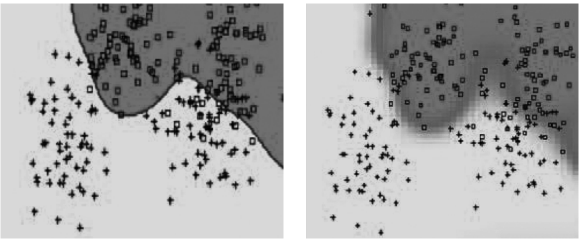

Fig. 1. Classification image of T(B) and T(B+Context+MI+ F_F +Rule)

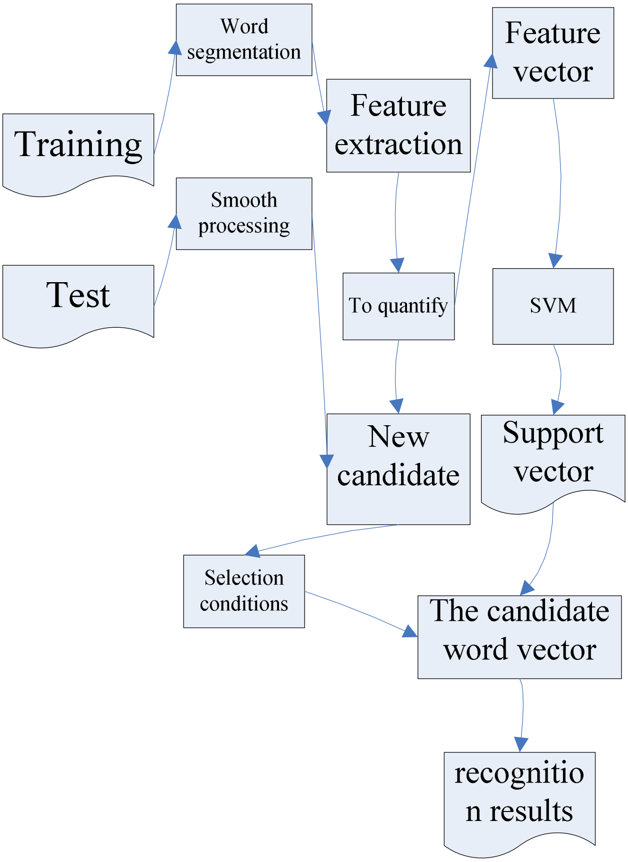

Fig. 2. The flow chart of the system

4 Conclusion

This article proposed one kind based on the SVM and word features a method of Chinese new word identification, by modifying the experimental conditions, different experimental conditions the results show, this method can improve the correct rate of word recognition and recall rate, but there are still some deficiencies, such as for string comparison long new word the new part of the false identification of words into new surrounding words affect the calculation, future work can be added to the training algorithm analysis to identify the long string and improve word recognition accuracy.

References

1. Chen, K., Bai, M.H.: Unknown word detection for Chinese by a corpus- based learning method. Computational Linguistics and Chinese Language Processing 3(1), 27–44 (1998)
2. Ning, S.: Based on word features and search engine for Chinese new word identification. Journal of Wuhan University (Science Edition) 56(6), 704–710 (2010)
3. Qian, Q., Zhang, Z.: A method based on multiple SVM classification method of relevance feedback image retrieval. Computer Technology and Development 19(8), 66–69 (2009)
4. Huang, X., Wang, Y.: SVM in unbalanced data set. Computer Technology and Development 19(6), 190–193 (2009)
5. Yong, F., Hua, L.: Based on Adaptive Chinese word segmentation and approximation of SVM text classification algorithm. Computer Science 37, 251–254, 293 (2010)
6. Cao, B., Han, Z.: ASP.NET database system project development practice. Science Press, Beijing (2005)
7. Wang, B.: Database access technology based on ASP.NET. Computer Application and Software 21(2), 120–122 (2004)
8. Jeroslow, R., Wang, J.: Solving propositional satisfiability problems. In: Annals of Mat Hematics and Artificial intelligence. Springer (1990)
9. Nie, J.-Y.: Unknown Word Detection and Segmentation of Chinese using Statistical and-heuristic Knowledge. Communications of COLIPS 5(I&2), 47–57 (2008)
10. Luo, Z., Song, R.: The adaptive method for Chinese new word identification based on multiple feature. Journal of Beijing University of Technology 23(7), 718–725 (2007)
11. Li, Y., Wang, H.: Intelligent computer assisted instruction system of knowledge ambiguity elimination. Computer Technology and Development 19(4), 220–223 (2009)