Journal of Southeast University (English Edition) Vol. 23, No. 2, pp. 179 – 184 June 2007 ISSN 1003—7985

Feature study for improving Chinese overlapping ambiguity resolution based on SVM

Xiong Ying Zhu Jie

(Department of Electronic Engineering, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: In order to improve Chinese overlapping ambiguity resolution based on a support vector machine, statistical features are studied for representing the feature vectors. First, four statistical parameters—mutual information, accessor variety, two-character word frequency and single-character word frequency are used to describe the feature vectors respectively. Then other parameters are tried to add as complementary features to the parameters which obtain the best results for further improving the classification performance. Experimental results show that features represented by mutual information, single-character word frequency and accessor variety can obtain an optimum result of 94.39%. Compared with a commonly used word probability model, the accuracy has been improved by 6.62%. Such comparative results confirm that the classification performance can be improved by feature selection and representation.

Key words: support vector machine; Chinese overlapping ambiguity; Chinese word segmentation; word probability model

Unlike English and other western languages, Chinese is character based, not word based. There are no "blank spaces" serving as word boundaries in Chinese sentences. Word segmentation is the first step for Chinese information processing, which is one of the bottlenecks restricting the rapid development of Chinese language understanding. An important factor which influences the accuracy of word segmentation is ambiguity segmentation.

Generally speaking, ambiguity segmentation can be classified into two classes: overlapping ambiguity (OA) and combination ambiguity (CA)[1]. OA accounts for the greater part of the ambiguity segmentation[2]. Especially, the overlapping ambiguity strings (OAS) with the length of three characters is in the majority. Here, the 3-character OAS is used as an example to describe the definition.

Given a Chinese character string $c_1c_2c_3$, if $c_1c_2 \in W$ and $c_2c_3 \in W$, then $c_1c_2c_3$ is called a 3-character OAS, where c_1, c_2 and c_3 are all Chinese characters, W is a Chinese lexicon.

1 Previous Work

Previous methods can be classified into two types: the rule-based approach and the statistical-based approach. Although rule-based approaches may obtain better performance, it takes more time and effort to build rule sets. Furthermore, it is impossible for a human to enumerate all possible instances. Hence, statistical based approaches have become the mainstream used in resolving overlapping ambiguity.

Ref. [3] adopted the traditional word probability model (WPM) to resolve the problem of OAS in closed sets and obtained an accuracy of 85.46%. Then it was combined with the rule based method, and the results of up to 92.07% could be achieved. Although a significant improvement is gained, it takes more time and effort for a human to build rule sets. Ref. [4] viewed the overlapping ambiguity resolution as a classification problem and used a support vector machine (SVM) combined with a K-NN classifier. The choice of classifier is determined by the distance of the sample to the optimal hyperplane. When the distance exceeded the threshold ε, the SVM classifier was adopted. Otherwise, the K-NN classifier was used. The method took the mutual information (MI) between each character pair as a feature and obtained a result of 91.65%. However, the value of the threshold is mainly decided by the experiment. And only MI is used as features of SVM. Ref. [5] presented an ensemble of adapted Naïve Bayesian classifiers which used context words within windows of different sizes as features to resolve ambiguities. The approach obtained a result of 94.3%. This method should make independency assumptions and take measures for smoothing. Ref. [6] employed a maximum entropy (ME) model as a classifier to deal with overlapping ambiguity. The experimental results

Received 2006-11-06.

Biographies: Xiong Ying (1977—), female, graduate; Zhu Jie (corresponding author), male, doctor, professor, zhujie@sjtu.edu.cn.

showed that this method could obtain an accuracy of 95.01% in open test set. From all the above results, it is difficult to estimate which method is better because each method uses different data sets.

Our work is based on Ref. [4]. We still view the overlapping ambiguity resolution as a classification problem, but only the SVM classifier is adopted for its excellent learning ability, classification ability and high generalization performance[7]. In this paper, we mainly focus on the features for improving the classification accuracy. The statistical characteristics inside the overlapping ambiguity strings are exploited. Other than Ref. [4], which combines different classifiers for further improving the accuracy, this paper makes use of the SVM ability of multi-feature fusion.

2 Features for Improving Chinese Overlapping Ambiguity Resolution Based on SVM

Since most overlapping ambiguity strings are context-free, overlapping ambiguity resolution can be regarded as a binary classification problem[4].

For example, given a 3-character Chinese OAS $c_1c_2c_3$, there are two segmentation paths (seg_1 and seg_2) for $c_1c_2c_3$ to be segmented into words. We define "$c_1c_2 \mid c_3$" as positive segmentation (seg_1) and define "$c_1 \mid c_2c_3$" as negative segmentation (seg_2). Which path should be selected mostly depends on the statistical characteristics among the OAS.

2.1 Feature selection criterion

In order to improve the classification accuracy, it is very important to select features which can clearly describe the differences between the two segmentation paths. The combination strength of neighbor characters, word frequency and word independency can reflect the differences in different aspects.

Taking the above ambiguity string $c_1c_2c_3$ as an example, if one of the following four conditions is satisfied, it is more possible for the string $c_1c_2c_3$ to choose the positive segmentation.

① Character c_1 and character c_2 combine more strongly than characters c_2 and c_3;

② Word c_1c_2 appears more frequently than word c_2c_3;

③ Character c_3 serves as a single character word more often than character c_1;

④ Word c_1c_2 can appear in more different context than word c_2c_3, i. e. word c_1c_2 is more independent than word c_2c_3.

For example, given a Chinese overlapping ambiguity string "出具有(chu1-ju4-you3)", when only considering conditions ①, ② and ④, it will be segmented into "出/具有". After adding condition ③, the right segmentation form ("出具/有") is obtained.

2.2 Feature representation

Several parameters (MI, accessor variety (AV), single-character word frequency (SWF), and two-character word frequency (TWF)) are used to represent the features based on the SVM.

- Mutual information

MI was used as an index to measure the absolute combination strength between two Chinese characters in Refs. [4, 8]. In this paper, MI is also adopted to represent features of the SVM. The representation method is the same as in Refs. [4, 8].

For example, given a 3-character OAS $c_1c_2c_3$, MI_1 means the mutual information between the two characters c_1 and c_2, MI_2 means the mutual information between character c_2 and character c_3. The formula of MI can be defined as

$$\text{MI}(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)} = \log_2 \frac{Nr(x, y)}{r(x)r(y)} \quad (1)$$

where $p(\cdot)$ represents the probability; $r(\cdot)$ is the occurrence frequency; and N is the total number of Chinese characters in the corpus.

In the experiment, data sparseness problems still exist. The method proposed in Ref. [8] is used for smoothing.

Suppose that $r(\cdot)$ is the actual occurrence frequency; $r^*(\cdot)$ is the frequency after smoothing, then

$$r^* = \frac{(r+1)N}{N+S} \quad (2)$$

where S is a constant value ($S = 6\,775$), representing the number of Chinese characters which are used most frequently.

- Accessor variety[9-10]

AV is used to measure the independence of a string. The bigger the value, the greater the independence. The AV of a string s is defined as

$$\text{AV}(s) = \min\{L_{\text{av}}(s), R_{\text{av}}(s)\} \quad (3)$$

where $L_{\text{av}}(s)$ means the left accessor variety of s, which is defined as the number of its distinct preceding characters; $R_{\text{av}}(s)$ means the right accessor variety of s, which is defined as the number of its distinct succeeding characters. If s appears at the beginning of a sentence more than once, the value of $L_{\text{av}}(s)$ should add 1 each time when s appears at the beginning of a sentence again, so as to obtain the value of $R_{\text{av}}(s)$.

To illustrate the definitions of these novel statistical criteria clearly, let us take the following six Chinese sentences as an example.

① 门把手弄坏了. (“The door hurts the hand” or “The door handle is broken.”)

② 小明修好了门把手. (“Xiao Ming fixed the door handle.”)

③ 这个门把手很漂亮. (“This door handle is very beautiful.”)

④ 这个门把手坏了. (“This door handle is broken.”)

⑤ 门把手坏了吗? (“Is the door handle broken?”)

⑥ 谁修好了门把手?(“Who fixed the door handle?”)

Considering the string “门把手” in these six sentences, it has 3 distinct preceding characters, i. e., “B (which represents the beginning of a sentence)”, “了”, “个”, and 4 distinct succeeding characters, i. e., “弄”, “E (which means the end of a sentence)”, “很”, “坏”. As the string “门把手” appears twice at the beginning and ending of sentences, the values of L_{av}(“门把手”) and R_{av}(“门把手”) should add 1 respectively. Therefore, the value of L_{av}(“门把手”) is 4, and the value of R_{av}(“门把手”) is 5. And the AV value of “门把手” is $\min(4,5)=4$.

From the definition of AV, it is clear that the AV value of a string can reflect the independence of the string in a given corpus. The more appearances under different linguistic environments, the more independence. As to the problem of overlapping ambiguity resolution, if $\mathrm{AV}(c_1c_2) > \mathrm{AV}(c_2c_3)$, which means the word c_1c_2 can appear under more different context environments than the word c_2c_3 in the corpus, then the OAS $c_1c_2c_3$ tends to choose positive segmentation as the correct path. Here, we set

$$\mathrm{AV}_1 = \mathrm{AV}(c_1c_2) \tag{4}$$

$$\mathrm{AV}_2 = \mathrm{AV}(c_2c_3) \tag{5}$$

- Single-character word frequency

Given a 3-character ambiguity string $c_1c_2c_3$, if the occurrence frequency of single character word c_3 is larger than that of c_1 in the training corpus, then positive segmentation is more possible than negative segmentation.

Here we set

$$\mathrm{SWF}_1 = \mathrm{SWF}(c_3) = r^*(c_3) \tag{6}$$

$$\mathrm{SWF}_2 = \mathrm{SWF}(c_1) = r^*(c_1) \tag{7}$$

where the meaning of $r^*(\cdot)$ is the same as in Eq. (2).

- Two-character word frequency

Supposing a 3-character ambiguity string $c_1c_2c_3$, if word c_1c_2 appears more frequently than word c_2c_3 in the training corpus, then the proper path may trend to select positive segmentation.

Here we set

$$\mathrm{TWF}_1 = \mathrm{TWF}(c_1c_2) = r^*(c_1c_2) \tag{8}$$

$$\mathrm{TWF}_2 = \mathrm{TWF}(c_2c_3) = r^*(c_2c_3) \tag{9}$$

where the meaning of $r^*(\cdot)$ is the same as Eq. (2).

3 Experiments

3.1 Experimental condition

The Lancaster corpus of Mandarin Chinese (LCMC)[11] is used as the training set. 220 articles about education are used as test data. The forward maximum match (FMM) and the backward maximum match (BMM) methods are used for determining the overlapping ambiguity strings. The dictionary used in the FMM and the BMM contains about 108 750 words. Each OAS has been manually proofread with the original segmented training corpus. There are 8 594 and 1 194 3-character overlapping ambiguity strings in the training set and the test set, respectively.

In the experiment, LIBSVM[12] is adopted to train the SVM classifier.

3.2 Kernel function and parameter selection

In order to measure the influence of the kernel function on the classification accuracy, three common kernel functions are tested in the experiment.

- Linear

$$K(x, x_i) = (x \cdot x_i) \tag{10}$$

- The d-th polynomial

$$K(x, x_i) = [(x \cdot x_i) + 1]^d \tag{11}$$

- The Gaussian radial basis function (RBF)

$$K(x \cdot x_i) = \exp\left(-\frac{|x - x_i|^2}{\delta^2}\right) = \exp(-\sigma \cdot |x - x_i|^2) \tag{12}$$

In this experiment, parameters are set to the default values in the LIBSVM. All the overlapping ambiguity strings in the training set are used as training samples. The 5-fold cross validation method is used for parameter optimization. The features are two dimensional vectors which are composed of MI_1 and MI_2 as mentioned in section 2.2, i. e. $\langle \mathrm{MI}_1, \mathrm{MI}_2 \rangle$. The experimental results are shown in Tab. 1.

Tab. 1 Performance comparison with different kernel functions

Kernel function	Number of support vector	Accuracy/%
Linear	3 233	88.78
The 3rd polynomial	4 762	83.33
RBF	3 207	89.28

From the results, we can see that the kernel function of the RBF performs best. Therefore, the kernel functions in the following experiments will adopt the

RBF. There are two parameters while using the RBF kernels: penalty constant C and kernel parameter σ. Fig. 1 describes the relationship between the classification accuracy and the SVM parameters (C, σ). Fig. 2 depicts the relationship between the number of support vectors and the SVM parameters (C, σ). From Fig. 1, we can see that the classification accuracy attains a maximum value when $C = 1$, $\sigma = 0.5$; when $C = 10$, $\sigma = 10$, the accuracy is a bit lower than the maximum value, but the number of support vectors is much smaller than the former, as shown in Fig. 2. However, the greater the number of support vectors, the lower the speed of the classification. So the parameters C and σ will be set to $C = 10$, $\sigma = 10$.

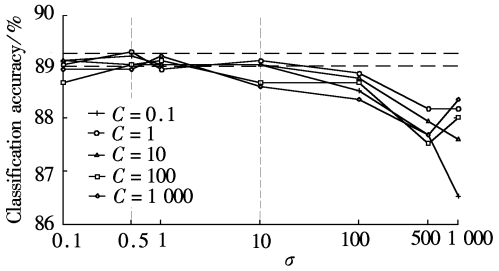

Fig. 1 Classification accuracy under different SVM parameters

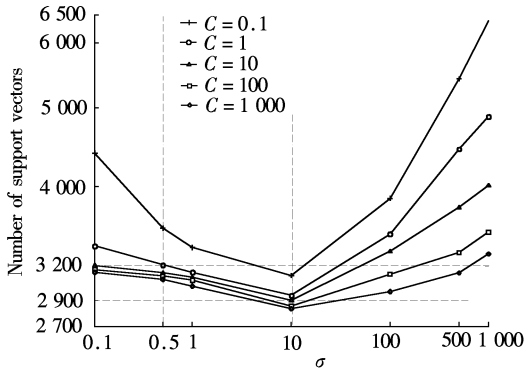

Fig. 2 Number of support vectors under different SVM parameters

3.3 Performances of different feature representations

In order to evaluate the performances when different statistical criteria are used to represent the features of SVM, four statistical criteria (MI, AV, SWF, TWF) are used to represent the feature vectors of two dimensions respectively, such as $\langle AV_1, AV_2 \rangle$. The results are shown in Tab. 2.

Tab. 2 The performance comparison with different statistical criteria

Statistical criterion	Number of support vectors	Accuracy/%
MI	2 906	89.03
AV	6 227	69.51
SWF	3 681	87.35
TWF	6 504	67.84

Tab. 2 shows that features represented by MI perform best. Features represented by TWF have the lowest accuracy of classification.

Since the features represented by MI obtain promising results, three other statistical parameters (AV, SWF, TWF) are added as complementary features to the SVM, respectively, so as to further improve the classification performance. Then features can be represented as vectors of four dimensions, such as $\langle MI_1, MI_2, AV_1, AV_2 \rangle$. The training samples are the same as in section 3.2. Tab. 3 shows the results when different features are added to MI.

Tab. 3 Feature complementarity comparison with MI

Statistical criterion	Number of support vector	Accuracy/%
MI	2 906	89.03
MI + AV	2 714	91.29
MI + TWF	2 646	91.29
MI + SWF	2 436	92.96

It is shown in Tab. 3 that the accuracy has improved 3.93% after the feature SWF is added. While the performances achieve improvements of 2.26% after adding AV and TWF respectively. Hence, compared with the two other parameters AV and TWF, SWF gets the best complementarity with MI.

As the SVM can be suitable for dealing with high dimensional vectors, other two features (AV, TWF) are added to MI + SWF so as to try the possibility for further improving the performance. The results are shown in Tab. 4.

Tab. 4 Feature complementarity comparison with MI + SWF

Statistical criterion	Number of support vectors	Accuracy/%
MI + SWF	2 436	92.96
MI + SWF + TWF	2 303	93.47
MI + SWF + AV	2 260	94.14
MI + SWF + AV + TWF	2 298	93.80

From Tab. 4, we can see that MI + SWF + AV obtains the best performance. The accuracy can achieve up to 94.14%. Compared with MI and MI + SWF, the result has improved 5.11% and 1.18%, respectively. However, when TWF is added to MI + SWF + AV, the accuracy declines a little. That is because TWF cannot serve as a complement to AV. So it is important for the SVM to add complementary features for improving the classification accuracy.

In the experiment, the computational cost increases little when multi-features are added to the SVM. Because the computational cost is mostly dependent on the number of training samples not the dimension of the features.

3.4 Classification performance vs. number of training samples

Finally the relationship between the classification performance and the number of training samples is tested. We compare the performances of SVM classifiers represented by different dimensional features with word probability model methods, respectively. The word probability model method is viewed as a baseline which can be described as follows:

Suppose that $W = c_1 c_2 \ldots c_n$ is an OAS, where $c_i (i = 1, 2, \ldots, n)$ are Chinese characters. There are two segmentation paths $(\text{seg}_1, \text{seg}_2)$ for W to be segmented into words. The task of WPM is to choose the path with the maximum product of word unigram probabilities.

$$w^* = \arg\max_{\text{seg}} p(c_1 c_2 \ldots c_n) = \arg\max_{w} \prod_{i=1}^{k} p(w_i) \tag{13}$$

where w_i is the i-th word, and k is the number of words in W.

In the experiment, the accuracy of WPM is 87.77%, which gives no change when the number of training samples is increased. Because the performance of WPM is only related to the size of the training set, not the number of training samples.

Fig. 3 to Fig. 5 show the relationship between the accuracy of the SVM classifiers represented by different features and the number of training samples. From Fig. 3, we can see that the SVM classifier with features represented by MI obtain better performance than that of the baseline, while others represented by AV, TWF and SWF are lower than that of the baseline. Furthermore, the performances of all these four SVM classifiers change slightly when the number of the training samples increases. After AV, TWF and SWF are added as complementary features to MI respectively, the accuracy improves rapidly when the number of training samples is smaller than 2 000. After that, the classification accuracy changes with the increment of the training samples. When the number of training samples reaches 8 000, the SVM classifier represented by MI and SWF attains its best performance of 93.05%, as shown in Fig. 4.

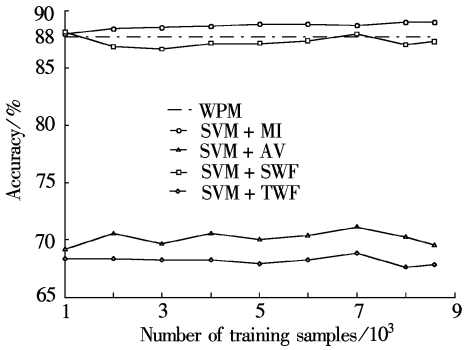

Fig. 3 Accuracy with two dimensional vectors vs. number of training samples

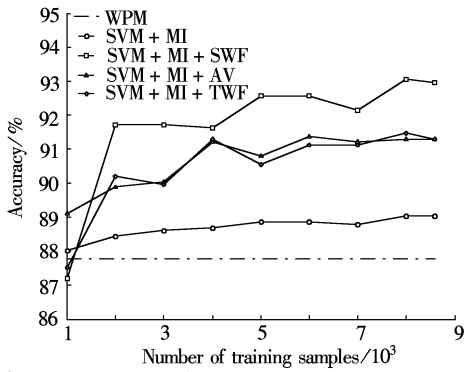

Fig. 4 Accuracy with four dimensional vectors vs. number of training samples

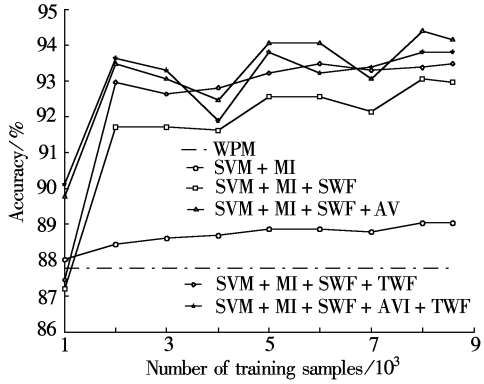

Fig. 5 Accuracy with six/eight dimensional vectors vs. number of training samples

Fig. 5 shows that the performances of the SVM classifiers represented by MI + SWF + AV, MI + SWF + TWF and MI + SWF + AV + TWF are better than that of MI + SWF. The SVM classifier with features of MI + SWF + AV achieves a global maximum value of 94.39% when the number of the training sample is 8 000. Compared with the baseline, the accuracy has improved 6.62%.

4 Conclusion

This paper studies the features for improving the overlapping ambiguity resolution based on SVM. Four statistical criteria are employed to represent the feature vectors. First, two-dimensional vectors are represented as features. MI outperforms other three statistical criteria. After that, these three statistical criteria are added to MI as complementary features, the performance of the SVM classifier represented by MI + SWF improves notably. Finally, the SVM classifier represented by MI + SWF + AV gets the optimal performance of

94.39%. The classification accuracy has enhanced 6.62% when compared with the baseline.

The experimental results prove that the performance of the SVM classifiers can be improved by feature selection and representation. Future work will focus on the stability of the SVM classifiers when different data are used.

References

[1] Liang Nanyuan. CDWS—the modern printed Chinese distinguishing word system[J]. *Journal of Chinese Information Processing*, 1987, **1**(2): 44 - 52. (in Chinese)

[2] Sun Maosong, Zuo Zhengping, Tsou B K. The role of high frequent maximal crossing ambiguities in Chinese word segmentation[J]. *Journal of Chinese Information Processing*, 1999, **13**(1): 27 - 34. (in Chinese)

[3] Sun Maosong, Zuo Zhengping, Huang Changning. Algorithm for solving 3-character crossing ambiguities in Chinese word segmentation[J]. *Journal of Tsinghua University Science and Technology*, 1999, **39**(5): 101 - 103. (in Chinese)

[4] Li Rong, Liu Shaohui, Ye Shiwei, et al. A method of crossing ambiguities in Chinese word segmentation based on SVM and K-NN[J]. *Journal of Chinese Information Processing*, 2001, **15**(6): 13 - 18. (in Chinese)

[5] Li Mu, Gao Jianfeng, Huang Changning, et al. Unsupervised training for overlapping ambiguity resolution in Chinese word segmentation[C]//*Proceeding of the Second Sighan Workshop on Chinese Language Processing*. Sapporo, 2003: 1 - 7.

[6] Zhang Feng, Fan Xiaozhong. Resolution of overlapping ambiguity strings based on maximum entropy model[J]. *Transactions of Beijing Institute of Technology*, 2005, **25**(7): 590 - 593. (in Chinese)

[7] Vapnik V N. *The nature of statistical learning theory*[M]. New York: Springer Verlag, 1995: 1 - 188.

[8] Sun Maosong, Huang Changning, Tsou B K, et al. Using character bigram for ambiguity resolution in Chinese word segmentation[J]. *Computer Research and Development*, 1997, **34**(5): 332 - 339. (in Chinese)

[9] Feng Haodi, Chen Kang, Kit Chunyu, et al. Unsupervised segmentation of Chinese corpus using accessor variety[C]//*Lecture Notes in Artificial Intelligence*. Berlin: Springer Verlag, 2005, **3248**: 694 - 703.

[10] Feng Haodi, Chen Kang, Deng Xiaotie, et al. Accessor variety criteria for Chinese word extraction[J]. *Computational Linguistics*, 2004, **30**(1): 75 - 93.

[11] McEnery Tony, Xiao Richard. The lancaster corpus of Mandarin Chinese[EB/OL]. (2004-09-15)[2005-12-07]. http://bowland-files.lancs.ac.uk/corplang/lcmc/.

[12] Chang Chih-Chung, Lin Chih-Jen. LIBSVM—a library for support vector machines[EB/OL]. (2005-11-30)[2006-04-01]. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

基于支持向量机的改进中文交集型歧义消解特征研究

熊 英 朱 杰

(上海交通大学电子工程系,上海 200240)

摘要: 为了提高支持向量机用于解决中文交集型歧义的能力,研究了用于表示特征向量的4种统计量.首先,分别用互信息、附属种类、二字词频和单字词频单独描述特征向量,然后将其他的统计量分别与分类性能最佳的统计量结合以进一步提高分类的正确率.实验结果表明,采用互信息、单字词频和附属种类表示的特征向量所取得的分类性能最优,正确率可达94.39%.与常用的词概率模型相比,正确率提高了6.62%.由此证明了特征的选择和表示方法对提高分类性能的有效性.

关键词: 支持向量机;中文交集型歧义;中文词语切分;词概率模型

中图分类号: TP391