中 文 信 息 学 报
第 18 卷 第 2 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 18 No. 2

文章编号:1003 - 0077(2004)02 - 0001 - 07

基于 SVM 的中文组块分析*

李 珩,朱靖波,姚天顺

(东北大学 计算机软件与理论研究所,辽宁 沈阳 110004)

摘要:基于 SVM(support vector machine)理论的分类算法,由于其完善的理论基础和良好的实验结果,目前已逐渐引起国内外研究者的关注。和其他分类算法相比,基于结构风险最小化原则的 SVM 在小样本模式识别中表现较好的泛化能力。文本组块分析作为句法分析的预处理阶段,通过将文本划分成一组互不重叠的片断,来达到降低句法分析的难度。本文将中文组块识别问题看成分类问题,并利用 SVM 加以解决。实验结果证明,SVM 算法在汉语组块识别方面是有效的,在哈尔滨工业大学树库语料测试的结果是 F = 88.67%,并且特别适用于有限的汉语带标信息的情况。

关键词:计算机应用;中文信息处理;支持向量机;结构风险最小化;文本组块

中图分类号:TP391 **文献标识码**:A

SVM Based Chinese Text Chunking

LI Heng, ZHU Jing-bo, YAO Tian-shun

(Institute of Computer Software and Theory, Northeastern University, Shenyang, Liaoning 110004, China)

Abstract: The classification algorithm based on SVM (support vector machine) attracts more attention from researchers due to its perfect theoretical properties and good empirical results. Compared with other classification algorithms, structural risk minimizations based SVM achieve high generalization performance with small number of samples. The text chunking, as a preprocessing step for parsing, is to divide text into syntactically related non-overlapping groups of words (chunks), reducing the complexity of the full parsing. In this paper, we treat Chinese text chunking as a classification problem, and apply SVM to solve it. The chunking experiments were carried out on the HIT Chinese Treebank corpus. Experimental results show that it is an effective approach, achieving an F score of 88.67%, especially for a small number of Chinese labeled samples.

Key words: computer application; Chinese information processing; support vector machine; structural risk minimization; text chunking

1 引言

近几年来,机器学习方法在自然语言处理中的应用获得很大的关注。传统的机器学习方法,如隐马尔科夫模型(HMM),最大熵(maximum entropy),naïve bayes,决策树(decision tree),神经网络(neural network)等虽然取得了一定的成功,但这些方法仍然存在一些问题,一是容易产生对训练数据的过适应问题,二是需要采用启发式学习进行特征的选取。基于结构

* 收稿日期:2003 - 07 - 25
基金项目:国家自然科学基金资助项目(60083006);国家重点基础研究发展规划 973 资助项目(G19980305011);国家自然科学基金和微软亚洲研究院联合资助项目(60203019)
作者简介:李珩(1975—),男,东北大学博士生,主要研究方向为统计语言处理.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

风险最小化归纳原则的 SVM(support vector machine, 支持向量机)由于可以控制整个样本集期望风险,因此可以避免对训练数据产生过适应的缺点,而且由于引入核函数,在解决线性不可分以及高维特征空间稀疏问题表现出很好的性能。

组块分析作为一种预处理手段,可以大大降低进行短语划分和短语分析处理的复杂性,为进一步对句子的深层次分析提供了基础,使得句法分析任务在某种程度上得到简化,同时对机器翻译、信息提取、信息检索、专有名词识别等都具有非常重要的意义。目前,应用于组块分析的机器学习方法有,基于转换的学习[1],基于记忆的学习[2],隐马尔科夫模型[3],最大熵[4]等。

在英文组块(chunk)识别方面,文献[5][6]使用支持向量机的机器学习方法进行了 chunk 的识别,在 CoNLL - 2000 的共享任务组块处理中取得了最好的性能。

在汉语组块识别方面,国内已有人做过一些有益的工作。例如张昱琪等[18]利用短语内部结构和词汇信息对预测中出现的边界歧义和短语类型歧义进行了排歧处理。周强等[19]介绍的汉语句子的组块分析体系,通过引入词界块和成分组概念,将成分识别问题从完整的句法分析任务中分离出来。赵军等[20]从语言学的系统角度定义了汉语基本名词短语,提出了将汉语基本名词短语的结构模板和其上下文环境特征结合的汉语基本名词短语识别模型。以上的研究都取得了很好的结果。

SVM 是一个学习分类器的有效方法,这种方法已经成功地运用在自然语言处理的任务中,如英语组块识别[5,6],英语词性标注[7],英语专名识别[8]和文本分类[9]等方面,并在这些领域都取得了不错的效果,但是在汉语组块分析方面却没有相关的报道,所以本文使用了 SVM 来进行汉语组块的识别。

本文第 2 节介绍了 SVM 学习算法,第 3 节描述了基于 SVM 的中文组块分析,第 4 节给出了实验结果并得出一些重要的结论,最后总结全文,并指出进一步研究的方向。

2 SVM 学习算法

首先,我们先简单的回顾一下 SVM 理论,这构成了本文的理论基础。

2.1 SVM 理论概述

支持向量机的理论最初来自于对数据二值分类问题的处理。其机理可以简单地描述为:寻找一个满足分类要求的最优超平面,使其在保证分类精度的同时最大化超平面两侧的空白区域。这使得 SVM 分类器的结果不仅在训练集上得到优化,而且在整个样本集上的风险也拥有上界,这就是 SVM 的结构风险最小化的思想。

对于线性不可分的问题,Vapnik[15,16]等人成功地引入了核空间理论,将低维的输入向量通过非线性映射函数映射到高维特征空间。可以证明,如果选用适当的映射函数,大多数输入空间线性不可分问题在特征空间可以转化为线性可分问题。

2.2 多分类支持向量机(Multi-class SVM)

SVM 构建了一个二值分类器,仅能够对两个类别进行分类,在多类别的情况下需要将 SVM 扩展到多个类别的分类器。对于 n 个类别的分类问题,目前构造多分类器的方法有如下两种:

1. One-against-all:从 n 个类别中为每一个类别创建二值分类问题,也就是,对于每一个类别 $i(1 \leq i < n)$,构造 n 个 SVM 二值分类器 $c_i, i \in \{1,2 \cdots n\}$,这里标记为 $y = i$ 的实例被认为是正例,所有其余标记的实例被认为是反例。

2. Pairwise:在任意类别 i 和类别 $j(1 \leq i, j \leq n, i \neq j)$ 之间构造一个 SVM 二值分类器,从

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

而生成了 $n(n-1)/2$ 个 SVM 二值分类器 $c_i(i \in \{1,2 \cdots n(n-1)/2\})$,使用给定的二值学习算法来区别所有类别的每一对,这里标记为 $y=i$ 的实例被认为是正例,标记为 $y=j$ 的实例被认为是反例,对于一个未知样本每个分类器都有一个选票,其结果是具有选票最多的类别。

3 基于 SVM 的中文组块分析

文本组块是指将一个输入文本划分成一组互不重叠的片断,这些片断是非递归的,即片断不能嵌套,这些片断定义为 Chunk[10]。请看一个文本组块的例子:

[BNT 10 分钟]后队伍就[BVP 分散开了]。

当然,也可以通过为 Chunk 加标记来表示文本组块。本文采用 IOB2[11]的标注集合,该标注集合包含三种类型的标记:B-X 表示 Chunk 类型为 X,并且是该 Chunk 的起始词,I-X 表示 Chunk 类型为 X,并且是该 Chunk 的非起始词,O 表示不在任何 Chunk 内的词。于是,上述的例子也可以表示如下:

10/ B-BNT 分钟/ I-BNT 后/ O 队伍/ O 就/ O 分散/ B-BVP 开/ I-BVP 了/ I-BVP。/ O

这样,文本组块分析过程也可以看成对文本进行 Chunk 分类过程。传统的 SVM 算法解决的是数据的二值分类问题,因此,为了应用于组块分析,我们必须将其扩展为能够识别多个类别。在特征选取上,本文采用了和文献[6]相同的上下文作为特征,即当前词及前两个词和后两个词,当前词词性及前两个词词性和后两个词词性,前两个词的组块类型标记。

3.1 特征向量

在实验中,我们并没有如传统方法(像最大熵、naïve bayes 等)进行特征选取,主要考虑到 SVM 在处理高维特征空间仍能具有良好的泛化能力,并能够在具有多种特征组合的情况下进行训练。所以我们将出现在训练数据中不同位置的所有单词 w、词性标记 p 和组块类型标记 t 作为特征,充分利用当前标记位置的上下文信息,将每一个样本 x 用 12 个特征 f 来表示:

$$x = (w_{-2}, p_{-2}, t_{-2}, w_{-1}, p_{-1}, t_{-1}, w_0, p_0, w_{+1}, p_{+1}, w_{+2}, p_{+2});$$

在本文中,假设分类过程是从左到右进行的,这个可以从下面的特征定义中看到。

w_0:表示当前位置的单词,p_0 表示 w_0 的词性标记;

w_{-i}:表示从当前位置往前数第 i 个的单词,p_{-i} 表示 w_{-i} 的词性标记,t_{-i} 表示 w_{-i} 的组块类型标记。

w_{+i}:表示从当前位置往后数第 i 个单词,p_{+i} 表示 w_{+i} 的词性标记。

对于特征 SVM 二值分类器仅接受数字化的值,为了满足这个限制,通过构建一个关于特征的倒排索引表 InvTab,其中的每个记录为二元组$\langle f, index_w \rangle$,其中 $index$ 是特征 f 在的特征列表中的位置。如$\langle w_{-2}=$队伍,1001$\rangle$,表示"w_{-2}队伍"这个特征是特征列表中的第 1001 个元素。特征倒排索引表按特征以散列方式组织,能够实现快速查找,从而很方便地将每一个样本的特征用一系列的数字来表示。由于测试集中并没有组块类型标记,因此,在标注过程中需要动态确定,可以用动态规划技术来解决。

3.2 多分类

设任意一个有 n 个词的汉语句子表示为:$W^T = w_1 w_2 \cdots w_n$, $w_i(1 \leq i \leq n)$,表示第 i 个单词,该汉语句子的词性序列为 $P^T = P_1 P_2 \cdots P_n$, $P_i(1 \leq i \leq n)$ 表示 w_i 的词性,该汉语句子的组块类型标记序列为 $T^T = t_1 t_2 \cdots t_n$, $t_i(1 \leq i \leq n)$ 表示 W^T 所对应的组块类型标记。则汉语组块识别问题,可以看作是在给定词序列 W^T 和其所对应的词性序列 T^T 的情况下,将每一个

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$w_i(1 \le i \le n)$所对应的组块类型标记 $t_i(1 \le i \le n)$进行分类的过程。由文献[12]可知,Pairwise 方法取得比 One-against-all 方法更好的性能,因此本文采用的是 Pairwise 方法。

4 实验结果及分析

我们采用哈工大公开的中文树库语料:Chinese Treebank[2] 作为我们的训练语料和测试语料。它包含 2000 个句子,21498 个词。目前可以识别的组块类型包括:BDP, BAP, BMP, BNT, BNS, BNP, BVP。

本文取前 1500 个句子作为训练集,后 500 个句子作为测试集。实验中的性能指标定义如下:

$$\text{组块准确率}(Precision) = \frac{\text{正确标注的组块的个数}}{\text{标注的组块的总个数}} \times 100\ \%$$

$$\text{组块召回率}(Recall) = \frac{\text{正确标注的组块的个数}}{\text{正确的组块的总个数}} \times 100\ \%$$

$$F\beta = \frac{(\beta^2 + 1) \times Recall \times Precision}{\beta^2 \times Recall + Precision}$$ 其中取值 $\beta = 1$

4.1 选择最佳阶次多项式

表 1 各个阶次多项式 SVM 的组块分析结果,最好的结果用黑体表示

组块类型	Precision Recall FB						
	d = 1 dim = 16091	d = 2 $dim10^7$	d = 3 $dim10^{11}$	d = 4 $dim10^{14}$	d = 5 $dim10^{17}$	d = 6 $dim10^{20}$	d = 7 $dim10^{23}$
ALL	87.02 85.00 86.00	90.82 **86.61** **88.67**	90.20 81.34 85.54	90.01 73.21 80.75	**91.87** 60.54 72.98	89.50 43.39 58.45	69.82 21.07 32.37
BAP	71.43 70.83 71.13	80.70 **76.67** **78.63**	79.05 69.17 73.78	83.33 62.50 71.43	**89.55** 50.00 64.17	89.19 27.50 42.04	88.89 20.00 32.65
BDP	90.00 75.00 81.82	**100.00** **75.00** **85.71**	100.00 33.33 50.00	100.00 33.33 50.00	0.00 0.00 0.00	0.00 0.00 0.00	0.00 0.00 0.00
BMP	93.78 **95.05** **94.41**	**94.12** 93.69 93.91	92.17 90.09 91.12	89.05 84.23 86.57	92.47 77.48 84.31	88.08 59.91 71.31	56.60 27.03 36.59
BNP	86.35 87.73 87.03	90.16 **89.55** **89.85**	91.00 85.00 87.90	90.81 74.09 81.60	**93.36** 57.50 71.17	90.63 39.55 55.06	57.61 12.05 19.92
BNS	50.00 **22.22** **30.77**	**100.00** 11.11 20.00	0.00 0.00 0.00	0.00 0.00 0.00	0.00 0.00 0.00	0.00 0.00 0.00	0.00 0.00 0.00
BNT	74.07 68.97 71.43	**80.00** **68.97** **74.07**	72.00 62.07 66.67	53.85 24.14 33.33	40.00 13.79 20.51	40.00 6.90 11.76	50.00 6.90 12.12
BVP	91.22 82.99 86.91	**94.25** **85.42** **89.62**	93.55 80.56 86.57	94.04 76.74 84.51	92.65 65.63 76.83	91.14 50.00 64.57	88.99 33.68 48.87

本实验采用 d 阶多项式作为核函数:$K(x, x_j) = [(x \cdot x_j) + 1]^d$,通过逐渐增大多项式阶数 d,来观察组块识别的效果,从而选择最佳阶次多项式。需要指出的是,本文并没有选取其

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

他核函数如径向基、多层感知机进行实验,基于两点考虑,一是采用 d 阶多项式容易比较线性($d=1$)和非线性($d\geq 2$)两种情况的分类效果以及在高维特征空间的泛化能力。二是通过选取适当的参数,所有三种类型(多项式、径向基、多层感知机)的 SVM 表现出大致相同的性能,在文本分类[9]实验中,径向基比多项式仅高 0.4 %,而在数字识别[16]中,三者的粗错误率相差也不过 0.1 %。

dim 为特征空间的维数,其计算公式为:$(n/d)^d$,n 为输入空间的维数,我们对训练集进行了特征抽取,共抽取了 16091 个特征,因此 $n=16091$。

由实验结果可以得出以下结论:

1) 对多项式 SVM,分类器的性能并不随着特征空间维数的增高而有大的变化,说明不存在过适应问题。

2) 即使没有进行特征选取,SVM 仍然保持良好的性能,可见 SVM 的泛化能力并不严格地依赖于特征的数量。

3) 在大量相关特征的特征空间中,样本对应的特征值不为零的特征却很稀疏(本实验中当前词对应的特征值不为零的特征为 12),显然,SVM 对处理这类情况非常适合。

4) 对从 2 阶到 7 阶的多项式,即使在最好和最差解之间差别小的情况下,SVM 也给出了一个逼近最佳解的方法。

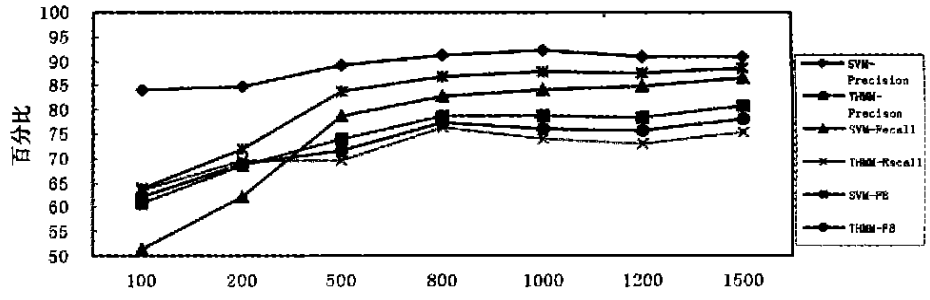
实验结果显示,各种类型组块识别结果差别较大。其中 BNS 和 BNT 的识别效果相对不理想。由于 BNS 和 BNT 是 BNP 的子类,这种语法结构上的易混淆性反映在实验结果里。针对这种情况,我们尝试利用能够对语义信息有一定反映的词汇信息对这种类型错误进行纠正。BNS 和 BNT 的主要标志性词语见表 2:

表 2　BNS, BNT 两种类型的主要标志性词语

	标志性词语
BNS	地区、省、市、县
BNT	年代、时、时候、时期、时代、阶段、时刻

4.2　训练集规模与组块识别结果的关系

为了观察训练集规模对组块识别的影响,我们逐渐增大训练集中句子的个数(样本的个数),并和基于增益的 HMM 算法进行了比较,关于增益的 HMM 算法参见文献[14]。



训练语料(句子数)

图 1　不同训练集规模下,多项式 SVM(d = 2)和增益的 HMM 的组块分析结果的比较

由实验结果可以得出以下结论:

1) 与增益的 HMM 相比,SVM 提高了约 7 个百分点,可见,SVM 在解决小样本统计模式识别问题中表现出特有的优势。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

2) 随着训练语料规模的不断加大,增益的 HMM 仍比 SVM 性能略差,但不会像在小规模的语料测试中表现得这样明显,这在 CoNLL - 2000 的英语组块识别共享任务得到了验证,增益的 HMM 只比 SVM 低了约 1.5 个百分点。具体的数据对比参见文献[17]。

我们同时注意到这样一个有趣的现象,训练集为 1000 个句子的分类效果比 1200 个句子更好,也就是说,不是训练集规模越大,其分类效果就越好,有选择的训练集要比随机选取的训练集更能得到好的性能。这是主动学习 (active learning)[13]的问题,即通过人工标注尽量少的数据,使得分类器具有相当的性能。这也是我们下一步研究的方向。

4.3 错误分析

通过对实验结果的分析,我们发现组块识别的错误主要有以下几个方面:

1) 训练语料的组块标注错误,例如"你们/ r 这些/ r 年轻人/ nc"应该标注为"你们/ r BNP[这些/ r 年轻人/ nc]",而不是"BNP[你们/ r 这些/ r 年轻人/ nc]"。

2) 某些组块类型如 BAP、BDP、BNS 和 BNT 的召回率很低,是由于特征选取加入了词汇信息,虽然词汇信息能够保证这些组块类型的准确率,但有时词性信息对于一些类型的组块识别已经足够,例如:BNS[西岳/ nd 华山/ nd],词性 nd 表示该词为地名,如果再加入词汇信息的话,反而碰到类似的情况"苏/ nd 杭/ nd"不能标注为 BNS。于是,我们可以考虑针对不同类型的组块,采用不同长度的特征表示。

3) 训练语料规模较小,很多组块实例只出现一次甚至没有出现,这也是 BNS 和 BNT 召回率低的一个主要原因。

5 结论

SVM 和其他的分类方法相比具有较高的分类精度。本文将中文组块识别问题看成是一个分类问题,并提出了基于 SVM 的组块识别算法,实验结果表明,SVM 算法取得了较好的分类效果,尤其是在小样本的情况下,表现得更为明显。进一步的工作是改进 SVM 算法,比如直推式 SVM 算法,SVM 的多类识别算法等,从而进一步提高它的分类能力。

参 考 文 献:

[1] Ramshaw L, Marcus M. Text Chunking Using Transformation-Based Learning[A]. Proceedings of third Workshop on Very Large Corpora[C]. Massachusetts: Association for Computational Linguistics, 1995. 82 - 94.

[2] Daelemans W, Buchholz S, Veenstra J. Memory-Based Shallow Parsing[A]. Proceedings of CoNLL[C], Bergen: Association for Computational Linguistics, 1999. 53 - 60.

[3] Pla Ferran, Molina Antonio, Prieto Natividad. Improving chunking by means of lexical-contextual information in statistical language models[A]. Proceedings of CoNLL - 2000 and LLL - 2000[C], Lisbon: Association for Computational Linguistics, 2000. 148 - 150.

[4] Koeling Rob. Chunking with maximum entropy models[A]. Proceedings of CoNLL - 2000 and LLL - 2000[C], Lisbon: Association for Computational Linguistics, 2000. 139 - 141.

[5] Taku Kudo and Yuji Matsumoto. Chunking with Support Vector Machines [A]. In: Proceedings of NAACL 2001[C], Pittsburgh, USA, 2001. Morgan Kaufman Publishers.

[6] Taku Kudo and Yuji Matsumoto. Use of Support Vector Learning for Chunk Identification [A]. In: Proceedings of CoNLL - 2000 and LLL - 2000[C], Lisbon, Portugal, September 2000.

[7] Tetsji Nakagawa, Taku Kudoh, and Yuji Matsumoto, Unknown word guessing and part-of-speech tagging using support vector machines [A], In: Proceedings of the Sixth Natural Language Processing Pa-

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

cific Rim Symposium[C], 2001, 325 - 331.
[8] Hiroyasu Yamada, Taku Kudoh, and Yuji Matsumoto, Japanese named entity extraction using support vector machines (in Japanese) [A], In: IPSJ SIG Notes NL - 142 - 17[C], 2001.
[9] T. Joachims. Text categorization with support vector machines: learning with many relevant features [A]. In: European Conference on Machine Learing, ECML98[C], pages 137 - 142, 1998.
[10] Steven Abney. Parsing by chunk [J]. Berwick, A. and Tenny, editors, Principle-Based Parsing. Kluwer. 1991.
[11] Ratnaparkhi A. Maximum Entropy Models for Natural Language Ambiguity Resolution[D]. Pennsylvania: University of Pennsylvania, 1998. 55 - 61.
[12] Kreβel. U. Pairwise Classification and Support Vector Machines [J]. B. Schölkopf, C. J. C. Burges, and A. J. Smola (Eds.), Advances in Kernel Methods —Support Vector Learning, Cambridge, MA, 255 - 268. MIT Press. 1999.
[13] Tong S. & Koller D. Support vector machine active learning with applications to text classification [A]. Seventeenth International Conference on Machine Learning[C]. 2000.
[14] Joachims, T. Making large scale svm learning practical [J]. Schölkopf, B., Burges, C., and Smola, A., editors, Advances in Kernel Methods - Support Vector Learning. MIT Press. 1999.
[15] Vapnik, V. Statistical Learning Theory [M]. Wiley. 1998.
[16] Vapnik V. The Nature of Statistical Learning Theory [M]. New York: Springer-Verlag, 1995.
[17] 李珩,等. 基于增益的隐马尔科夫模型的文本组块分析[J],计算机科学,已录用.
[18] 张昱琪,周强. 汉语基本短语的自动识别[J],中文信息学报,2002,16(6):1 - 8.
[19] 周强,孙茂松,黄昌宁. 汉语句子的组块分析体系[J],计算机学报, 1999,22(11):1158 - 1165.
[20] 赵军,黄昌宁. 基于转换的汉语基本名词短语识别模型[J],中文信息学报,1998,13(2):1 - 7.
[21] 姚天顺,等. 自然语言理解 ——一种让机器懂得人类语言的研究(第二版)[M],北京:清华大学出版社,2002,10.

[会议消息]

少数民族文字信息处理基础软件研发学术研讨会

随着国家西部开发战略的实施,民族语言文字信息处理技术研发与产业化工作越来越受到重视,从事民文信息处理研发的科研单位和企业在国家项目的支持下,对民文信息技术进行了更加深入地研究和开发,取得了一批成果,形成了一批适用于不同民族语言的软件产品。为了交流民族语言文字信息处理的最新研究成果,中国中文信息学会、中国科学院软件研究所及西北民族大学将在甘肃省兰州市联合举办"少数民族文字信息处理基础软件研发学术研讨会"。

会议将选取优秀论文推荐在《中文信息学报》发表。

时间(暂定):2004 年 9 月 19 日 - 23 日,会期 4 天。

征文内容(但不限于此):民族文字字符集编码与字型标准与实现;操作系统、办公套件、网络软件等基础软件多民族语言处理体系结构;复杂文本处理与 OpenType 字型;民文输入输出;多民族文字混合处理;辅助翻译与检索;民族文字文献数字化与电子出版等。

论文截止日期:2004 年 7 月 1 日

电子版论文发至:wujian@iscas.cn(请注明会议论文),论文必须有中文和英文的题目与摘要;论文提交形式要求电子版,使用 pdf 格式或 rtf 格式

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net