计算机模拟汉字字形认知过程的研究

陈 静,穆志纯,孙筱倩
(北京科技大学 信息工程学院,北京 100083)

摘 要:对汉字的认知研究不仅是认知科学、也是计算机科学特别是人工智能领域中的一个研究热点.但是,目前汉字认知的计算机模拟研究还相对滞后,其在认知科学研究中的作用还无法和行为实验研究等同.从认知科学的角度出发,建立汉字字形表征库,构建模型,确定训练和测试方式等,对汉字字形认知过程(学习发展历程)中汉字聚类与部件拆分意识进行了计算机模拟,以便研究汉字字形学习中的某些认知规律.通过对模型的训练与测试,得到了输入汉字的聚类效果图、部件拆分情况,以及对模型进行生字测试的结果.得出的结果能够反映某些汉字认知的规律,所以模型在一定程度上模拟了汉字字形的认知过程.

关键词:认知科学;人工智能;汉字认知;计算机模拟;自组织模型
中图分类号: TP391.1 **文献标识码**: A **文章编号**: 1673-4785(2008)03-0216-06

Computer simulation of the cognition of Chinese characters

CHEN Jing, MU Zhi-chun, SUN Xiao-qian
(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China)

Abstract: Research on the cognition of Chinese characters is a hotspot in both cognitive science and computer science, and is an especially lively field among those investigating artificial intelligence. In spite of this, research using computer simulations to analyze Chinese character cognition remains relatively backward, and its value in cognition studies has not been comparable with experimental research on behavior. In this paper, starting from the viewpoint of cognition science, a representative database of Chinese characters was set up, a cognitive model constructed, and training and testing modes determined. Computer simulations were made of the clustering and splitting of Chinese characters in the course of cognition, so that cognition rules for the perception of Chinese characters may be better understood. The model was based on a multi-layer self-organizing neural network. This training and testing method ensured that we knew how the Chinese characters were clustered and split during analysis so that the recognition of unknown words could be achieved. The research outcome suggests cognition rules for recognizing Chinese characters, implying that the proposed model does simulate the cognition process for Chinese characters.

Keywords: cognitive science; artificial intelligence; Chinese characters cognition; computer simulation; self-organized model

21世纪被认为是生命科学的世纪,生命科学的核心内容之一是对大脑的研究和探索.语言是反映人脑信息处理能力的高级功能,因此,阐明语言加工的信息处理机制对揭示人脑的奥秘具有重要意义.汉字是中国特有的表意文字,在形、音、义加工方面与西方拼音文字有很大不同,研究汉字认知就是应用认知科学的观点和方法,研究语言习得中汉字信息的输入、储存、内部加工和输出等过程.不少学者认为,利用汉字的一些特点进行相应的研究有可能澄清目前国际上关于言语加工机制中一些重要的争论,对认知科学的发展具有重要意义[1].目前,随着认知科学的发展,汉字的认知研究取得了一些新的进步,比如从认知神经科学的角度,采用脑成像技术对汉字认知脑机制的研究等[2].但是,将计算机科学与认知心理学相结合所进行的汉字认知研究并不多见.

收稿日期: 2007-10-26.
基金项目:北京市教委重点学科共建基金资助项目(XK100080537);
北京语言大学规划资助项目(04GH01).
通讯作者:陈 静. E-mail: heart931@163.com.

尽管有关汉字认知的研究已经取得了一些成果,但也有不少问题有待深入.这些问题不仅有汉字认知规律各方面的内容,也包括了研究的方法和研究的角度.目前,在从认知心理学角度出发的汉字认知研究中,大多采用归纳式或经验式的行为实验方法,虽然能得到某些认知规律,但这些方法对数据需求量大、实验时间长、重复性差、局限性明显、且缺乏对复杂认知规律的预测作用。随着计算机科学的发展,从认知心理学的角度出发,构建汉字认知过程的计算机模型,来研究汉字认知规律是汉字认知研究的一个新途径.但是,目前相关的计算机模拟研究还相对滞后,其在认知科学研究中的作用还无法与行为实验等同.因此,根据汉字的认知心理学特点,建立计算机模型对汉字认知的信息加工过程进行深入研究是必要的.

综上,汉字认知是现代认知科学的一个重要研究领域.本研究通过建立计算机认知模型的方法,研究汉字认知过程中的字形认知问题,期望能从新的角度揭示汉字认知过程中的信息加工机制和规律.

1　研究意义

1)认知心理学领域的传统研究方法多为归纳实验和经验实验,这些实验局限性明显,需要大量的人员配合,大量的实验和统计时间,而且实验的可重复性差.计算机模拟的研究方法不同,可以根据已掌握的认知心理学知识构建计算机模型,在较短的时间内,模拟出人类需要在较长时间内才能获得的知识与技能,以便于探索认知过程的规律.而通过得到的计算机模拟结果,还可以更加深入地对认知行为实验的结果进行分析并给出合理的解释.计算机模拟还可以对某些认知现象(如不同汉字类型的自组织现象)的发生进行预测.

2)汉字字形认知过程的计算机模拟研究,对反映字形的心理过程进行模拟,来表现汉字的认知特征.而结合认知心理学的研究成果,提取汉字识别特征,可以节省整字匹配处理的时间,也将有利于机器自动识别汉字技术的发展.

3)行为实验在汉字认知研究中目前仍占有很重要的地位,但是计算机模拟研究不仅在某种程度上可以等同于行为实验研究,而且可能摆脱以行为实验为主的研究方法的局限性.比如说可以随时改变训练方法,检测不同训练方法对学习结果的影响;可以损伤任何一个部分的表征或形、音、义之间任何的联结来模拟汉字认知过程中的阅读困难现象等,这些都是行为实验无法实现的.因此,进行计算机模拟研究的目的不仅仅在于验证行为实验的结果,更重要的是获得一些行为实验无法获得的结果.

从语言习得的应用角度来看,计算机模拟汉字认知过程的研究,能够发现汉字认知过程中的规律,对汉字、汉语的教学及促进中外文化的交流也有积极的意义.

2　研究内容

本课题旨在从认知心理学的角度出发,研究汉字字形认知过程的计算机模拟问题.建立模型要体现对汉字字形认知过程研究的特点,不像只是一般的简单映射,对过程、中间结果等问题也要研究,所以在建立模型时要考虑到这些情况.

根据汉字字形认知本身的特点,采用无监督学习的自组织特征映射网络(self-organizing feature map),建立了汉字聚类及部件拆分模型.不只研究汉字认知的结果,还要研究模拟汉字认知的信息加工过程,从而对汉字认知心理机制进行描述和刻画.

2.1　汉字聚类及部件拆分模型

2.1.1　汉字及部件的表征

如何对汉字字形进行表征,是汉字认知过程的计算机模拟研究的关键,因为它是给模型提供输入信息的方法和途径.表征方法不仅能够表征汉字的结构规律,而且能够体现学习者汉字认知过程的特征.但是目前这方面的研究很少,这可能是由于汉字字形的复杂性造成的.

本研究中,汉字字形表征采用文献[3]和文献[4]中的表征方案,其汉字与部件的表征架构如图 1所示.此表征方案的特点有:在结构上充分体现了在整字表征框架下的部件表征,表现了汉字结构的层次性;充分考虑到汉字认知的特点,如汉字字形的视觉特征以及部件构字位置的特征,有利于研究汉字字形认知的特点和过程;基于汉字字形信息统计的表征方法,是在客观分析汉字字形特征分布的基础上表征汉字字形的.由此提取和处理上述表征方案的表征库中的数据,并建立汉字与其对应部件表征中的对应关系,作为输入数据进行计算机模拟汉字字形认知过程的研究.

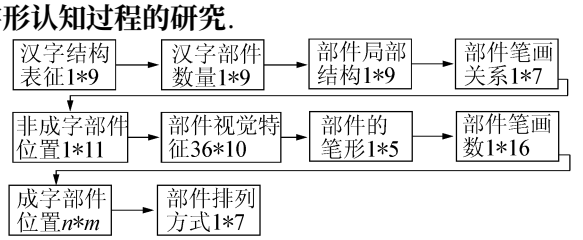


图 1　汉字与部件表征架构示意图

Fig 1 Representations of Chinese characters and components

图 1中,每一个方框代表在汉字字形中需要表示出来的特征,可以视为一个表征环节,以反映每个汉字的独有的特征. 每个表征环节下面都有一个数据量,要用到向量维数 n与该表征划分的等级 m,它充分表示了该环节的表征. 在图中的 $n*m$,如"1*9",表示"汉字结构表征"可以用 1个维度来表示,在表示时可以把它划分成 9个等级,如用 0.1111～0.9999分别来表示. 每一个表征环节的向量维数与等级的划分都有其汉字字形统计信息依据.

以"扒"字为例,汉字和部件具体的表征向量如下所示:

扒 0.286 (部件数) 1.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.300 0.000 0.000 0.000 0.0000.000 0.200 1.000 0.200 0.000 0.188 0.333 0.636 (第 1个部件)... 1.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 1.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.600 0.800 0.000 0.000 0.125 0.500 0.091 (第 7个部件).

扌 a1 0.700 (左右结构) 0.100 1.000 1.000 (部件位置) 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.300 0.000 (视觉特征) 0.200 1.000 0.200 0.000 (笔形特征) 0.188 (笔画数) 0.333 (笔画结构关系) 0.636 (位置特征).

八 a7 0.700 0.700 0.000 ... 0.125 0.500 0.091.

2.1.2 汉字聚类及部件拆分模型结构及算法

目前已经有人研究出一些语言模型,但是这些模型采用的算法中大多是有监督的学习算法,例如典型的反馈式学习方法. 而从语言材料和语言学习的关系来看,语言认知过程基本是一个类似于无监督的学习过程,而教师的指导只是一种促进. 所以,采用无监督的自组织特征映射网络学习算法可能将作为实现这一模型的方法之一.

研究发现[5],大脑是由大量协同作用的神经元群体组成的,大脑的神经网络是一个十分复杂的反馈系统;这个系统含有各种反馈作用,有整体反馈、局部反馈;在大脑处理信息的过程中,聚类(Clustering)是极其重要的功能,大脑通过聚类过程从而识别外界信号,并产生自组织现象. 由此可以总结出,要选用的计算机模型须具备以下几个特征:具有与大脑类似的拓扑结构;具有内反馈的功能;具有无监督的学习功能;具有对知识自组织的过程;能够实现聚类.

神经网络中的自组织特征映射模型,是 Kohonen依据大脑对信号进行处理的特点提出的一种神经网络模型[6]. 自组织特征映射模型是由输入层(模拟视网膜神经元)和竞争层(模拟大脑皮层神经元,也叫输出层)构成的网络(如图 2所示):它的输出层以二维阵列的形式输出获胜神经元,这种结构能够较好地模拟大脑皮层神经元的拓扑结构;2层之间的各神经元实现双向全连接,且输出层的获胜神经元能影响其邻域内神经元的连接权值,网络中没有隐含层,在模拟过程中不断调整 2层神经元全连接的权值和输出层获胜神经元邻域内的神经元权值,来模拟大脑认知过程中的反馈作用;自组织特征映射模型可以反映自组织特征,自组织的过程实际上就是一种无指导的学习过程;自组织特征映射网络可实现从一组表征数据中提取有意义的特征或者一些内在的规律性,它通过自身训练,自动对输入模式进行分类实现聚类功能.

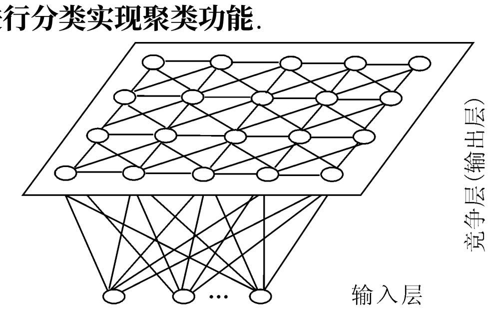


图 2 SOM网络结构示意图

Fig 2 SOM network structure

由此可见,自组织特征映射网络接近生物神经系统,其工作过程比较符合人类大脑认知的过程的特点.

由于模拟汉字认知过程中汉字聚类和部件拆分情况,涉及到整字与部件 2个方面以及它们之间的关系,因此要研究汉字和部件的认知过程,描述汉字与部件之间的关系,就需要对传统的 SOFM网络进行改进,使模型能够分别实现汉字的聚类和部件的拆分,建立了汉字聚类与部件拆分模型(双层双向网络). 要求训练好的网络既能够描述由汉字到部件的学习过程,即部件的拆分情况;又能够描述由部

件到汉字的学习过程,即部件的构字情况.模型结构如图 3所示,将模型分为输入层、汉字聚类层和部件拆分层.其中汉字聚类层与部件拆分层之间互相都有连接关系,从而体现出汉字的部件构成情况以及部件的构字情况.由图 3可知,在汉字聚类与部件拆分模型的训练阶段,将所选汉字样本的表征向量输入汉字层,同时将对应部件的表征向量输入部件层.这是模型的输入,模拟汉字认知过程中的视觉刺激.模型通过自组织形成获胜神经元,并不断调整 2层间的权连接,建立汉字与部件之间的关系,模拟汉字认知过程中汉字聚类与部件拆分意识的形成过程.

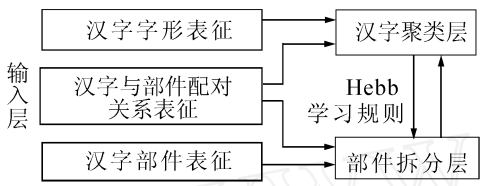


图 3　模型结构图

Fig 3　Architecture of the model

完成自组织特征映射的算法较多,常用的自组织算法可描述为[7]

1)权值初始化并选定领域的大小;

2)输入模式;

3)计算空间距离:

$$d_i = \sum_{i=0}^{N-1} [x_i(t) - w_{ij}(t)]^2. \quad (1)$$

式中:$x_i(t)$是 t时刻 i节点的输入,$w_{ij}(t)$是输入节点 i与输出节点 j的连接强度,N为输入节点的数目;

4)选择节点 j^*,满足$\min_j d_j$;

5)按式(2)改变 j^*和其他领域节点的连接强度:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)[x_i(t) - w_{ij}(t)]. \quad (2)$$

式中:$j \in j^*$的领域,$0 \leq i \leq N-1$,$\alpha(t)$为衰减因子;

6)返回到 2),直至满足 $[x_i(t) - w_{ij}(t)]^2 < \varepsilon$($\varepsilon$为给定的误差)或学习次数大于预定值.

通过这种无监督的学习,稳定后的网络输出就对输入模式生成自然的特征映射,从而达到自动聚类的目的.

此外,本文借鉴了李平在文献[8]中提出的方法汉字聚类层和汉字部件拆分层两层之间的连接关系是双向的,并且同时接受输入层输入的表征信息(如图 1所示). 2层之间节点的权值更新使用 Hebb学习规则:

$$\Delta w_{lp} = \alpha(t) a_l^S a_p^D. \quad (3)$$

式中:w_{lp}是从源网络节点 l到目标网络节点 p的连接权值;a_l^S和 a_p^D分别是 2层连接中从源网络到目标网络的输出值.

2层之间的连接关系如图 4所示,在各层网络中,每一个单元都与该层网络的输入向量相连,由权值向量来表达它们之间的连结强度;而在 2层网络之间,各层的每个单元又与另一层的每个单元相连,由连接权值表达连结强度.这里只画出了 2层之间其中一个单元与另一层的连结关系.通过对配对输入的学习,只有少数连结被认为是有效连结,如果最大激活部件单元被找到了,那么这个汉字相应的部件单元就被确定下来了.反之道理是一样的.

通过训练和不同的测试项,并对计算机模拟结果加以分析,得到了与认知科学中行为实验研究相似的结果,而且还可以对行为实验结果做出合理的解释.

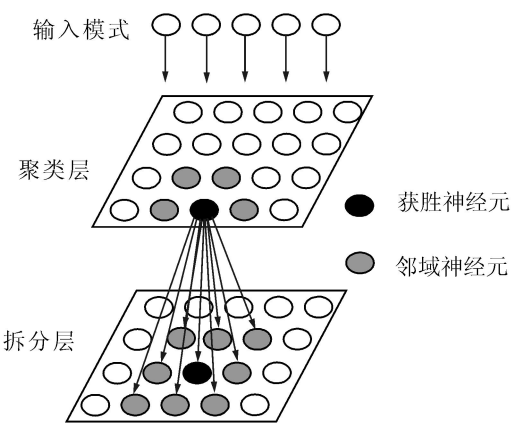


图 4　聚类层与拆分层连接关系示意图

Fig 4　The relation of clustering layer and splitting layer

3　模型的训练

以图 5和图 6所示部分的汉字聚类结果为例.

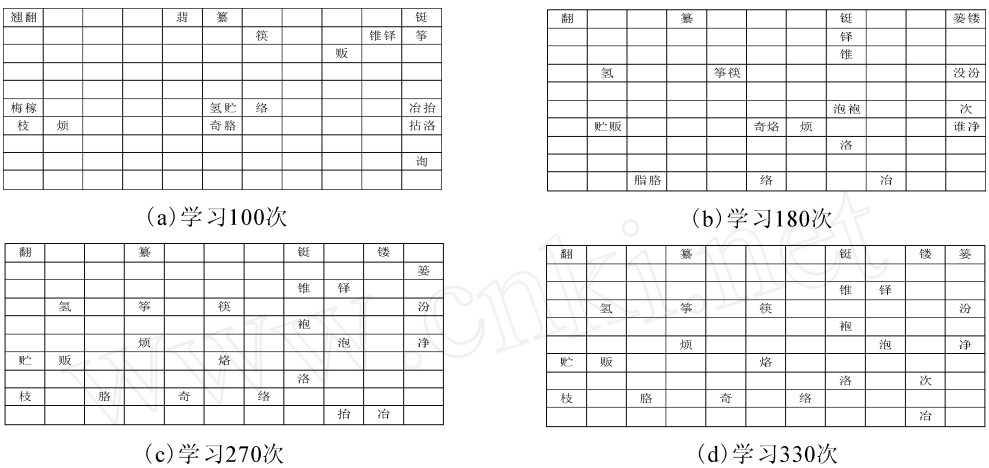
翻			篡				铤		镂	篓	
							锥	铎			
	氢		筝		筷					汾	
							袍				
			烦					泡		净	
贮	贩				烙						
							洛		次		谁
枝		胳		奇		络					
								抬	冷		淹
脂			晚		拈						
					呵				询		绳
类		晦					啦				
					嘈						吵
秋		暗						咽			

图 5　汉字的聚类结果(部分)

Fig 5　Chinese characters clustering

图 5是训练完成后的汉字聚类情况,从图中的阴影部分可以看出,网络对形似的汉字能够进行聚类,如“铤”、“锥”和“铎”,以及“烙”、“络”和“洛”等

都放在了相邻近的位置上. 图 6中显示了训练学习100次、180次、270次、330次的汉字的聚类结果. 通过汉字聚类结果的实验,可以得出以下结论:汉字聚类与部件拆分模型训练过程可以体现自组织聚类现象,具有相似特征的汉字和部件分布比较接近,这种聚类结果是网络算法拓扑性质的体现,同时表现出在汉字字形认知过程中自组织分类学习的认知心理学特征;而且从图 5和图 6中可以发现随着学习次数的增加,模型的汉字聚类效果也增强,并逐渐趋于稳定,这也大致符合汉字认知过程中随着学习量的增加掌握的知识量和知识结构趋于稳定的现象.



(a)学习100次 (b)学习180次

(c)学习270次 (d)学习330次

图 6 汉字的聚类动态过程 (部分)

Fig 6 The dynamic process of Chinese characters clustering

此外,研究工作还包括部件拆分 (汉字的部件组成)的研究,也得到了一些结果. 汉字聚类层与部件拆分层之间的关系,表示某个汉字由哪几个部件组成,如图 7所示. 在批注中显示出这个单元格对应激活的另外一层中的单元格的结果. 图中显示出“呵”字的部件“口 a7”、“丁 b6”、“口 a1”,部件后面的字母及数字组合表示的是该部件在不同汉字中的不同位置信息. 部件拆分层与汉字聚类层也有相似的连接关系,表示的是某个部件对应其激活了的那些汉字,如图 8所示. 由所得到的结果,可以总结汉字拆分成部件的学习效果,并可以归纳出 2种情况:

1) 网络的汉字学习结果是完全正确的. 如"呵"字的激活部件是“口 a7”、“丁 b6”、“口 a1”,部件层的激活神经元所代表的部件与汉字的实际部件拆分结果相吻合.

2) 网络的汉字学习结果是不完全正确的. 而这种情况下,又分为部件混淆和部件冗余 2种出错类型. 部件混淆是指部件层激活的部件序列中,部件数目与实际相符,但是存在一个或几个部件与相应的实际部件不符. 例如"唤"字的激活部件中的“厂b1”和“冂a5”,在实际序列中找不到完全吻合的部件,但是可以找到与之形似的部件“厂a5”和“冂a6”. 部件冗余则是指激活的部件多于实际部件数目,如“柴”字,在学习结果中,除了正确激活的“匕 b2”

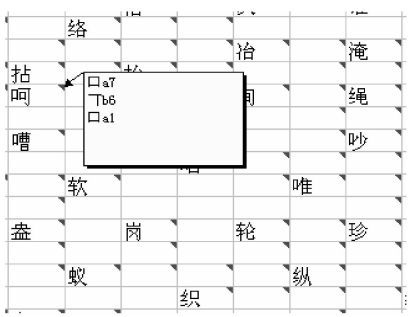


图 7 汉字聚类层与部件拆分层连接关系图 (部分)

Fig 7 The connection relation of Chinese characters clustering layer and components splitting layer

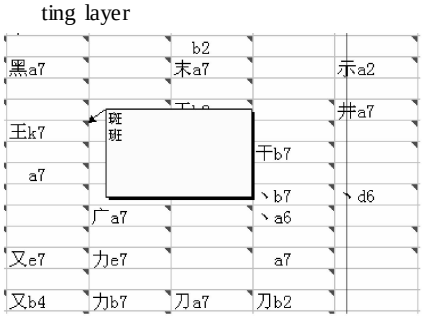


图 8 部件拆分层与汉字聚类层连接关系图 (部分)

Fig 8 The relation connection of components splitting layer and Chinese characters clustering layer

外,还多出了“匕 /5 ”这个部件.但是多出来的部件并不是同相应的汉字完全无关,它们总是与实际的部件有很大的相似度.

模型通过训练得到的这些结论与心理学研究中行为实验研究的结果相似,从不同的方面反映了汉字字形认知过程中的某些规律,体现了汉字字形认知过程中的部分特点.

4　模型的测试

对训练好的模型进行测试,以输入生字来考察模型测试效果为例,取生字 50个字,选用包括了左右、上下、包围 3个结构的字.测试结果如图 9所示.图中阴影中的字为测试字,可以看出网络能够根据之前学到的知识对输入的生字进行推测识别,测试字根据与已学过汉字有相似的结构或部件信息放在了这些汉字的附近,可以看出测试结果图中汉字的聚类效果仍然存在,例如将“怅 ”放在了“张 ”的附近,“训 ”放在了“计 ”的附近.

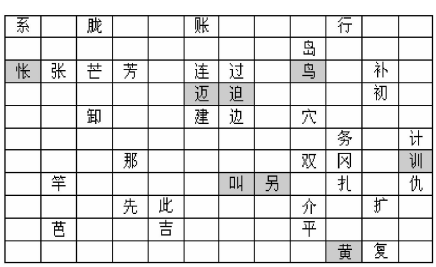


图 9　生字测试结果图 (部分)

Fig 9　Testing result of new Chinese character

5　结束语

从模型模拟过程中可以看出,对模型进行训练,网络通过对汉字及其部件信息的学习,对汉字的构形方式、结构规则等都有了一定的认识,能够发现其中的规律,在对学习的汉字进行其部件的拆分的同时,还能将有相似结构或部件的汉字聚类,在一定程度上模拟了汉字字形的认知过程.

参考文献

[1]周志华,曹存根. 神经网络及其应用 [M]. 北京:清华大学出版社, 2004: 366.

[2]唐一源, 张武田, 马　林, 翁旭初, 李德军, 何　华, 贾富仓. 默读汉字词的脑功能偏侧化成像研究 [J]. 心理学报, 2002, 34 (4): 333-337.
TANG Yiyuan, ZHANG Wutian, MA Lin, WENG Xuchu, LI Dejun, HE Hua, JIA Fucang. The laterality of brain function in silent reading of Chinese words revealed by fMRI [J]. Acta Psychological Sinica, 2002, 34(4): 333-337.

[3]王建勤. 外国学生汉字构形意识发展的模拟研究 ——基于自组织特征映射网络的汉字习得模型 [D]. 北京:北京语言大学, 2005.
WANG Jianqin. Simulating studies of CFL learners' Chinese orthographic awareness development based on self-organizing feature map network[D]. Beijing Language and Culture University, 2005.

[4]邢红兵. 小学语言教材形声字表音情况统计分析及小学生形声字命名的自组织模型 [D]. 北京:北京师范大学, 2002.
XING Hongbing. Analysis of phonetics of semantic-phonetic compound characters in elementary school textbooks and a self-organizing connectionist model of character acquisition in Chinese[D]. Beijing Language and Culture University, 2002.

[5]舒　华,韩在住,许忠宝. 认知神经心理学的基本假设和研究方法 [J]. 心理科学, 2002 (6): 721-722.
SHU Hua, HAN Zaizhu, XU Zhongbao. The basal hypotheses and research methods in cognitive neuropsychology[J]. Psychological Science, 2002 (6): 721-722.

[6]边肇祺,张学工. 模式识别 [M]. 北京:清华大学出版社, 2000: 259.

[7]KANAGA S J, KOHONEN T. Developments and applications of the self-organizing map and related algorithms[J]. Mathematics and Computers in Simulation, 1996, 41: 3-12.

[8]LI P, FARKAS I, MACWHINNEY B. Early lexical development in a self-organizing neural network[J]. Neural Networks, 2004 (17): 1345-1362.

作者简介:

陈　静,女, 1979年生,博士研究生,主要研究方向为人工智能、模式识别.

穆志纯,男, 1952年生,教授,博士生导师,主要研究方向为人工智能及其应用、模式识别、图像处理、生物特征识别、复杂系统的建模与控制. 1989～1991年和 1997～1999年间在英国进行访问研究.曾主持、参加国家自然科学基金项目 4项、青年“863”项目 1项、国家科技攻关和国际合作项目多项,并获部级科技进步二等奖 1项、三等奖 2项.已发表论文 90余篇,其中被 SCI和 EI检索 40余篇.

孙筱倩,女, 1983年生,硕士研究生,主要研究方向为人工智能、模式识别.