第30卷 第20期 计 算 机 工 程 2004年10月
Vol.30 №20 Computer Engineering October 2004
·人工智能及识别技术· 文章编号：1000—3428(2004)20—0133—03 文献标识码：A 中图分类号：TP391

基于神经元网络的汉语组块自动划分

王荣波，池哲儒
(香港理工大学电子及资讯工程系，香港)

摘 要：介绍一种基于三层神经元网络的汉语组块自动划分方法。输入信息为句子中每一个字本身及与前后字组合的划分情况，输出为句子中每个字的划分结果。对于一个新输入的汉语句子，在该方法中，并不对句子进行切词，这是与别的组块分析方法的不同之处。实验表明，该方法是可行的，也是有效的。

关键词：组块分析；神经元网络；中文信息处理

Automatic Segmentation of Chinese Chunks Using a Neural Network

WANG Rongbo, CHI Zheru
(Department of Electronic and Information Engineering, HongKong Polytechnic University, Hongkong)

【Abstract】 This paper presents a method for automatic segmentation of Chinese chunks based on 3-layer neural networks. The corpus has been processed with Chinese word segmentation and phrase identification and tagging. In the neural networks model, the input data is the segmentation situation of every character and its combinations with neighbor characters in a Chinese sentence. The output is the segmentation results of every character in a Chinese sentence. The preliminary results show that the method is feasible and effective.

【Key words】 Chunk analysis；Neural networks；Chinese information processing

1 概述

自从计算机产生那天起，人们一直希望计算机能实现自然语言文本的自动翻译，从那时起，有了机器翻译(Machine Translation)研究。到目前为止，计算机在硬件和软件上的发展，应该是快速和令人满意的，而且这个发展势头还会保持下去。但是，对于机器翻译研究，可以说是充满着曲折和困难。目前还没有一个非常实用的，令人满意的翻译系统，研究者还在机器翻译研究道路上苦苦摸索着。这主要有两个原因。一方面，因为自然语言本身的复杂性和多样性。另一方面，人们还没有完全清楚认识人脑的运作机制，不清楚人脑是如何学习语言和进行不同语言之间的翻译。要找到这两方面原因的答案，还需要很长的时间，需要很大的努力。

基于实例机器翻译(Example-Based Machine Translation, 简称EBMT)方法最初提出于20世纪80年代末期。该方法的主要思想是：首先，需要一个进行过一定加工处理的双语句子实例库。当输入一个新的源语言句子时，从双语库中找到与输入句子最相似的源语言句子。再以对应于该句子的目标语言句子为模板，进行一些必要的组合和替换，最后得到对应于输入句子的翻译结果。

应该说，汉语组块分析思想(又称为浅层句法分析，shallow parsing)跟基于实例的机器翻译方法联系最紧密。因为，基于实例的机器翻译方法中，如果不对句子进行分析，会有匹配命中率低，所需实例库庞大等问题。如果进行完全分析，又违背了EBMT思想的初衷。基于规则的机器翻译(Rule-Based Machine Translation, 简称RBMT)方法需要对源语言句子进行比较完全的句法分析，但这个分析过程过于复杂详尽，而且语言知识规则的提取太难，不适合应用于机器翻译当中。且目前缺乏高效准确的自动句法分析器。为了解决或者避免完全句法分析带来的问题，在寻求机器翻译新方法的过程中，出现了一种新思路，也就是基于组块的处理方法。基于组块的EBMT方法对源语言句子不进行完全的句法分析，这是一种折中的方法。目前它已成为该领域的其中一个研究热点[1-4,8,12]。其主要思想是把原先作为基本处理单元的多义性词语扩大到单义性的组块,并给以中心词的标注。

众所周知，机器翻译的最大问题是歧义现象。如果加大信息处理的粒度，可以大大减少机器翻译过程中的歧义现象，同时又可以简化源语言句子的句型，有利于语料库加工过程中短语级的双语对应。可以肯定，这是个很好的设想，越来越多的研究者正被它所吸引，投入到对它的研究中。

例如：英文字“make”，查WordNet，有50种不同的意思。其中作为名词“make”有2个意思，作为动词“make”有48个意思。取语块“make a call”,“make a decision”，“make a phone call”,那它们都只有一个意思,十分明确。

又如：汉语语块“进一步考虑之后”，比较好的英语翻译是“On second thoughts”。如果按词翻译，那就不可能有正确的翻译结果。

所以，基于组块(Chunk-Based)分析方法的提出，有利于解决机器翻译中的歧义现象。而汉语组块的划分是基于组块方法的基本环节。

本文利用神经元网络来实现汉语句子中组块的自动划分。之所以用神经元网络，有两个原因：(1) 神经元网络具有比较强的非线性分类能力，同时又有比较强的抗干扰能力；(2) 与概率统计方法相比，神经元网络的方法在上下文取字个数方面可以突破二元语法的限制，同时又不会引起空间的过度膨胀。实验结果表明，该方法是可行的，并且是有效的[5]。

2 组块的定义

语料库在机器翻译过程中扮演着重要的角色[9]。在本实验中，采用了清华大学周强博士提供的部分汉语单语语料。该部分语料库总共包含3 556个句子，已经完成词和短语的

作者简介：王荣波(1978－)，男，硕士生，主研方向：计算语言学，机器翻译，模式识别；池哲儒，博士生、副教授、博导

收稿日期：2003-09-05 **E-mail：**wangrbo@163.com

划分及标注。在实验中，我们将用它来进行训练和学习。

下面给出语料库中的几个例句。

19 [dj [np 本/r 书/n] [vp 是/v [np [vp [pp 为/p [np 语言/n 的/u 初学者/n]] 编写/v] 的/u [np [mp 一/m 本/q] 教材/n]]]]

30 [dj [np 病人/n 的/u 妻子/n] [ap 伤心/a 得/u 很/d]]

41 [zj [dj [vp 采用/v [np [mp 这/r 种/q] 方法/n]] [ap [dp 不如/v [vp 采用/v [np [mp 那/r 种/q] 方法/n]]] [ap 省时/a ，/w 省力/a]]] 。/w]

74 [zj [pp [pp 从/p 北京/n] [pp 到/p 那里/r]] ,/w [vp 大约/d [vp 有/v [mp 二/m 百/m 多/m 公里/q]]] 。/w]

说明：每个句子前面的数字表示序号。符号“ [”和“] ”分别表示左边界和右边界。句子中的标记包括词类标记和短语类标记。限于篇幅，具体的词类和短语类标记定义，本文中不做说明，可以参考文献[6]。

基于以上的语料库，构造了一个词库和短语库。词库由语料库中所有的词语组成。短语库由语料库中所有的短语组成。其中，有些短语之间存在包含关系。例如，在以上第19句子中，同时把名词短语(np)“ 语言的初学者 ”和介词短语(pp)“ 为语言的初学者 ”收录到短语库中。

如同在汉语句法分析中存在不同的方法(主要有句子成分分析法，也叫中心词分析法和层次分析法)，文献[7]和对某些语言现象存在争议，我们认为汉语组块的定义既要吸收语言学研究成果，也要充分考虑组块分析的应用目标。在文献[10,11]中，作者从汉语句法分析角度研究了汉语短语的自动划分和标注问题，但考虑到具体的翻译实现，认为不能把现有的短语定义当作是组块定义。在本文中，为了探索神经元网络方法是否能用来汉语组块的自动划分，并结合EBMT的需要，对组块的定义作了一定的简化。

组块: 介于任何“ [”和“] ”之间的汉字片断，除去单句(dj)或者整句(zj)的情况。组块的组成可能是一个词语，也可能是一个短语。例如，对于上面的例句30，认为它是有两个组块组成的，分别是“ 病人的妻子 ”和“ 伤心得很 ”，而我们认为“ 病人的妻子伤心得很 ”不是一个组块，虽然它也是介于符号“ [”和“] ”之间，它已经形成了单句。对于例句74,它包含有5个组块，分别是“ 从北京 ”，“ 到那里 ”，“ 大约 ”，“ 有 ”和“ 二百多公里 ”。可以看出，这些组块的划分对汉语句子的组块分析是很有用的，也有利于双语语料库的对齐加工处理。

很多研究人员对汉语短语的边界识别进行了研究[10,11]，但边界识别的前提是对汉语句子进行切词处理，再利用词的信息和词之间的位置关系，进行短语的边界识别。但这样会有两个问题：第一，切词的错误会被传递并且扩大到短语边界识别；第二，短语本身的定义不适合于翻译这一目标。到目前，前人还没有对本文提出的方法进行过研究，我们也只是做了初步的探索。

3 神经元网络拓扑结构设计

选用BP网络作为模型。BP网络是一种有指导学习的前向网络，它能学习复杂的非线性映射，在工程领域应用非常广泛。一般的BP网络有3层，分别是：输入层、输出层和隐含层(又称中间层)。而每层又都有模拟神经元的网络节点组成，其中各个节点都有一个激活值，该激活值通过权值传播给下层节点。每个节点还有一个阈值，用于衡量和调整每个神经元的兴奋水平。一般来说，各层之间存在每个神经元的完全连接，而在层内部不存在连接。在网络学习过程中，输出层的输出数据与期望数据进行比较，根据比较结果进行逐层调整连接权值，进行多次循环，直到最后的输出结果满足误差要求。隐含层可有多层结构，其中每层的节点数可通过实验确定,理论上并没有确定隐含层节点数的一般性方法[5]。

3.1 输出层设计

输出层设计比较简单。在我们的模型中，最终目标是识别出每个字的左右边界是否存在划分，即句子中的一个字到底是处于组块的左边界、右边界或是中间位置，或者既是左边界又是右边界。因此，将输出层设计为3个节点，因为用3个节点可以表示出上述4种可能的情况。输出层每个节点对应一个激活值，该值的大小反映了当前字取相应边界划分的可能性。同时每个节点对应一个阈值，这些阈值可以相同。在确定划分边界时，就根据每个节点对应的激活值和阈值，确定该节点对应的划分是否存在。如果激活值大于阈值，则存在划分，否则不存在划分。下面用表格来说明，表1中列出了所有的4种情况，其中符号“ > ”表示激活值大于阈值，符号“ < ”表示激活值小于阈值。

表1 输出结果与划分情况对应关系

第一节点	第二节点	第三节点	划分情况
>	<	<	左边有划分
<	<	>	右边有划分
>	<	>	左右有划分
<	>	<	左右无划分

3.2 输入层设计

假设输入句子为$S = c_1c_2\cdots c_n$。其中c_i表示句子中的每个汉字字符，包括标点符号。

在模型中，假设词类的标记集$X=\{x_1,x_2, \cdots,x_{|X|}\}$。其中|X|=31，因为语料库中共有31词类。词类标记定义如表2。

表2 词类定义标记集合

基本词类						扩展词类	
名 词	n	数 词	m	标点符号	w	专有名词	ng
时间词	t	代 词	r	后接成分	k	指人的专有名词	ngp
处所词	s	语 气 词	y	语 素 字	g	名词性语素字	Ng
方位词	f	象 声 词	o	非语素字	x	形容词性语素字	Ag
量 词	q	叹 词	e	介 词	p	动词性语素字	Vg
区别词	b	助 词	u	副 词	d		
形容词	a	简称略语	j	连 词	c		
状态词	z	习 用 语	l	成 语	i		
动 词	v	前接成分	h				

在对当前字进行组块边界识别时，考虑其左边L个字，右边R个字的影响，共L+R+1个字，用C_L, $C_{-(L-1)}$, $\cdots$, C_{-1}, C_0, C_1, $\cdots$,C_R来表示，可以认为这是个窗口，在分析一个句子时，通过移动该窗口来确定句子中每个字的划分情况。

在输入层，共设置(L+1)*(R+1)*13个节点。下面解释该表达式。因为考虑的窗口大小是(L+R+1)，所以，总共有(L+1)*(R+1)种包含当前字的字串组合。如表3所示。

表3 窗口中包含当前字的组合集合

序号	组合情况	左边字数	右边字数
1	C_0	0	0
2	$C_{-1}C_0$	1	0
3	$C_{-2}C_{-1}C_0$	2	0
---		---	---
L+1	C_{-L}---$C_{-2}C_{-1}C_0$	L	0
L+2	C_0C_1	0	1
L+3	C0C1C2	0	2
---	---		
L+R+1	$C_0C_1C_2$...C_R	0	R
L+R+2	$C_{-1}C_0C_1$	1	1
L+R+3	$C_{-2}C_{-1}C_0C_1$	2	1
---	---	---	---
2L+R+1	C_{-L}---$C_{-2}C_{-1}C_0C_1$	L	1
2L+R+2	$C_{-1}C_0C_1C_2$	1	2
---	---	---	---
3L+R+1	C_{-L}---$C_{-2}C_{-1}C_0C_1C_2$	L	2
---	---	---	---
(L+1)*(R+1)-L+1	$C_{-1}C_0C_1C_2$...C_R	1	R
---	---	---	---
(L+1)*(R+1)	C_{-L}---$C_{-2}C_{-1}C_0C_1C_2$...C_R	L	R

接着说明这里13的含义。在该模型中，用10个节点的二进制数来表示31个词类。也就是说，随机地产生31个非零的10位二进制数，每个二进制数对应一个词类。在此，为了尽量避免神经网络训练过程中输入之间的偏差，采用了两个措施：第一，用10位二进制数来表示31个词类，而不是用5位二进制数，虽然5位二进制数可以表示32个类型。第二，随机产生31个不同的二进制数来表示31个词类，而不是人为地确定。对应每一个10位的二进制表示，用另外3个节点来表示对应该词类的字串的划分情况，即左划分，右划分，左右都有划分，左右都没有划分。所以，对应于每个字串，用13个节点来表示它的输入信息。如果字串属于某个词类，那么对应于它的13个节点中，前面10位为对应于该词类的一个二进制数，而后面的3位是对应于该字串的划分概率信息，也就是对应于每一个划分的概率数据。如果字串不属于任何词类，那么所有的13个节点都赋值为0。在这里，需要事先得到词库，在该词库中，包含有对应于语料库所有的词和该词的划分统计信息。

3.3 隐含层设计及实验结果

表4 实验结果

句子数	隐含层节点数	训练次数	正确率(%)
5	10	500	86.76
20	30	5000	92.49
20	20	5000	91.70
20	20	10000	90.51
20	30	10000	90.12
20	10	1000	85.77
508	30	1000	70.03
508	50	1000	74.48
508	80	1000	74.46

在隐含层，采用只有一层的结构。在理论上，还没有确定隐含层节点数的一般性方法,只能根据实验结果来确定[5]。

在实验中，取窗口大小为5，即取当前字左边右边都是2个字的情况。不同情况的实验结果如表4所示。

4 实验结果分析

在实验中，设置窗口大小为5个汉字的长度，这是因为考虑到一般的汉字组块长度不超过5个字左右。从以上的实验结果可知，在对5个句子进行测试时，由于存在数据稀疏问题，识别率也不是很高，能达到86.76%。在进行20个句子进行测试时，识别率比较好，能达到92.49%。在对508个句子进行测试时，识别率有所下降，但也可以达到75%左右。这说明，我们的神经元网络模型是可以用来进行汉语组块的自动划分的。

识别率计算方法：识别率 ＝ 正确划分的汉字字符数/总的汉字字符数(%)

5 进一步的工作

在神经元网络模型中，一个很重要的问题是输入信息的提取，使输入参数能更好地反应识别特征。神经元网络模型一个明显的优势是可以很方便地改变上下文取字个数，可以通过实验来确定最后的取字个数，也就是窗口的大小。对已有实验结果进行分析可知，输入信息的提取还是不够的。因此，在接下去的实验中，我们要提取更多的特征作为输入信息，其中包括利用短语库(现阶段只用到了词库)信息，对一些特殊的词或字进行特殊的处理，包括“了”，“的”，“着”等，因为这些字的划分情况比较固定。还有，我们需要进一步分析现有的错误情况，总结出错误规律和调整方法。也需要探索新的ANN模型，例如回馈的引入等等。

本文的语料库由周强博士提供。特此感谢！

参考文献

1 Abney, Steven. Chunks and Dependencies: Bring Processing Evidence to Bear Syntax. In Computational Linguistics and the Foundation of Linguistic Theory, CSLI, 1995

2 Abney, Steven. Parsing by Chunks. In Berwick, Abney, and Tenny, Editors,Principle-based Parsing. Kluwer Academic Publishers,1991

3 Abney,Steven.Chunk Stylebook.University of Tuebingen,Germany, 1996

4 Ciravegna F.Controlling Bottom-Up Parser Through Text Chunking. IRST Technical Report No. 9607-11,1996

5 奚晨海,孙茂松.基于神经元网络的汉语短语边界识别.中文信息学报, 2002,16(2):20-26

6 周　强. 汉语语料库的短语自动划分和标注处理[博士论文].北京: 北京大学, 2003

7 吴竞存, 侯学超. 现代汉语句法分析. 北京: 北京大学出版社, 1982

8 周　强,孙茂松, 黄昌宁. 汉语句子的组块分析体系. 计算机学报, 1999,22(11):1158-1165

9 常宝宝, 詹卫东, 柏晓静等.服务于汉英机器翻译的双语对齐语料库和短语库建设.第2届中日自然语言处理专家研讨会文集,2002

10 周　强, 张　伟.一个改进的汉语短语自动界定模型.中文电脑国际会议, ICCC'96 (新加坡),1996

11 周　强. 汉语短语的自动划分和标注.中文信息学报,1999,(1):1-10

12 孙宏林, 俞士文. 浅层句法分析方法概述. 当代语言学,20002,(2): 74-83