一种基于神经网络的特定文本信息提取方法

李帅 黄玺瑛 董家瑞

（装甲兵工程学院装备作战仿真中心，北京 100072）

摘要：信息提取已成为自然语言处理的一个重要的分支，文本信息提取是从大量的各类文本中提取出指定的相关信息。特定文本作为一种载有特殊的信息载体，其提取方式和方法有其特殊性，本文以军事文本信息作为研究试验的对象，运用特征提取和神经网络相结合的技术，用神经网络进行分词，继而根据军事领域特征词和前后缀进行军事文本中的特定信息的提取。神经网络方法具有学习功能强、开放性好以及分词速度快、精确度高等特点，可以大大提高军事信息提取的召回率和精确度。

关键字：信息提取；神经网络；BP 模型；特征词

An Extracting Measure of the Specific Text Information Based on Neural-Network

LI Shuai, HUANG Xiying, DONG Jiarui

（Academy of Armored Force Engineering, Beijing 100072,China）

Abstract: Information Extraction (IE) has been an important branch in the field of Natural Language Processing (NLP), Text Extraction is to extract specific information from all sorts of text documents. Military Text is a special carrier of information, the information's extracting speed and precision will sometimes determine the victory or defeat of a battle even a war. In the paper, feature extraction technique and Neural-Network technique are used, Neural-Network is used to segment words, and then extracting the military specific text information from military text according to characteristic word，Prefix and suffix. Neural-Network has some advantages such as strong studying ability, good opening property, high segmenting rate, and high precision, so it will improve the recall and precision.

Keywords: Information Extraction, Neural-Network, BP model, characteristic word

1 引言

信息提取是指从给定的文本中提取指定的某一类信息，并将其形成结构化数据，它的主要任务是识别命名实体，确定语义关系。针对文本中特定信息的识别问题，根据其使用方法的不同，现有的方案分为三种类型：规则法、统计法[1][2]以及规则统计相结合[3]的方法，这些方法都是基于大规模语料库[4]的。

军事领域是一个极其特殊的领域，军事文本所承载的信息的提取速度和精度有时能决定一场战争的胜负，因此对军事文本中相关信息的提取方法的不断改进也是军事信息技术进步的一个重要表现形式。军事

作者简介:李帅(1980-)，男，河北安国人，硕士研究生,主要从事自然语言处理，地理信息系统及其应用，装备作战与保障仿真的研究，E-Mail：lishuai6699@126.com。

信息具有其自身的特殊性，其词法、句法及章法都蕴于普通文本的构词成句，组句成章的一般规律而又有其自身的特点：首先，各专业领域的专业词汇构成该领域的文本材料，军用词的核心词汇也具有专业性，军事文本绝大部分是由军事专用语构成的；其次，军事文本的句子结构的简练，军事文本中的句子没有散乱的词句，总体上讲可以说是言简意赅，没有繁言碎语，此外，军事文本中的用语风格比较单一，不用过多的表达思想感情的用语，也不用遣词造句和修辞手法，以陈述事实的陈述和下达命令指示性的祈使句为主。因此，对军事文本中特定信息的提取比起对一般性文本中大海捞针般的提取要简单一些。

本文针对军事文本的性质和特点，运用特征提取和神经网络相结合的技术，用神经网络进行分词，继而根据军事领域一级、二级特征词和特定词前后缀进行军事文本中的特定信息的提取。神经网络方法具有学习功能强、开放性好以及分词速度快、精确度高等特点，可以大大的提高军事信息提取的召回率和精确度。

2 神经网络简介

2.1 神经网络原理

神经网络是在模拟人脑结构和行为的基础上，用大量简单的处理单元广泛连接组成的复杂网络，其研究成果显示了人工神经网络的主要特征为连续时间非线性动力学，网络的全局作用，大规模并行处理及高度的鲁棒性和学习联想能力。BP 算法也称为逆向传播算法，是目前广泛应用于前馈多层网络的学习算法，它含有输入层、输出层、中间层，整体上由三个神经元层次组成，各层次的神经元之间全部构成相互连接，各层次内的神经元之间没有连接。BP 模型的结构如图所示：

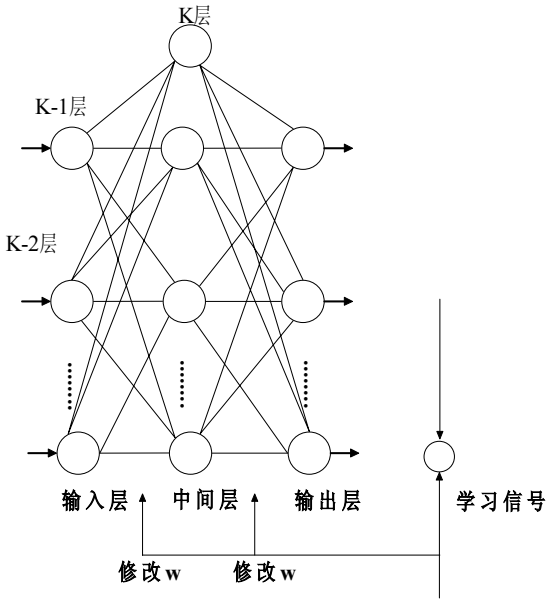

图1　BP 模型图

BP 算法分两步进行，即正向传播和逆向传播。

正向传播：输入的样本从输入层经过中间层一层一层进行处理，通过所有的中间层之后，传向输出层；在逐层处理的过程中，每一层神经元的状态只对下一层神经元的状态产生影响，在输出层把现行输出和期望输出进行比较，如果现行输出不等于期望输出，则进入逆向传播过程。

逆向传播：逆向传播时，把误差信号按原来正向传播的通路反向传回，并对每个中间层的各个神经元的权系数进行修改，以望误差信号趋向最小。

2.2 基于 BP 算法的神经网络的学习训练

BP 算法的的执行步骤如下：设每层有 n 个神经元，即有 $i=1,2,...,n,\quad j=1,2,...,n,$ 对于第 K 层的

第i个神经元则有n个权系数$c_{i1},c_{i2},...,c_{in}$，另外多取一个c_{in+1}用于表示阈值θ_i，并在输入样本X时，取$X=(x_1,x_2,...,x_n,1)$以及对应期望输出$Y=(y_1,y_2,...,y_n)$。

(1) 对各层的权系数c_{ij}置一个较小的非零随机数，其中$c_{in+1}=-\theta_i$；

(2) 读入一个向量对(X,Y)；

(3) 计算各层的输出，对于第K层第i个神经元的输出x_1^k，有：

$$U_i^k=\sum_{j=1}^{n+1}c_{ij}x_j^{k-1}$$

$$x_i^k=f(u_i^k)$$

$$f(x)=1/1+\exp(-x);$$

(4) 求各层的学习误差d_i^k：

对于输出层$K=m$，有：

$d_i^m=x_i^m(1-x_i^m)(x_i^m-y_i)$；

对于其它各层，有：

$d_i^k=x_i^k(1-x_i^k)\sum_l c_{lj}d_1^{k+1}$；

(5) 修正权系数c_{ij}和阈值θ_i分别为：

$c_{ij}(t+1)=c_{ij}(t)-\eta d_i^k x_j^{k-1}$　　$\theta_i=-c_{ij}$

其中η为学习效率，t为时间标志；

(6)当求出了各层各个权系数之后，可按给定品质指标判断是否满足要求，若满足，则算法结束，否则返回(3)执行。

3 军事文本特定信息提取[5]

3.1 分词

在本文中，我们以常见非保密的军用文书作为训练和测试的语料，对其进行统计，找出尾字和下文成词的特定词和其后置词组成的短语，放入表中，形成样本空间表。例如：名词“加榴炮”的尾字“炮”和下文“位”组成词“炮位”，我们将短语“自行加榴炮”放入样本空间表中，作为训练神经网络的语句。我们用来训练神经网络的样本空间示例如下：

①自行加榴炮②基本指挥所③陆军航空兵④作战飞机⑤化学武器⑥234.89 高地⑦预备役部队⑧“阿帕奇”攻击直升机

(1) 语句预处理

为了使神经网络能够接受外部数据，首先要对汉字进行编码，即将从军事文本中读入的句子进行预处理，把句子中的每一个汉字映射成神经网络模型能够接受的数字化的输入形式。编码方式是将汉字的计算机内码作为多个汉字的输入神经元初值，一个汉字的机内码占两个字节，将其化成二进制形式，变成神经网络可以识别的格式。

(2) 建立样本表、一级特征词表、二级特征词表和特定词前后缀表

对样本空间表中的每一个样本语句均进行汉字编码，并给出其切分的期望输出，形成输入输出向量对，存入样本表中；统计样本空间表中的每一条语句，把特征词分为一级特征词如“中国人民解放军”、“陆军”、“海军”、“空军”、“第二炮兵”、“二炮”、“战争”、“装备”等和二级特征词 “集团军”、“炮”、“车”、“员”等分别放入一级特征词表和二级特征词表中，特征词的前置、后置特征词放入另外的前缀、后缀词表中。例如：基本指挥所的汉字编码为如下向量X：

1011101111111001
1011000110111110
1101011010111000
1011101111010011
1100101111111001

对应的期望输出 Y 为：01001000。把(X,Y)向量对存入样本表，作为神经网络训练的输入，将特征词“指挥所”存入二级特征词表，用来进行特征提取。

(3) 训练神经网络

BP 模型的主要参数如下：

①输入层结点数：每个汉字用 16 位表示，若限定句子的长度为 n 个字，则神经元的输入结点数为 16n，在本文样本空间中，最长句子的汉字个数为 8，因此输入层结点数为 128 个；

②中间层结点数：一般比输入层神经元数目少，但是不能过少，否则将限制神经网络存储各种模式的能力，可以考虑设为：16na(a 为 0~1 之间的一个参量)

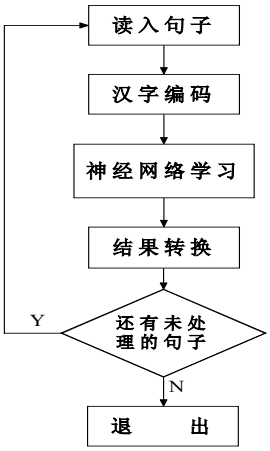

图 2　分词模型

③输出层结点数：输出层结点数与输入句子的汉字最大个数相等。这里我们选 8 个结点。

神经网络的训练过程如下：

训练开始时，将内部连接权，阈值初始化，并根据实验结果确定隐含层数及 a 的值，随机地给网络各单元之间的关联权及单元阈值赋值。

每给网络提供一向量对，首先进行正向传播并计算出各单元的实际输出，求出各单元的相应的误差。当各单元的误差都求出后，进行关联权和各单元阈值的调整，从而完成一次迭代。

取下一向量对，重复上述过程。当样本表中的所有向量对对各自的迭代都完成后,又重复对第一向量对的迭代。这样循环下去,直到输出层单元的误差满足要求为止，并把这时的权值输出形成一个权值表。

(4) 切分过程和规范化

对读入的句子，先查询特征词表，找到特征词以后截取 3～8 个汉字，进行汉字编码，然后根据权值表和神经网络模型进行切分。

例如：截取后的短语为：陆军航空兵

其汉字编码为：

1100001010111101
1011111011111100
1011101010111101
1011111111010101
1011111111010101

切分结果的规范化输出为：01001000，其中的 0 为不切分，1 为切分，也就是陆军/航空兵。

3.2 特征提取

首先运用神经网络对含有一级特征词的短语进行切分，之后查询二级特征词表。如果切分出的第二个词出现在特征词表中，则第一个词为一级特征词，将其提取出来，例如：陆军/航空兵，在二级特征词表中进行查询，找到“航空兵”，则第一个词“陆军”即为一级特征词；若在特征词表中，未能找到切分出的第二个词，继续寻找句子中含有一级特征词的短语，重复以上的分词和特征提取过程，直到输入的这个句子结束。

4 结论

本文运用神经网络进行军事文本的句子进行分词处理，继而根据一级、二级特征词，进行尾字和下文成词的军事信息的担取。要保持军事文本的特定信息提取的较高的召回率和精确度必须不断的扩大样本表和特征词表，这样才能保证神经网络能够进行正确的分词，进而能正确的提取指定信息。

神经网络的分词系统所具有的学习机制,使它可根据用户的要求随意地增添或删除某些关联的权重值,以达到维护信息库的目的。另外，神经网络允许输入学习样本,输入模式越接近于某一学习样本的输入模式,则其输出便越会接近学习样本的输出模式,这使得神经网络系统具有联想记忆的功能。当然，神经网络学习的过程是一个由简到繁、逐渐完成知识积累的过程。

下一步工作中，我们会尝试把神经网络和当前自然语言处理领域占主导地位的统计技术相结合，引入军事语料库，在此基础上进行军事文本信息的提取，利用神经网络分词较为精确的特性和军事特征词的概率统计，建立一个军事文本的信息提取系统，并进一步改进 BP 算法，提高指定的军事文本信息识别、提取的召回率和精确度。

参考文献

[1] 刘秉伟,黄萱菁等.基于统计方法的中文姓名识别[J].中文信息学报,1999,14(3): 16-24.
[2] 庄明,老松杨,吴玲达.一种统计和词性相结合的命名实体发现方法[J].计算机应用,2004,24(1):22-24.
[3] 季妲,罗振声. 基于统计和规则的中文姓名自动辨识[J].语言文字应用,2001,2(1):14-18.
[4] 郑家恒,李鑫等. 基于语料库的中文姓名识别方法研究[J].中文信息学报,1999,14(1):7-12.
[5] 吴芬芬. 信息抽取算法研究. 吉林大学硕士学位论文，2006.

..