神经网络自动分词原理

贺前华　　徐秉铮
(华南理工大学　无线电与自动控制研究所　510641)

如何利用文本中上下文知识解决歧义切分字段是自动汉语分词中的真正困难所在，利用知识消除分词歧义实际上是求解约束满足问题。用传统方法求解约束满足问题是很困难的，但这正好是 PDP 系统能以很自然的方式予以解决的问题。只要用一个单元表示一个假设，用一个联接表示假设之间的一个约束就行了。如果约束很重要，它的联接权重就大；反之，约束不那么重要，权重就小。

分词的模型如图 1 所示。网络由四部分组成，F_1 是输入单元集，F_2 是输出单元集，F_3 是自学习机制，F_4 是结构自组织机制。输入单元与国标 GB2312-80 中的汉字有一一对应关系，每个单元代表一个汉字，当对应汉字出现在输入模式中时，该单元被激活，置激活值为 1，否则置激活值为 0。输出单元的激活函数在训练阶段和测试阶段采用不同的形式。学习阶段为恒同函数，而测试阶段为线性阈值函数。F_2 是 F_1 的映象，有一一对应关系。输出模式代表输入字串的切分方式，某单元被激活表示该单元对应的汉字是输入字串中词与词的分界线。词与词之间的约束关系由联接矩阵 W表示。

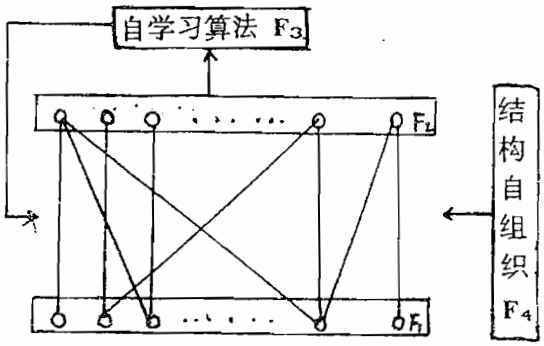

图 1　分词神经网络模型

该分词模型实际上是一模式联想机，输入模式可以是词也可以是句子，采用δ学习规则进行训练。设 $a_1 \ldots a_{n-1} a_n$ 为一词汇词，对应的输入模式为 1...11 (n 个 1)，网络的正确输出模式为 0...01 (n-1 个 0)。图 2 是词的网络映象之一。

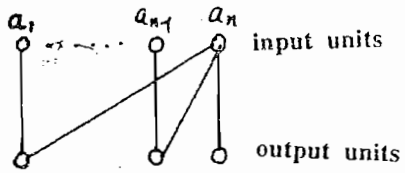

图 2　汉语词的网络映象

系统在分词时以一个语句为处理单位，首先对语句进行必要的预处理，分词网络以知识库为依据，根据预定规则动态生成。动态网运行后得到输出模式，该模式由确定的规则解释成语句的切分方式。应用神经网络进行分词，分词知识以统一的形式——联接权重存于网络的联接模式中，不必组建大量的产生式规则，且系统具有良好的学习能力和自适应能力，这是其它分词系统所不具有的性质。

歧义切分是汉语自动分词的真正困难所在，应用神经网络进行分词，总的处理原则是：分析影响歧义切分字段切分形式的约束因素，在歧义切分字段与这些控制因素的载体(上下文中的字或词)之间建立约束联接。系统经过学习后这些约束联接可保证该歧义字段在一类语境中得到正确切分。

分词神经网络分层次进行训练，经过训练后，分词网络具有了一定量的词汇知识、语法知识、语义及语用知识。目前系统的知识量为 22000 多词的词汇知识和 200 多句典型歧义句的切分知识，占 60KB 多的空间，系统对测试样本可达 100%的正确切分，切分速度达 650 字／秒。网络是开放的，当遇到新问题时，可以学习新知识，由自组织算法 F_4 和自学习算法 F_3 共同完成。