词性标注实现方法

# 简介

本文采用基于神经网络和规则相结合的方法，用于消减多词性词词性标注的不确定现象，进而提高分词准确程度。由于神经网络方法具有很好的概括能力，适合于较小的训练数据集，而且其基于上下文的处理能力较强。但是由于神经网络方法存在的一些缺陷，由于神经网络内部方法的不确定性，我们很难分辨其显示的最好和次好的词性标注在现实中的好坏，这就产生了多词性词的不确定线像。于是我们在采取读取神经网络输出，判断近似程度，进而用规则的方法进行再次确定的方法进行词性标注。 实验表明，仅基于神经网络的正确率为89%左右（由于训练较少），采取规则修正后为91%左右。

# 神经网络算法

## 神经网络参数

①、网络节点

输入层神经元个数由输入词组的个数确定，125\*3=375，即输入由待标记次的前后各两个词的词性概率及待标记次本身的词性概率表达。

输出层神经元（输出待标记词——中间词）个数由词的标注集大小确定，标注集大小为125，故输出层神经元个数为125，输出为0~1，之间的数，代表中间词在被标记为每个词性的概率，便于后面HMM的处理。

隐含层神经元个数，经验公式s=sqrt(0.43mn+0.12nn+2.54m+0.77n+0.35)+0.51，其中m为输入层个数，n为输出层个数；s = 152，但是由于神经网络过大，训练时间过长，在试验后修改为100

②、初始权值的确定

初始权值是不应完全相等的一组值。已经证明，即便确定  存在一组互不相等的使系统误差更小的权值，如果所设*W*ji的的初始值彼此相等，它们将在学习过程中始终保持相等。故而，在程序中，我们设计了一个随机发生器程序，产生一组一0.2~+0.2的随机数，作为网络的输入层与隐含层之间的初始权值，产生一组一2.0~+2.0的随机数，作为网络的隐含层与输出层之间的初始权值。

④、学习参数确定，本文采取学习率为0.05，修正率为0.1的参数。

⑤、Sigmoid函数，本文采用较为简单的1/(1+e-t)为激励函数

# 基于规则的HMM算法

本文，采取最为简单的三词的HMM算法，详见课件，在此不再赘述。

# 融合

## 融合方法

本文采取直接检测神经网络输出神经元输出的最大及次大两个词性的概率，如果两个概率相差小于0.85（参数调节后的结果），就采用基于规则的方法进行重新词性标注。训练发现，有9%左右的数据采取神经网络词性标注会得到相近的两个结果。

## 不足之处

不得不说，这次实验有很多不足之处。

①首先由于神经网络构建过大、本身电脑性能约束，使得训练时间大大提升，使得我们不得不采用削减输入词数、削减隐含神经元个数、减小训练集（本身训练集是两百多万的数据，但是我们仅仅训练了前三十万，并且并没有全部利用这三十万数据，最重要的是我们仅在三十万数据上训练了一次，这还花费三个小时）的方法这就直接导致了神经网络词性标注仅有89%的正确率。

②HMM算法截取的句子过短问题，我们仅仅截取了三个词作为一个句子进行计算概率，这会导致HMM算法对上下文的利用程度不够，但是为了与神经网络算法的融合，减少文本的读取时间，不得不如此。

③融合问题，我们的参数为人为调整。同时本应存在比这种‘硬链接’式方法更好地融合方法。

注：全文中我们没有进首词和尾词的词性标注

参 考 文 献：

[1] 刘伟权, 王明会,钟义信. 应用两种神经网络模型自动标 注汉语词类[ J ] .北京邮电大学学报, 1997, 34( 6):421 -426.

[2] Qing Ma, Masaki Murata, Kiyotaka Uchimoto, et al.Hybrid neuro and rule-based part of speech taggers[ J ] .Association for Computational Linguistics, 2000, 1( 4): 509-515.