Foolnltk阅读笔记

包的结构分析：

* Data文件夹里是所有的文档测试中要调用的文件
* Fool文件夹中主要用于测试fool模块 \_\_init.py\_\_ 用来定义各函数的基本操作 里面的dictionary.py lexicial.py 都是foolnltk的应用 通过分词，写入词典来生成用户自己的词典
* Test 文件夹内主要是用来测试 通过给与一个txt文档，在文档中设置各个词语的权重从而实现用户自定义词典
* Train 中也是基于foolnltk 模块的基本功能进行简单的操作和应用

Foolnltk 本质是一种中文处理工具包

## 特点

\* 可能不是最快的开源中文分词，但很可能是最准的开源中文分词

\* 基于[BiLSTM模型](http://www.aclweb.org/anthology/N16-1030 )训练而成

\* 包含分词，词性标注，实体识别,　都有比较高的准确率

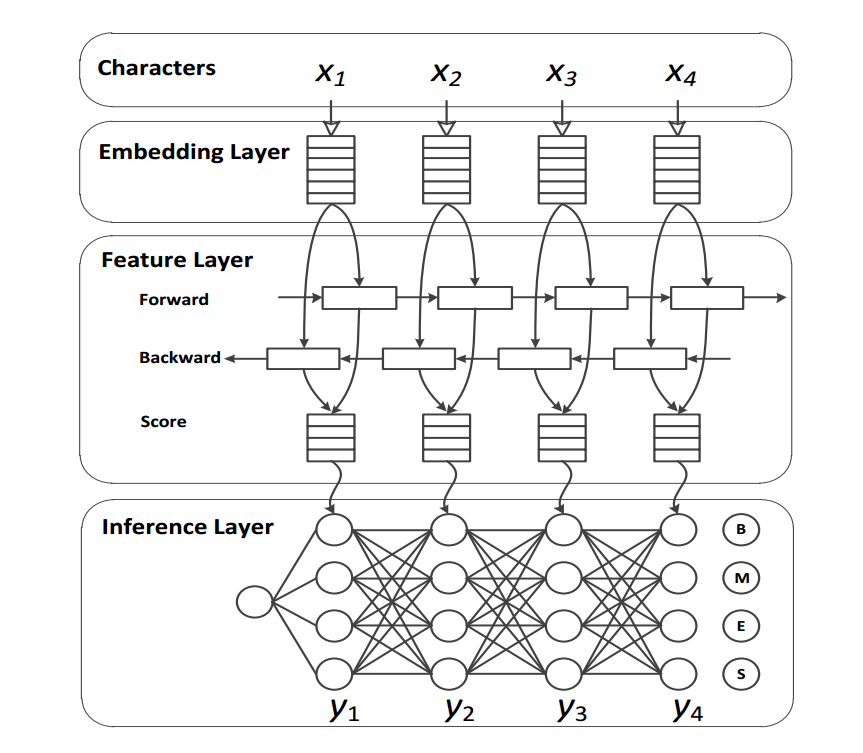
\* 用户自定义词典

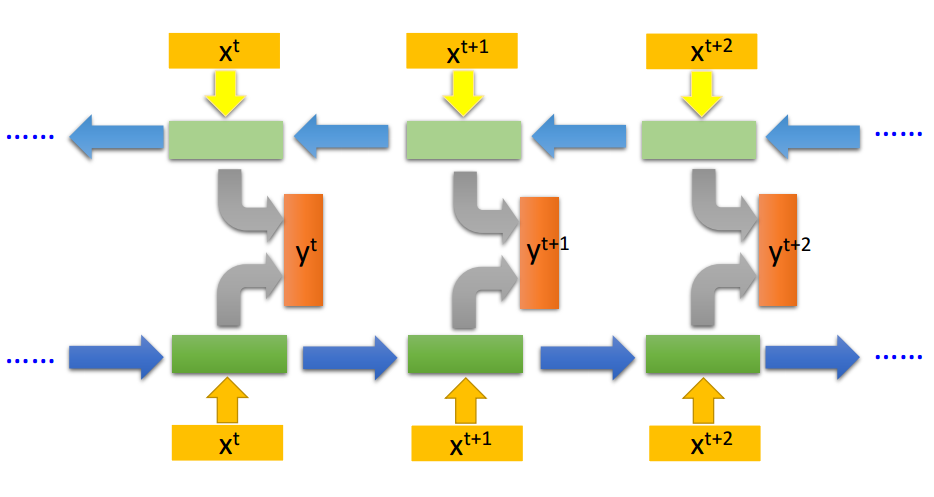
\* 可训练自己的模型

\* 批量处理

BiLSTM 模型：

在中文分词上，基于**神经网络**的方法，往往使用「字向量 + 双向 LSTM + CRF」模型，利用神经网络来学习特征，将传统 CRF 中的人工特征工程量将到最低。





将两个反向读取的 LSTM 网络连接就成了 BiLSTM。

特点：使用这种 BiLSTM 的好处是模型的观察范围比较广。这种分词工具更倾向于将其分割为更为离散的词组。但是可以将专业词汇添加到词组中加强分词的效果。

依赖：python3

安装：

pip install foolnltk

使用：

python -m fool [filename]

功能：

* 分词功能 （命令行分词）
* 用户自定义词典
* 删除词典
* 词性标注
* 实体识别

用户自定义词典

词的权重越高，词的长度越长就越越可能出现, 权重值大于 1

#add\_argument 读入命令行参数 ArgumentParser.add\_argument(name or flags…[, action][, nargs][, const][, default][, type][, choices][, required][, help][, metavar][, dest])

关于命令行分词

>>> parser.add\_argument('-f', '--foo') #选项参数

>>> parser.add\_argument('bar') #位置参数

不知道命令行和参数是什么意思

命令行：

1. 命令行的含义：

1.1在字面意思讲，命令行就是指一个终端命令输入软件，系统会自动编译并执行您所输入的命令，别称“命令提示符”或“终端”。

什么是元组：

Python的元组与列表类似，不同之处在于元组的元素不能修改。

元组使用小括号，列表使用方括号。

元组创建很简单，只需要在括号中添加元素，并使用逗号隔开即可。

