**统计自然语言处理**

**中文分词项目**

姓名：张奇

学号：1132130124

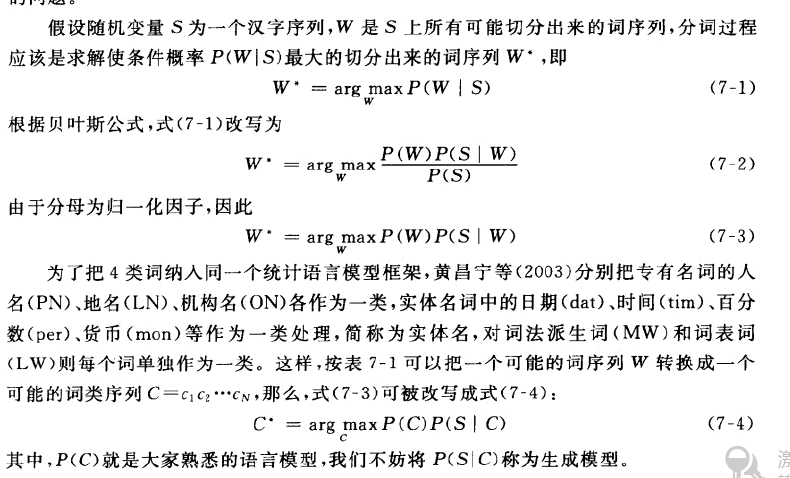
中文分词的概念：

中文分词(Chinese Word Segmentation) 指的是将一个汉字序列切分成一个一个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。我们知道，在英文的行文中，单词之间是以空格作为自然分界符的，而中文只是字、句和段能通过明显的分界符来简单划界，唯独词没有一个形式上的分界符，虽然英文也同样存在短语的划分问题，不过在词这一层上，中文比之英文要复杂的多、困难的多。

摘自百度百科

**算法：**

1. 分词任务的形式化定义



1. 分词任务的建模方法

本项目采用最大匹配法：

最大匹配是指以词典为依据，取词典中最长单词为第一个次取字数量的扫描串，在词典中进行扫描（为提升扫描效率，还可以跟据字数多少设计多个字典，然后根据字数分别从不同字典中进行扫描）。例如：词典中最长词为“中华人民共和国”共7个汉字，则最大匹配起始字数为7个汉字。然后逐字递减，在对应的词典中进行查找。

以 “我是一个坏人” 为例

正向的顺序为  
我是一  
我是  
我 ===> 得到一个词  
是一个  
是一  
是 ===>得到一个词  
一个坏  
一个===> 得到一个词  
坏人===>得到一个词  
结果 我、是、一个、坏人

反向算法  
个坏人  
坏人==> 坏人  
是一个  
一个==> 一个  
我是  
是==> 是  
我==> 我  
结果 我、是、一个、坏人

3) 算法伪代码

逆向最大匹配：  
(1)将文章分成句子(通过标点符号来实现)；  
(2)循环的读入每一个句子S，设句子中的字数为n；  
(3)设置一个最大词长度，就是我们要截取的词的最大长度max；  
(4)从句子中取n-max到n的字符串subword，去字典中查找是否有这个词。  
如果有就走(5),没有就走(6)；  
(5)记住subword，从n-max付值给n，继续执行(4)，直到n=0。  
(6)将max-1，再执行(4)。

伪代码如下：  
max=4;  
for(i..N){  
　　　sentence = 读入下一个句子；  
　 n = 取字数(sentence)；  
 for(j=n;j>0;){  
 subword = 截取字符串(j-max,j)；  
 flag=去字典中查找(subword);  
 if(flag){//找到这个词  
 记住(subword);  
 j=n-max;  
 }else{  
 for(k=max;k>0;k--){  
 tempWord=截取字符串(subword.length-max,subword.length，subword)  
 temFlg=去字典中查找(subword);  
 if(temFlg)break;  
 }  
 记住(tempWord)  
 j=n-max+k;  
 }  
}  
}

**分词算法设计中的几个基本原则：**

1、颗粒度越大越好：用于进行语义分析的文本分词，要求分词结果的颗粒度越大，即单词的字数越多，所能表示的含义越确切，如：“公安局长”可以分为“公安 局长”、“公安局 长”、“公安局长”都算对，但是要用于语义分析，则“公安局长”的分词结果最好（当然前提是所使用的词典中有这个词）

2、切分结果中非词典词越少越好，单字字典词数越少越好，这里的“非词典词”就是不包含在词典中的单字，而“单字字典词”指的是可以独立运用的单字，如“的”、“了”、“和”、“你”、“我”、“他”。例如：“技术和服务”，可以分为“技术 和服 务”以及“技术 和 服务”，但“务”字无法独立成词（即词典中没有），但“和”字可以单独成词（词典中要包含），因此“技术 和服 务”有1个非词典词，而“技术 和 服务”有0个非词典词，因此选用后者。

3、总体词数越少越好，在相同字数的情况下，总词数越少，说明语义单元越少，那么相对的单个语义单元的权重会越大，因此准确性会越高。

**实验结果：**

1. 训练，测试数据规模

共有7000句，3527字

IMG_256

1. 评价方法

1、准确率与召回率（Precision & Recall）

       精度和召回率是广泛用于信息检索和统计学分类领域的两个度量值，用来评价结果的质量。其中精度是检索出相关文档数与检索出的文档总数的比率，衡量的是检索系统的查准率；召回率是指检索出的相关文档数和文档库中所有的相关文档数的比率，衡量的是检索系统的查全率。

一般来说，Precision就是检索出来的条目（比如：文档、网页等）有多少是准确的，Recall就是所有准确的条目有多少被检索出来了。

正确率、召回率和 F 值是在鱼龙混杂的环境中，选出目标的重要评价指标。不妨看看这些指标的定义先：

    1. 正确率 = 提取出的正确信息条数 /  提取出的信息条数

    2. 召回率 = 提取出的正确信息条数 /  样本中的信息条数

两者取值在0和1之间，数值越接近1，查准率或查全率就越高。

    3. F值  = 正确率 \* 召回率 \* 2 / (正确率 + 召回率) （F 值即为正确率和召回率的调和平均值）

不妨举这样一个例子：某池塘有1400条鲤鱼，300只虾，300只鳖。现在以捕鲤鱼为目的。撒一大网，逮着了700条鲤鱼，200只虾，100只鳖。那么，这些指标分别如下：

正确率 = 700 / (700 + 200 + 100) = 70%

召回率 = 700 / 1400 = 50%

F值 = 70% \* 50% \* 2 / (70% + 50%) = 58.3%

不妨看看如果把池子里的所有的鲤鱼、虾和鳖都一网打尽，这些指标又有何变化：

正确率 = 1400 / (1400 + 300 + 300) = 70%

召回率 = 1400 / 1400 = 100%

F值 = 70% \* 100% \* 2 / (70% + 100%) = 82.35%

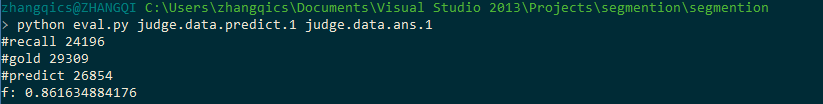
 由此可见，正确率是评估捕获的成果中目标成果所占得比例；召回率，顾名思义，就是从关注领域中，召回目标类别的比例；而F值，则是综合这二者指标的评估指标，用于综合反映整体的指标。

        当然希望检索结果Precision越高越好，同时Recall也越高越好，但事实上这两者在某些情况下有矛盾的。比如极端情况下，我们只搜索出了一个结果，且是准确的，那么Precision就是100%，但是Recall就很低；而如果我们把所有结果都返回，那么比如Recall是100%，但是Precision就会很低。因此在不同的场合中需要自己判断希望Precision比较高或是Recall比较高。如果是做实验研究，可以绘制Precision-Recall曲线来帮助分析

评价脚本为eval.py

3) 准确率

在验证集jugde.data.1上分词输出为jugde.data.predict.1，与标准结果jugde.data.ans.1比较



在验证集jugde.data.2上分词输出为jugde.data.predict.2，与标准结果result.dev比较

