# 基于HMM的命名实体识别

## 一、学习任务

### 1.1、命名实体识别

在文本中对实体的引用可以有三种形式：命名性指称、名称性指称、代词性指称。例如在句子“中国乒乓球男队主教练刘国梁出席了会议，他布置了备战世乒赛的具体安排”中“刘国梁”的指称项有三个，其中“中国乒乓球男队主教练”是名词性指称，“刘国梁”是命名性指称，“他”是代词性指称。

命名实体识别任务是识别出文本中实体的命名性指称项，并标明其类别。一般来说，命名实体识别任务中，实体类别一般分为三大类（实体类、时间类、数字类）、七小类（人名、机构名、地名、时间、日期、货币和百分比）。不同任务对命名实体类别粒度的需求不同，例如在有些任务中只要识别出一个实体是人，有些任务则需要识别出一个实体是教师、学生或者医生。

### 1.2、命名实体识别标注方法

目前主流的命名实体标注方法分为两种，一种是BIO方法,一种是BIOES方法。BIO方法中B标注命名实体的开始词、I标注命名实体识别的中间词、O标注不属于命名实体的其他词。BIOES方法中B、I、O的意义与BIO标注方法中的相同，但增加了E和S，E标注命名实体的结束词、S标注只包含一个词的命名实体。通常在标注的过程中，会加上命名实体的类别信息。如下是分别使用BIO方法和BIOES方法对同一个句子进行标注的结果。

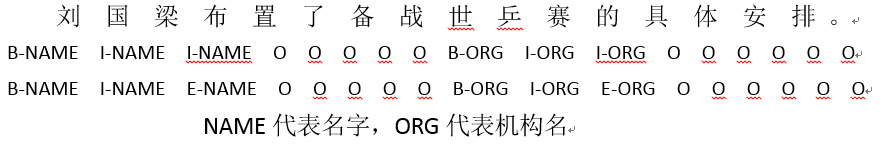


图1.1

### 1.3、命名实体识别评估标准

在命名实体识别任务中，只有当预测的命名实体边界和类别与实际完全相同时，才可以认为识别成功。准确率、召回率，F1值是评估命名实体识别系统性能的三大指标。式（1-1）、式（1-2）、式（1-3）分别是它们的计算公式。其中TP(True Positive)表示预测的实体中，正确实体的数量、FP(False Positive)表示预测的实体中错误实体的数量。FN(False Negative)表示语料库中未被预测出的实体数量。

 式（1-1）

 式（1-2）

 式（1-3）

## 二、学习算法

### 2.1、隐马尔可夫模型概念

隐马尔可夫模型是关于时序的概率模型，描述由一个隐藏的马尔可夫链随机生成不可观测的状态随机序列，再由各个状态生成一个观测而产生观测随机序列的过程。隐马尔可夫随机生成的状态的序列，称为状态序列；每个状态生成一个观测，而由此产生的观测的随机序列，称为观测序列。

隐马尔可夫模型由初始状态概率向量、状态转移概率矩阵和观测概率矩阵确定。 的形式化定义如下。

设是所有可能的状态集合，是所有可能的观测的集合



其中是可能的状态数，是可能的观测数

设是长度为的状态序列，是对应的观测序列



是状态转移概率矩阵：

 式（2-1）

其中 是时刻t处于状态的概率

是观测概率矩阵：

 式（2-2）

其中 是时刻t处于状态的概率

是初始状态概率向量

 式（2-3）

其中  

### 2.2、概率计算算法

概率计算算法是计算已知隐马尔可夫模型的参数（），求观测序列出现的概率，即求，采用直接计算的方法时间复杂度是，实际完全不可行，前向算法、后向算法是时间复杂度为，是计算有效算法。

#### 2.2.1、前向算法：

前向概率：给定隐马尔可夫模型，定义到时刻部分观测序列为的概率，记作

 式（2-4）

前向算法求概率的步骤如下

1. 初值

 式（2-5）

1. 递推

 式（2-6）

1. 终止

 式（2-7）

#### 2.2.2、后向算法

后向概率：给定隐马尔可夫模型，定义在时刻状态为的条件下，从到的部分观测序列为的概率为后向概率，记作

 式（2-8）

后向算法求概率的步骤如下

1. 初值

 式（2-9）

1. 递推

 式（2-10）

1. 终止

 式（2-11）

### 2.3、学习算法

#### 2.3.1、极大似然估计

假如训练数据既包括观测序列、状态序列，可以利用极大似然来估计隐马尔可夫模型的参数。

1. 估计转移概率，设样本中时刻处于状态时刻转移到状态的频数为，那么状态转移概率的估计是

 式（2-12）

1. 估计观测概率，设样本中状态为并且观测为的频数是，那么状态为观测为的概率的估计是

 式（2-13）

1. 初始状态概率的估计为 为样本中初始状态为的频率

#### 2.3.2、Baum-Welch算法

给定模型和观测，在时候处于状态的概率记为

 式（2-14）

给定模型和观测，在时刻处于状态且在时刻处于状态的概率，记

 式（2-15）

Baum-Welch算法估计模型参数步骤如下

1. 初始化

对,选取得到模型

1. 递推

 式（2-16）

 式（2-17）

 式（2-18）

1. 终止，得到模型，

### 2.4、预测算法

维特比算法是解决隐马尔可夫模型预测问题的有效方法之一。其利用动态规划方法求概率最大路径，这时一条路径对应着一个状态序列。

定义两个变量。定义在时刻t状态为i的所有单个路径中概率最大值为

 式（2-19）

定义在时刻状态为的所有单个路径中概率最大的路径的第个节点为

 式（2-20）

维特比算法具体步骤如下：

1. 初始化

 式（2-21）

 式（2-22）

1. 递推

 式（2-23）

 式（2-24）

1. 终止

 式（2-25）

 式（2-26）

1. 最优路径回溯

 式（2-27）

## 三、数据集

数据集是从网上获取的一个简历数据集，分为三个文件，分别是train.char.bmes、test.char.bmes、dev.char.bmes文件。三个文件统一用BIOES标注方法标注。如图3.1所示，每一行为一个中文字符及其对应的标记标注，中文字符和标记用空格隔开。句子与句子之间用换行符分割。train.char.bmes有将近10万个中文字符，test.char.bmes、dev.char.bmes分别有大约1.5万个中文字符

数据集中的实体类别包括NAME(名字)、CONT(国籍)、EDU(学历)、TITLE(头衔)、ORG(组织)、RACE(民族)、PRO(专业)、LOC(籍贯)。命名实体识别的任务是同时准确识别命名实体的边界和类别

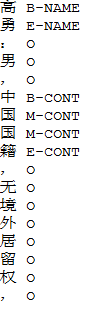


图3.1

## 四、学习结果

train.char.bmes文件中的数据作为训练集，test.char.bmes文件中的数据作为测试集，采用有监督的学习算法进行实验。最终实验结果如图4.1所示。测试集中共有1630个命名实体，TITLE、ORG、LOC、CONT、NAME、EDU、RACE、PRO的实体个数分别为772、553、6、28、112、112、14、33。其中F1值最高的两个类别是国籍和民族，F1值最低的两个类别是专业和籍贯。忽略类别，求平均，可以得到HMM模型在测试集的性能：准确率为0.822287、召回率为0.860123、F1值0.840780

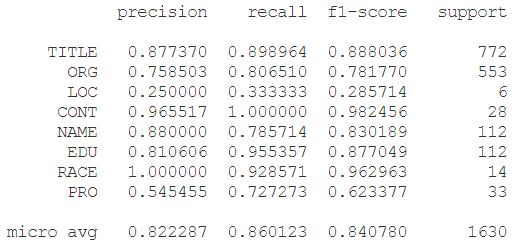


图4.1

## 五、理论分析

观察实验结果，可以发现对于不同的实体类别隐马尔可夫模型预测出来的结果相差较大。通过分析，发现主要是两个因素造成的，一个是类别本身的特性，国籍、民族这些类别的命名实体都很固定、规范，而专业、籍贯这些类别的命名实体多种多样。另一个是训练集中类别所包含的实体数量，在训练集中，类别包含的实体越多，其召回率、f1值，相应得也越高。

隐马尔可夫模型的学习算法有两种，一种是有监督的学习算法，主要是利用最大似然估计。另一种是无监督的学习算法Baum-Welch算法，其基本原理是期望极大算法（EM）。经过实验发现，无监督的学习算法严重依赖于初值的选择，而命名实体识别要求不仅仅是标出单个字符的标记，而是必须识别实体的边界和类别，所以无监督的学习算法在测试集上效果不理想。利用带标签的数据、使用有监督的学习算法，在命名实体识别的测试集上，F1值能达0.84。

序列标注的机器学习模型，不仅仅包含隐马尔可夫模型，条件随机场也是

一种有效得模型。未来的话，考虑实现条件随机场模型，进一步提高命名实体

识别的性能。