实验二 汉语命名实体自动识别系统

一、实验目的

1. 熟悉国内外汉语命名实体自动识别技术的进展
2. 独立完成汉语命名实体自动识别系统

二、命名实体识别方法综述

命名实体是命名实体识别的研究主体，一般包括三大类（实体类、时间类和数字类）和七小类（人名、地名、机构名、时间、日期、货币和百分比）命名实体。评判一个命名实体是否被正确识别包括两个方面：实体的边界是否正确和实体的类型是否标注正确。

命名实体识别的主要技术方法分为：基于规则和词典的方法、基于统计的方法、二者混合的方法等。

1、基于规则和词典的方法

基于规则的方法多采用语言学专家手工构造规则模板，选用特征包括统计信息、标点符号、关键字、指示词和方向词、位置词（如尾字）、中心词等方法，以模式和字符串相匹配为主要手段，这类系统大多依赖于知识库和词典的建立。

2、基于统计的方法

基于统计机器学习的方法主要包括隐马尔可夫模型（HiddenMarkovMode，HMM）、最大熵（MaxmiumEntropy，ME）、支持向量机（Support VectorMachine，SVM）、条件随机场

（ConditionalRandom Fields，CRF）等。

3、混合方法

(1)统计学习方法之间或内部层叠融合。(2)规则、词典和机器学习方法之间的融合，其核心是融合方法技术。在基于统计的学习方法中引入部分规则，将机器学习和人工知识结合起来。(3)将各类模型、算法结合起来，将前一级模型的结果作为下一级的训练数据，并用这些训练数据对模型进行训练，得到下一级模型。

三、基于条件随机场(CRF)汉语命名自动识别系统实验原理

实验可以采用基于条件随机场(CRF)来实现汉语命名自动识别系统，也可以采用上述的其它方法。

条件随机场(CRF)：设 X 与 Y 是随机变量，P(Y|X)是给定 X 的条件下 Y 的条件概率分布，若随机变量 Y 构成一个由无向图 G=(V,E)表示的马尔科夫随机场。则称条件概率分布 P(Y|X)为条件随机场。因为是在 X 条件下的马尔科夫随机场，所有叫条件随机场。

linear chain CRF 的公式如下：

$$P(Y \mid X) = \frac{1}{Z} \prod_{i=0}^{n-1} \psi_i(Y_i, Y_{i+1} \mid X)$$

再详细一些如下：

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} u_l s_l(y_i, x, i)\right)$$

t 和 s 都是特征函数，一个是转移特征，一个状态特征，x=（x1,x2,...,xn）为观察变量,y=(y1,y2,...,yn)为隐含变量。所以，CRF 也就是直接预测 p（y|x）,属于判别式模型。注意一个细节，特征函数里面的观测变量为 x，而不是 xi，这也就是说你可以前后随意看观测变量，所以特征模板里面可以随意定义前后要看几个观测值。亦或表示如下：

$$P(I|O)=\frac{1}{Z(O)}\prod_{i}\Psi_i(I_i|O)=\frac{1}{Z(O)}\prod_{i}e^{\sum_k\lambda_k f_k(O_i,I_{i-1},I_i,i)}=\frac{1}{Z(O)}e^{\sum_i\sum_k\lambda_k f_k(O_i,I_{i-1},I_i,i)}$$

O 为观察序列，I 为预测的隐变量序列。

模型训练过程：CRF 模型的训练主要训练特征函数的权重参数 λ，一般情况下不把两种特征区别的那么开，合在一起如下：

$$P(I|O)=\frac{1}{Z(O)}e^{\sum_i^T\sum_k^M\lambda_k f_k(O,I_{i-1},I_i,i)}$$

每个 token 会对应多个特征函数，特征函数 f 取值为 0 或者 1，在训练的时候主要训练权重 λ，权重为 0 则没贡献，甚至你还可以让他打负分，充分惩罚。利用极大似然估计寻找最优参数解。

四、实验步骤

1. 采用人民日报 1998 中文标注语料库(群文件)
2. 通过合适的文本预处理方法得到 CRF++的数据格式(训练文件

和测试文件都需要写成特定的格式)

CRF 中通常定义的词位信息如下：

(1) 实体首部，常用 B 表示。

(2) 实体中部，常用 M 表示。

(3) 实体尾部，常用 E 表示。

(4) 单子词，常用 S 表示。

(5) 单个实体，用 W 表示。

(6) 非实体，用 O 表示。

3. 模型训练(需要自己了解 CRF 算法的实现)以及测试

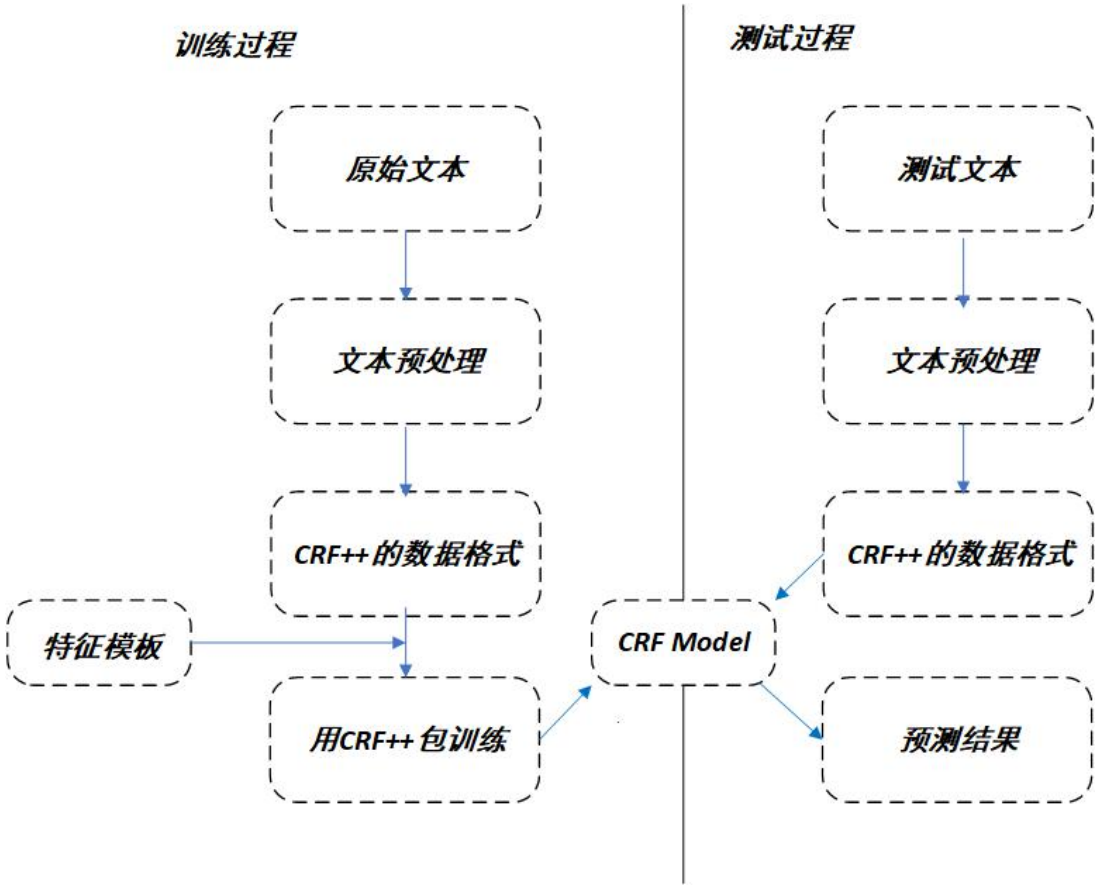
4. 最后通过准确率、召回率和 f 值来判断命名实体识别的准确度。

准确率 = 交集/模型抽取出的实体

召回率 = 交集/数据集中的所有实体

f 值 = 2 × (准确率 × 召回率) / (准确率 + 召回率)

实验流程图如下：



五、实验要求

1. 本次实验，两节课完成，第二节实验课交由老师检查实验结果，并在一周内按时提交实验报告。

2. 实验报告统一格式：学号+姓名+第*次实验.pdf(doc)