北京航空航天大學

BEIHANG UNIVERSITY

自然语言处理实验一

基于 HMM 的命名实体识别

院 系 名 称	国际通用工程学院
任 课 老 师	丁嵘
学 生 姓 名	林威
学 生 学 号	19351024

2021 年 10 月 21 日

（1）模型原理简介

1. HMM 模型

隐马尔可夫模型（HMM）是统计模型，它用来描述一个含有隐含未知参数的马尔可夫过程。根据定义，HMM 要求存在一个可观察的过程 Y，其结果以已知的方式收到 X 结果的影响。因为 X 不能被直接观察，所以我们的目标是通过观察 Y 来学习到 X。HMM 还要求 Y 在 $t=t_0$ 时的结果只会被 X 在 $t=t_0$ 时的结果影响。

对于命名实体识别来说，“BIO”的实体标注就是一个不可观察的隐状态。而文本内容就是一个可观察状态。HMM 模型描述的就是由“BIO”实体标注（即隐状态）生成文本（可观察状态）的过程。

我们的可观察状态序列是由所有汉字以及中文标点符号组成的集合，用V_{obs}来表示：$V_{obs} = \{v_1, v_2, \dots, v_M\}$。v 表示字典中的单个字，M 表示整个中文集合的大小。

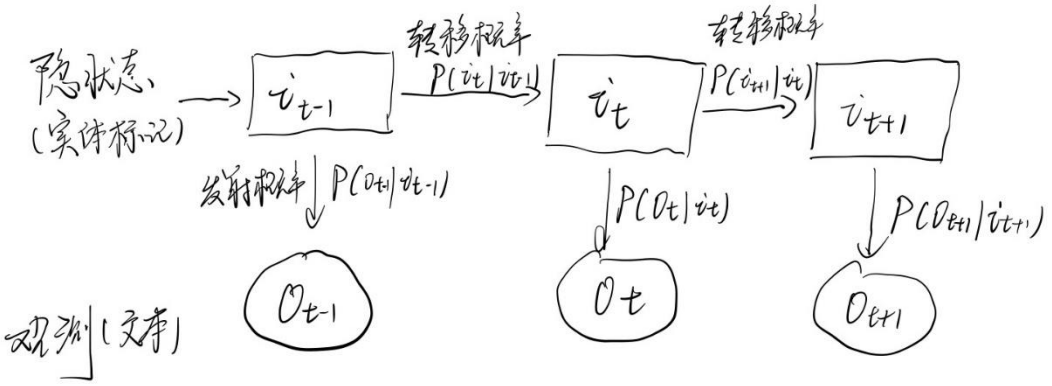
我们的隐藏状态集合时所有的“BIO”命名标签，用Q_{hid}来表示：$Q_{hid} = \{q_1, q_2, \dots, q_N\}$。其中 N 为标签的数量。

对于训练集来说，我们观察到一串中文文本 O，一共有 T 个汉字，我们也知道这段文本对应的实体标签，也就是隐状态 I：

$$I = \{i_1, i_2, \dots, i_T\} \text{ (实体标签，隐状态)}$$

$$O = \{o_1, o_2, \dots, o_T\} \text{ (文本，可观察状态)}$$

用 HMM 模型来描述，即为下图的过程：



正如图中所描述的，HMM 有两个重要的假设：

1. 第 t 个隐状态 i_t（实体标签）只和第 t-1 个隐状态 i_{t-1} 有关。
2. 第 t 个可观察状态 o_t 只和第 t 个隐状态 i_t 有关。

2. HMM 的参数

转移概率：图中的$P(i_t|i_{t-1})$)是指隐状态从i_{t-1}转移到i_t的概率。我们的隐状态一共有 N 种，因此可以用一个 N*N 的矩阵 A_{N*N} 来表示所有的隐状态转移概率:

$$A_{ij} = P(i_t = q_j | i_{t-1} = q_i) \quad q_i, q_j \in Q_{hid}$$

我们可以对训练集进行统计来得到转移概率矩阵的参数估计：

$$\widehat{A_{ij}} = \frac{\text{count}(q_i\text{后面出现}q_j\text{的次数})}{count(q_i\text{出现的次数})}$$

发射概率：图中的$P(o_t|i_t)$是指隐状态生成观察结果的概率。我们一共有 N 个隐状态，M 个观察结果，因为可以用一个 N*M 的矩阵 B_{N*M} 来表示所有的发射概率：

$$B_{jk} = P(o_t = v_k | i_t = q_j) \quad q_j \in Q_{hid}, v_k \in V_{obs}$$

我们可以对训练集进行统计来得到发射概率矩阵的参数估计：

$$\widehat{B_{ij}} = \frac{count(q_j\text{与}v_k\text{同时出现的次数})}{count(q_j\text{出现的次数})}$$

初始隐状态概率：对于自然语言序列中的第一个字o_1的实体标记是q_i的概率，也

就是初始隐状态概率，我们用一个向量$\vec{\pi}$表示：

$$\pi_i = P(i_1 = q_i)\ \ q_i \in Q_{hid}$$

我们可以对训练集进行统计来得到初始隐状态概率向量的参数估计：

$$\widehat{\pi_{q_1}} = \frac{count(q_i^1)}{count(o_1)}$$

即q_i第一次出现的次数占总第一个字o_1观测次数的比例。

3. 维特比算法

维特比算法使用了动态规划来解决 HMM 的预测问题，用维特比算法可以找到概率最大路径。在命名实体识别当中，我们也可以用维特比算法来找到文本对应的最有的实体标注序列。

简单来说，维特比算法就是在每一时刻，计算当前时刻落在每种隐状态的最大概率，并记录这个最大概率是从前一个时刻的哪一种隐状态转移过来的，最后从结尾回溯最大概率，也就是最有可能的路径。

计算过程：

一开始我们要计算初始隐状态的生成概率:

$$P\left(q_j^{t=0}\right) = \pi_j Bjk$$

对于接下来的某一时刻 t，我们要计算到每个隐状态的最大概率：

$$P\left(q_j^{t=t'}\right) = \max\left(P\left(q_i^{t=t'-1}\right)A_{ij}B_{jk}\right)$$

并且记录下转移路径：

$$step^{t=t'} = \operatorname{argmax}_i\left(P\left(q_i^{t=t'-1}\right)A_{ij}B_{jk}\right)$$

当我们遍历完整个文本时，就可以回溯转移路径来得到最有可能的隐状态序列了。

（2） 实验步骤与结果介绍

1. 数据预处理

{"text": "浙商银行企业信贷部叶老桂博士则从另一个角度对五道门槛进行了解读。叶老桂认为，对目前国内商业银行而言，", "label": {"name": {"叶老桂": [[9, 11]]}, "company": {"浙商银行": [[0, 3]]}}}
{"text": "生生不息CSOL生化狂潮让你填弹狂扫", "label": {"game": {"CSOL": [[4, 7]]}}}
{"text": "那不勒斯vs锡耶纳以及桑普vs热那亚之上呢？", "label": {"organization": {"那不勒斯": [[0, 3]], "锡耶纳": [[6, 8]], "桑普": [[11, 12]], "热那亚": [[15, 17]]}}}
{"text": "加勒比海盗3：世界尽头》的去年同期成绩死死甩在身后，后者则即将赶超《变形金刚》，", "label": {"movie": {"加勒比海盗3：世界尽头》": [[0, 11]], "《变形金刚》": [[33, 38]]}}}
{"text": "布鲁京斯研究所桑顿中国中心研究部主任李成说，东亚的和平与安全，是美国的“核心利益”之一。", "label": {"address": {"美国": [[32, 33]]}, "organization": {"布鲁京斯研究所桑顿中国中心": [[0, 12]]}, "name": {"李成": [[18, 19]]}, "position": {"研究部主任": [[13, 17]]}}}

将如上图所示的训练集进行 BIO 标注，得到的结果如下图所示：

{'text': ['浙', '商', '银', '行', '企', '业', '信', '贷', '部', '叶', '老', '桂', '博', '士', '则', '从', '另', '一', '个', '角', '度', '对', '五', '道', '门', '槛', '进', '行', '了', '解', '读', '。', '叶', '老', '桂', '认', '为', '，', '对', '目', '前', '国', '内', '商', '业', '银', '行', '而', '言', '，'], 'label': ['B-COM', 'I-COM', 'I-COM', 'I-COM', 'O', 'O', 'O', 'O', 'O', 'B-NAME', 'I-NAME', 'I-NAME', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']}
{'text': ['生', '生', '不', '息', 'C', 'S', 'O', 'L', '生', '化', '狂', '潮', '让', '你', '填', '弹', '狂', '扫'], 'label': ['O', 'O', 'O', 'O', 'B-GAME', 'I-GAME', 'I-GAME', 'I-GAME', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']}
{'text': ['那', '不', '勒', '斯', 'v', 's', '锡', '耶', '纳', '以', '及', '桑', '普', 'v', 's', '热', '那', '亚', '之', '上', '呢', '？'], 'label': ['B-ORG', 'I-ORG', 'I-ORG', 'I-ORG', 'O', 'O', 'B-ORG', 'I-ORG', 'I-ORG', 'O', 'O', 'B-ORG', 'I-ORG', 'O', 'O', 'B-ORG', 'I-ORG', 'I-ORG', 'O', 'O', 'O', 'O']}
{'text': ['加', '勒', '比', '海', '盗', '3', '：', '世', '界', '尽', '头', '》', '的', '去', '年', '同', '期', '成', '绩', '死', '死', '甩', '在', '身', '后', '，', '后', '者', '则', '即', '将', '赶', '超', '《', '变', '形', '金', '刚', '》', '，'], 'label': ['B-MOVIE', 'I-MOVIE', 'I-MOVIE', 'I-MOVIE', 'I-MOVIE', 'I-MOVIE', 'I-MOVIE', 'I-MOVIE', 'I-MOVIE', 'I-MOVIE', 'I-MOVIE', 'I-MOVIE', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-MOVIE', 'I-MOVIE', 'I-MOVIE', 'I-MOVIE', 'I-MOVIE', 'I-MOVIE', 'O']}
{'text': ['布', '鲁', '京', '斯', '研', '究', '所', '桑', '顿', '中', '国', '中', '心', '研', '究', '部', '主', '任', '李', '成', '说', '，', '东', '亚', '的', '和', '平', '与', '安', '全', '，', '是', '美', '国', '的', '“', '核', '心', '利', '益', '”', '之', '一', '。'], 'label': ['B-ORG', 'I-ORG', 'I-ORG', 'I-ORG', 'I-ORG', 'I-ORG', 'I-ORG', 'I-ORG', 'I-ORG', 'I-ORG', 'I-ORG', 'I-ORG', 'I-ORG', 'B-POS', 'I-POS', 'I-POS', 'I-POS', 'I-POS', 'B-NAME', 'I-NAME', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-ADDRESS', 'I-ADDRESS', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']}

2. 模型实现

然后根据上文描述的模型，将其用 python 实现。具体主要包含三个大块：a) 根据训练集统计计算 HMM 的三个参数：初始概率矩阵、发射概率矩阵和转移概率矩阵。b) 用维特比算法对输入的文本进行标注预测 c)用 seqeval 对预测结果进行评估。

3. 实验结果

	precision	recall	f1-score	support
ADDRESS	0.35	0.32	0.33	373
BOOK	0.33	0.26	0.29	154
COM	0.46	0.45	0.45	378
GAME	0.55	0.66	0.60	295
GOV	0.37	0.47	0.41	247
MOVIE	0.45	0.50	0.47	151
NAME	0.54	0.56	0.55	465
ORG	0.46	0.47	0.47	367
POS	0.57	0.62	0.59	433
SCENE	0.41	0.38	0.39	209
micro avg	0.47	0.49	0.48	3072
macro avg	0.45	0.47	0.46	3072
weighted avg	0.46	0.49	0.47	3072

具体的实验结果如上图所示，可以看到单纯的用 HMM 统计方法进行命名实体识别的准确率并不是很高。