自然语言处理课程实验报告

Koji Tadokoro
Shimokitazawa College

1. 背景

命名实体识别（named entity recognition, NER）是一个经典任务，它旨在识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等，以及时间、数量、货币、比例数值等文字。举个如下的例子：

美国 *(LOC)* 的 华莱士 *(PER)* ，我 跟 他 谈笑风生。

这句话中含有地名“美国”和人名“华莱士”，命名实体识别的任务即是从文本中识别出这些实体。这次作业中采用条件随机场（Conditional Random Fields, CRF）模型来处理命名实体识别任务。

2. 要点简述

本小节就学习课程内容时和实验中遇到的一些要点做出和讨论。

生成式序列标注与判别式序列标注. 序列标注（Sequence Labeling）是自然语言处理中最基础的任务，应用十分广泛，如分词、词性标注、命名实体识别、语义角色标注等实质上都属于序列标注的范畴。可以将序列标注任务形式化描述为：对模型输入序列 $W_{1:n} = w_1 w_2 \cdots w_n$，模型给出最有可能的标注序列 $T_{1:n} = t_1 t_2 \cdots t_n$。

生成式模型和判别式模型都可以用于序列标注任务。生成式模型从数据中学习联合概率 $P(T_{1:n}, W_{1:n})$，再根据贝叶斯公式求解条件概率 $P(T_{1:n}|W_{1:n})$ 进行预测，例如隐马尔科夫模型（HMM）朴素贝叶斯都等都属于生成式模型；判别式模型直接根据数据学习条件概率 $P(T_{1:n}|W_{1:n})$，典型的判别式模型包括感知机、支持向量机（SVM）、对数线性模型（log-linear models）如最大熵模型等，判别式模型的一个优点是可以对数据进行各种程度上的抽象，定义并使用丰富的特征，学习的是条件概率 $P(T_{1:n}|W_{1:n})$ 或决策函数 $f(W_{1:n})$，直接面向预测，简化了学习问题，往往准确率较生成式模型更高。

最大熵马尔科夫模型与标签偏置. 最大熵马尔科夫模型（Maximum Entropy Markov Model, MEMM）是一种对数线性模型，它结合了 HMM 和最大熵模型的特点，采用了 HMM 中的解码算法和前向-后向算法，也使用了最大熵模型中特征提取和归一化部分的方法。首先根据链式法则和马尔科夫假设对给定序列的条件概率进行分解：

$$P(T_{1:n}|W_{1:n}) = \prod_{i=1}^{n} P(t_i|T_{1:i-1}, W_{1:n}) \approx \prod_{i=1}^{n} P(t_i|T_{i-k:i-1}, W_{1:n}) \tag{1}$$

其中，k 代表模型采用 k 阶马尔科夫假设，即每个位置的标签只依赖于序列中前 k 个标签的值。

最大熵马尔科夫模型根据最大熵原理建模 $P(t_i|T_{i-k:i-1}, W_{1:n})$，如式 2 所示。

$$P(t_i = t|T_{i-k:i-1}, W_{1:n}) = \frac{\exp\left(\vec{\theta} \cdot \vec{\phi}(t_i = t, T_{i-k:i-1}, W_{1:n})\right)}{\sum_{t' \in L} \exp\left(\vec{\theta} \cdot \vec{\phi}(t_i = t', T_{i-k:i-1}, W_{1:n})\right)} \tag{2}$$

其中，$\vec{\phi}$ 表示特征向量，L 表示含有所有可能标签取值的集合，$\vec{\theta}$ 表示参数向量即分配给不同特征函数的权重。

HMM 与 MEMM 的对比如图 1 所示（采用一阶马尔科夫假设）。在 HMM 的概率图中，蓝色虚线框中的是转移概率 $P(t_i|t_{i-1})$，生成标签序列；黄色虚线框中的是发射概率 $P(w_i|t_i)$，生成词序列。在 MEMM 中也有生成式的过程，但这里是给定 w_i 之后生成标签序列的生成过程，与 HMM 中将一个时间步的状态和观测同时生成出来的做法是有区别的。

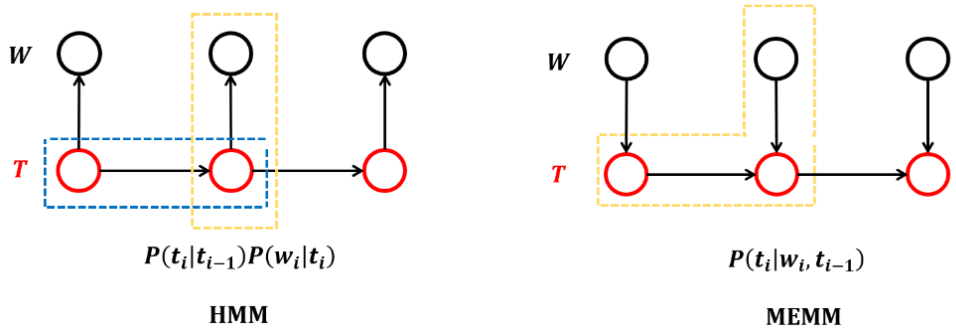

图 1. HMM 和 MEMM 的概率图（一阶马尔科夫假设）。

如式 1 所示，MEMM 通过对局部的条件概率的建模，间接地对全局的标签序列进行了概率估算，但局部的概率归一化和全局建模存在着矛盾：如果存在某一个标签非常倾向于另一个特定的标签作为其转移的后继，那么这样一个标签转移对会被局部模型过多地选择到模型预测出的标签序列里，也就是标签偏置（label bias）问题。

条件随机场. 为了解决标签偏置问题，很自然的做法就是将整个标签序列作为单独的单元考虑，在序列上进行概率分布的归一化。条件随机场是一个对数线性模型，它通过聚合局部特征向量得到输入序列的全局特征向量，直接建模整个序列的条件概率 $P(T_{1:n}|W_{1:n})$，在整个序列上进行概率归一化。

与之前的 case study 中用平均感知机模型处理中文分词问题 [1]中的特征提取方法一样，输入序列的全局特征向量都是由聚合所有的局部特征向量得到：

$$\vec{\phi}(T_{1:n}, W_{1:n}) = \sum_{i=1}^{n} \vec{\phi}(t_i, T_{i-k:i-1}, W_{1:n}) \tag{3}$$

其中，k 代表模型采用 k 阶马尔科夫假设，$\vec{\phi}$ 表示特征向量。

CRF 求解条件概率的方法如式 4 所示。

$$P(T_{1:n}|W_{1:n}) = \frac{\exp\left(\sum_{i=1}^{n} \vec{\theta} \cdot \vec{\phi}(t_i, T_{i-k:i-1}, W_{1:n})\right)}{\sum_{\overline{T}_{1:n}} \exp\left(\sum_{i=1}^{n} \vec{\theta} \cdot \vec{\phi}(\overline{t}_i, \overline{T}_{i-k:i-1}, W_{1:n})\right)} = \frac{\exp\left(\vec{\theta} \cdot \vec{\phi}(T_{1:n}, W_{1:n})\right)}{\sum_{\overline{T}_{1:n}} \exp\left(\vec{\theta} \cdot \vec{\phi}(\overline{T}_{1:n}, W_{1:n})\right)} \tag{4}$$

3. 实验设置与实验结果

实验设置. 数据集采用人民日报语料（1998 年 1 月），使用了 7:3 的训练集-测试集划分。选取的实体类型有 4 种：地点（LOC）、组织（ORG）、人名（PER）、时间（T），标注体系选择 BIO，只标出每种实体类型的起始 token（B-X）和非起始 token（I-X）以及非实体（O）。选用了较简单的特征模板 $\{w_{-1}, w_0, w_{+1}, w_{-1}w_0, w_0w_1\}$ 进行特征提取。调用了 sklearn-crfsuite 库实现 CRF 模型的定义，学习算法选择默认的 L-BFGS 算法。

实验结果. 用训练完的模型对测试集进行预测，得到的结果如表 1 所示。

	Precision	Recall	F1-score
B-LOC	0.925	0.865	0.894
I-LOC	0.906	0.853	0.879
B-ORG	0.937	0.906	0.921
I-ORG	0.923	0.885	0.904
B-PER	0.963	0.907	0.934
I-PER	0.955	0.922	0.938
B-T	0.979	0.961	0.970
I-T	0.984	0.975	0.979
micro avg	0.944	0.908	0.926
macro avg	0.947	0.909	0.927
weighted avg	0.944	0.908	0.925

表 1. 模型评估结果。

实验分析. 通过声明 CRF 模型时将传入的 all_possible_transitions 参数设置为 True，模型会生成所有标签之间的转移概率。我们加载训练完之后的模型，可以观察到模型认为更可能发生的和更不可能发生的转移，发生可能性最高的转移如表 2 所示，发生可能性最低的一些转移则如表 3 所示。

top likely transitions
I-LOC → I-LOC
B-LOC → I-LOC
B-ORG → I-ORG
B-PER → I-PER
B-T → I-T
O → O

表 2. 更可能发生的状态转移对。（按可能性由高到低排列）

结合表 2 和表 3 可以看出，从一个地名的起始 token（B-LOC）或中间 token（I-LOC）发生的转移更倾向于选择地名的中间 token（I-LOC）作为其后继，而从其他标签（如表 3 中的 O → I-LOC 和 I-ORG → I-LOC）跳转到 I-LOC 的转移会受到惩罚。

top unlikely transitions
I-T → B-T
O → I-LOC
O → I-T
O → I-ORG
I-ORG → I-LOC
O → I-PER

表 3. 更不可能发生的状态转移对。（按可能性由低到高排列）

另外，可以读取模型观察特征-标签对的权重高低，做了排序之后得到的权重最高的一些特征-标签对的信息如表 4 所示。可以观察到：对顿号（"、"）和句号（"。"），模型几乎可以断定它对应的标签是非实体（O），这也是从语料中很容易学习到的信息；对" 刘"、" 吴"、" 李" 这些人名的姓氏，模型倾向于分配给它们 B-PER 的标签。

label	attribute
O	w_0: 、
B-PER	w_0: 刘
O	w_0: ，
O	w_0: 的
B-PER	w_0: 吴
O	w_0: 。
B-PER	w_0: 袁
O	w_0: 与
B-PER	w_0: 郭
O	w_0: 对
I-T	$w_{-1}w_0$: 晚上
B-PER	$w_{-1}w_0$: 江主

表 4. 权重最高的一些特征-标签对。（按权重由高到低排列）

参考文献

[1] Y. Zhang and S. Clark. Chinese segmentation with a word-based perceptron algorithm. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 840–847, 2007. 2