汉语命名实体自动识别系统

条件随机场

条件随机场(CRFs) 是用来标注和划分序列结构数据的概率化结构模型，就是对于给定的输出标识序列 Y 和观测序列 X，条件随机场通过定义条件概率 P（Y|X），而不是联合概率分布 P（X，Y）来描述模型。 CRF 也可以看作一个无向图模型或者马尔可夫随机场（Markov random field）。

设 G＝（V，E）为一个无向图，V 为结点 集合，E 为无向边的集合。Y＝{Yv|v∈V}，即 V 中的每个结点对应于一 个随机变量 Yv，其取值范围为可能的标记集合{y}。如果以观察序列 X 为 条件，每一个随机变量 Yv 都满足以下马尔可夫特性：

$$p(Y_v \mid X, Y_w, w \neq v) = p(Y_v \mid X, Y_w, w \sim v)$$

其中，w~v 表示两个结点在图 G 中是临近结点，那么(X,Y)是一个条件随机场。

CRF 的链式结构图如 1-1 所示：

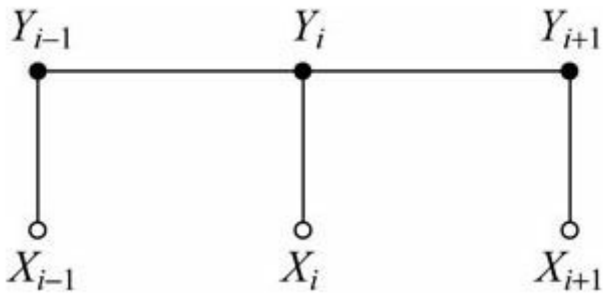

图 1-1

在给定观察序列 X 时，某个特定标记序列 Y 的概率可以定义为：

$$\exp\left(\sum_j \lambda_j t_j(y_{i-1}, y_i, X, i) + \sum_k \mu_k s_k(y_i, X, i)\right)$$

其中，tj（yi-1，yi，X，i）是转移函数，表示对于观察序列 X 其标注序列 在 i 及 i－1 位置上标记的转移概率；sk（yi，X，i）是状态函数，表示对于观察序列 X 其 i 位置的标记概率； λ j 和 μ k 分别是 tj 和 sk 的权重，需要从训练样本中估计出来。

条件随机场模型需要解决三个基本问题：特征的选取、参数训练和解码。其中，参数训练过程可在训练数据集上基于对数似然函数的最大化进行。

相对于 HMM，CRF 的主要优点在于它的条件随机性，只需要考虑当前已经出现的观测状态的特性，没有独立性的严格要求，对于整个序列内部的信息和外部观测信息均可有效利用，避免了 MEMM 和其他针对线性序列模型的条件马尔可夫模型会出现的标识偏置问题。CRF 具有 MEMM 的一切优点，两者的关键区别在于，MEMM 使用每一个状态的 指数模型来计算给定前一个状态下当前状态的条件概率，而 CRF 用单个指数模型来计算给定观察序列与整个标记序列的联合概率。因此，不同状态的不同特征权重可以相互交替代换。

基于 CRF 的命名实体识别方法

原理 基于 CRF 的命名实体识别原理就是把命名实体识别过程看作一个序列标注问题。其基本思路是（以汉语为例）：将给定的文本首先进行分词处理，然后对人名、简单地名和简单的组织机构名进行识别，最后识别复合地名和复合组织机构名。它是属于有监督的学习方法，因此需要利用已标注的大规模语料对 CRF 模型的参数进行训练。

步骤 训练阶段：首先需要将分词语料的标记符号转化成用于命名实体 序列标注的标记，如本实验用 B 表示地名的首部，M 表示地名中部，E 表示地名尾部，S 表示单字词，W 表示单个实体，O 表示非实体。确定特征模版：特征模板一般采用当前位置的前后 n（n≥1）个位置上的字（或词、字母、数字、标点等，不妨统称 为“字串”）及其标记表示，即以当前位置的前后 n 个位置范围内的字串及其标记作为观察窗口。如果窗口开得较大时，算法的执行效率会太低，而且模板的通用性较差，但窗口太小时，所涵盖的信息 量又太少，不足以确定当前位置上字串的标记，因此，一般情况下将 n 值取为 2～3，即以当前位置上前后 2～3 个位置上的字串及其标记作为构成特征模型的符号。最后一步就是训练 CRF 模型参数。