知识图谱

技术分享

参考项目：https://github.com/buppt/ChineseNER.git

目 录

C O N T E N T S

3

the part one

命名实体识别——序列标注

3.0、命名实体识别

本节课目标：

1、理解NER任务，能自己做简单的NER任务

2、理解概率图模型，能自己做简单的推算，可以通过课后的自学达到面试水平

3.0、命名实体识别

任务简介

标注规范

常用数据集

开源工具

基于概率图的方法

HMM

MEMM

CRF

基于深度学习的方法

LSTM

Transformer

BERT

基于词典和规则的方法

Trim树

3.1、序列标注—规范，数据集，工具

1、序列标注任务

NLP中一个重要的任务，它包括分词、词性标注、命名实体识别等等

2、命名实体识别

命名实体识别(Named Entity Recognizer,NER)在第六届信息理解会议(MUC-6)被提出，是信息抽取、本体构建、问答系统等自然语言处理任务的基础

3、标注规范

这段标注采用的是BIO标注方式，即Begin，Intermediate，Other，针对不同的标注任务标注方式也各不相同。常用规范还有：BIOE，BIOES

他	和	爸爸	去	电影	院	看	哈利	波特
B-Per	O	B-Per	O	B-Loc	I-Loc	O	B-Per	I-Per

他	和	爸	爸	去	电	影	院	看
B-Per	O	B-Per	E-Per	O	B-Loc	I-Loc	E-Loc	O
哈	利	波	特					
B-Per	I-Per	I-Per	E-Per					

他	和	爸	爸	去	电	影	院	看
S-Per	O	B-Per	I-Per	O	B-Org	I-Org	I-Org	O
哈	利	波	特					
B-Per	I-Per	I-Per	I-Per					

3.1、序列标注—规范，数据集，工具

1、人民日报数据集

人名、地名、组织名三种实体类型。

2、New York Times 数据集

NYT数据集是通过将freebase中的关系与纽约时报（NYT）语料库对齐而生成的。纽约时报New York Times数据集包含150篇来自纽约时报的商业文章。抓取了从2009年11月到2010年1月纽约时报网站上的所有文章。

3、MSRA微软亚洲研究院数据集

5 万多条中文命名实体识别标注数据（包括地点、机构、人物）

4、CoNLL 2003

这个数据集包括1393篇英语新闻文章和909篇德语新闻文章。英语语料库是免费的，德国语料库需要收钱(75美元)。英语语料实际上是RCV1(Reuters Corpus, Volume 1， https://trec.nist.gov/data/reuters/reuters.html)，路透社早些年公开的一些数据集。

3.1、序列标注—规范，数据集，工具

1、HanLP

HanLP（汉语言处理包）是一款开源的使用Java进行开发的中文自然语言处理工具，提供的功能包括中文分词、词性标注、命名实体识别、依存句法分析等。

2、哈工大 HIT LTP

LPT(Language Technology Platform)是哈尔滨工业大学开发的中文自然语言处理工具。代码开源，商业使用付费。支持中文分词、词性标注、命名实体识别、依存句法分析、语言角色标注（中文）

3、清华 THU LAC

THULAC (THU Lexical Analyzer for Chinese) 是由清华大学自然语言处理与社会人文计算实验室研制推出的一套中文词法分析工具包，具有中文分词和词性标注功能。

4、斯坦福 Stanford CoreNLP

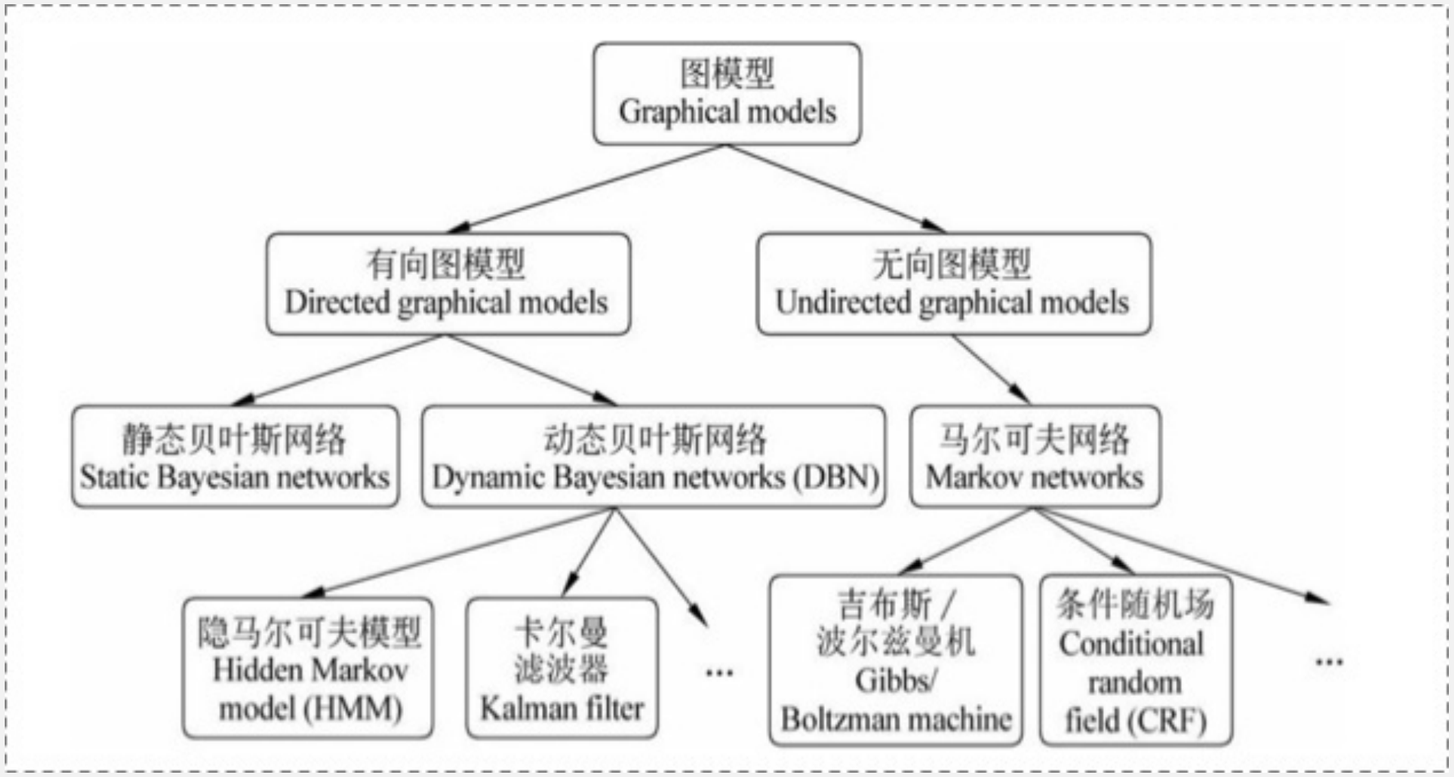
自然语言处理工具包，能实现对自然语言文本的文本分析，包括词形还原，词性标注、命名实体标注、共指消解、句法分析以及依存分析等功能

3.1、序列标注—规范，数据集，工具

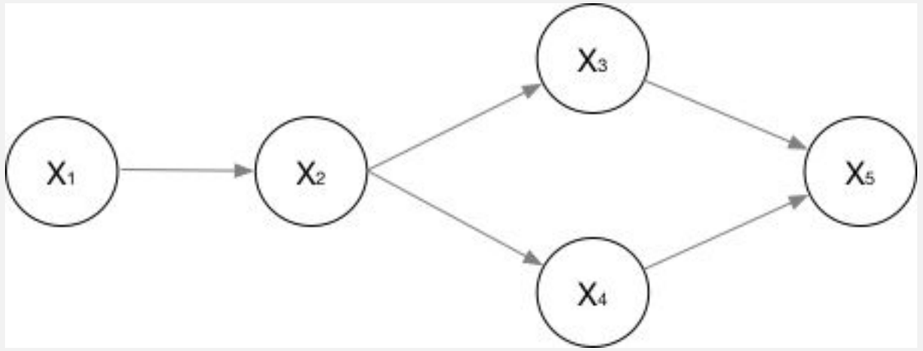
看一个例子

https://bosonnlp.com/demo?source=home-banner

3.2、序列标注-基于概率图的方法



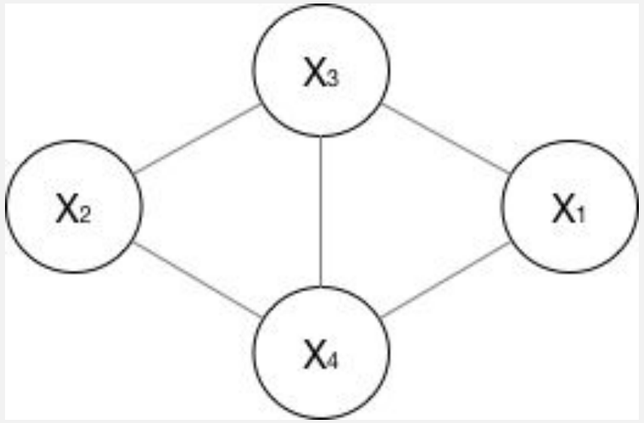
3.2、序列标注—有向图和无向图



公式1 $$P(x_1,\cdots,x_n)=\prod_{i=0} P(x_i|\pi(x_i))$$

公式2

$$P(x_1,\cdots,x_n)=P(x_1)\cdot P(x_2|x_1)\cdot P(x_3|x_2)\cdot P(x_4|x_2)\cdot P(x_5|x_3,x_4)$$



概率无向图模型，又称为马尔可夫随机场

它假设随机场中任意一个结点的赋值，仅仅和它的邻结点的取值有关，和不相邻的结点的取值无关。

无向图G中任何两个结点均有边连接的结点子集称为团。若C是无向图的一个团，且不能再加进任何一个G的结点使其成为更大的一个团，则此C为最大团。

它Y的联合概率可以表示为其最大团C上随机变量的函数的乘积的形t式。

公式3 $$P(Y)=\frac{1}{Z(x)}\prod_{c}\psi_c(Y_c)$$

公式4 $$Z(x)=\sum_{Y}\prod_{c}\psi_c(Y_c)$$

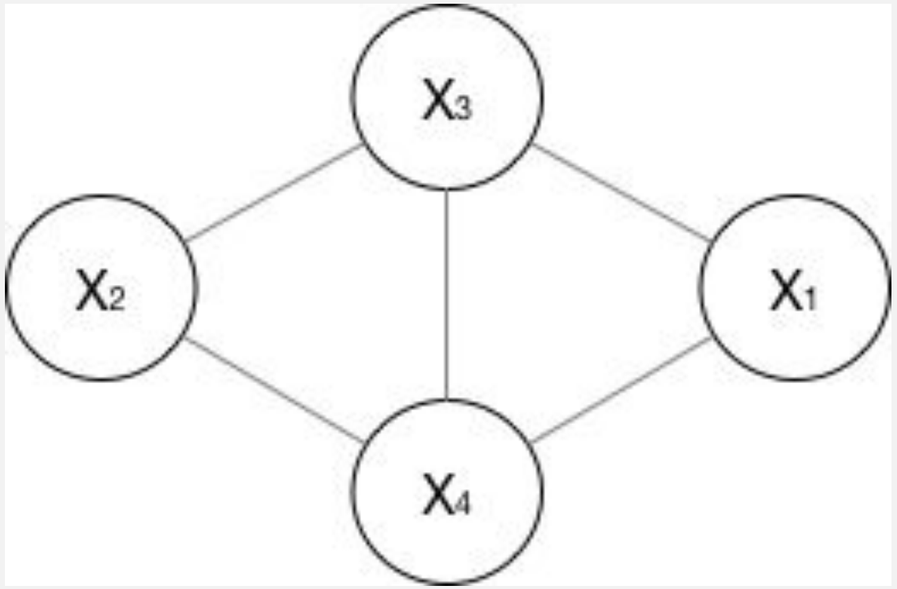
公式5 $$\psi_c(Y_c)=e^{-E(Y_c)}$$

Hammersley-Clifford定理

3.2、成对马尔可夫性

- 设 u 和 v 是无向图G中任意两个没有边连接的结点，结点u和v分别对应随机变量 Yu 和 Yv。其他所有结点为 O（集合），对应的随机变量组是 Yo。成对马尔可夫性是指给定随机变量组 Yo 的条件下随机变量 Yu 和 Yv 是条件独立的，其实意思就是说没有直连边的任意两个节点是独立的，即

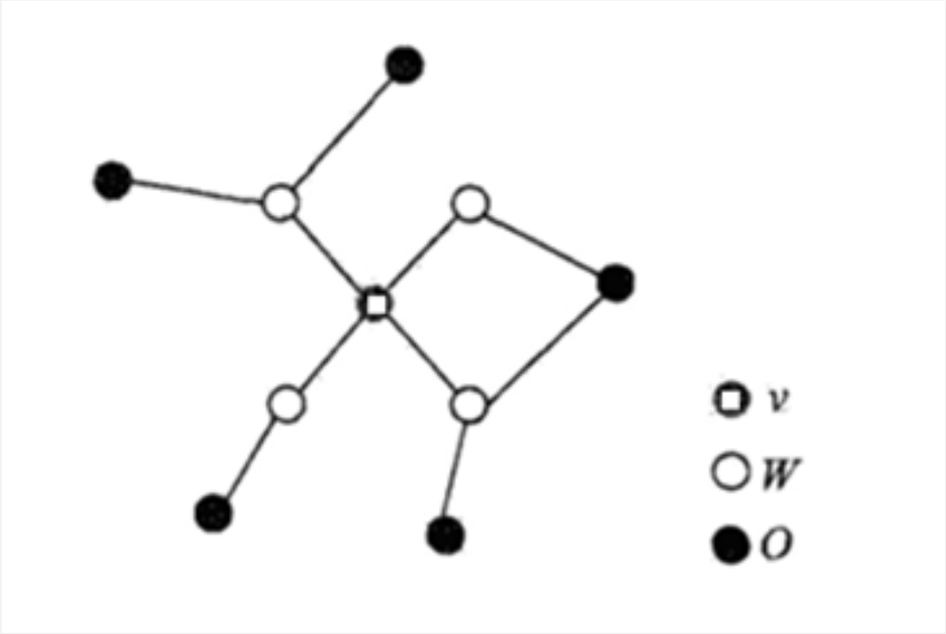
$$P(Y_u, Y_v|Y_O) = P(Y_u|Y_O)P(Y_v|Y_O)$$



3.2、局部马尔可夫性

- 设 v \in V 是无向图 G 中任意一个结点，W 是与 v 有边连接的所有结点，O 是 v,W 以外的其他所有结点。v 表示的随机变量是 Yv ,W 表示的随机变量组是 Y_w，O 表示的随机变量组是 Y_o。局部马尔可夫性是指在给定随机变量组 Y_w 的条件下随机变量 v 与随机变量组 Y_o 是独立的，即

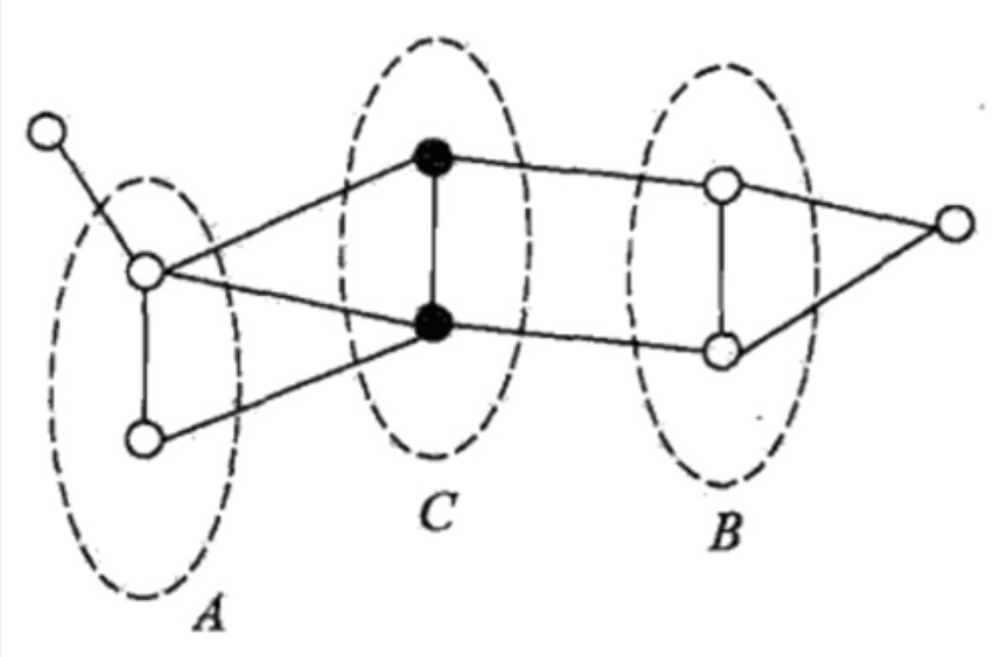
$$P(Y_v, Y_O|Y_W) = P(Y_v|Y_W)P(Y_O|Y_W)$$



3.2、全局马尔可夫性

- 设结点集合 A,B 是在无向图 G 中被结点集合 C 分开的任意结点集合，如图所示。结点集合 A,B 和 C 所对应的随机变量组分别是 YA,YB 和 YC。全局马尔可夫性是指给定随机变量组条件下随机变量组 YA 和 YB 是条件独立的，即

$$P(Y_A, Y_B|Y_C) = P(Y_A|Y_C)P(Y_B|Y_C)$$



3.2、序列标注

判别式（discriminative）模型 vs. 生成式(generative)模型

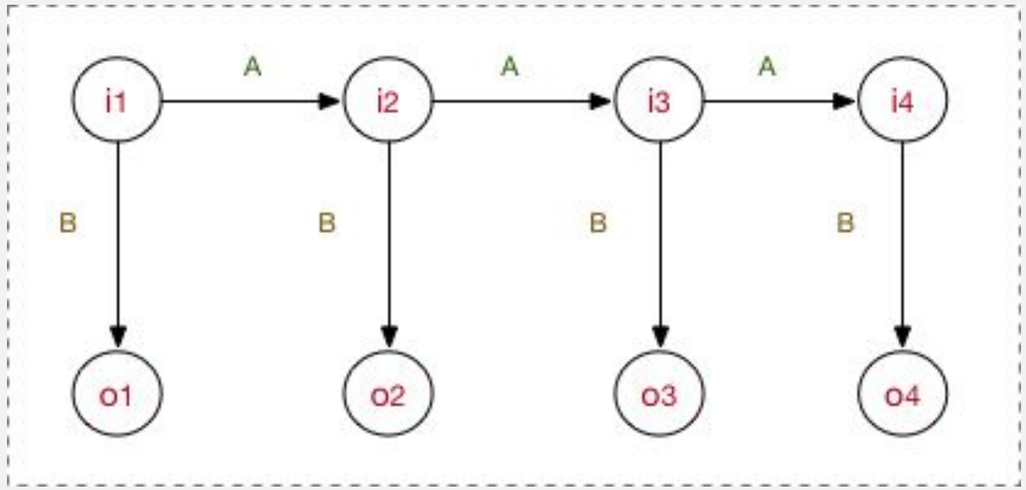
对于输入x，类别标签y：

产生式模型估计它们的联合概率分布P(x,y)
判别式模型估计条件概率分布P(y|x)

判别式模型常见的主要有：
- Logistic Regression
- SVM
- CRF
- Linear Regression

生成式模型常见的主要有：
- Naive Bayes
- HMMs
- Markov Random Fields

3.2、HMM隐马尔可夫模型

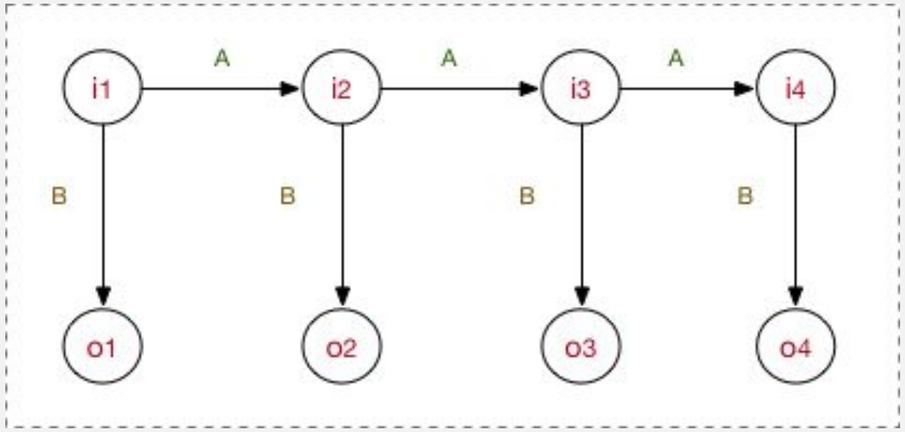


- 小明所在城市的天气有{晴天，阴天，雨天}三种情况，小明每天的活动有{宅，打球}两种选项。
- 作为小明的朋友，我们只知道他每天参与了什么活动，而不知道他所在城市的天气是什么样的。
- 这个城市每天的天气情况，会和前一天的天气情况有点关系。譬如说，如果前一天是晴天，那么后一天是晴天的概率，就大于后一天是雨天的概率。
- 小明所在的城市，一年四季的天气情况都差不多。
- 小明每天会根据当天的天气情况，决定今天进行什么样的活动。
- 我们想通过小明的活动，猜测他所在城市的天气情况。

那么，城市天气情况和小明的活动选择，就构成了一个隐马尔科夫模型HMM

3.2、HMM隐马尔可夫模型

- HMM的基本定义：**HMM是用于描述由隐藏的状态序列和显性的观测序列组合而成的双重随机过程**。在前例中，城市天气就是隐藏的状态序列，这个序列是我们观测不到的。小明的活动就是观测序列，这个序列是我们能够观测到的。这两个序列都是随机序列。
- HMM的假设一：**马尔可夫性假设**。当前时刻的状态值，仅依赖于前一时刻的状态值，而不依赖于更早时刻的状态值。每天的天气情况，会和前一天的天气情况有点关系。
- HMM的假设二：**齐次性假设**。状态转移概率矩阵与时间无关。即所有时刻共享同一个状态转移矩阵。小明所在的城市，一年四季的天气情况都差不多。
- HMM的假设三：**观测独立性假设**。当前时刻的观察值，仅依赖于当前时刻的状态值。小明每天会根据当天的天气情况，决定今天进行什么样的活动。
- HMM的应用目的：**通过可观测到的数据，预测不可观测到的数据**。我们想通过小明的活动，猜测他所在城市的天气情况。

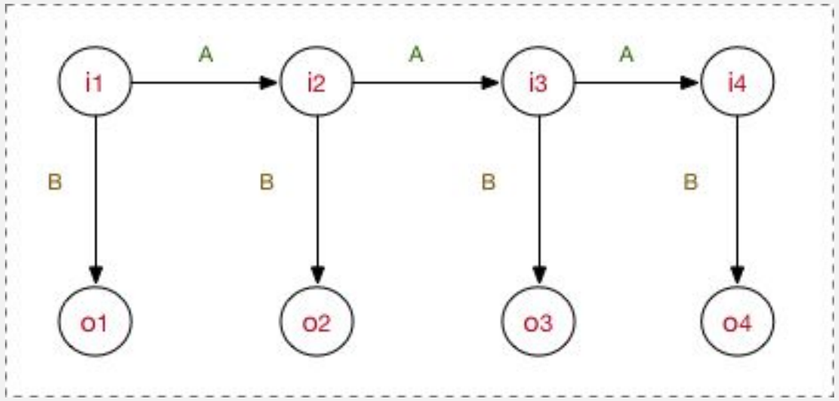


注一：马尔可夫性：随机过程中某事件的发生只取决于它的上一事件，是“无记忆”过程。
注二：HMM被广泛应用于标注任务。在标注任务中，状态值对应着标记，任务会给定观测序列，以预测其对应的标记序列。

他	和	爸爸	去	电影	院	看	哈利	波特
B-Per	O	B-Per	O	B-Loc	I-Loc	O	B-Per	I-Per

注三：HMM属于生成模型，是有向图。

3.2、HMM隐马尔可夫模型



三个基本问题

- 概率计算问题：给定模型参数和观测序列，计算该观测序列的概率。是后面两个问题的基础。
- 学习训练问题：给定观测序列，估计模型参数。
- **解码预测问题：给定模型参数和观测序列，求概率最大的状态序列。**

状态转移概率矩阵A：

	晴天	阴天	雨天
晴天	0.5	0.2	0.3
阴天	0.3	0.5	0.2
雨天	0.2	0.3	0.5

观测概率矩阵B：

	宅	打球
晴天	0.5	0.5
阴天	0.4	0.6
雨天	0.7	0.3

初始概率状态向量π：

晴天	0.2
阴天	0.4
雨天	0.4

3.2、HMM隐马尔可夫模型

约定一下HMM的标记符号，并通过套用上文的例子来理解：

- **状态值集合**（一共有N种状态值）： $\{s_1, s_2, \ldots\ldots, s_N\}$

  天气的状态值集合为{晴天，阴天，雨天}。

- **观测值集合**（一共有M种观测值）： $\{o_1, o_2, \ldots\ldots, o_M\}$

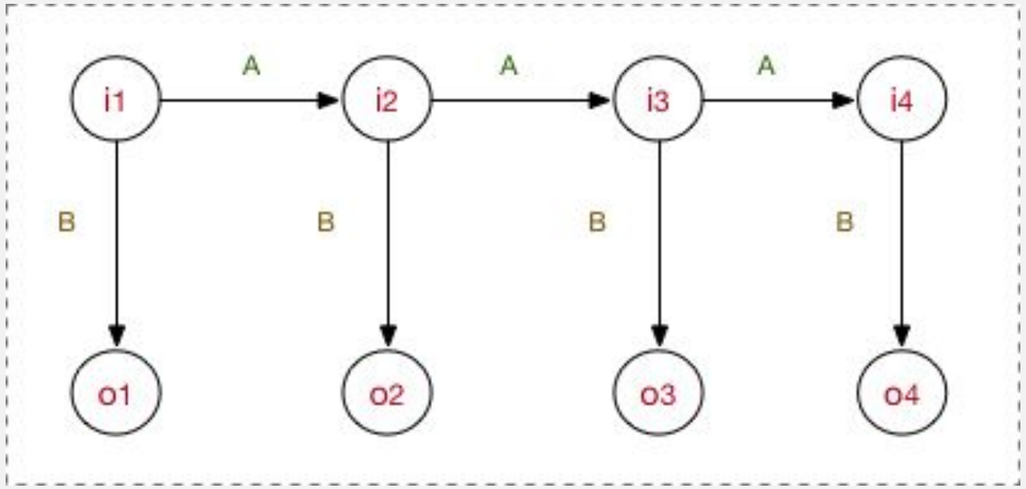
  小明活动的观测值集合为{宅，打球}。

- **状态值序列**： $y_1, y_2, \cdots, y_t, \cdots, y_T$

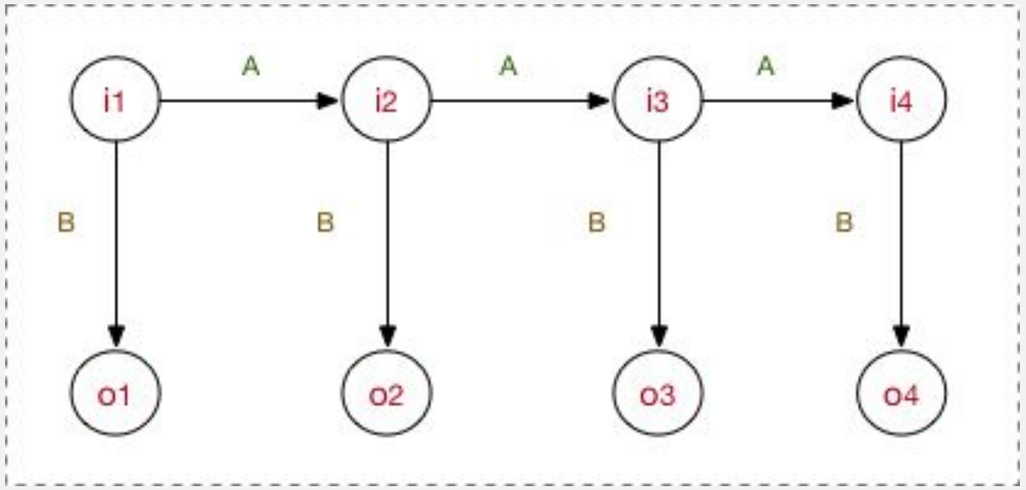
  每一天的城市天气状态值构成的序列{晴晴晴阴雨晴}

- **观测值序列**： $x_1, x_2, \ldots, x_t, \ldots, x_T$

  每一天的小明活动的观测值构成的序列{球宅宅球宅宅}



3.2、HMM隐马尔可夫模型



- HMM模型的三个参数：A，B，π。
- A：状态转移概率矩阵。表征转移概率，维度为N*N。
- B：观测概率矩阵。表征发射概率，维度为N*M。
- π：初始状态概率向量。维度为N*1。
- λ=(A,B,π)，表示模型的所有参数。

状态转移概率矩阵A：

	晴天	阴天	雨天
晴天	0.5	0.2	0.3
阴天	0.3	0.5	0.2
雨天	0.2	0.3	0.5

观测概率矩阵B：

	宅	打球
晴天	0.5	0.5
阴天	0.4	0.6
雨天	0.7	0.3

初始概率状态向量π：

晴天	0.2
阴天	0.4
雨天	0.4

3.2、HMM——概率计算问题

状态转移概率矩阵A：

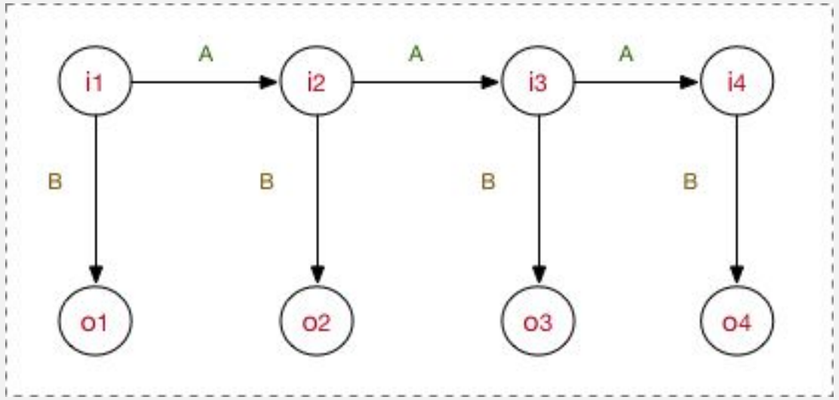
	晴天	阴天	雨天
晴天	0.5	0.2	0.3
阴天	0.3	0.5	0.2
雨天	0.2	0.3	0.5

观测概率矩阵B：

	宅	打球
晴天	0.5	0.5
阴天	0.4	0.6
雨天	0.7	0.3

初始概率状态向量π：

晴天	0.2
阴天	0.4
雨天	0.4



简单算一下，已知模型参数，观测序列是{宅球宅}的概率

天数	天气	行为	概率	累计概率
第一天	晴	宅	0.10	0.54
	阴	宅	0.16	
	雨	宅	0.28	
第二天	晴	打球	0.077	0.248
	阴	打球	0.1104	
	雨	打球	0.0606	
第三天	晴	宅	0.04187	0.13022
	阴	宅	0.03551	
	雨	宅	0.05284	

和N-Gram是什么关系?

3.2、（HMM——解码预测问题） && 维特比算法

状态转移概率矩阵A：

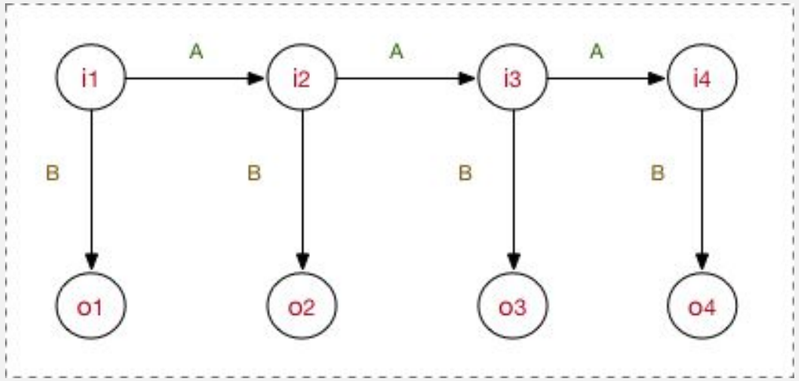
	晴天	阴天	雨天
晴天	0.5	0.2	0.3
阴天	0.3	0.5	0.2
雨天	0.2	0.3	0.5

观测概率矩阵B：

	宅	打球
晴天	0.5	0.5
阴天	0.4	0.6
雨天	0.7	0.3

初始概率状态向量π：

晴天	0.2
阴天	0.4
雨天	0.4



已知模型参数，观测序列是{宅球宅}，最有可能的天气状况?

实际就是用动态规划求解概率最大路径。这时一条路径对应着一个状态序列。

根据动态规划原理，最优路径必须满足这样的特性：如果最优路径在时刻t通过节点it*那么这一路径从it*到终点iT*的部分路径，对于从it*到iT*的所有可能来说，必须是最优的。因为假如不是这样，就会有一条新的最优路径。

所以只要递归的求在t时刻状态为i的各条路径的最大概率，直到时刻t=T，此时最大的概率就是最优路径的概率P，同时也得到最优路径的终点。从这个终点逐步反推，即可得到最优路径，这就是维特比算法。

3.2、（HMM——解码预测问题） && 维特比算法

状态转移概率矩阵A：

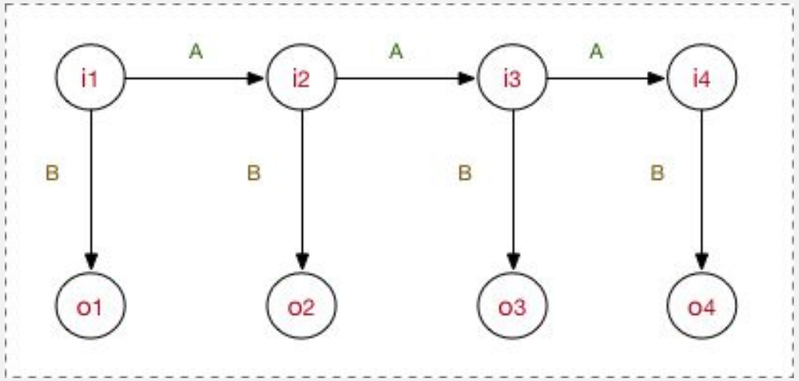
	晴天	阴天	雨天
晴天	0.5	0.2	0.3
阴天	0.3	0.5	0.2
雨天	0.2	0.3	0.5

观测概率矩阵B：

	宅	打球
晴天	0.5	0.5
阴天	0.4	0.6
雨天	0.7	0.3

初始概率状态向量π：

晴天	0.2
阴天	0.4
雨天	0.4



已知模型参数，观测序列是{宅球宅}，最有可能的天气状况?

- 定义在时刻t状态为i的所有单个路径（i1，i2，... it)中的概率最大值为：

$$\delta_t(s) = \max_{i_1,i_2,\ldots i_{t-1}} P(\, i_t = s, i_{t-1}, i_{t-2}, , , i_1, o_1, o_2, , , o_t \mid \lambda\,), \; s = 1,2, \ldots, N$$

那么t+1的时刻，就是：

$$\delta_{t+1}(s) = \max_{i_1,i_2,\ldots i_t} P(\, i_{t+1} = s, i_t, i_{t-1}, i_{t-2}, , , i_1, o_1, o_2, , , o_t, o_{t+1} \mid \lambda\,), \; s = 1,2, \ldots, N$$

$$= \max_{1 \le j \le N} [\, \delta_t(\, j\,)\, a_{js}\,]\, b_i(\, o_{t+1}\,)$$

再定义一个变量，用来回溯最大路径：在时刻t状态i的所有单个路径（i1，i2，，，it-1,it)中，概率最大的路径第t-1个节点为：

$$\varphi_t(\, i\,) = \arg \max_{1 \le j \le N} [\, \delta_{t-1}(\, j\,)\, a_{ji}\,], \; i = 1,2, , , N$$

3.2、（HMM——解码预测问题） && 维特比算法

状态转移概率矩阵A：

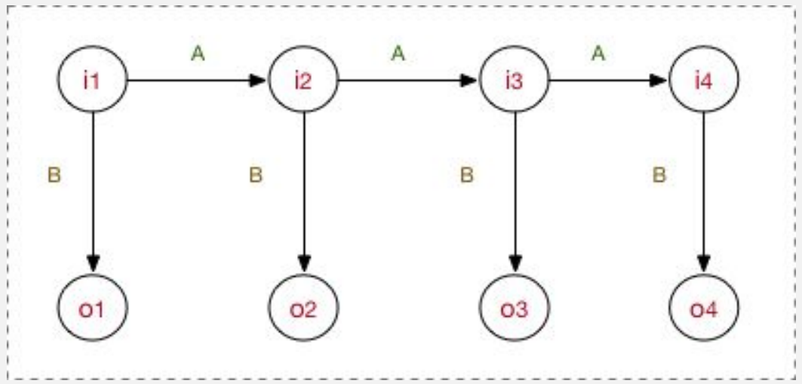
	晴天	阴天	雨天
晴天	0.5	0.2	0.3
阴天	0.3	0.5	0.2
雨天	0.2	0.3	0.5

观测概率矩阵B：

	宅	打球
晴天	0.5	0.5
阴天	0.4	0.6
雨天	0.7	0.3

初始概率状态向量π：

晴天	0.2
阴天	0.4
雨天	0.4



已知模型参数，观测序列是{宅球宅}，最有可能的天气状况？

维特比示例	第一天 宅	第二天 打球	第三天 宅
雨天	0.28 ←	0.042 ←	0.0147
阴天	0.16	0.0504 ←	0.01008
晴天	0.1	0.028	0.00756

对应到NLP情景中：
已知HMM模型参数——已知语料集
已知观测序列——“已知的句子”
最有可能的内部状态？——最有可能的词性序列？

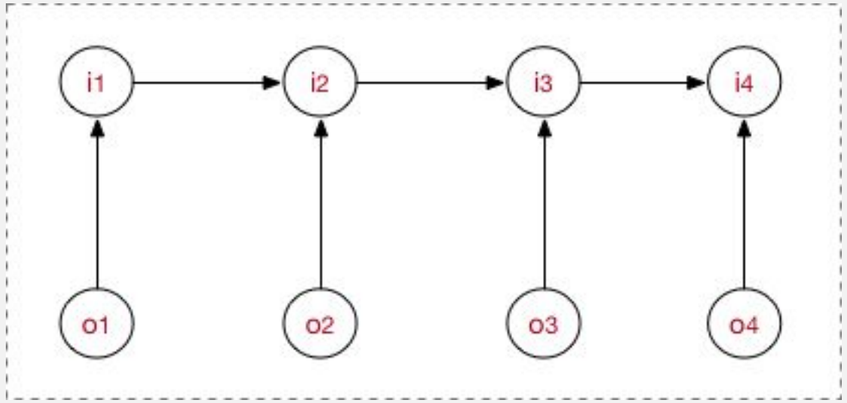
3.2、HMM的缺陷

- 马尔可夫性（有限历史性）：实际上在NLP领域的文本数据，很多词语都是长依赖的。
- 齐次性：序列不同位置的状态转移矩阵可能会有所变化，即位置信息会影响预测结果。
- 观测独立性：观测值和观测值（字与字）之间是有相关性的。
- 单向图：只与前序状态有关，和后续状态无关。在NLP任务中，上下文的信息都是必须的。
- 标记偏置LabelBias：若状态A能够向N种状态转移，状态B能够向M种状态转移。若N<<M，则预测序列更有可能选择状态A，因为A的转移概率较大。

苹果	，	对于	我	来说	并不	好吃	。
Fruit							
苹果	，	是	乔布斯	创建	的	。	
Company							
好用吗	，	这个	苹果	？			
			Product				
好吃吗	，	这个	苹果	？			
			Fruit				

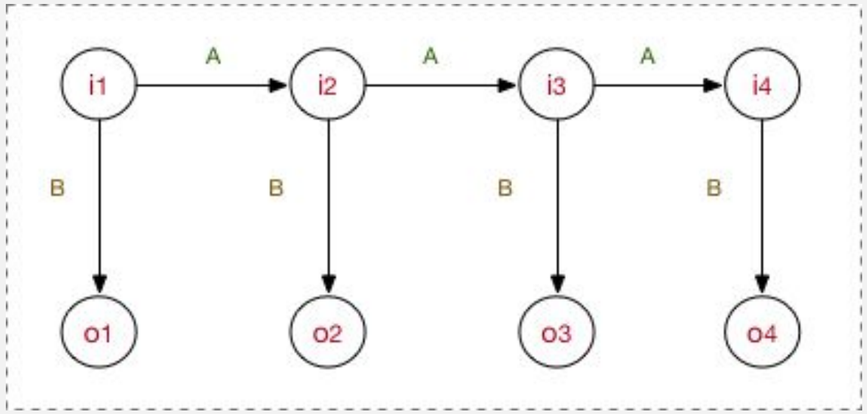
3.2、MEMM最大熵马尔可夫模型

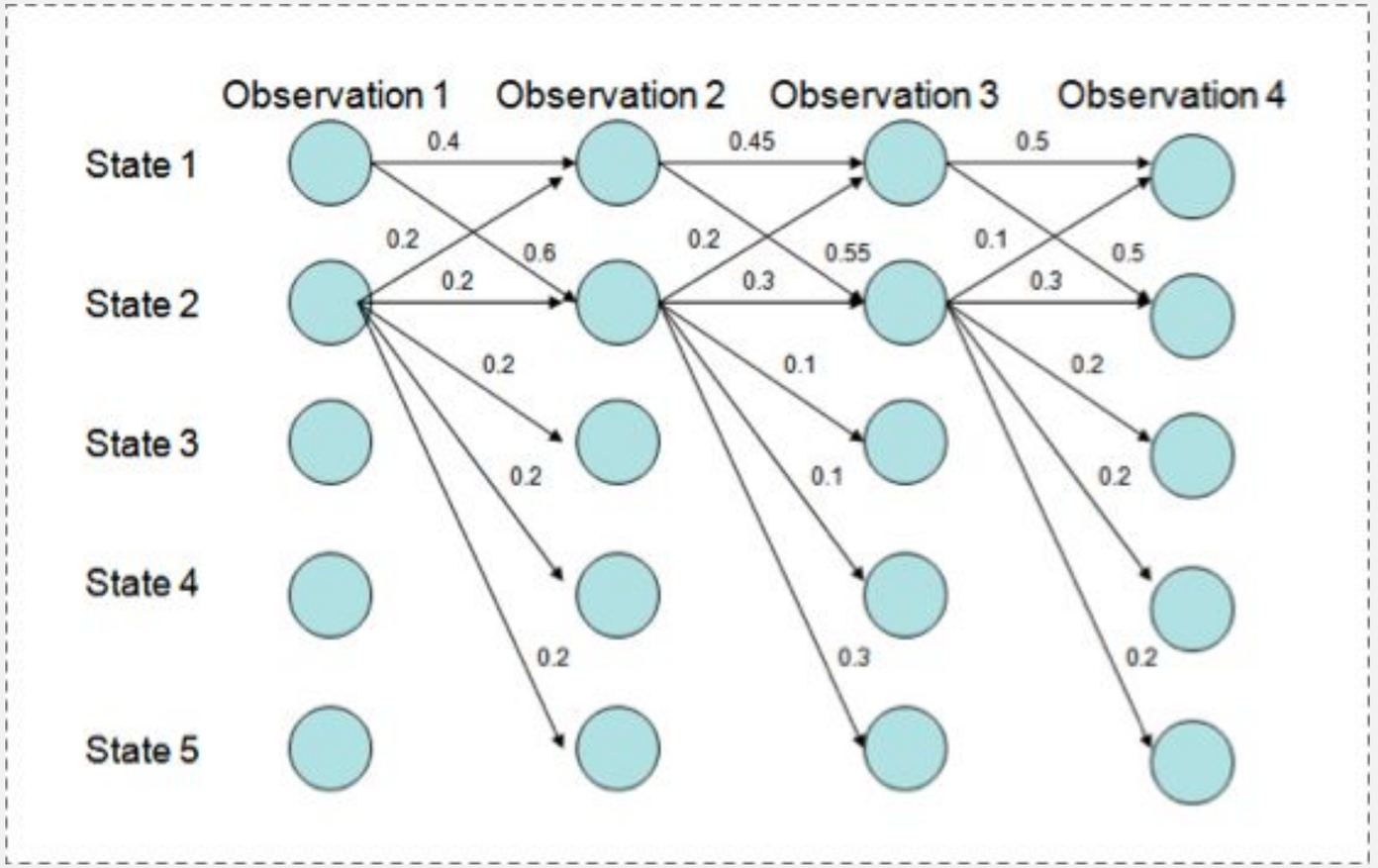
最大熵马尔可夫模型



$$P(i_i|i_{i-1},o_i)(i=1,\cdots,n)$$

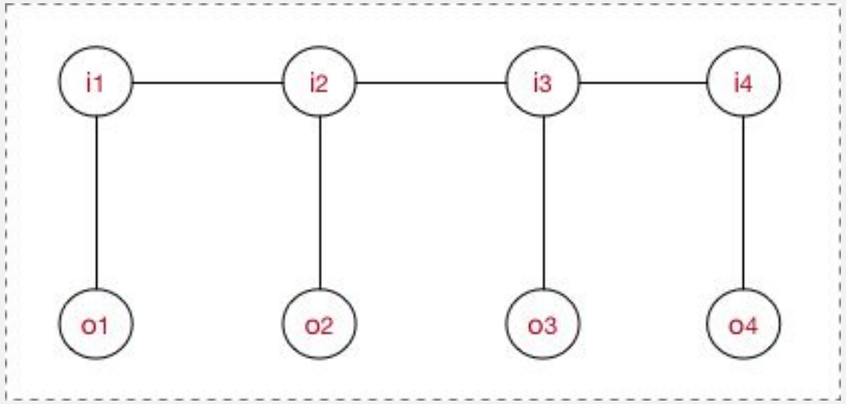
隐马尔可夫模型

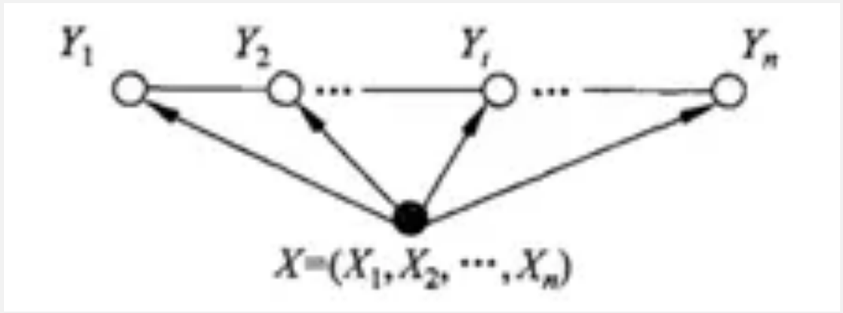


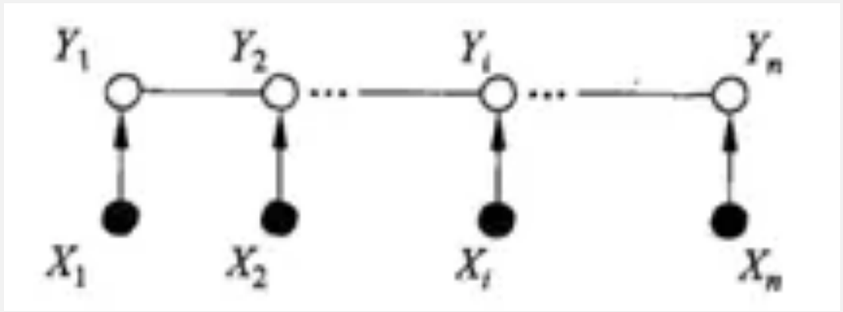


- P(1-> 1-> 1-> 1)= 0.4 x 0.45 x 0.5 = 0.09 ,
- P(2->2->2->2)= 0.2 X 0.3 X 0.3 = 0.018,
- P(1->2->1->2)= 0.6 X 0.2 X 0.5 = 0.06,
- P(1->1->2->2)= 0.4 X 0.55 X 0.3 = 0.066

3.2、CRF条件随机场

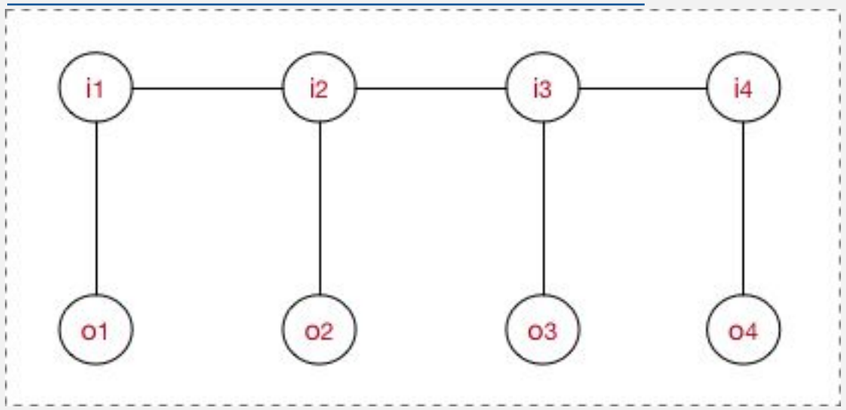


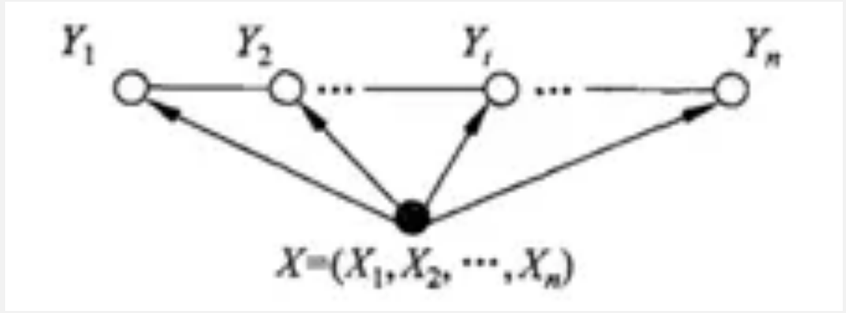


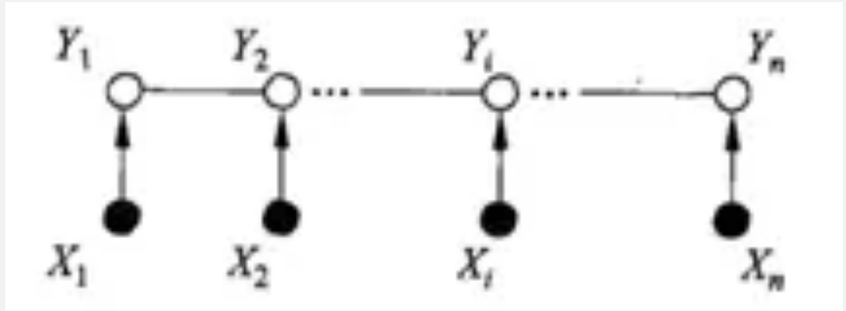


- 随机场是一个图模型，是由若干个结点（随机变量）和边（依赖关系）组成的图模型，当给每一个结点按照某种分布随机赋予一个值之后，其全体就叫做随机场。
- 马尔可夫随机场是随机场的特例，它假设随机场中任意一个结点的赋值，仅仅和它的邻结点的取值有关，和不相邻的结点的取值无关。用学术语言表示是：满足成对、局部或全局马尔科夫性。
- 条件随机场CRF是马尔可夫随机场的特例，它假设模型中只有X（输入变量，观测值）和Y（输出变量，状态值）两种变量。输出变量Y构成马尔可夫随机场，输入变量X不具有马尔可夫性。
- 线性链条件随机场，是状态序列是线性链的条件随机场。

3.2、CRF条件随机场

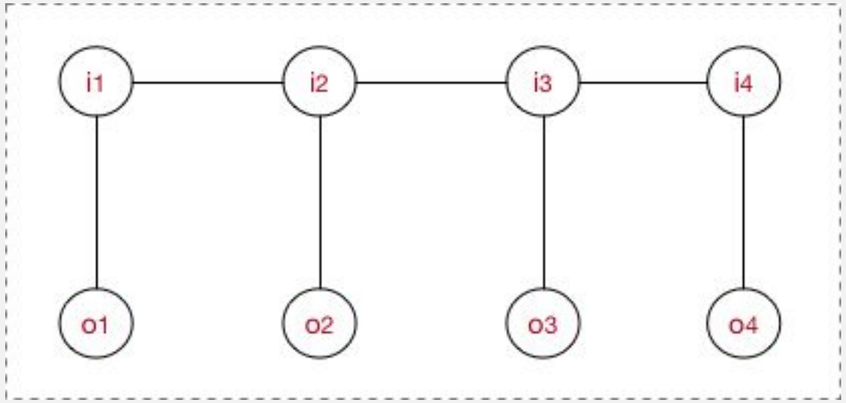


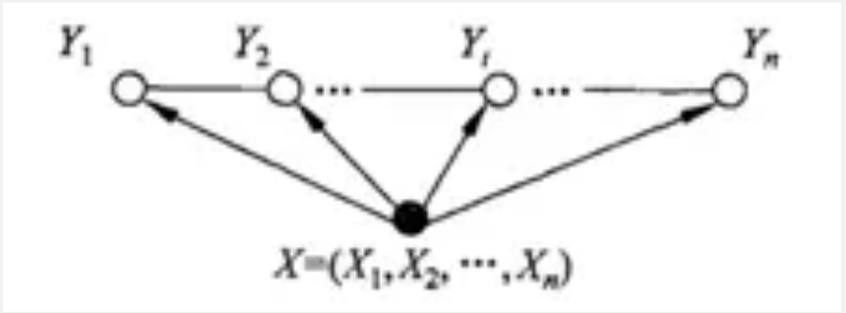


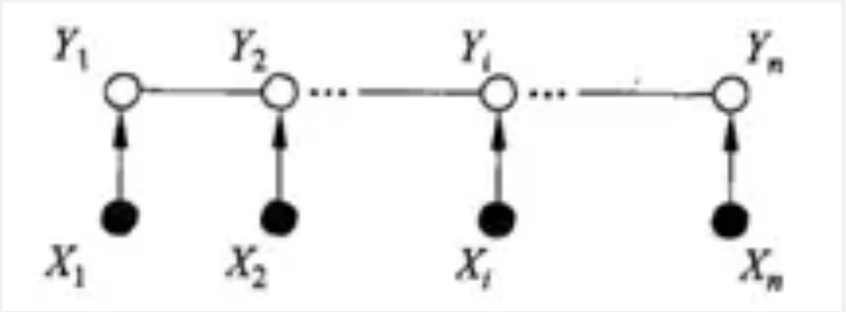


- 线性链条件随机场有以下性质：
  - 对于状态序列y，y的值只与相邻的y有关系，体现马尔可夫性。
  - 任意位置的y与所有位置的x都有关系。
  - 我们研究的线性链条件随机场，假设状态序列Y和观测序列X有相同的结构，但是实际上后文公式的推导，对于状态序列Y和观测序列X结构不同的条件随机场也适用。
  - 观测序列X是作为一个整体，去影响状态序列Y的值，而不是只影响相同或邻近位置（时刻）的Y。
  - 线性链条件随机场的示意图如左：

3.2、CRF条件随机场





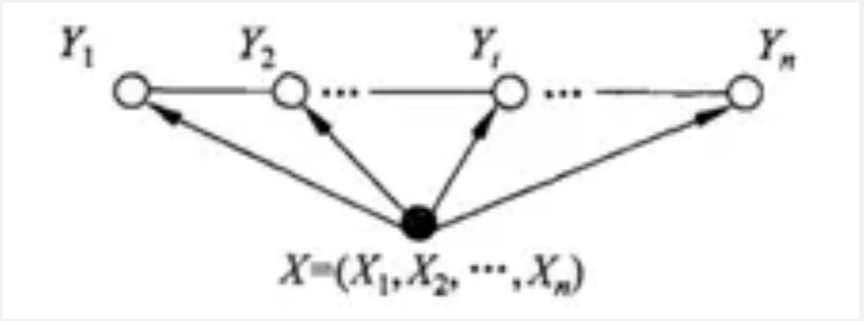


设X=(X1,X2,...Xn),Y=(Y1,Y2,...Yn)均为线性链表示的随机变量序列，在给定随机变量序列X的情况下，随机变量Y的条件概率分布P(Y|X)构成条件随机场，即满足马尔可夫性：

$$P(Yi|\ X,Y1,Y2,...Yn)\ =\ P(Yi|\ X,Yi-1,Yi+1)$$

则称P(Y|X)为线性链条件随机场。

3.2、CRF条件随机场公式推“倒”



在linear-CRF中，特征函数分为两类。
第一类是定义在Y节点上的节点特征函数，这类特征函数只和当前节点有关，记为：

$$s_l(y_i, x, i), \quad l = 1, 2, \ldots L$$

第二类是定义在Y上下文的局部特征函数，这类特征函数只和当前节点和上一个节点有关，记为：

$$t_k(y_{i-1}, y_i, x, i), \quad k = 1, 2, \ldots K$$

无论是节点特征函数还是局部特征函数，它们的取值只能是0或者1。即满足特征条件或者不满足特征条件。同时，我们可以为每个特征函数赋予一个权值，用以表达我们对这个特征函数的信任度。假设tk的权重系数是λk,sl的权重系数是μl,则linear-CRF由我们所有的tk,λk,sl,μl共同决定。此时我们得到了linear-CRF的参数化形式如下：

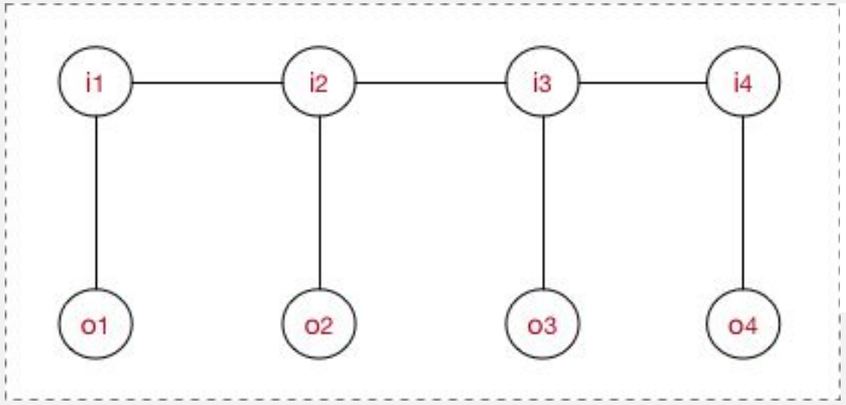
$$P(y|x) = \frac{1}{Z(x)} exp\Big(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\Big)$$

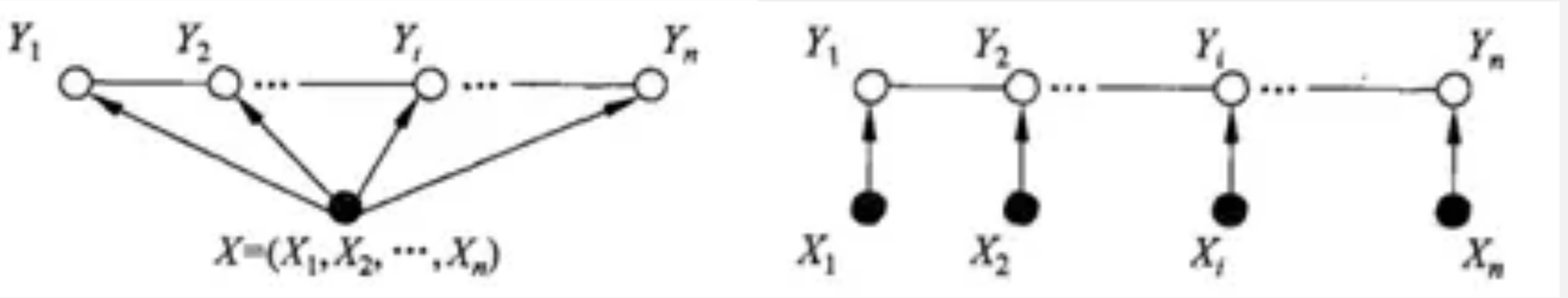
其中，Z(x)为规范化因子：

$$Z(x) = \sum_{y} exp\Big(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\Big)$$

- 参考文档：https://www.cnblogs.com/pinard/p/7048333.html

3.2、CRF条件随机场公式推“倒”





$$P(Y) = \frac{1}{Z(x)} \prod_c \psi_c(Y_c)$$

$$P(y|x) = \frac{1}{Z(x)} exp\Big(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i) \Big)$$

$$P(I|O) = \frac{1}{Z(O)} e^{\sum_i^T \sum_k^M \lambda_k f_k(O, I_{i-1}, I_i, i)} = \frac{1}{Z(O)} e^{[\sum_i^T \sum_j^J \lambda_j t_j(O, I_{i-1}, I_i, i) + \sum_i^T \sum_l^L \mu_l s_l(O, I_i, i)]}$$

3.2、CRF条件随机场

一个linear-CRF用于词性标注的实例，为了方便，我们简化了词性的种类。假设输入的都是三个词的句子，即X=(X1,X2,X3),输出的词性标记为Y=(Y1,Y2,Y3),其中Y∈{1(名词)，2(动词)}

这里只标记出取值为1的特征函数如下：

t1 = t1(yi−1=1,yi=2,x,i), i=2,3, λ1=1

t2 = t2(y1=1,y2=1,x,2) λ2=0.5

t3 = t3(y2=2,y3=1,x,3) λ3=1

t4 = t4(y1=2,y2=1,x,2) λ4=1

t5 = t5(y2=2,y3=2,x,3) λ5=0.2

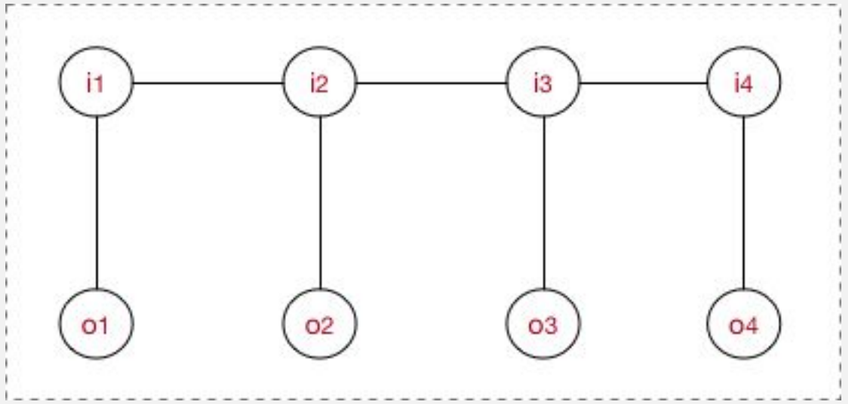
s1 = s1(y1=1,x,1) μ1=1

s2 = s2(yi=2,x,i), i=1,2, μ2=0.5

s3 = s3(yi=1,x,i), i=2,3, μ3=0.8

s4 = s4(y3=2,x,3) μ4=0.5

求标记(1,2,2)的非规范化概率。



$$P(y|x) \propto exp\Big[\sum_{k=1}^{5}\lambda_k \sum_{i=2}^{3} t_k(y_{i-1}, y_i, x, i) + \sum_{l=1}^{4}\mu_l \sum_{i=1}^{3} s_l(y_i, x, i)\Big]$$

$$P(y_1 = 1, y_2 = 2, y_3 = 2|x) \propto exp(3.2)$$

3.2、CRF条件随机场的优缺点

CRF相对于HMM的优点

（1）规避了马尔可夫性（有限历史性），能够获取长文本的远距离依赖的信息。

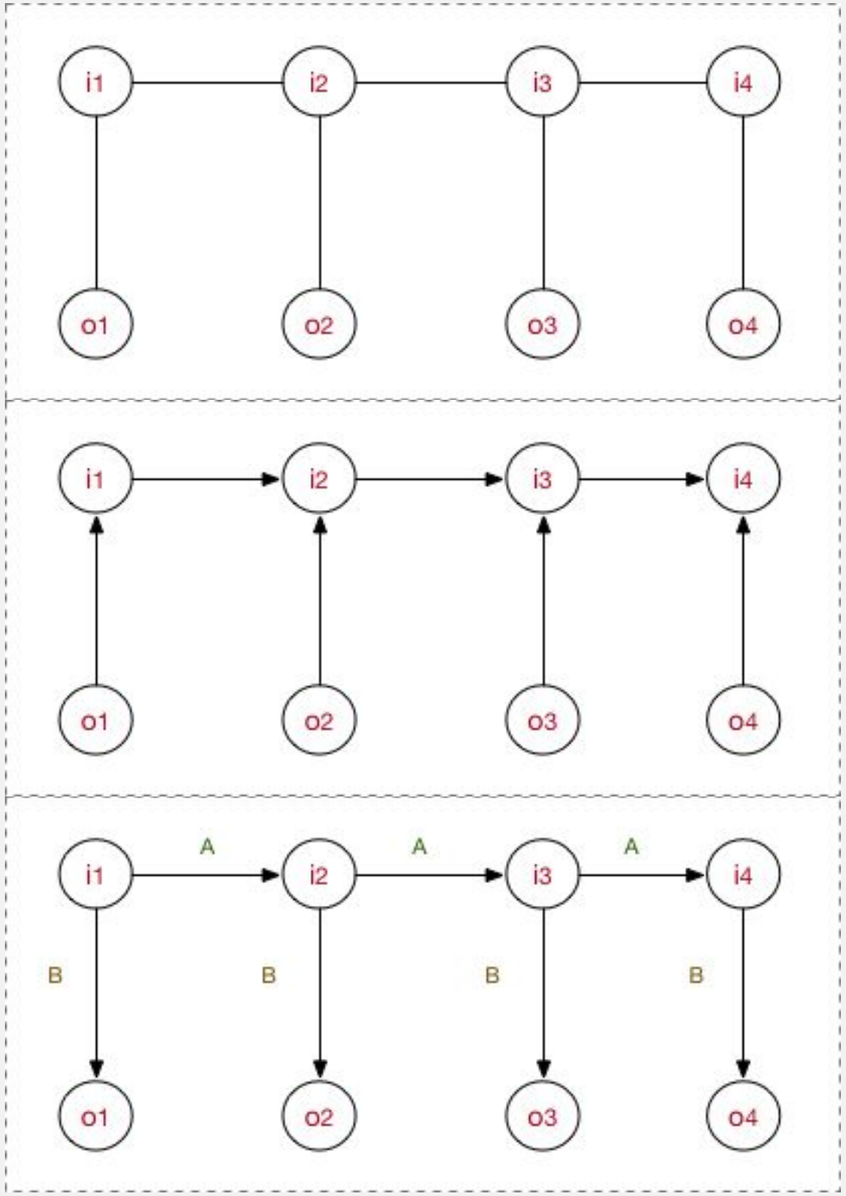
（2）规避了齐次性，模型能够获取序列的位置信息，并且序列的位置信息会影响预测出的状态序列。

（3）规避了观测独立性，观测值之间的相关性信息能够被提取。

（4）不是单向图，而是无向图，能够充分提取上下文信息作为特征。

（5）改善了标记偏置LabelBias问题，因为CRF相对于HMM能够更多地获取序列的全局概率信息。

（6）CRF的思路是利用多个特征，对状态序列进行预测。HMM的表现形式使他无法使用多个复杂特征。



3.2、CRF条件随机场的优缺点

条件随机场CRF的缺点

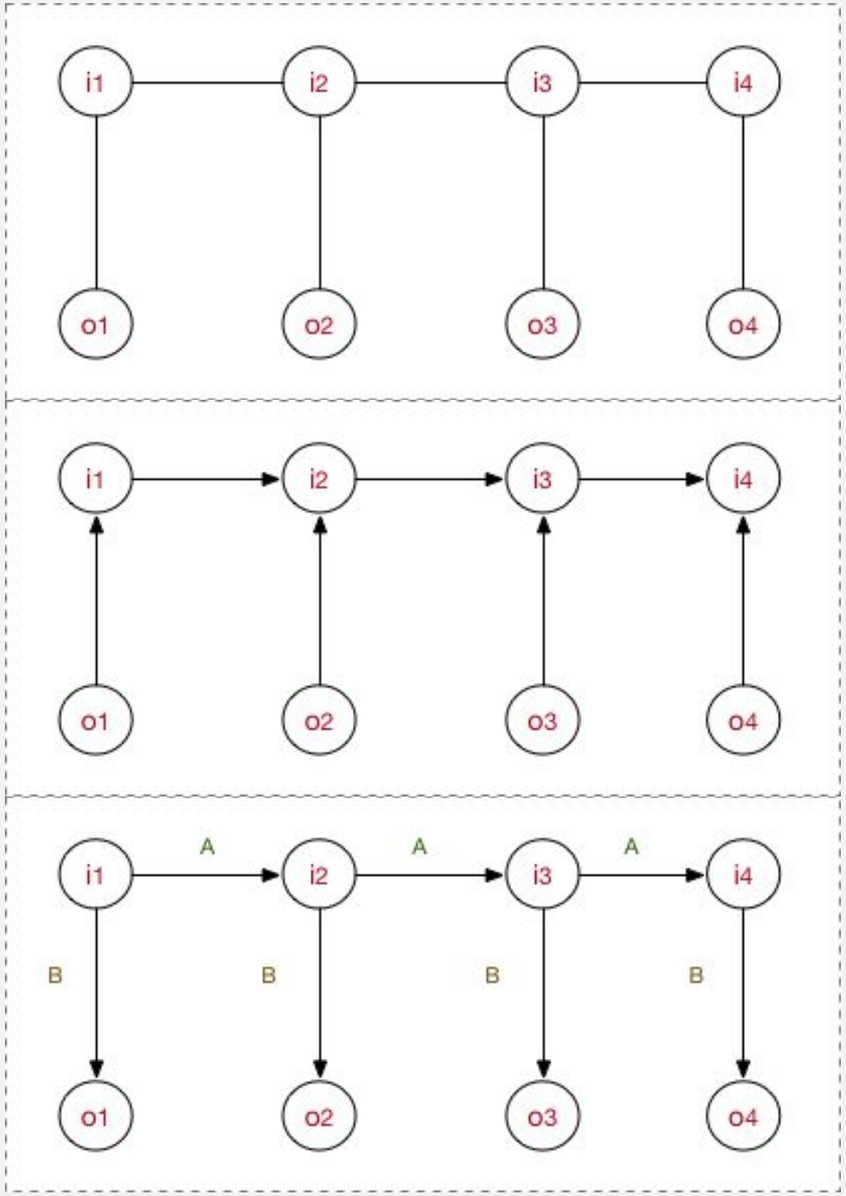
（1）CRF训练代价大、复杂度高。

（2）每个特征的权重固定，特征函数只有0和1两个取值。

（3）模型过于复杂，在海量数据的情况下，业界多用神经网络。

（4）需要人为构造特征函数，特征工程对CRF模型的影响很大。

（5）转移特征函数的自变量只涉及两个相邻位置，而CRF定义中的马尔可夫性，应该涉及三个相邻位置。



3.2、概率图模型课后复习问题清单

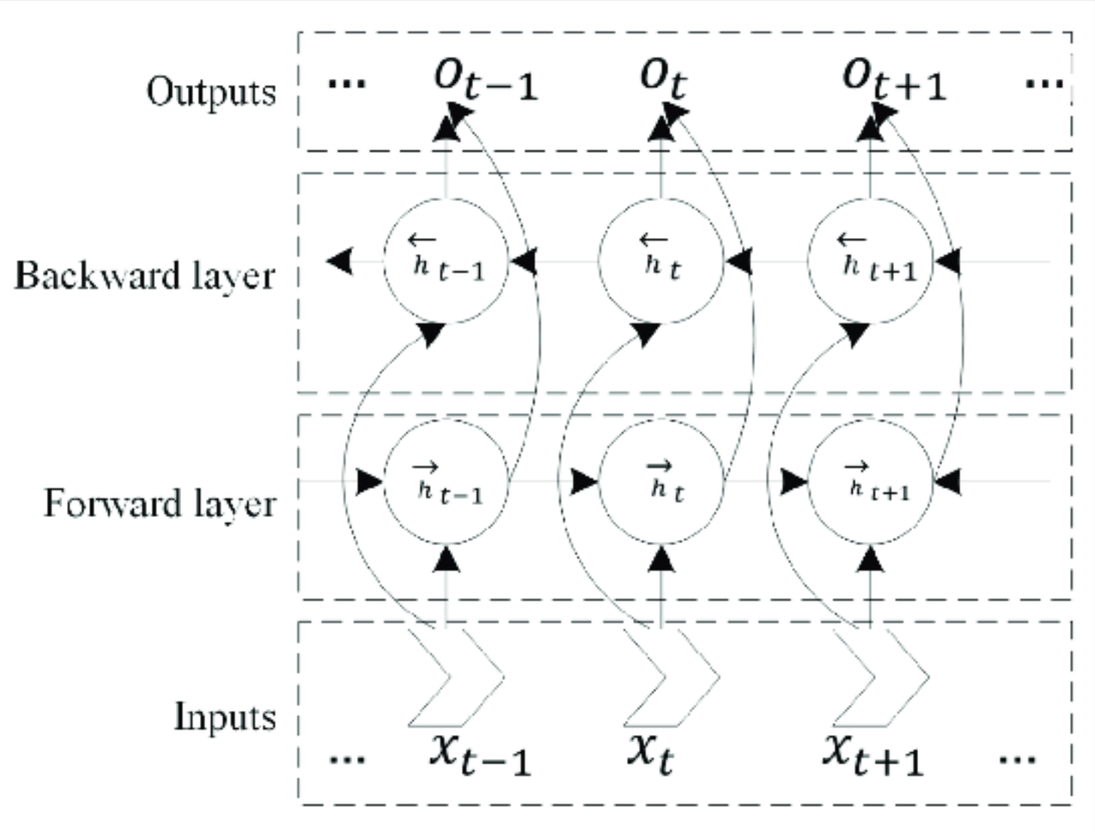
- 有向图和无向图的P（Y）分别如何计算?
- HMM，HEMM，CRF的定义是什么？两两之间的区别是什么？优缺点是什么?
- 前向算法，后向算法如何计算?
- 标注偏置的原因是什么？如何解决?
- 维特比算法如何求解？维特比算法的DP公式如何写?
- EM算法是什么？如何求HMM参数?

3.3、序列标注实验

准备动手

3.3、序列标注—深度模型

LSTM,Transformer,BERT

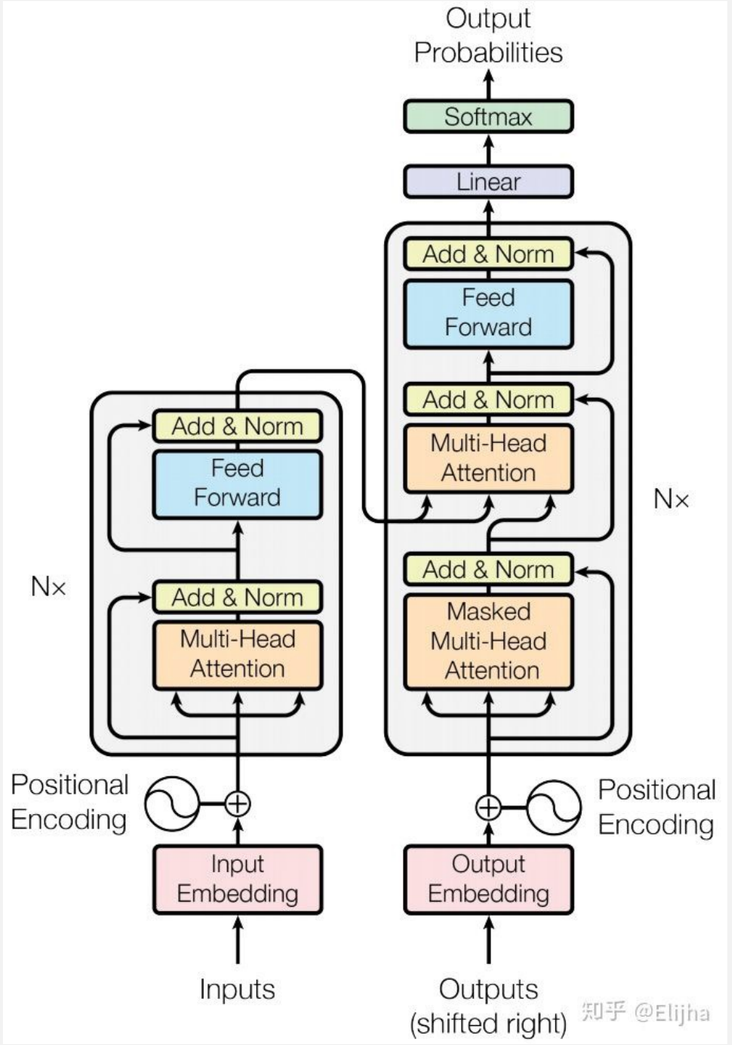


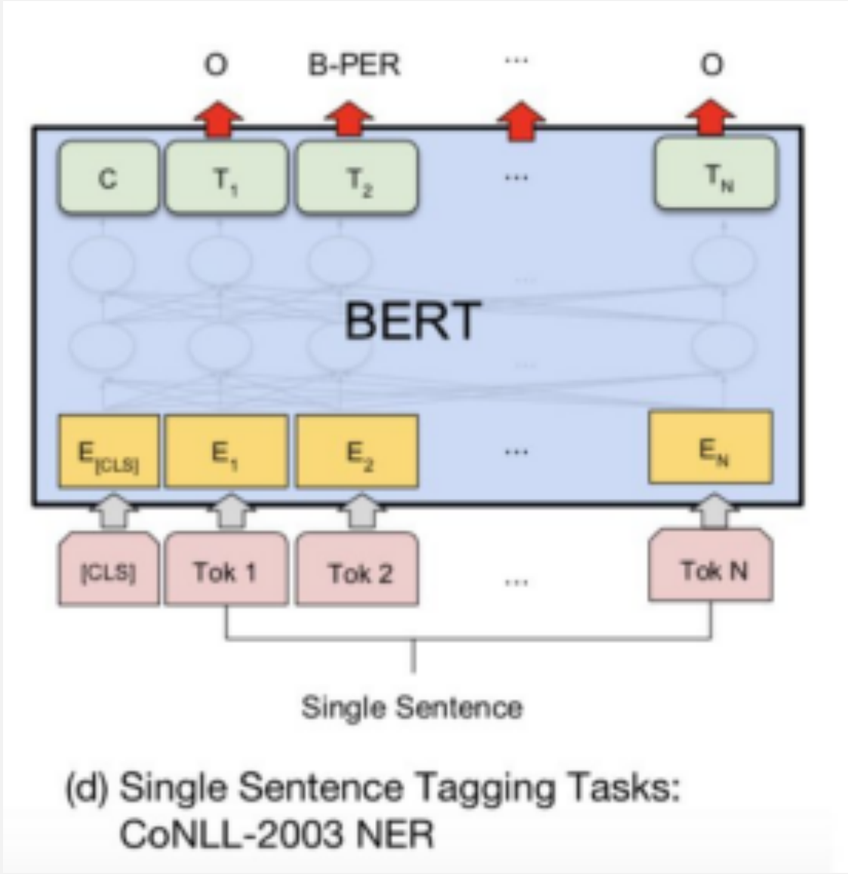
- LSTM已经可以胜任序列标注问题，为每个token预测一个label（LSTM后面接:分类器）；而CRF也是一样的，为每个token预测一个label。
- 但是他们的预测机理是不同的。CRF是全局范围内统计归一化的条件状态转移概率矩阵，再预测出一条指定的sample的每个token的label；LSTM（RNNs，不区分here）是依靠神经网络的超强非线性拟合能力，在训练时将samples通过复杂的高阶高纬度异度空间的非线性变换，学习出一个模型，然后再预测出一条指定的sample的每个token的label。

3.3、序列标注—深度模型

- input: "学习出一个模型，然后再预测出一条指定"
- expected output: 学/B 习/E 出/S 一/B 个/E 模/B 型/E ，/S 然/B 后/E 再/E 预/B 测/E ……
- real output: 学/B 习/E 出/S 一/B 个/B 模/B 型/E ，/S 然/B 后/B 再/E 预/B 测/E ……

3.3、序列标注—深度模型


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)
知乎 @Elijha


O
B-PER
...
O
C
T1
T2
...
TN
BERT
E[CLS]
E1
E2
...
EN
[CLS]
Tok 1
Tok 2
...
Tok N
Single Sentence
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

谢 谢 大 家

Multipurpose Presentation