南京農業大學

NANJING AGRICULTURAL UNIVERSITY

数字人文下的命名实体识别

目录
CONTENT

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

命名实体识别
概念与基本原理

1、命名实体

◆实体识别所指的实体，是文本中的实体，是对广义实体的指称（Mention）。

◆指称的方式一般有三种

- **命名性指称（Name Mention）：通过名字来指称实体，比如“乔布斯”。**
- 名词性指称（Nominal Mention）：通过名词或名词性短语来表示实体，比如“苹果公司CEO”。
- 代词性指称（Pronoun Mentions）：通过代词来指代实体，比如“大家都很想念他”中的“他”。

誠 樸 勤 仁

1、命名实体

◆命名实体是对实体的命名性指称，是一种专指性词项。

◆命名实体具有五种特性

- **指称性**：用来指示或称说某些事物，以便将这些事物跟其他事物区分开来。不是所有的词语都有指称性，例如形容词表示事物的性质，动词表示动作或行为。代词、名词通常都有指称性。
- **专门性**：专门用来指示或称说某一个事物，以便将这个事物跟同类的其他事物区分开来。例如，“教授”、“年轻的教授”都是对一类人的指称，而“李教授”则是对某一个姓李的教授的指称。（注意，“李教授”绝不是对所有姓李的教授的指称）。

誠樸篤仁

1、命名实体

- **词汇性：**命名实体属于词汇，词汇成员包括词和固定词组。组织名通常是固定词组，固定词组中一般不含虚词。凭句法手段构造的自由词组也可用来指称某个个体，例如，“这粒沙子”。这些自由词组不属于词汇，当然也不是命名实体。
- **开放性：**命名实体是词汇中最直接反映客观世界变化的部分。新事物不断产生，而且往往对我们特别有重要性，需要命名，所以命名实体的数量往往非常庞大，而且层出不穷，难以胜数。
- **可替换性：**每一类（或每一小类）中的命名实体之间是可以替换的。替换之后语法上、语义上仍然是成立的，尽管可能不符合事实。

誠樸篤仁

1、命名实体

◆人名

- 中国人名：马云、刘强东、张近东
- 外国人名译名：阿兰·图灵、冯·诺依曼

◆地名

- 中国地名：北京、上海、广州
- 外国地名译名：纽约、伦敦、巴黎

◆机构名

- 国务院、教育局、南京农业大学

◆商标、品牌名

- 苹果、华为、金拱门

誠樸篤仁

2、命名实体识别

- ◆命名实体识别（Named Entity Recognition，NER）是指通过设计相应的算法以序列和分类的思路实现对命名实体的识别。
- ◆其目的是识别出文本中表示命名实体的成分，并对其进行分类，因此有时也称为命名实体识别和分类（Named Entity Recognition and Classification，NERC）
- ◆命名实体识别任务早期源于信息抽取和信息检索的需求，因为命名实体经常成为检索关键词，并且是事件和关系中的重要结构项。后来逐渐成为自然语言处理中一项独立的重要任务。
- ◆未登录词识别是命名实体识别中的难点，而做好命名实体识别，也有助于提高未登录词识别的准确率和召回率。

誠樸勤仁

3、序列化标注

◆自然语言处理中常使用序列化标注的方法来完成命名实体识别任务。

◆在序列化标注中，首先将文本表示成词语或汉字的序列，然后使用机器学习模型对该序列中的每个词语或汉字进行分类。

◆序列化标注的类别有BIOES、BIO等模式。在BIOES类别模式下，B表示命名实体的开头，I表示命名实体内部，O表示命名实体之外的文本，E表示命名实体的结尾，S表示单独的命名实体。通过这样的类别模式完成了词语或汉字序列成分的分类之后，就相当于完成了命名实体识别的任务。

誠樸勤仁

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

命名实体识别评测会议

命名实体识别评测

- MUC-6（1995）、MUC-7（1998）
  - 华尔街日报语料（规模较小）
  - 规则方法
    - 字符序列规则：前后提示词、上下文语境
    - 词汇规则：词形、词性
    - 短语规则
  - F1值96.42%（MUC-6）、93.39%（MUC-7）
- MET（1996）、MET-2（1998）
  - 增加了中文
  - F1值84.51%（MET）、86%（MET-2）
- 出现了**ME和HMM**的初步尝试

誠樸勤仁

命名实体识别评测

- ◆CoNLL由ACL下的SIGNLL举办，1999年开始有Shared Task 评测
- ◆CoNLL-2002和CoNLL-2003主要任务
  - ➢ 独立于语言的命名实体识别（Language-Independent Named Entity Recognition）
- ◆语言
  - ➢ 西班牙语、荷兰语（CoNLL-2002）
  - ➢ 英语、德语（CoNLL-2003）

誠樸勤仁

命名实体识别评测

◆CoNLL的主要方法（机器学习为主）

- 隐马尔可夫HMM
- 最大熵ME
- 支持向量机SVM
- 条件随机场CRF
- AdaBoost
- 长短记忆网络LSTM

◆评测结果

- CoNLL-2002：AdaBoost.MH，F1值 81.39%和77.05%
- CoNLL-2003：最大熵ME，F1值88.76%和72.41%。

誠樸篤仁

中文命名实体识别评测

◆863评测（2003-2005）

- 汉语NER作为自动分词的子任务
- 总体F1值82.38%
- 地名0.83，人名0.86，机构名0.61，日期0.85，时间0.85，数字0.93

◆SIGHAN Bakeoff（2006、2007）

- 从未登录词到命名实体识别
- 机器学习方法：CRF模型、ME模型
  - 未登录词的比例会有很大影响
  - 机构名识别最难
  - 训练语料之外增加外在数据会有很大影响

◆SIGHAN Bakeoff-2010：命名实体消歧

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

命名实体识别的方法

基于机器学习的命名实体识别

- NER研究中，机器学习方法成为主流。
- 三种思路
  - 新的模型、改进模型
    - 层叠马尔科夫
    - 多层条件随机场
  - 选择合适的特征
  - 多种方法的综合
    - 混合多个SVM
    - 混合HMM和ME
    - 统计和规则相结合

选择合适的特征

◆CRF、最大熵等机器学习模型完成序列化标注任务时，可以根据需要构建标注任务相关的特征函数，以提高标注任务的性能。

◆在命名实体识别任务中，特征函数的构建主要依据命名实体相关的语言知识，一般我们将这类知识叫做特征。

◆命名实体识别利用的特征可以分为全局特征和局部特征。

誠樸篤仁

选择合适的特征

◆全局特征

- 命名实体（例如人名）常常有连续出现的情况，如果其中某个已经被识别为命名实体，利用搭配约束可提高识别其余命名实体的效果。
- 一个命名实体往往在初次出现时具有较丰富的上下文特征，以后出现时则不一定总带着这些特征。利用篇章约束可以提高其后续出现的识别效果。

誠樸篤仁

选择合适的特征

◆局部特征

- **构成特征**（命名实体内部）：在某种基元（词或字符）的序列中，命名实体为一子序列，充当这个子序列的各个基元及其属性。
- **上下文特征**（命名实体之外）：在某种基元（词或字符）的序列中，命名实体为一子序列，这个子序列之前或之后的基元及其属性。
- **语序特征**：带有位置信息的特征。例如当前词的前一个词（-1）的属性，后一个词（+1）的属性。
- **结构特征**：这里“结构”不限于语法结构，可宽泛理解为几个特征的同现或复合。例如前两个词（-2和-1）的复合属性，后两个词（+1和+2）的复合属性。

誠樸勤仁

选择合适的特征

◆以词为基的识别特征

- 在一个词语序列中，命名实体的构成特征和上下文特征。已分词的中文文本宜采用这些识别特征。

◆以字符为基的识别特征

- 在一个字符序列中，命名实体的构成特征和上下文特征。在未分词的中文文本中识别单词型中文命名实体，宜采用这些识别特征。

选择合适的特征

◆以词为基的识别特征

- 人名最显著的识别特征是称谓语，此外有些动词、名词、形容词与人的相关性很强。
- 地名常出现在某些介词（如“至”、“从）或动词（如“坐镇”、“抵达”）之后，或出现在含地理实体通名的词语序列之前（如“等地”、“等岛屿”）
- 机构名除简称（如“安理会”）之外，通常由多个词语组成，应主要利用其构成特征。

中文机构名的构造模式

◆机构名是以机构通名为中心的偏正式复合词。

◆修饰部分一般只含名词（普名或专名）、序数词、形容词和动词，通常不含虚词。

◆构造模式：

- 地名＋团体＋序数词＋人名＋专造名＋产品/对象＋功能/方式/等级+学科/行业＋机构通名

◆机构通名必须出现。

◆地名、团体、人名、专造名，至少出现一个。

汉人姓名的构造模式和用字特征

◆完整形式：姓（1～2字）＋名（1～2字）。

◆不完整形式：小李、老王，陈总、袁某

◆姓氏：单姓、复姓（欧阳）、双姓（范徐）。

- 17万个姓名中，有729个姓氏，前365个姓氏的覆盖率达99%，分布相对集中，有利于识别姓名的左边界。

◆人名：17万个姓名中，有3345个用字，欲使覆盖率达99%，需取前1141个用字。

中国地名构造模式和用字特征

- 用字自由、分散，在17637个地名中，用汉字2595个；有相对集中的覆盖能力，前100字可覆盖50%，前900字可覆盖90%.
- 结构方式一般是专名＋通名。
- 大地名有完整资料可供参考，难点在小地名的识别。

译名的构造模式和用字特征

◆译名长度没有太多限制。

◆译名用字相对集中，首字、中字和尾字三个集合交叉不严重。

◆译名中一般不出现普通多字词，但会含某些专名或类似专名的多字词（170多个）：马克、西亚、印度、贝尔，福利、圣地（巧合）

◆资源：

- 《英语译名手册》38862条译名
- 《新英汉词典》 2400条译名

隐马尔可夫模型的定义

- ◆一个HMM模型　λ=(Q, V, A, B, π)
- ◆状态集合　Q = (q1, q2, . . . , qn)
- ◆观测集合　V = (v1, v2, . . . , vm)
- ◆隐马尔可夫模型的三组参数
  - ➢ 初始状态概率分布（π）
  - ➢ 状态转移概率分布（A）
  - ➢ 观测概率分布（B）

誠 樸 篤 仁

隐马尔可夫模型两个假设

◆齐次马尔科夫假设

- 隐藏序列中任意时刻状态只与前一时刻状态有关
- 与其他时刻状态和观测无关
- p(it|it-1,ot-1,...i1,o1) = p(it|it-1)

◆观测独立性假设

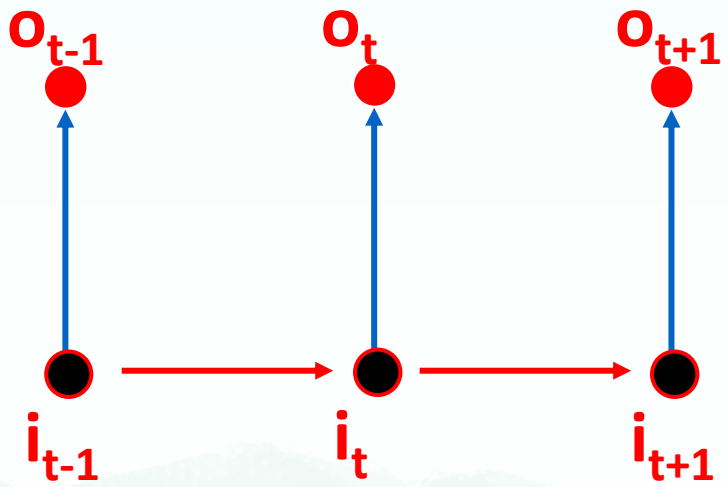
- 任意时刻的观测只依赖于该时刻的状态
- P(ot|iT,oT,iT-1,oT-1,...,it+1,ot+1,it,it-1,ot-1,...

i1,o1) = P(ot|it)

誠樸篤仁

隐马尔可夫模型两个假设

◆何谓“隐 Hidden”



誠 樸 篤 仁

隐马尔可夫模型的应用

◆1 概率计算问题：给定模型和观测序列，求观察序列出现的概率

◆2 学习问题：已知观测序列，估计模型参数，使得该模型下观测序列概率最大

◆3 预测问题（解码问题）：已知模型和观测序列，求最有可能对应的状态序列

誠 樸 篤力 仁

隐马尔可夫模型的应用

解码问题

$$I^* = \underset{I}{\operatorname{argmax}} \prod_{t=1}^{T} P(O_t|I_t)P(I_t|I_{t-1})$$

I_t表示状态，O_t表示观测

P(I_1):初始状态概率

P$(O_t|I_t)$:观测概率

P$(I_t|I_{t-1})$:状态转移概率

誠 樸 篤力 仁

隐马尔可夫模型用于命名实体识别

- 以字为元
- 字的类别有13种：
  - PER_S,PER_B,PER_E,PER_M
  - LOC_S,LOC_B,LOC_E,LOC_M
  - ORG_S,ORG_B,ORG_E,ORG_M
  - N
- 状态序列就是类别的序列I
  - 状态转移概率就是前后两个类别标记之间的条件概率
- 观测序列就是字的序列O
  - 观测概率就是从类别到字的条件概率

誠 樸 篤 仁

隐马尔可夫模型用于命名实体识别

◆ 何谓“隐 Hidden”

鄭	PER_SPER_B	PER_EPER_M	LOC_S ……
伯	PER_SPER_B	PER_EPER_M	LOC_S ……
克	PER_SPER_B	PER_EPER_M	LOC_S ……
段	PER_SPER_B	PER_EPER_M	LOC_S ……
于	PER_SPER_B	PER_EPER_M	LOC_S ……
鄢	PER_SPER_B	PER_EPER_M	LOC_S ……

观察序列

状态序列（隐藏）
Hidden

誠 樸 篤 仁

隐马尔可夫模型用于命名实体识别

$$I^* = \underset{I}{\operatorname{argmax}} \prod_{t=1}^{T} P(O_t|I_t)P(I_t|I_{t-1})$$

需要求解的参数

P(I_1):PER_S,PER_B...出现在句首的概率

P($I_t|I_{t-1}$):P(PER_S|PER_S),P(PER_B|PER_S)...

P ($O_t|I_t$) :P(郑|PER_S),P(郑|PER_B)...

求解

概率乘积最大

或转换为-log，费用之和最小

誠 樸 篤 仁

隐马尔可夫模型用于命名实体识别

$$I^* = \underset{I}{\operatorname{argmax}} \prod_{t=1}^{T} P(O_t|I_t)P(I_t|I_{t-1})$$

参数规模

初始概率向量π：13

状态转移矩阵A：13*13 = 169

观测矩阵B：6*13 = 78

直接计算量（鄭伯克段于鄢）

所有的状态序列数 AB = 136 = 4826809

计算量太大

誠樸勤仁

用Viterbi算法求解

- 思路：动态规划（全局最优、局部最优）
- 输入：模型λ=(A,B,π)，观测O=(o1,o2,...,oT)
- 输出：最优路径I=(i1,i2,...,iT)
- 定义1:在时刻t状态为i的所有单个路径$(i_1,i_2,...,i_t)$中，概率最大值为$\delta_t(i)$
  - $\delta_t(i)=\max_{i_1,i_2,\dots,i_{t-1}} P(i_t=i,i_{t-1},\dots,i_1,o_t,\dots,o_1|\lambda),\quad i=1,2,\dots,N$
  - $\delta_{t+1}(i)=\max_{i_1,i_2,\dots,i_t} P(i_{t+1}=i,i_t,\dots,i_1,o_t,\dots,o_1|\lambda)$

    $=\max_{1\le j\le N}\left[\delta_t(j)a_{ji}\right]b_i(o_{t+1}),\quad i=1,2,\dots,N;t=1,2,\dots,T-1$
- 定义2：在时刻t状态为i的所有单个路径$(i_1,i_2,...,i_t)$中概率最大的路径的第t-1个结点为$\psi_t(i)$
  - $\psi_t(i)=arg\max_{1\le j\le N}\left[\delta_{t-1}(j)a_{ji}\right],\quad i=1,2,\dots,N$

誠樸 篤力 仁

用Viterbi算法求解

- 初始化：
  - $\delta_1(i) = \pi_i b_i(o_1)$
  - $\psi_1(i) = 0$
- 递推：对t = 2,3,…,T
  - $\delta_t(i) = \max\limits_{1 \le j \le N} \left[\delta_{t-1}(j) a_{ji}\right] b_i(o_t)$
  - $\psi_t(i) = arg \max\limits_{1 \le j \le N} \left[\delta_{t-1}(j) a_{ji}\right]$
- 终止：
  - $\mathrm{P} * = \max\limits_{1 \le j \le N} \delta_{\mathrm{T}}(i)$
  - $i_T^* = arg \max\limits_{1 \le j \le N} [\delta_T(i)]$

誠 樸 篤 仁

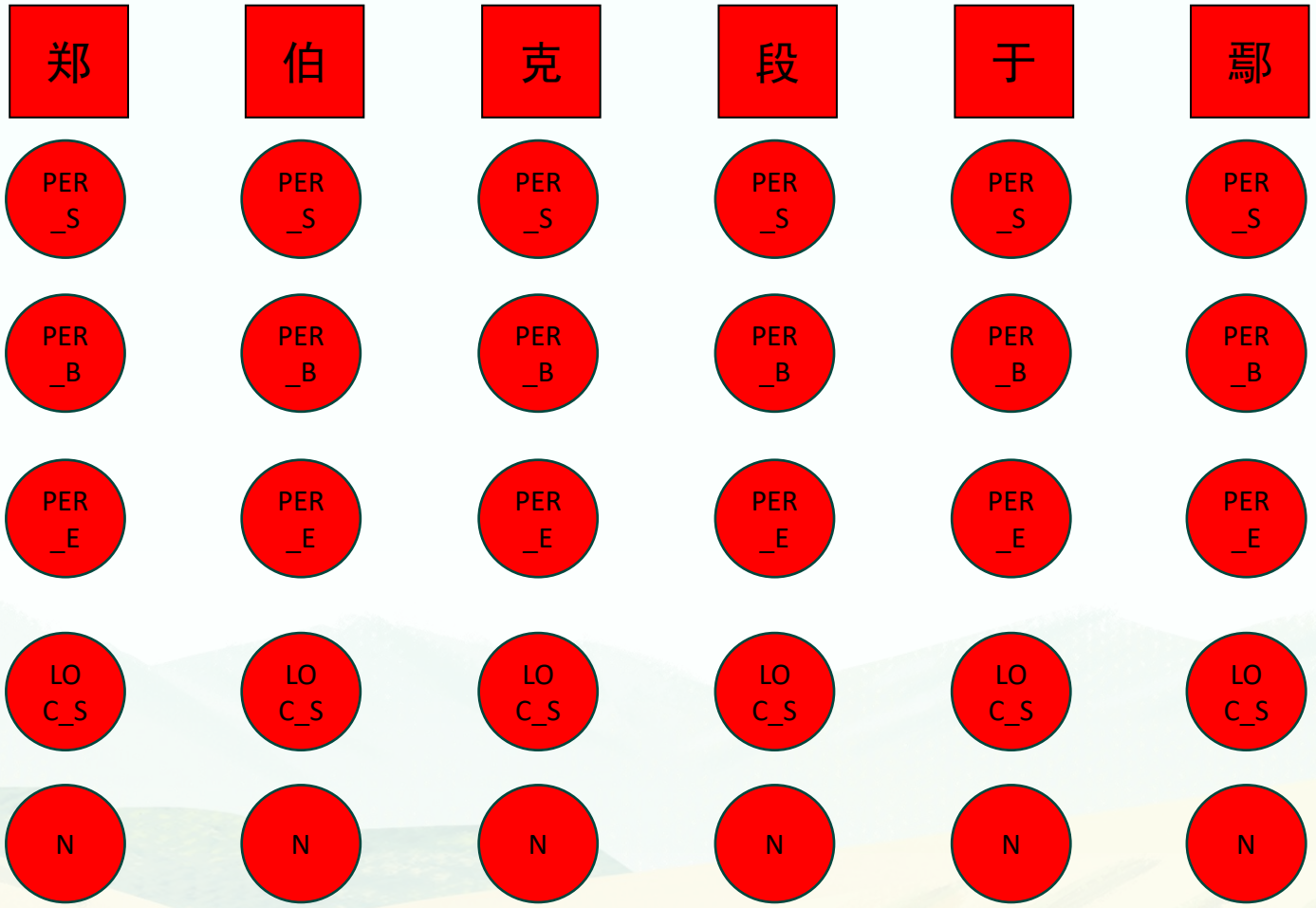
用Viterbi算法求解

- 最优路径回溯：
  - 对t=T-1,T-2,...,1
  - $i_T^* = \psi_{t+1}(i_{t+1}^*)$

- 求得最优路径：
  - $I^* = (i_1^*, i_2^*, \ldots, i_T^*)$

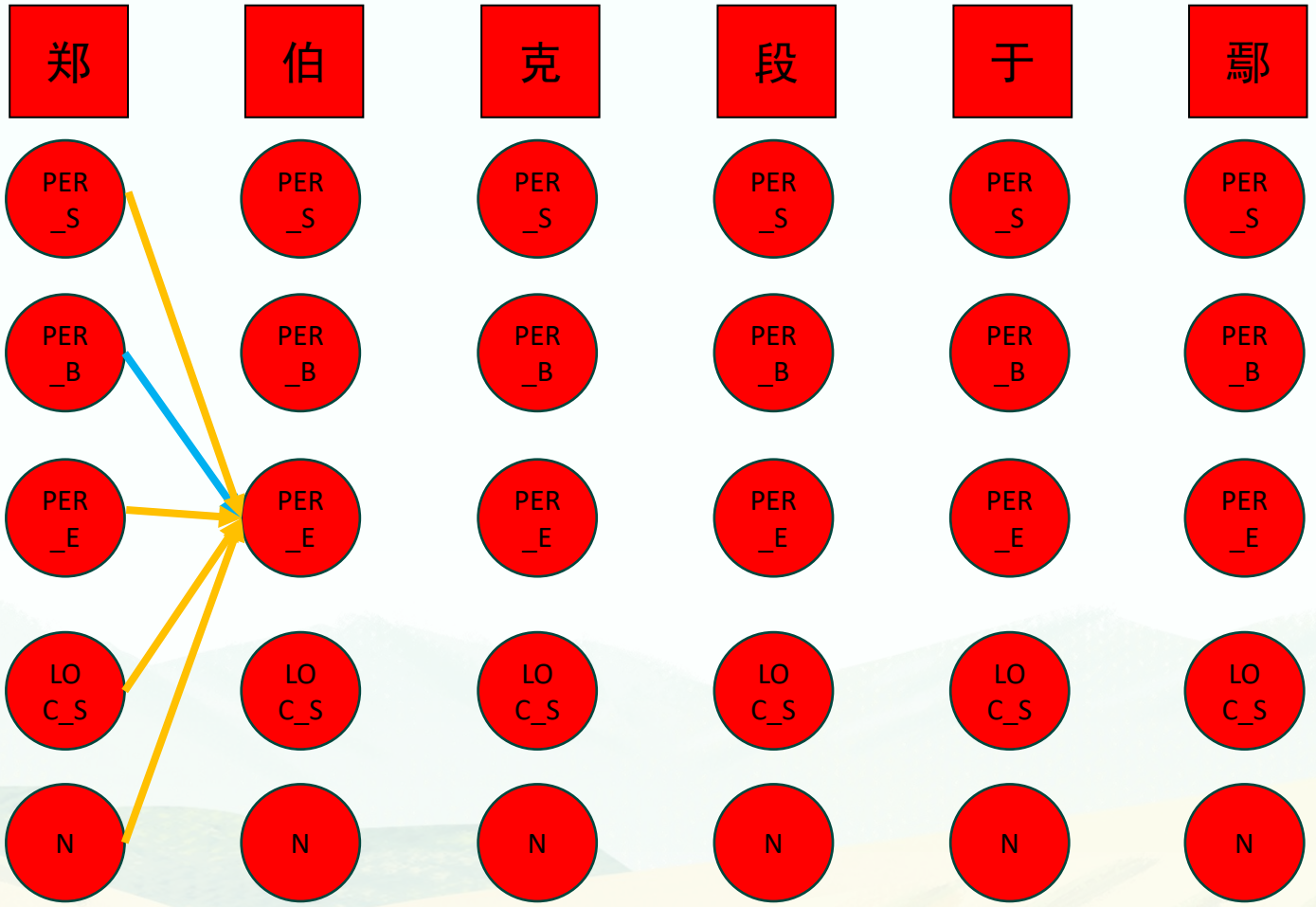
誠 樸 篤 仁

用Viterbi算法求解



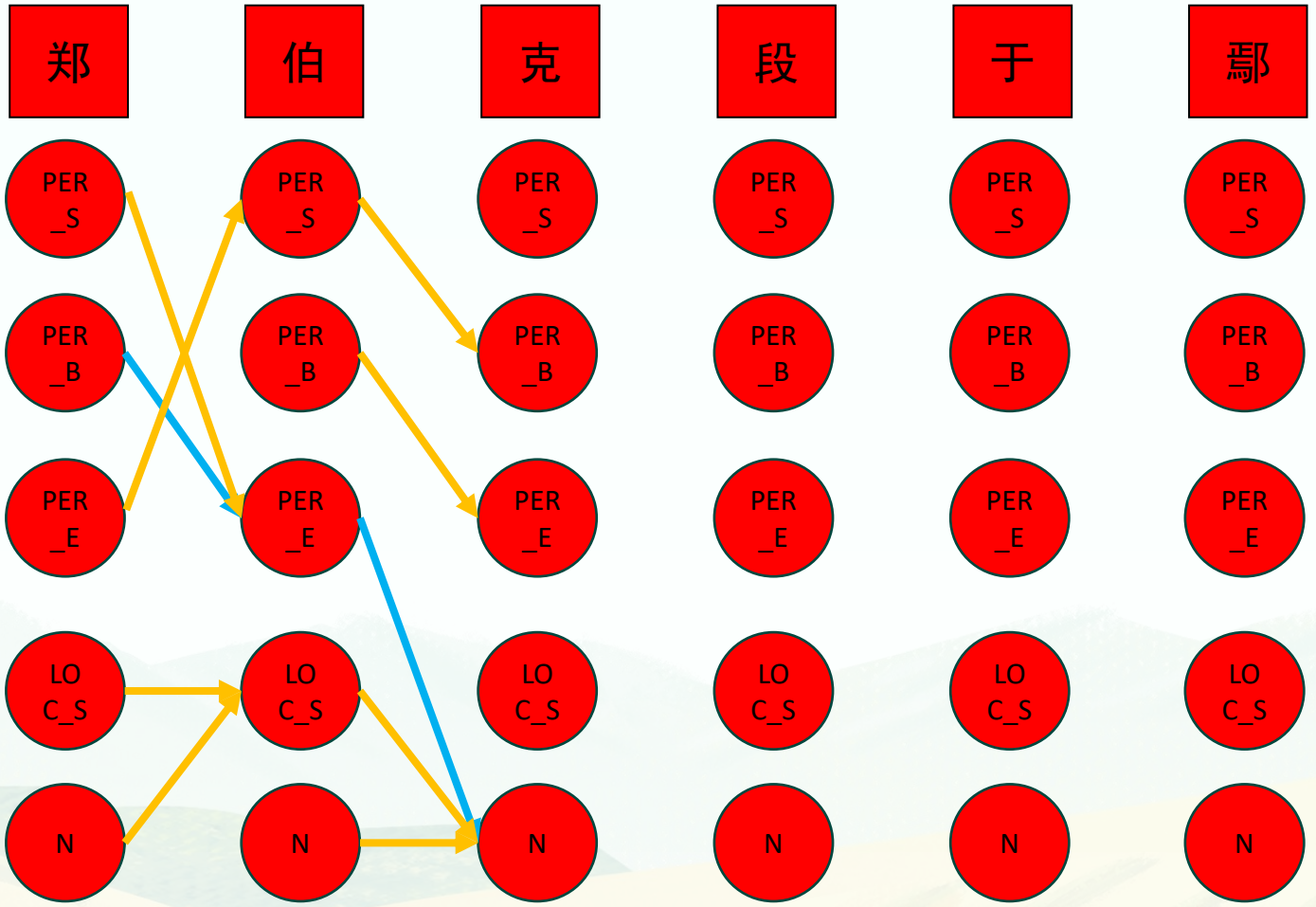
誠樸勤仁

用Viterbi算法求解



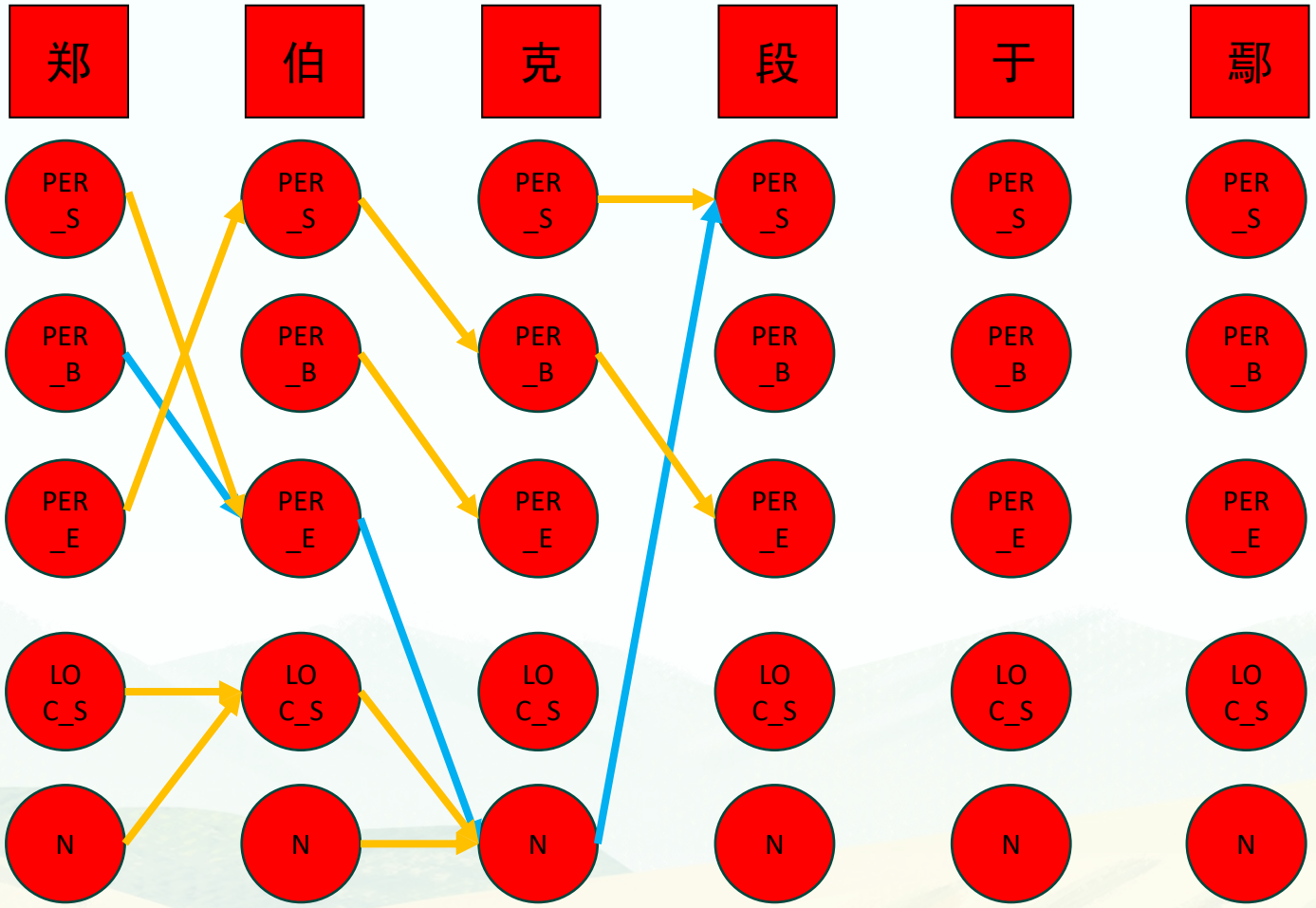
誠樸勤仁

用Viterbi算法求解



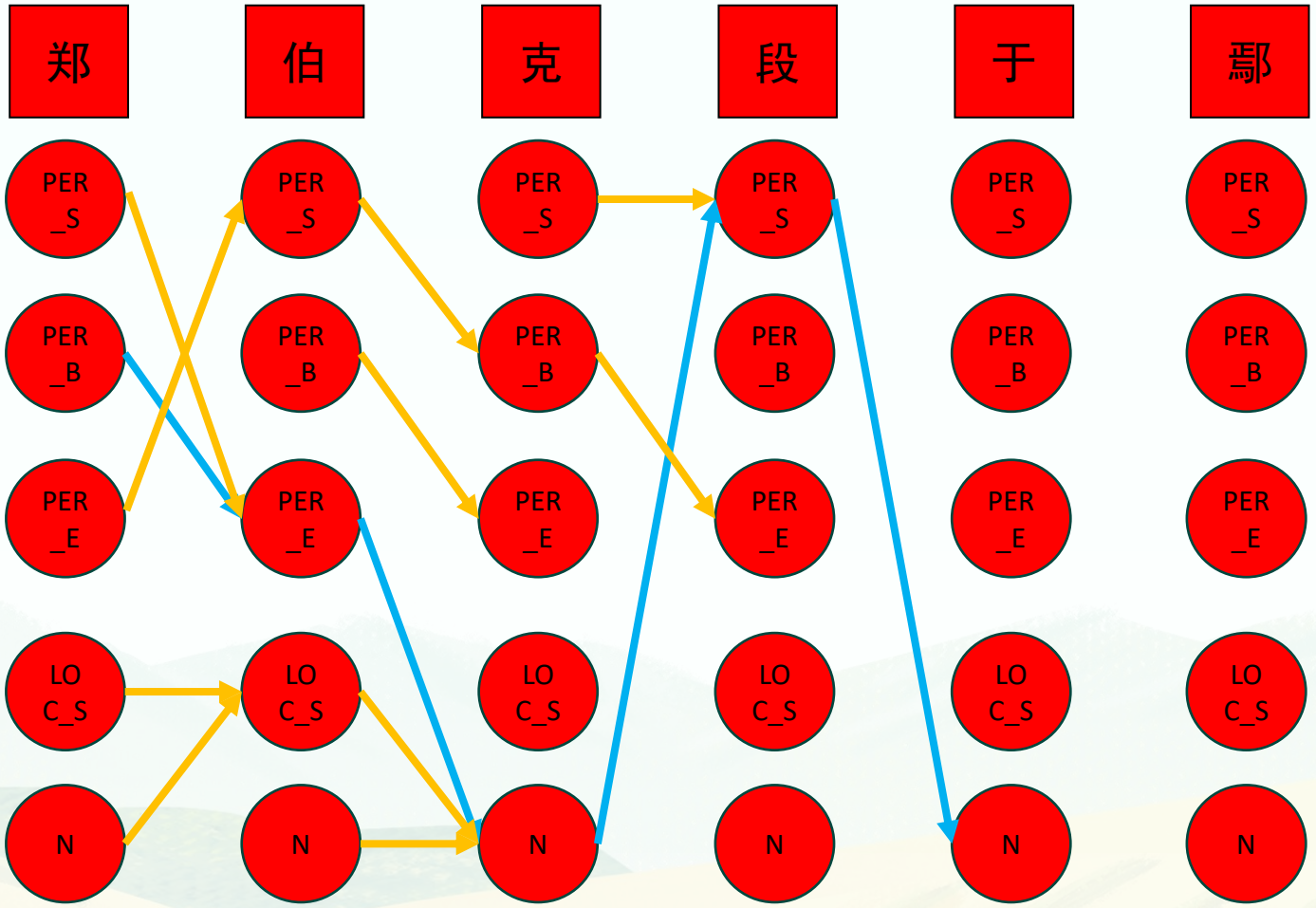
誠樸勤仁

用Viterbi算法求解



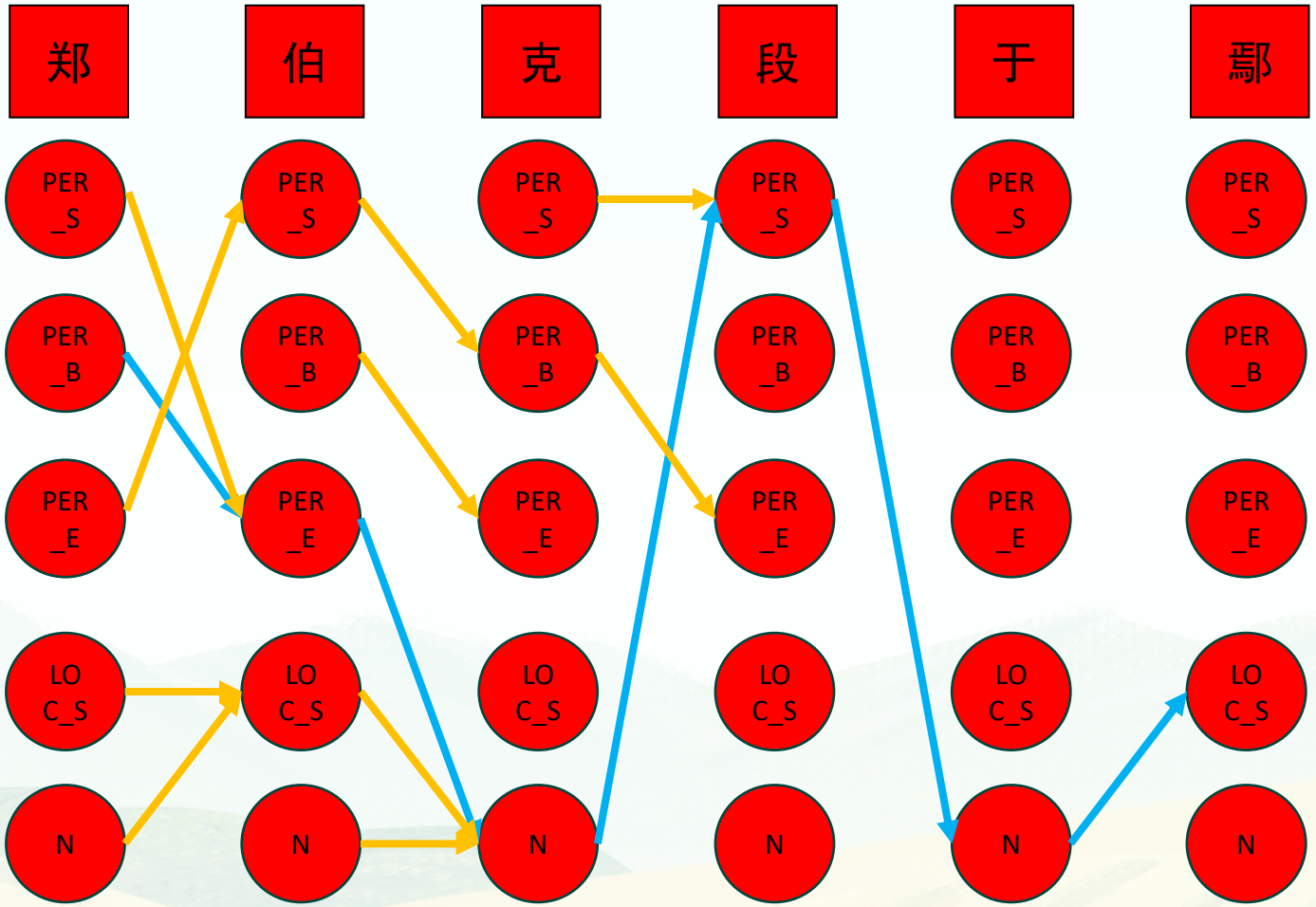
誠樸勤仁

用Viterbi算法求解



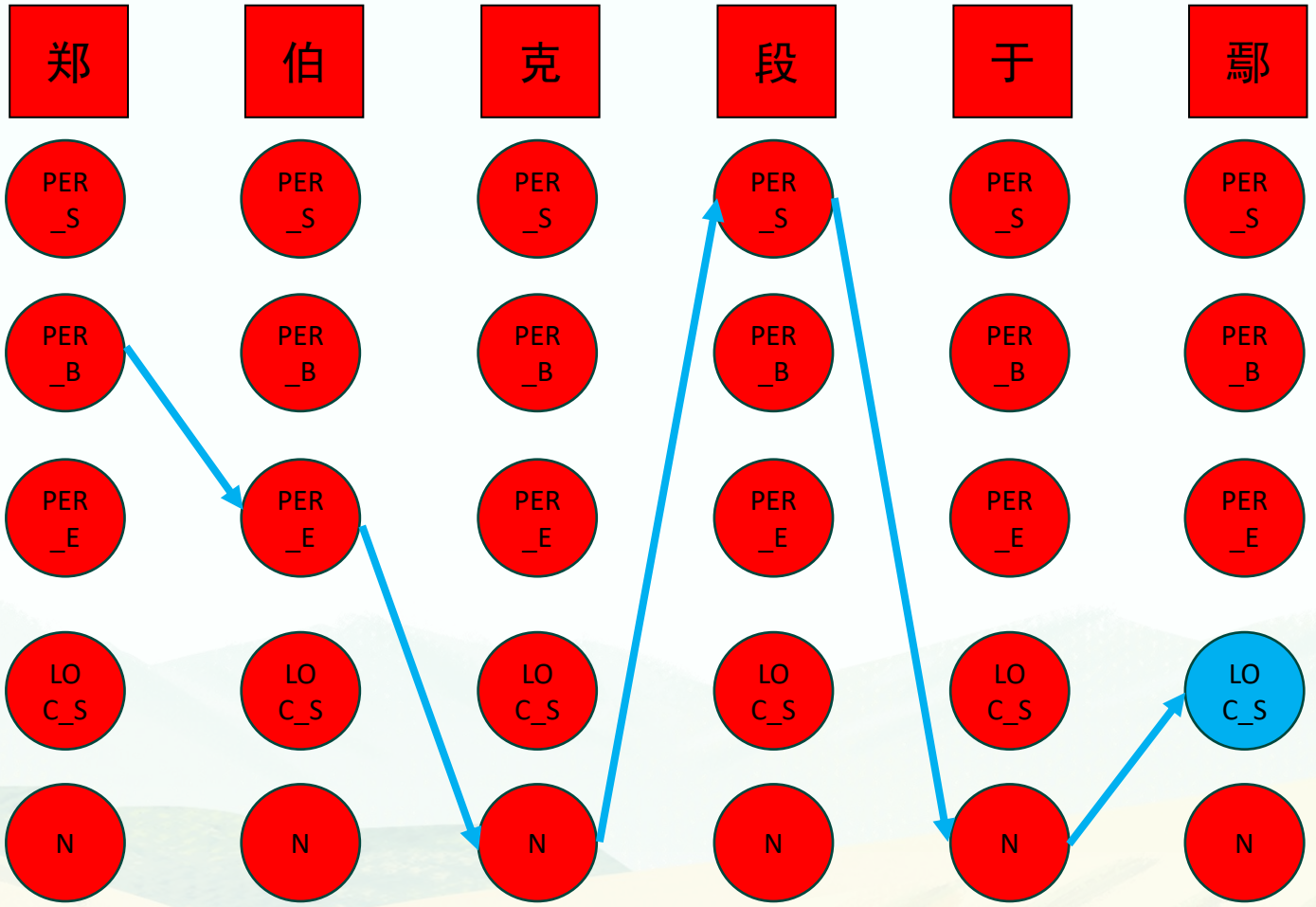
誠樸勤仁

用Viterbi算法求解



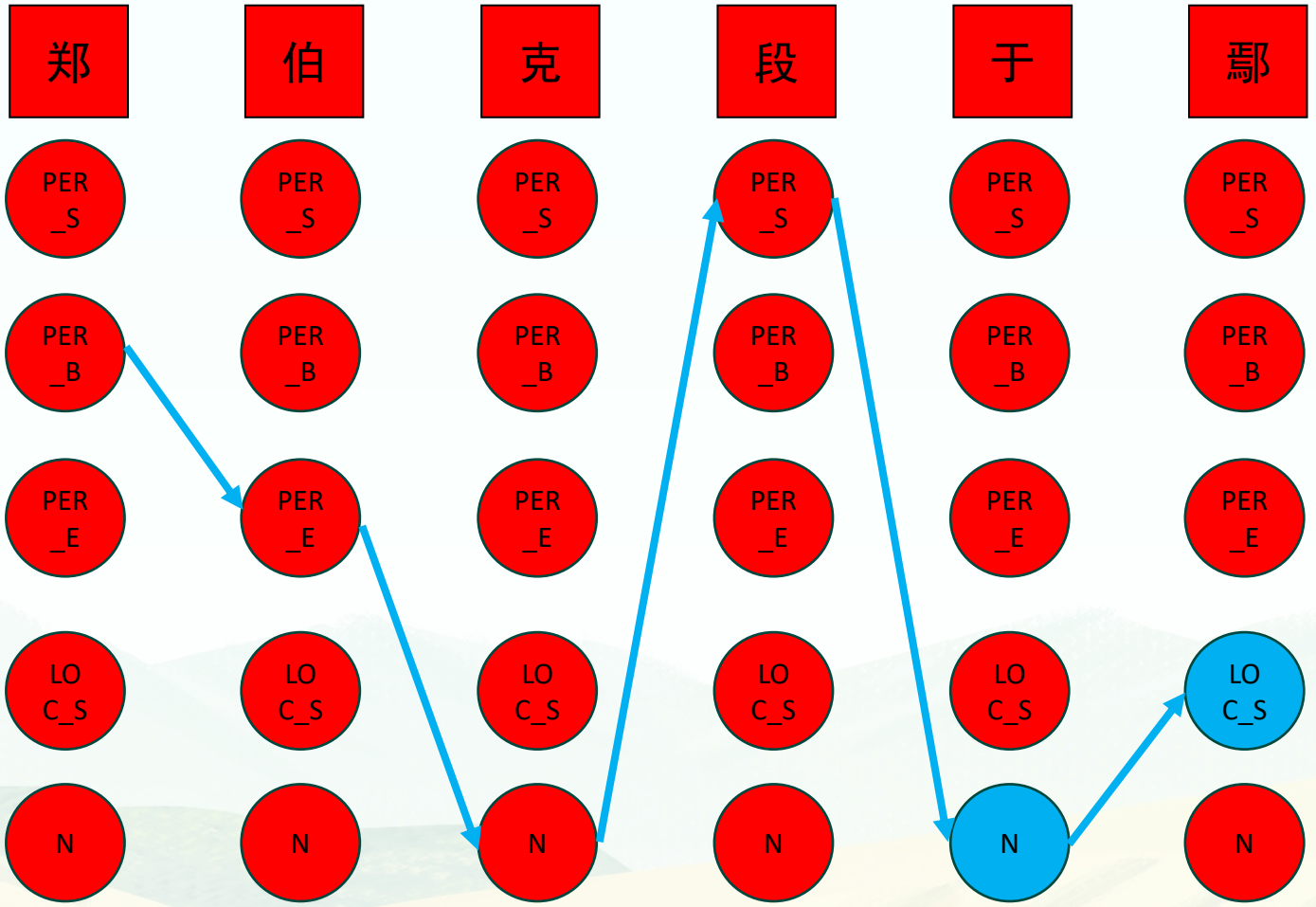
誠樸勤仁

用Viterbi算法求解



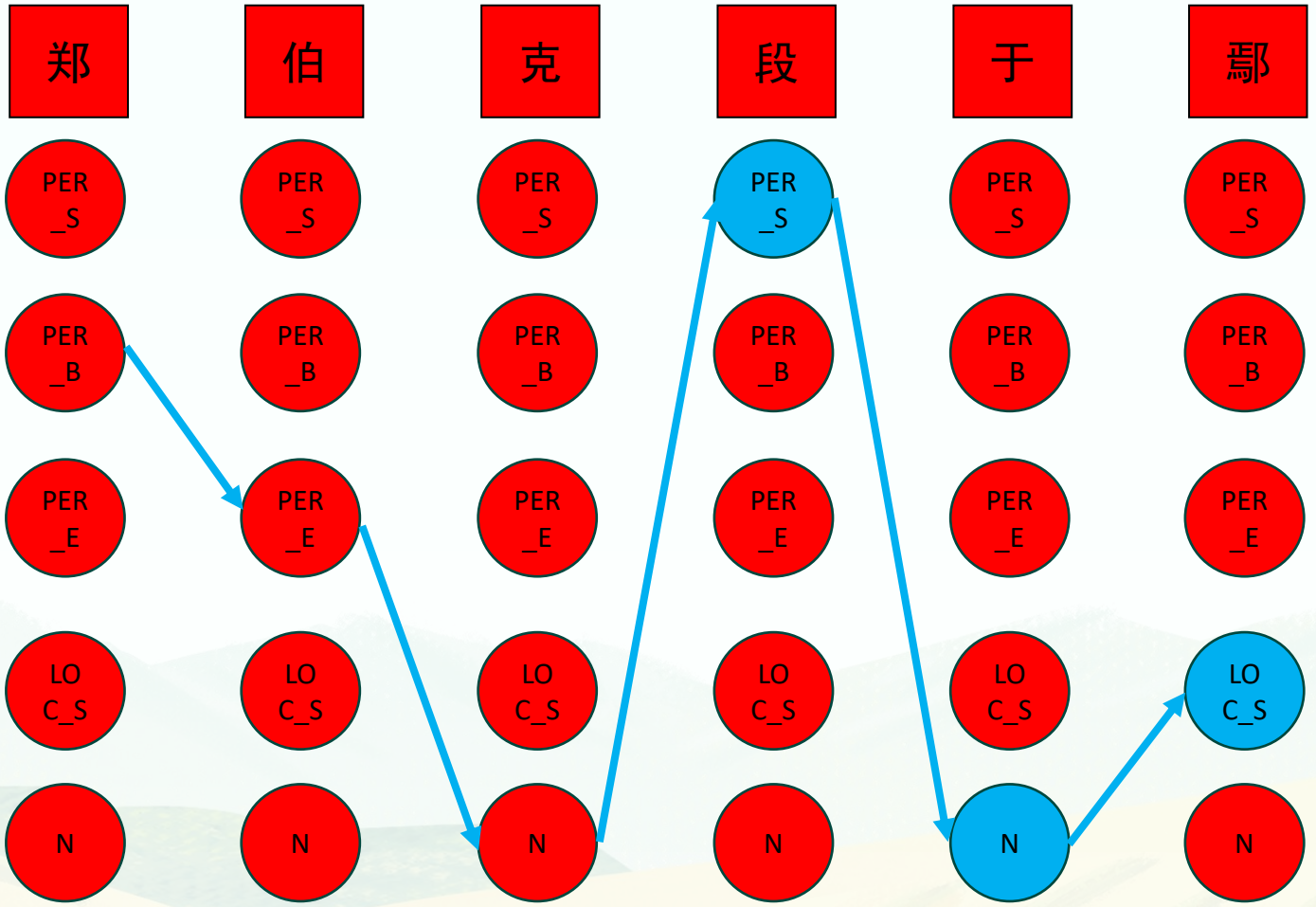
誠樸勤仁

用Viterbi算法求解



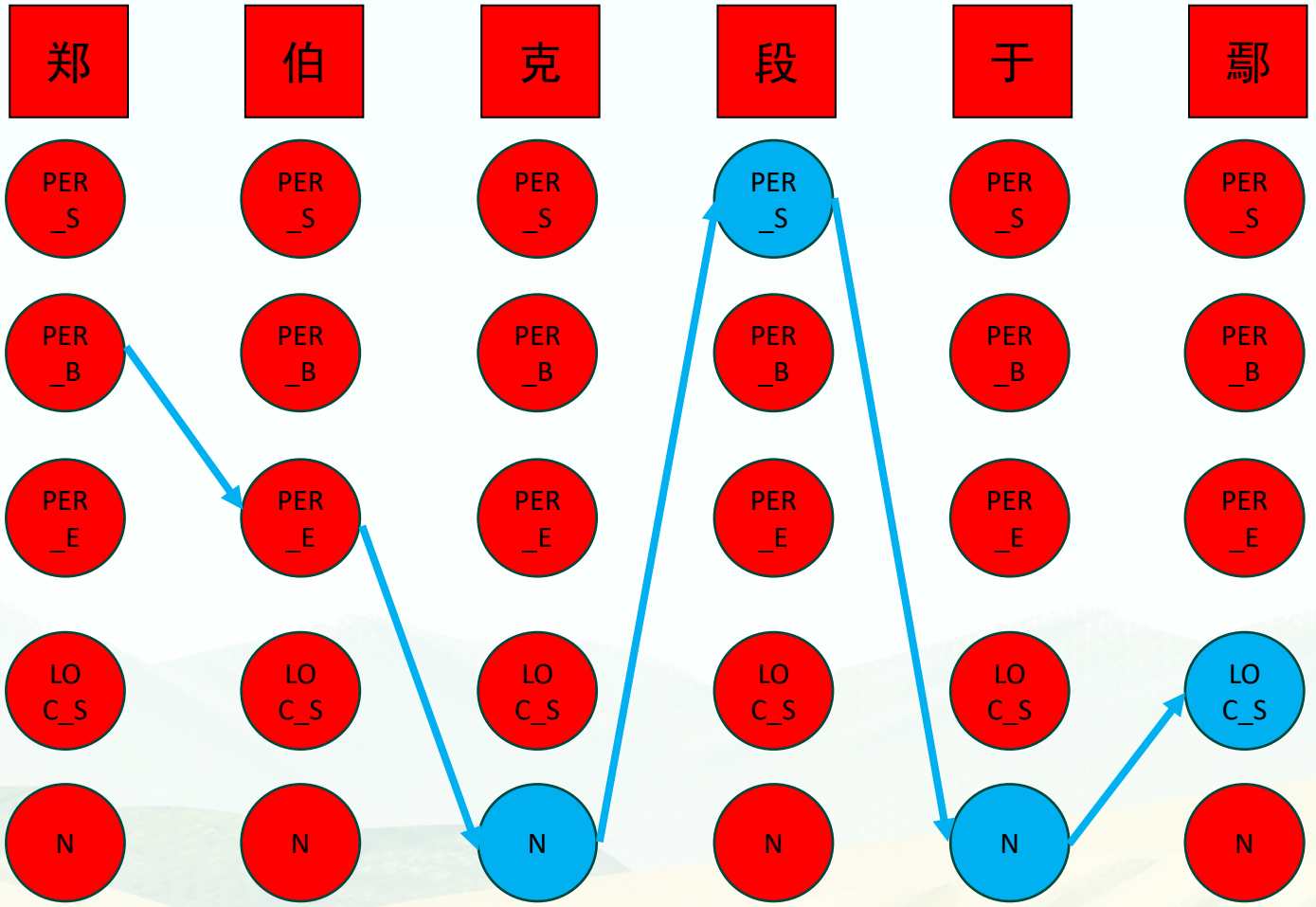
誠樸勤仁

用Viterbi算法求解



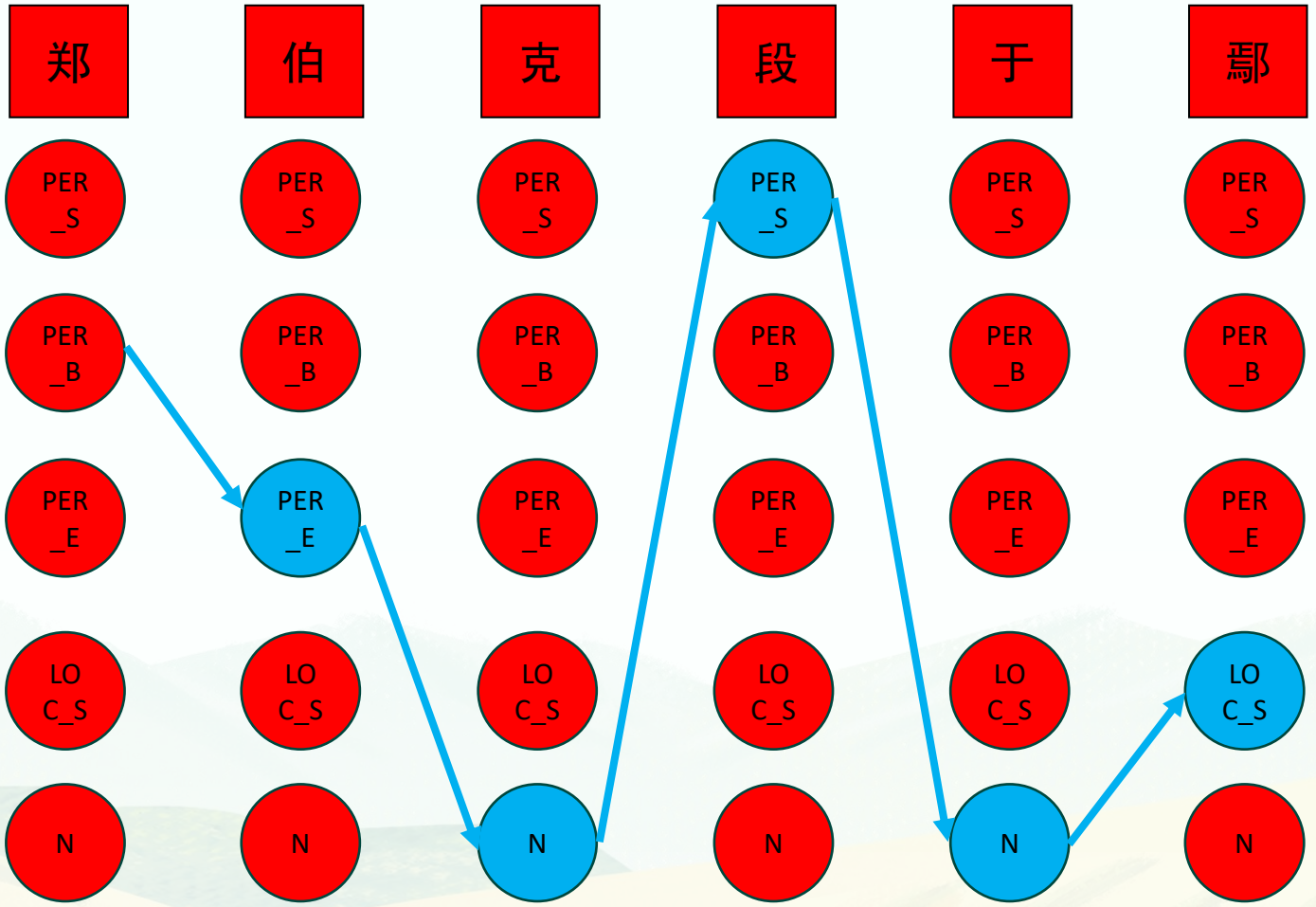
誠樸勤仁

用Viterbi算法求解



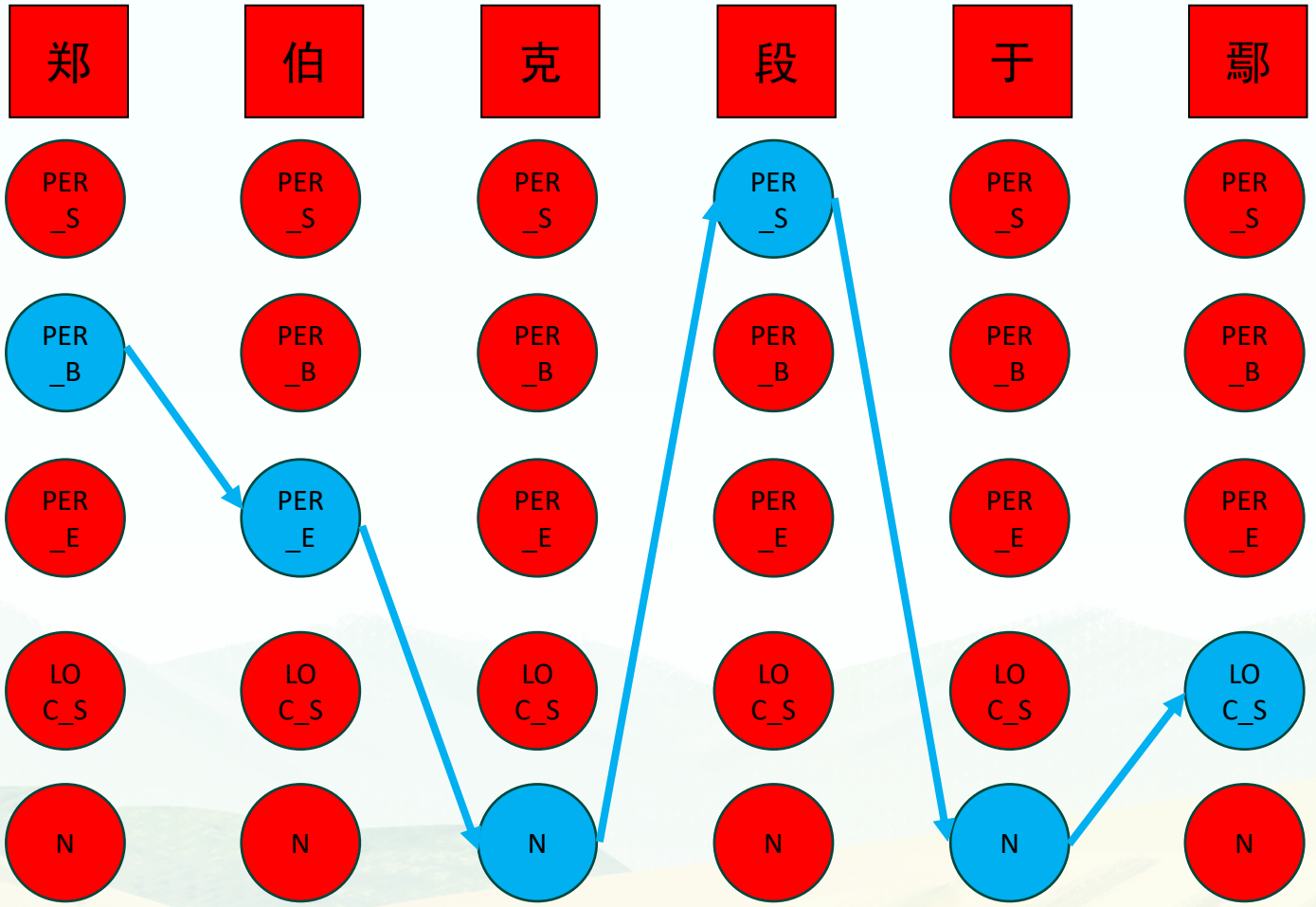
誠樸勤仁

用Viterbi算法求解



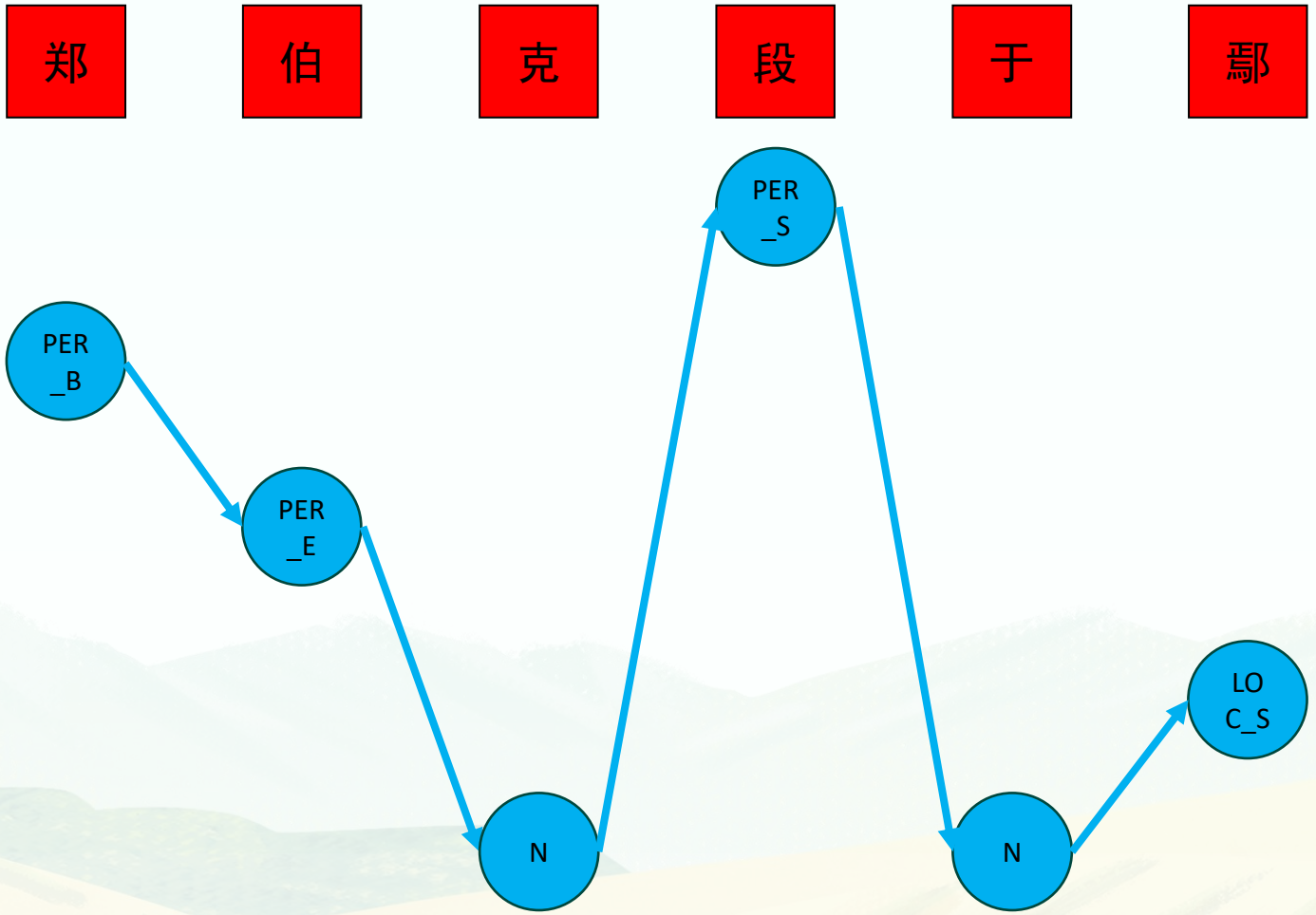
誠樸勤仁

用Viterbi算法求解



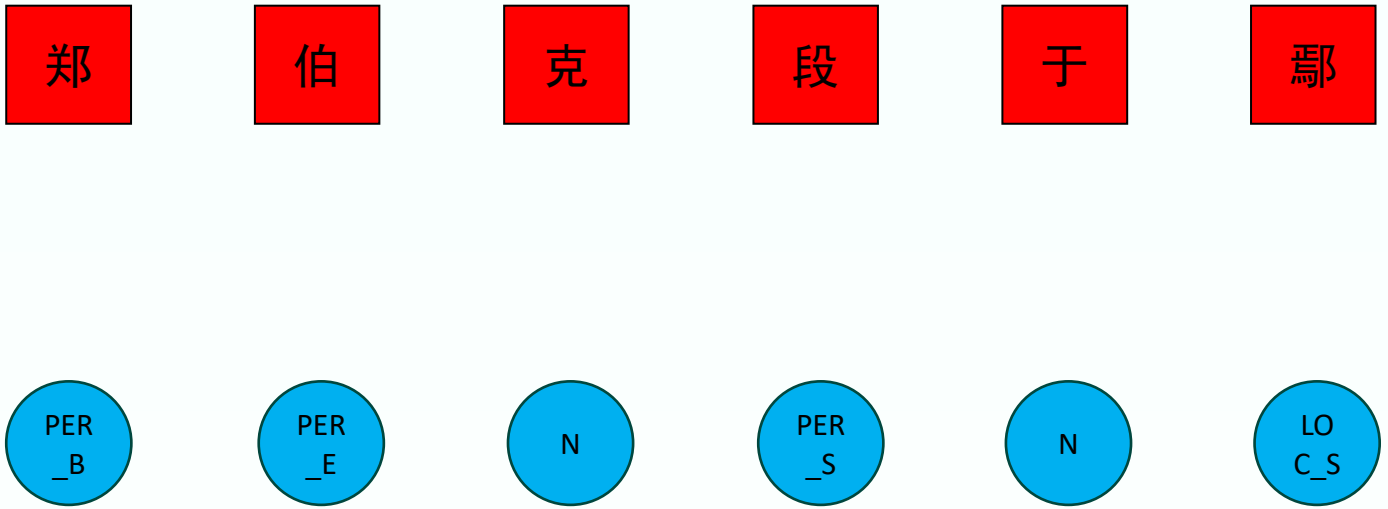
誠樸勤仁

用Viterbi算法求解



誠樸勤仁

用Viterbi算法求解



(PER 郑伯)克(PER 段)于(LOC 鄢)

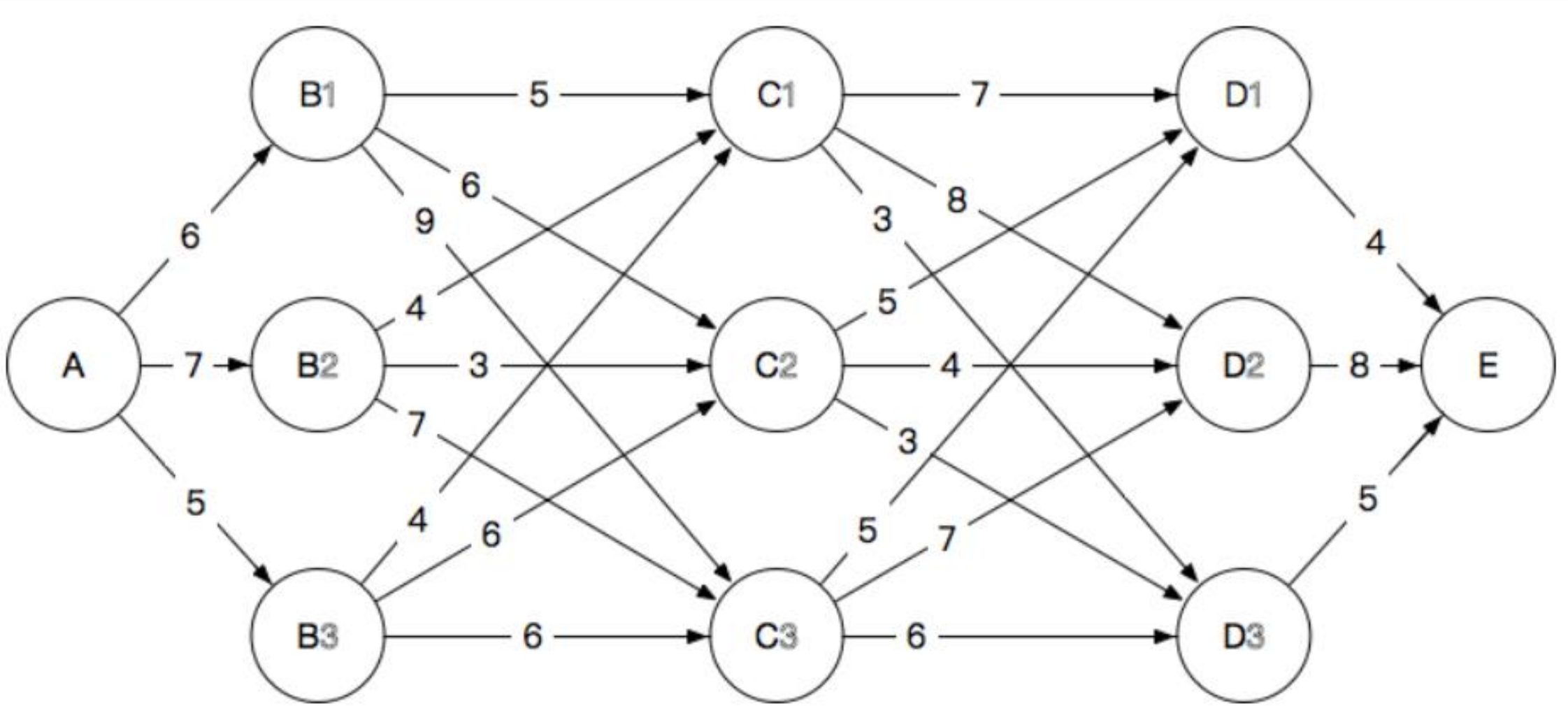
誠 樸 勤 仁

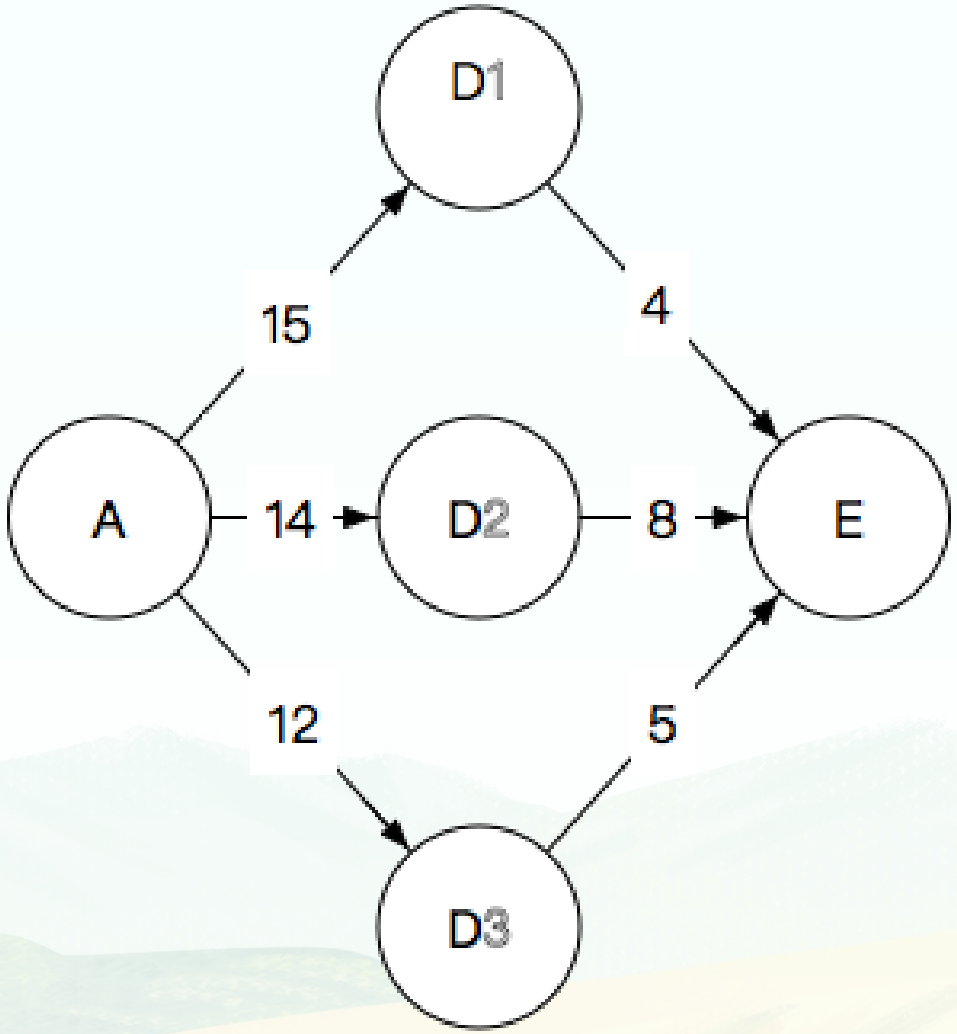
Viterbi算法的时间复杂度

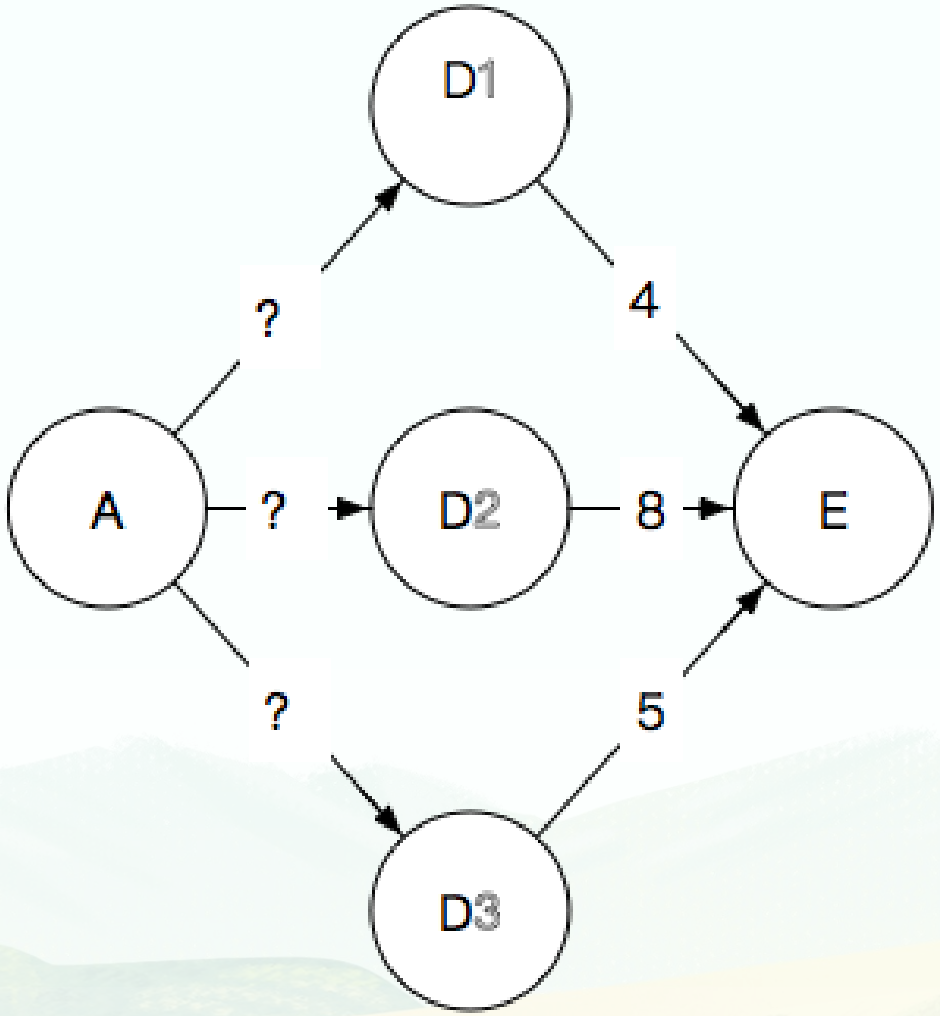
- 每计算一个 $\delta_t(i)$ ，必须考虑从t-1时刻所有的N个状态转移到状态的si概率，时间复杂度为O(N)
- 对应每个时刻t，要计算N个中间变量 $\delta_t(1)...\delta_t(N)$ ，时间复杂度为$O(N^2)$，又t从1…T，因此整个算法时间复杂度为$O(N^2T)$

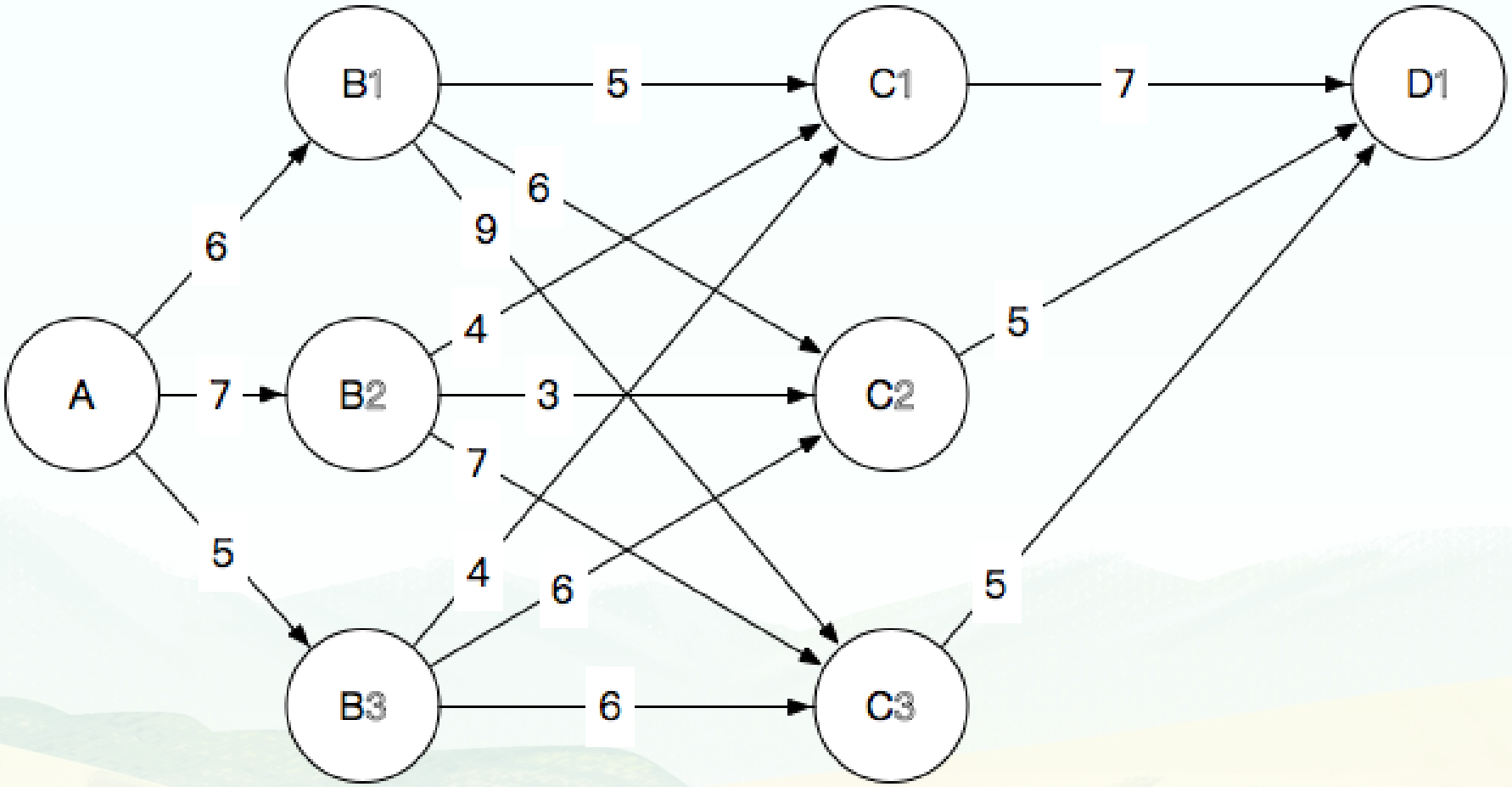
誠 樸 篤力 仁

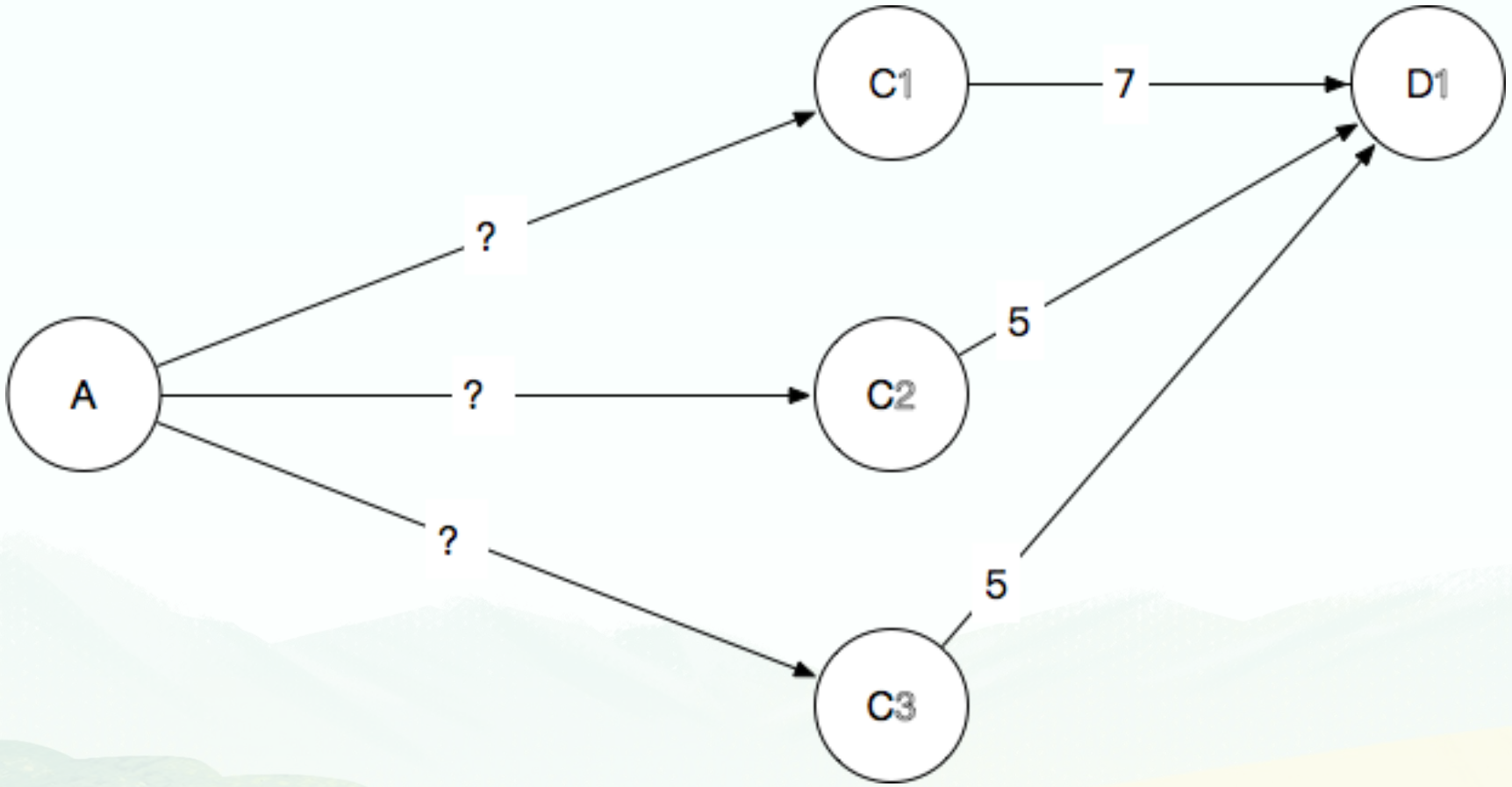
动态规划算法

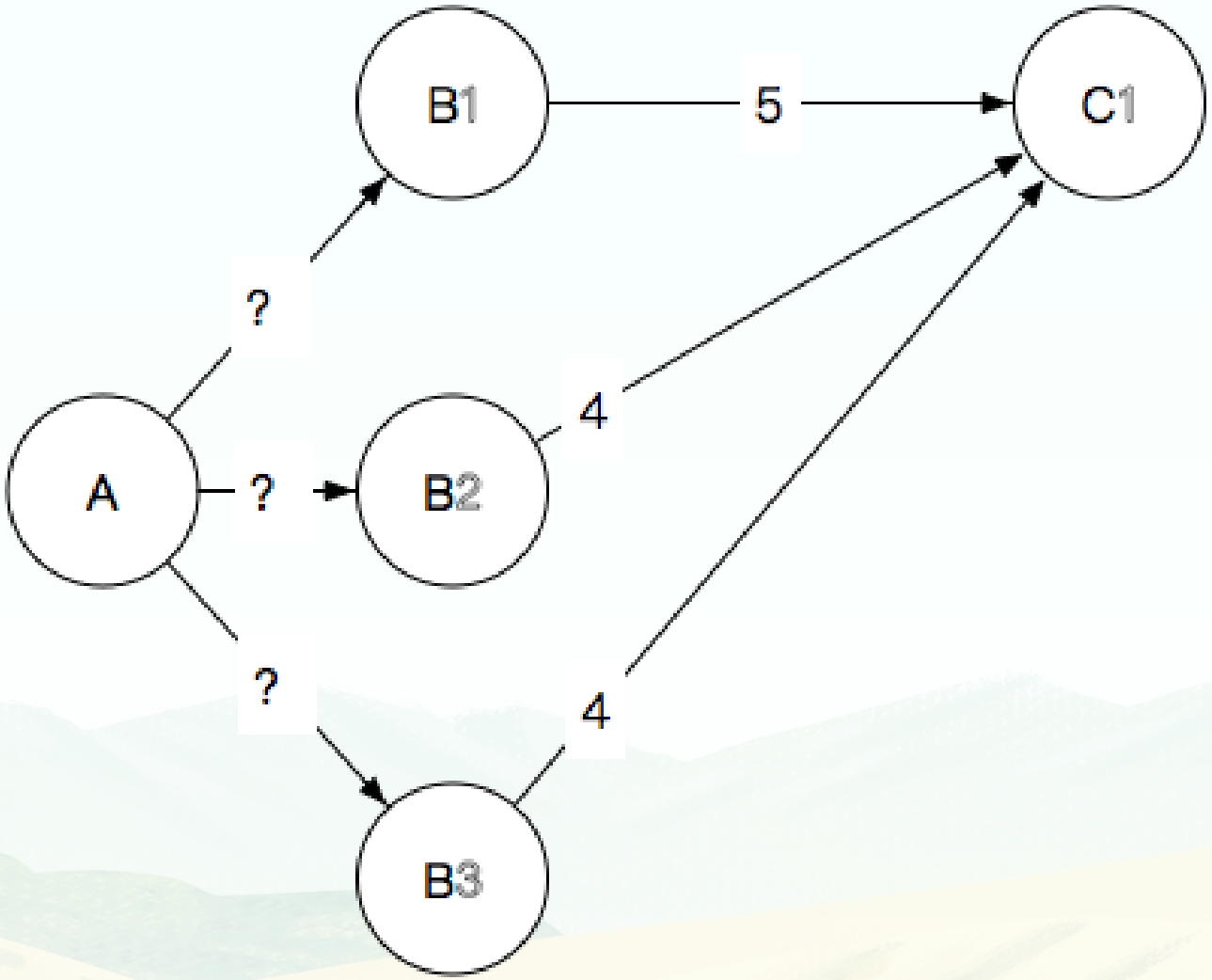


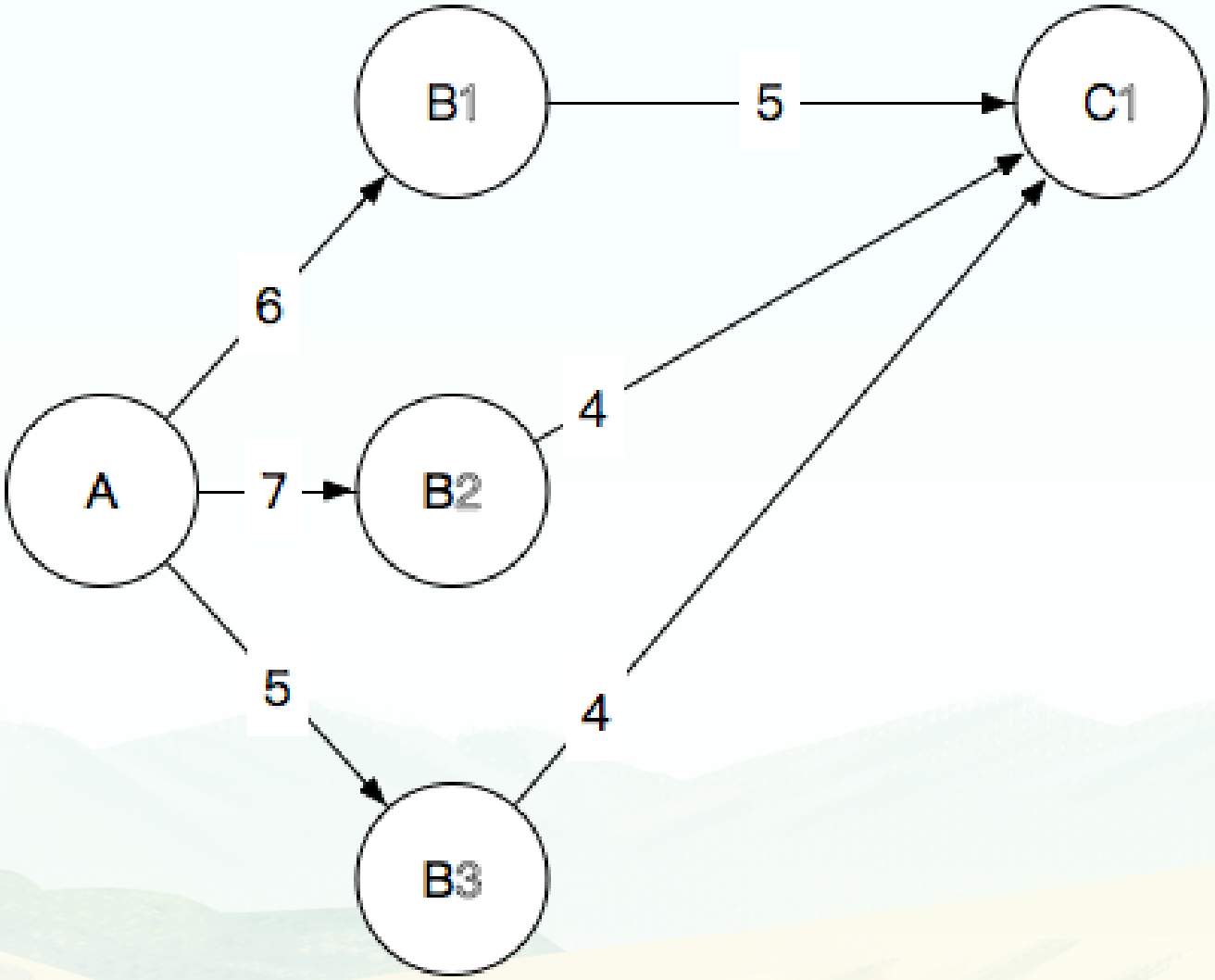

D1
15
4
A
14
D2
8
E
12
5
D3


D1
?
4
A
?
D2
8
E
?
5
D3

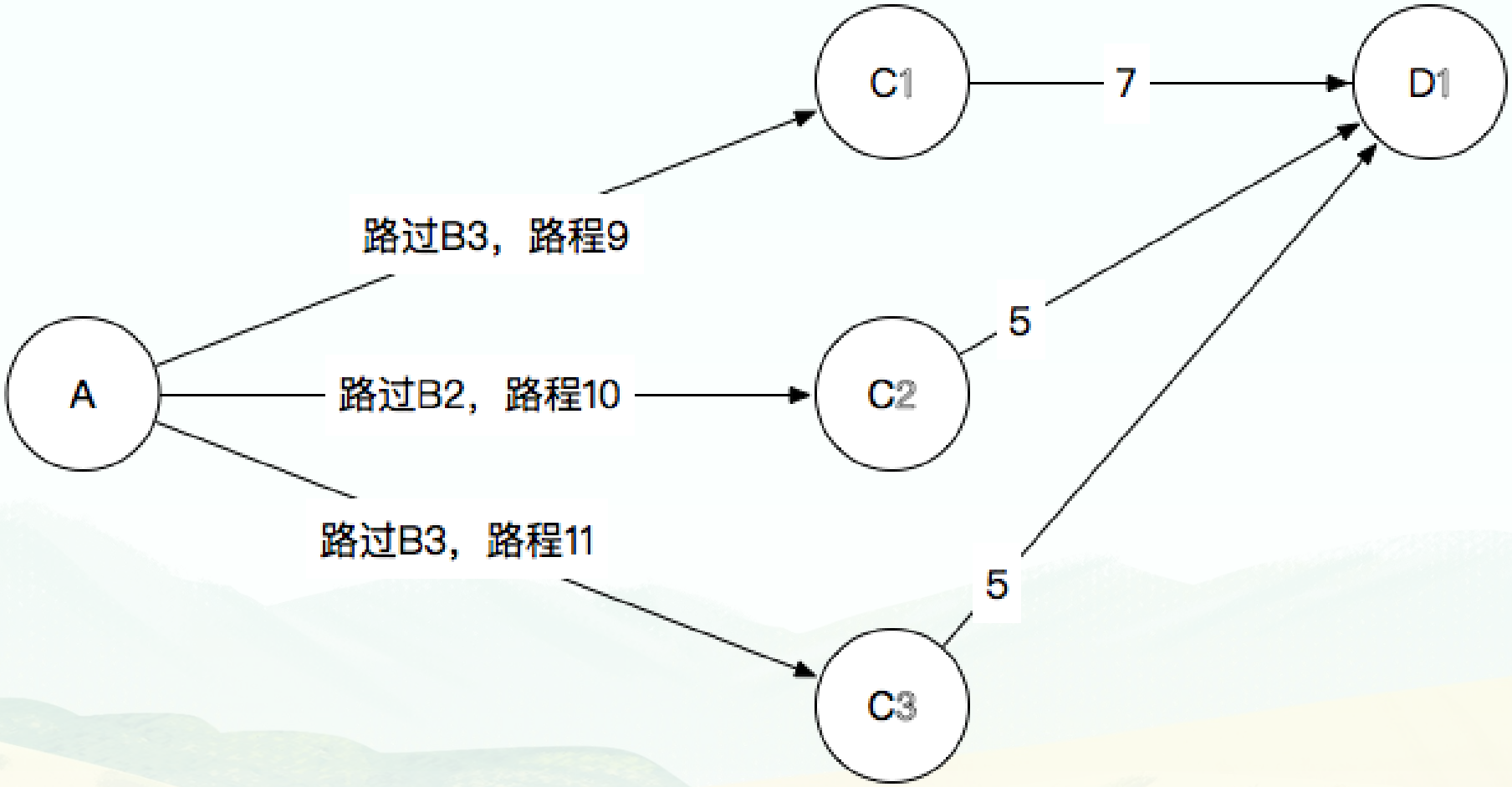

A
B1
B2
B3
C1
C2
C3
D1
6
7
5
5
4
6
9
4
3
7
4
6
6
7
5
5

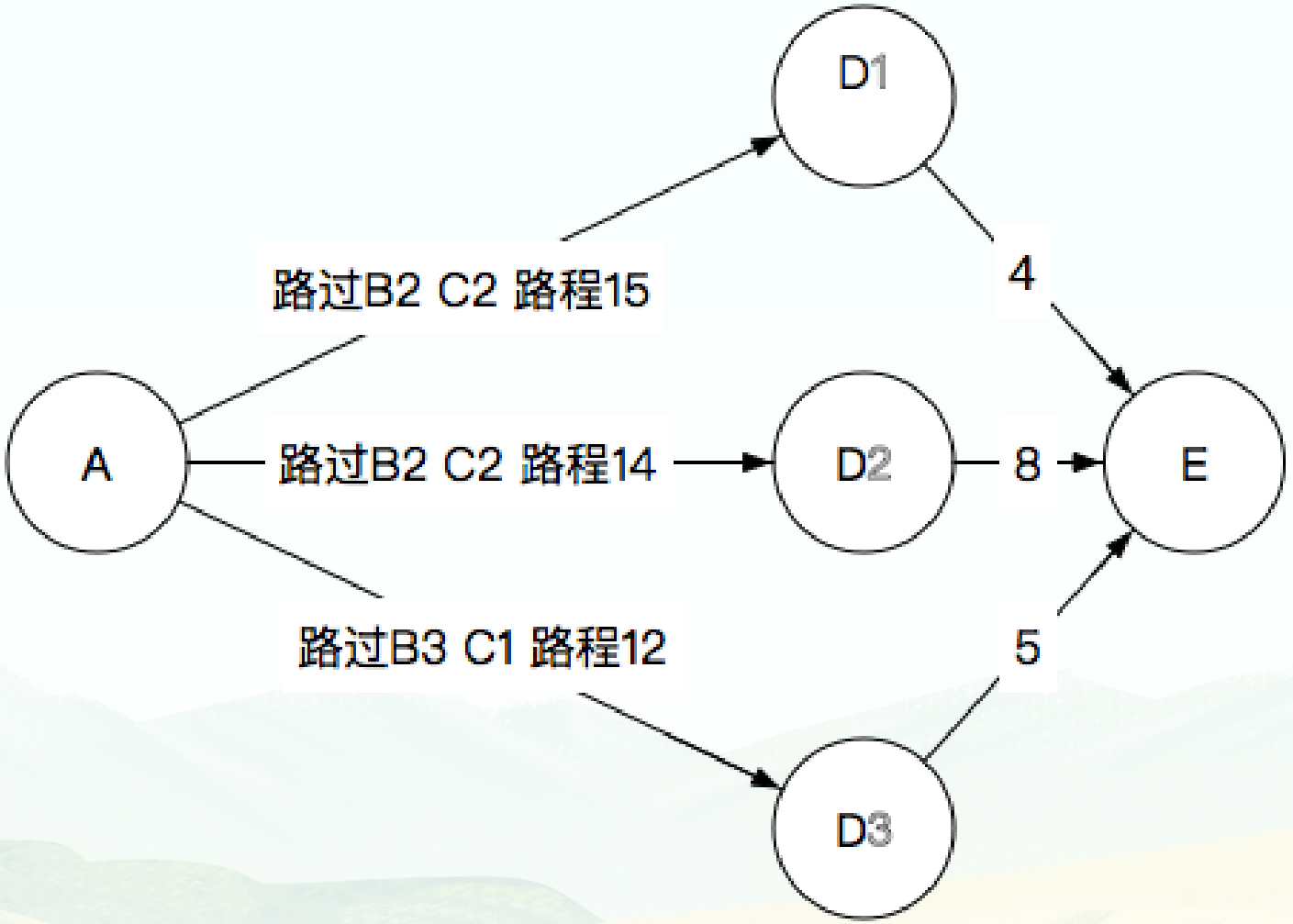

C1
7
D1
?
A
?
C2
5
?
5
C3


B1
5
C1
?
A
?
B2
4
?
4
B3



誠樸篤仁


C1
7
D1
路过B3，路程9
A
路过B2，路程10
C2
5
路过B3，路程11
5
C3



誠樸篤仁

命名实体识别的方法

◆探索、标注和测试各类特征知识，以及与此密切相关的特征工程，是机器学习时代命名实体识别中最重要的一项研究内容。

◆在深度学习兴起之后，知识的表示被词嵌入（word embedding）和预训练（pre-train）等表示学习方法替代，计算机可以自动从上下文中学到稠密低维的文本向量，削弱了特征知识标注的必要性，同时进一步提高了命名实体识别的效果。

◆深度学习模型与传统机器学习模型看待特征知识的视角截然不同，但这不意味着特征知识的探索的停止，有研究表明，利用好的特征知识的CRF模型能够获取比深度学习模型更好的性能，这在古汉语等语料规模较小的研究领域中尤其如此。

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

古文实体识别

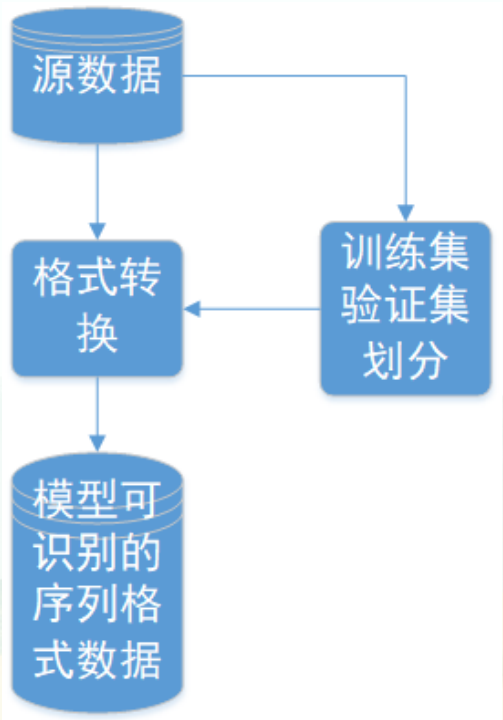
古文实体识别

◆古汉语命名实体研究较之现代汉语最大的差别在于，不仅关注以名字指称实体对象的“命名性指称”，还关注“名词性指称”，如“齐桓公”、“中行寅”、“聲孟子”等，这类人名中往往包含了谥号、官职等非名字成分，因此古汉语命名实体识别的难点之一就在于人名识别。

古文实体识别流程

◆1 数据的预处理

➢ 模型将实体识别问题看作一个序列标注问题，因而需要将标注好的典籍文本转换为模型能够识别的序列格式。



誠樸勤仁

古文实体识别流程

◆1 数据的预处理

➢ 格式转换

输入格式

惠公/nr 元妃/v 孟子/nr 。/w ↓
孟子/nr 卒/v ， /w 繼/v 室/n 以/p 聲子/n ， /w 生/v 隱公/nr 。/w ↓
宋武公/nr 生/v 仲子/nr ， /w 仲子/nr 生/v 而/c 有/v 文/n 在/p 其/r 手/n ， /w 曰/v 為/v 魯/ns 夫人/n ， /w 故/c 仲子/nr 歸/v 于/p 我/r 。/w ↓

输出格式

惠 B-nr↓
公 E-nr↓
元 O↓
妃 O↓
孟 B-nr↓
子 E-nr↓
。 O↓
↓
孟 B-nr↓
子 E-nr↓
卒 O↓
， O↓

↓
宋 B-nr↓
武 I-nr↓
公 E-nr↓
生 O↓
仲 B-nr↓
子 E-nr↓
， O↓
仲 B-nr↓
子 E-nr↓
生 O↓
而 O↓

古文实体识别流程

◆1 数据的预处理

- 训练集测试集划分：模型训练时需要将数据集划分为训练集和验证集，比例一般为9：1或99：1（数据较大时）。有时为了防止数据划分的偶然性，还需要进行**十折交叉验证**。

誠樸 篤力 仁

古文实体识别流程

◆2 基于BERT的典籍古文实体识别

- 创建虚拟环境
- 安装所需前置python包
- 放入本次实验所需数据
- 下载本次bert实验使用的预训练模型
- 修改run.sh中的参数、模型路径、数据路径等
- 修改settings.py中的标签类型
- 运行bert模型
- 查看模型预测结果

誠 樸 篤 仁