希望语义相近的词在编码空间中的位置相近 ⟹ 可以泛化

机器在学习过程中对原本固定的离散编码进行修改来达到目的

Chinese Word Segmentation

分割

- Although great progress has been made in improving the performance of ***Chinese word segmentation*** systems, Chinese word segmentation remains an ***extremely difficult problem***.
- There is no existing approach that is able to reliably segment ***unfamiliar types of texts*** before fine-tuning with massive training data, not to mention the ***open texts*** mixed with the texts from different topics, genres, and regions.
- ***Word segmentation*** still remains a ***main research topic*** in the field of Chinese language processing, which indicates that there may be unresolved theoretical and processing issues.

Word Boundary Ambiguities

歧义

- **Combination Ambiguity**

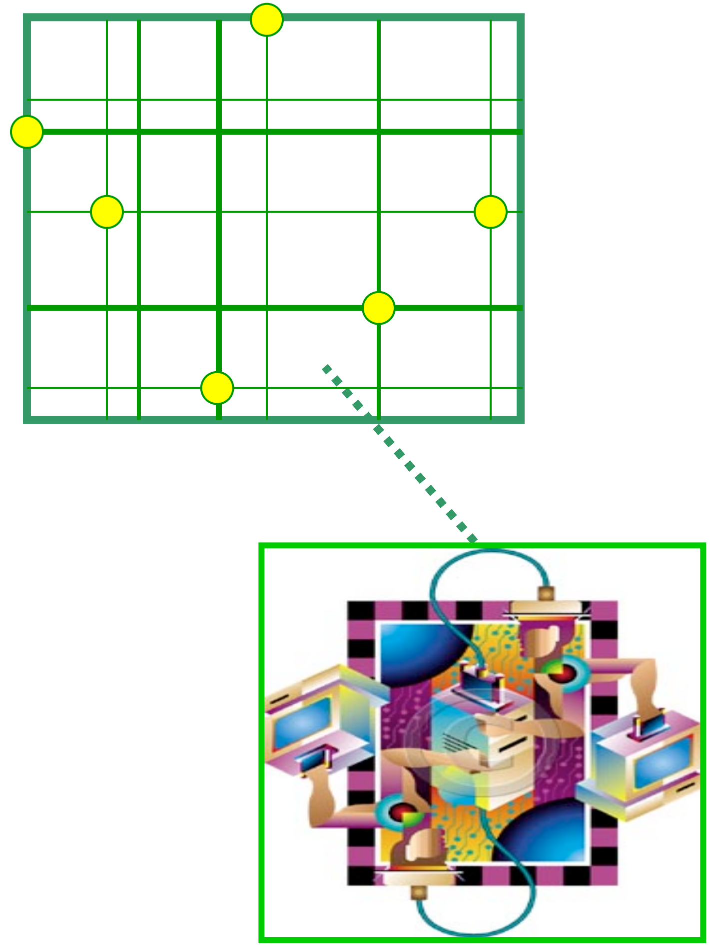
  以 / 我 / 个人 / 的名义
  他 / 一 / 个 / 人 / 在家

- **Overlap Ambiguity**

  从 / 小学 / 到 / 中学
  从小 / 学 / 计算机

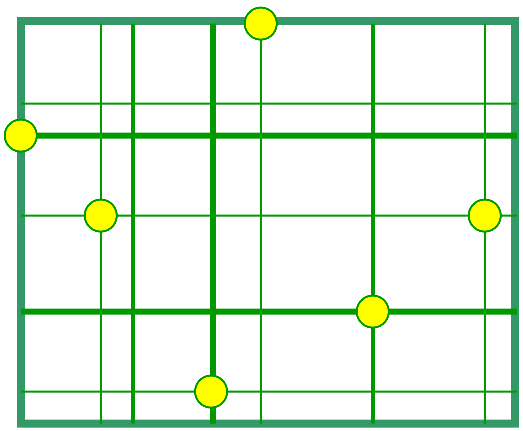
- **Global Ambiguity**

  美国 / 会 / 采取行动
  美 / 国会 / 采取行动



Unknown words

他从小学画画
他从小学开始画画
他从小学着画画



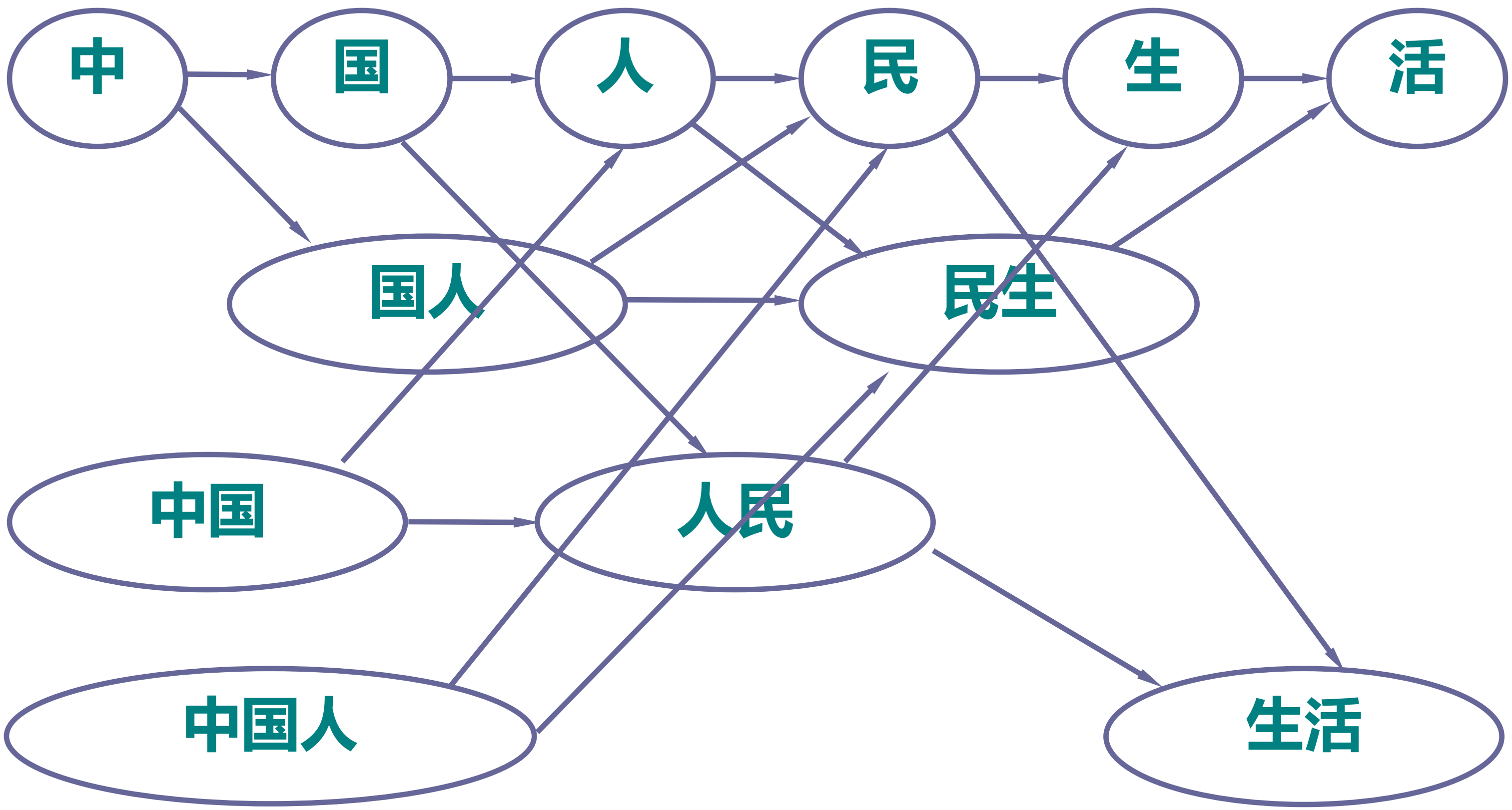
他 / 从小 / 学 / 画画
他 / 从 / 小学 / 开始 / 画画
他 / 从小 / 学着 / 画画

数量、时间： “2012年5月22日”、 “一市斤”、 “356克”
人名、机构名、地名： “李维汉”、 “阿凡提”、 “上海浦东张江镇”
外文翻译及缩写： “阿诺德·施瓦辛格”、 “萨默维尔”、 “FDA”
专用名称的缩写： “复旦”、 “中航”、 “央行”
新词： “安家费”、 “三通”、 “菜鸟”、 “八卦”

Word Formation

- **重叠词**
  AA式：爸爸、宝宝、星星
  AABB式：大大咧咧、形形色色、漂漂亮亮
- **派生词**
  前缀 + 词根：阿姨、老虎、老婆
  词根 + 后缀：椅子、木头、鸟儿
  词根 + 中缀 + 词根：对得起、来得及
- **词转化为短语**
  动宾结构式：鞠躬 → 鞠个躬 → 鞠个九十度的躬
  偏正或联合结构：同学 → 同过三年学
  补充结构式：达到 → 达得/不到
- **短语转化为词**
  ABCD → AC式：土地改革 → 土改、地下铁道 → 地铁
  ABCD → AD式：空中小姐 → 空姐、高等院校 → 高校
  截段简缩：中国南极长城站 → 长城站、复旦大学 → 复旦
  综合简缩：联合国安全理事会 → 安全理事会 → 安理会

Chinese word segmentation



Character-based Tagging Solution

Sentence: 从 小 学 计 算 机 。

Tags: B E S B I E S

词开始 词结束 词中间

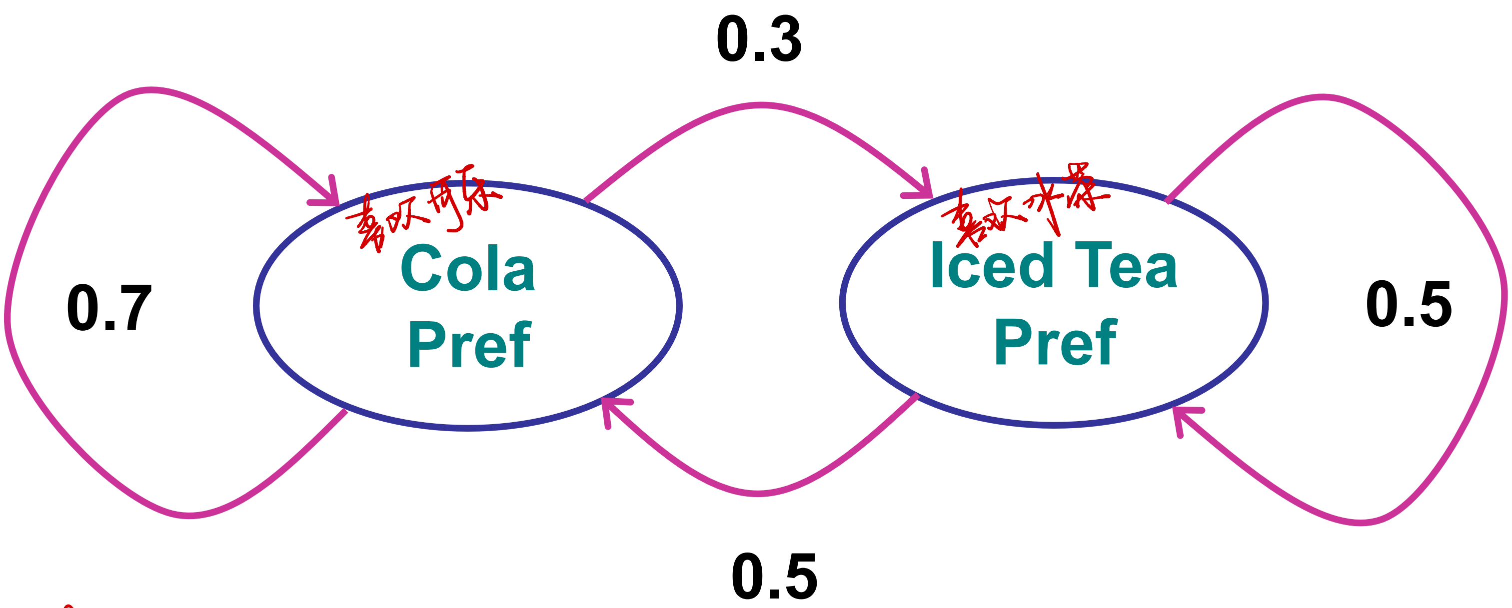
单独成词

Character-based models treat known and unknown words equally and no other exception processing is necessary.

Xue, CLCLP2003
(University of Pennsylvania)

隐式马尔克夫模型

The crazy soft drink machine



machine

	Cola	Iced tea	Lemonade
Cola Pref (CP)	0.6	0.1	0.3
Iced Tea Pref (IP)	0.1	0.7	0.2

隐马模型
五元组

General form of an HMM

An ***Hidden Markov Model*** (HMM) is specified by a five-tuple (S, K, Π, A, B)

各状态的分布律

Set of states	$S = \{S_1, \ldots, S_N\}$
Output alphabet	$K = \{K_1, \ldots, K_M\} = \{1, \ldots, M\}$
Initial state probabilities	$\Pi = \{\pi_i\},\ i \in S$
State transition probabilities	$A = \{a_{ij}\},\ i, j \in S$
Symbol emission probabilities	$B = \{b_{\boxed{ijk}}\},\ i, j \in S,\ k \in K$
State sequence	$X = \{X_1, \ldots, X_{T+1}\}\ X_t: S \rightarrow \{1, \ldots, N\}$
Output sequence	$O = \{O_1, \ldots, O_T\}\ o_t \in K$

$S_i \to S_j$ 转化概率

$S_i \to S_j$

output K

这次S_i，下一次S_j，吐出K

$$P(O_t = k \mid X_t = s_i, X_{t+1} = s_j) = b_{ijk}$$

Three fundamental questions for HMMs

- Given a model $\mu = \{A, B, \Pi\}$, how do we efficiently compute how likely a certain observations is, that is $P(O|\mu)$?
- Given the observation sequence O and a model μ, how do we choose a state sequence $\{X_1, \ldots, X_{T+1}\}$ that best explains the observations?
- Given an observation sequence O, and a space of possible models found by varying the model parameters $\mu = \{A, B, \Pi\}$, how do we find the model that best explains the observed data?

动态规划问题

Finding the probability of an observation

Given the observation sequence $O = \{O_1, \ldots, O_T\}$ and a model $\mu = \{A, B, \Pi\}$, we wish to know how to efficiently compute $P(O|\mu)$ – the probability of the observation given the model.

$$\alpha_i(t) = P(o_1 o_2 \ldots o_{t-1}, X_t = i | \mu)$$

Initialization

$$\alpha_i(1) = \pi_i ,\ 1 \le i \le N$$

第t次是status i的概率

Induction

$$\alpha_j(t+1) = \sum_{i=1}^{N} \boxed{\alpha_i(t)}\, \boxed{a_{ij}}\, \boxed{b_{ij o_t}},\ 1 \le t \le T,\ 1 \le j \le N$$

i→j下…的概率

i→j 的概率

Total

$$P(O|\mu) = \sum_{i=1}^{N} \alpha_i(T+1)$$

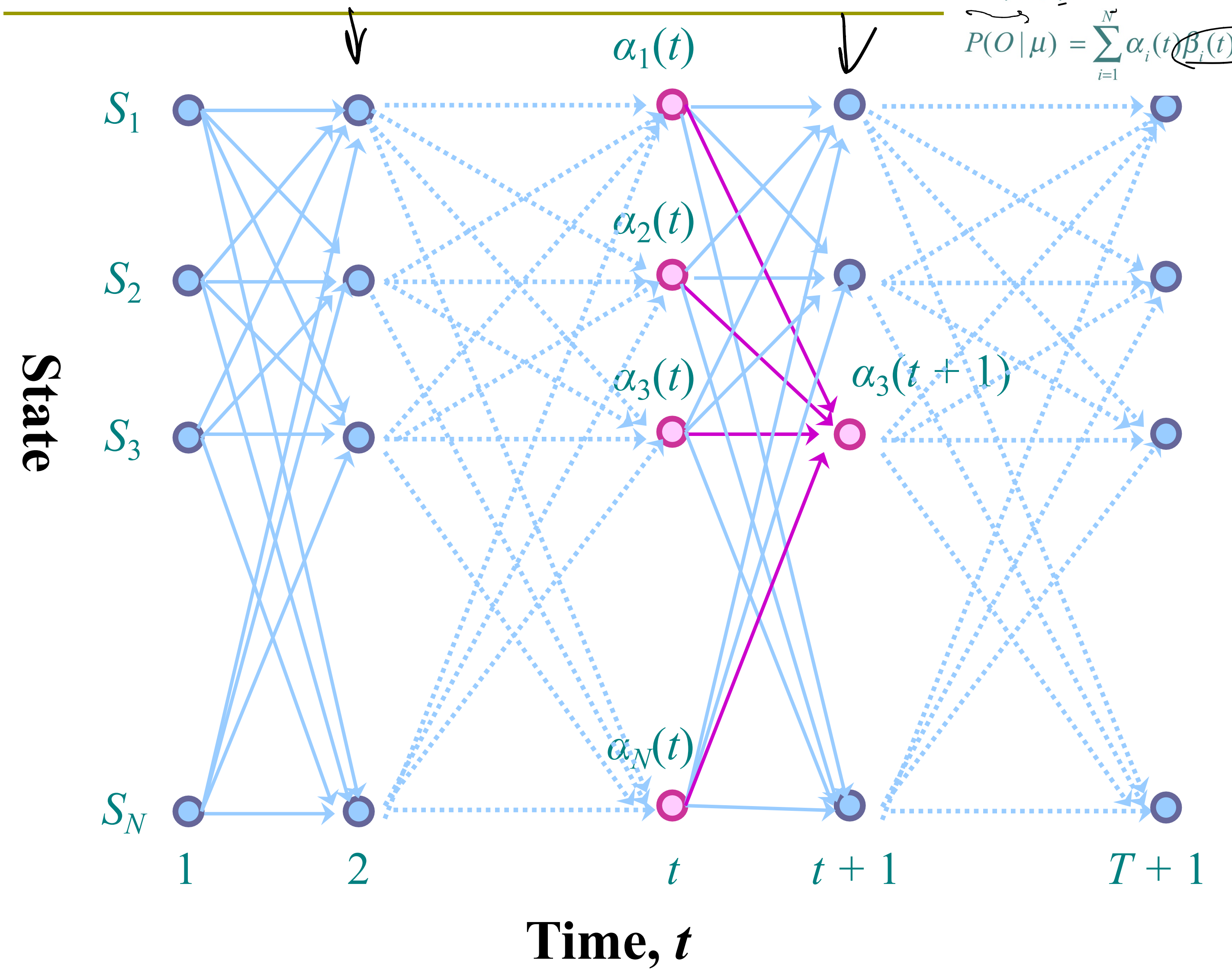
j j

Trellis (lattices)algorithm

$$\begin{aligned}
&P(O, \boxed{X_t = i} \mid \mu) \\
&= P(o_1 \cdots o_T, X_t = i \mid \mu) \\
&= P(o_1 \cdots o_{t-1}, X_t = i, o_t \cdots o_T \mid \mu) \\
&= P(o_1 \cdots o_{t-1}, X_t = i \mid \mu) \times P(o_t \cdots o_T \mid o_1 \cdots o_{t-1}, X_t = i, \mu) \\
&= P(o_1 \cdots o_{t-1}, X_t = i \mid \mu) \times P(o_t \cdots o_T \mid X_t = i, \mu) \\
&= \alpha_i(t)\beta_i(t)
\end{aligned}$$

$$P(O \mid \mu) = \sum_{i=1}^{N} \alpha_i(t)\beta_i(t), 1 \le t \le T+1$$

任何时刻



The backward procedure

Define backward variables

$\beta_i(t) = P(o_t \ldots o_{T+1}, X_{t=i} | \mu)$

Initialization

$\beta_i(T+1) = 1\, ,\; 1 \le i \le N$

Induction

$$\beta_i(t) = \sum_{j=1}^{N} a_{ij} b_{ij o_t} \beta_j(t+1), 1 \le t \le T, 1 \le j \le N$$

Total

$$P(O \mid \mu) = \sum_{i=1}^{N} \pi_i \beta_i(1)$$

Finding the best state sequence

Commonly we want to find the most likely complete path, that is:

$$\arg\max_{X} P(X \mid O, \mu)$$

To do this, it is sufficient to maximize for a fixed O:

$$\arg\max_{X} P(X, O \mid \mu)$$

An efficient trellis algorithm for computing this path is the ***Viterbi*** algorithm

$$\delta_j(t) = \max_{X_1 \cdots X_{t-1}} P(X_1 \cdots X_{t-1}, o_1 \cdots o_{t-1}, X_t = j \mid \mu)$$

Viterbi algorithm

Initialization

$$\delta_j(1) = \pi_j,\ 1 \le j \le N$$

Induction

$$\delta_j(t+1) = \max_{1 \le i \le N} \delta_i(t) a_{ij} b_{ijo_t},\ 1 \le j \le N$$

$$\psi_j(t+1) = \arg\max_{1 \le i \le N} \delta_i(t) a_{ij} b_{ijo_t},\ 1 \le j \le N$$

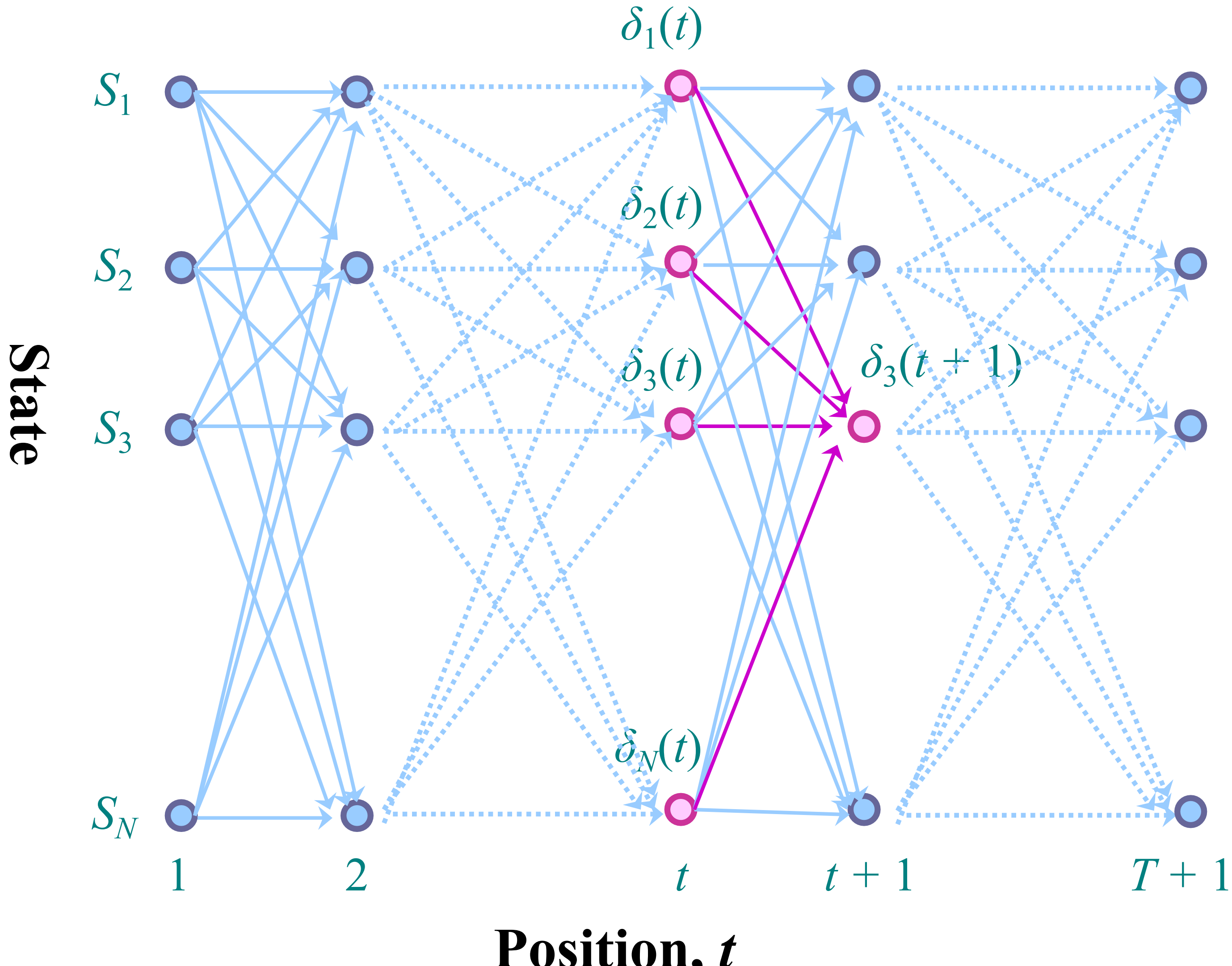
Termination

$$\hat{X}_{T+1} = \arg\max_{1 \le i \le N} \delta_i(T+1)$$

$$\hat{X}_t = \psi_{\hat{X}_{t+1}}(t+1)$$

$$P(\hat{X}) = \max_{1 \le i \le N} \delta_i(T+1)$$

Viterbi algorithm



The crazy soft drink machine

Transition	Cola Pref (CP)	Iced Tea Pref (IP)
Cola Pref (CP)	0.7	0.3
Iced Tea Pref (IP)	0.5	0.5

$P(O_k|S_i)$ 模型[illegible]

$b_{ijk} = P(O_k|S_{i\to j})$ [illegible]优化

Emission	Cola	Iced tea	Lemonade
Cola Pref (CP)	0.6	0.1	0.3
Iced Tea Pref (IP)	0.1	0.7	0.2

Initial	Cola Pref (CP)	Iced Tea Pref (IP)
π	0.4	0.6

Evaluation

Times	1	2	3
CP	0.6×0.4		
IP	0.1×0.6		

Transition	CP	IP
CP	0.7	0.3
IP	0.5	0.5

Emission	Cola	Iced tea	Lemonade
CP	0.6	0.1	0.3
IP	0.1	0.7	0.2

Initial	CP	IP
π	0.4	0.6

如果在第Times次是CP的
而当对应饮料的概率

$0.24 \times 0.7 \times 0.1$
$+ 0.06 \times 0.5 \times 0.1$

Evaluation

Times	1	2	3
CP	0.2400	0.0198	0.0149
IP	0.0600	0.0714	0.0083

0.0232



Transition	CP	IP
CP	0.7	0.3
IP	0.5	0.5

Emission	Cola	Iced tea	Lemonade
CP	0.6	0.1	0.3
IP	0.1	0.7	0.2

Initial	CP	IP
π	0.4	0.6

$= 0.2400 \times 0.7 \times 0.1 +$
$0.0600 \times 0.5 \times 0.1$
$= 0.0198$

How about    ?

bottom-up.

Decoding

0.24×0.7×0.1

阶段最大的

最后一个状态是cp.

top-down

Times	1	2	3
CP	0.2400	0.0168	0.0076
IP	0.0600	0.0504	0.0050

Times	1	2	3
CP		CP	IP
IP		CP	IP

CP

Transition	CP	IP
CP	0.7	0.3
IP	0.5	0.5

Emission	Cola	Iced tea	Lemonade
CP	0.6	0.1	0.3
IP	0.1	0.7	0.2

Initial	CP	IP
π	0.4	0.6

以前面一个状态

最大的

States sequence: CP IP CP

= 0.2400 × 0.7 × 0.1 = 0.0168

= 0.0600 × 0.5 × 0.1 = 0.0003

Parameter estimation

Given a certain observation sequence, we want to find the values of the model parameters $\mu = \{A, B, \Pi\}$ which best explain what we observed. Using ***Expectation Maximization*** (EM), that means we want to find the values that maximize $P(O|\mu)$:

t时刻是S_i

EM算法

$$
\begin{aligned}
& P(O, \boxed{X_t = i} \mid \mu) \\
& = P(o_1 \cdots o_T, X_t = i \mid \mu) \\
& = P(o_1 \cdots o_{t-1}, X_t = i, o_t \cdots o_T \mid \mu) \\
& = P(o_1 \cdots o_{t-1}, X_t = i \mid \mu) \times P(o_t \cdots o_T \mid o_1 \cdots o_{t-1}, X_t = i, \mu) \\
& = P(o_1 \cdots o_{t-1}, X_t = i \mid \mu) \times P(o_t \cdots o_T \mid X_t = i, \mu) \\
& = \alpha_i(t)\beta_i(t)
\end{aligned}
$$

$$
P(O \mid \mu) = \sum_{i=1}^{N} \alpha_i(t)\beta_i(t), 1 \le t \le T+1
$$

随机给定μ

↳ 算出 θ的期望

步骤相同 ⇒ 不断打乱随机，还要保持归一性

进行对μ的调整

θ的期望向φ之以收敛.

Parameter estimation

$$\gamma_i(t) = P(X_t = i \mid O, \mu)$$

$$= \frac{P(X_t = i, O \mid \mu)}{P(O \mid \mu)}$$

t

$$= \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)}$$

$$p_t(i,j) = P(X_t = i, X_{t+1} = j \mid O, \mu)$$

$$= \frac{P(X_t = i, X_{t+1} = j, O \mid \mu)}{P(O \mid \mu)}$$

$$= \frac{\alpha_i(t)\boxed{a_{ij}b_{ijo_t}}\beta_j(t+1)}{\sum_{m=1}^{N}\sum_{n=1}^{N}\alpha_m(t)\boxed{a_{mn}b_{mno_t}}\beta_n(t+1)}$$

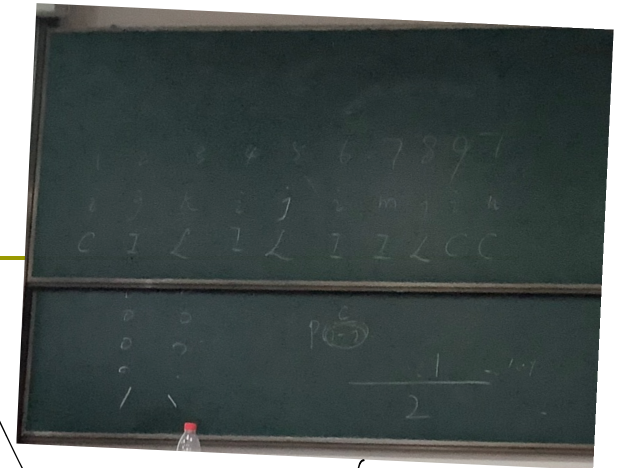
Note that:

在t次status处于i的概率

$$\gamma_i(t) = \sum_{j=1}^{N} p_t(i,j)$$

在t次status从i→j的概率

Parameter estimation



$$\hat{\pi}_i = \text{Expected frequency in state } i \text{ at time } t = 1$$

$$= \gamma_i(1)$$

$$\hat{a}_{ij} = \frac{\text{Expected number of transitions from state } i \text{ to } j}{\text{Expected number of transitions from state } i}$$

$$= \frac{\sum_{t=1}^{T} p_t(i,j)}{\sum_{t=1}^{T} \gamma_i(t)}$$

在t时从 status 为i —> j 的概率

在t时从 status 为i 的概率

$$\hat{b}_{ijk} = \frac{\text{Expected number of transitions from state } i \text{ to } j \text{ with } k \text{ observed}}{\text{Expected number of transitions from state } i \text{ to } j}$$

t时刻观测到的为 O_k

$$= \frac{\sum_{\{t:o_t=k,1\le t\le T\}} p_t(i,j)}{\sum_{t=1}^{T} p_t(i,j)}$$

$$P(O \mid \hat{\mu}) \ge P(O \mid \mu)$$