基于HMM的语音识别

洪青阳

纲要

6.1 建模单元

6.2 发音过程与HMM状态

6.3 串接HMM

6.4 固定语法的识别

6.5 随机语法的识别

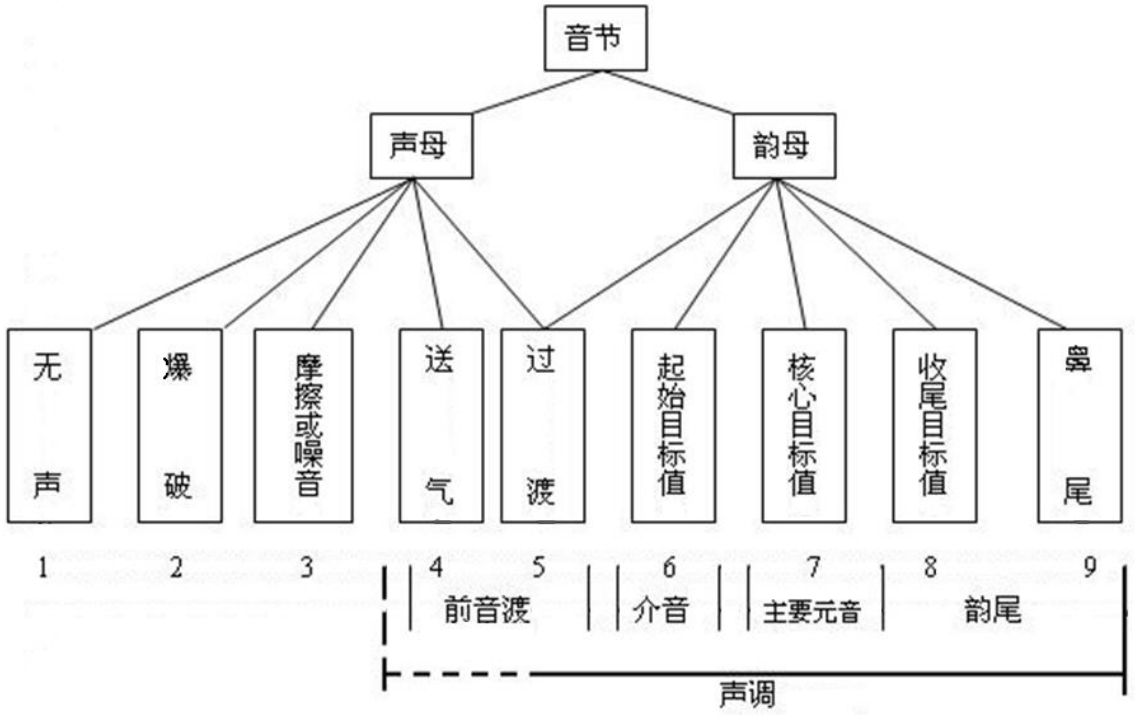
6.6 音素的上下文建模

6.7 本章小结

6.1 建模单元

- 整词(Word)
  - 词
  - 音节(单字)：汉语约400个

- 音素(phone)
  - 声母（辅音）
  - 韵母（元音）



声韵母建模—原始划分

声母集(21):

b p m f d t n l g k h j q x zh ch sh r z c s

注：y和w为韵头(yi-i, wu-u)，不划入声母集合。

韵母集(36):

单韵母6：a o e i u ü

复韵母14：ai ao ei er ia ie iao iou ou ua uai uei uo üe

鼻韵母16：an ian uan üan en in uen ün ang iang uang eng ing ueng ong iong

声韵母建模—改造划分

声母集(21+6)：

b p m f d t n l g k h j q x zh ch sh r z c s

新增：_a _o _e _y _w _v (用于没有声母的音节a o e yi wu yu)

韵母集(36+2)：

单韵母9：a o e i(ti/pi/bi) i1(zi/ci/si) i2(zhi/chi/shi) u v(ü)

复韵母14：ai ao ei er ia ie iao iou ou ua uai uei uo ve(üe)

鼻韵母16：an ian uan van(üan) en in uen vn(ün) ang iang uang eng ing ueng ong iong

声韵母建模（带声调）

声母集(27)：

b p m f d t n l g k h j q x zh ch sh r z c s _a _o _e _y _w _v

带声调韵母集(200)：

a1 a2 a3 a4 a5 ai1 ai2 ai3 ai4 ai5 an1 an2 an3 an4 an5 ang1 ang2 ang3 ang4 ang5 ao1 ao2 ao3 ao4 ao5 e1 e2 e3 e4 e5 ei1 ei2 ei3 ei4 ei5 en1 en2 en3 en4 en5 eng1 eng2 eng3 eng4 eng5 er1 er2 er3 er4 er5 i1 i2 i3 i4 i5 ia1 ia2 ia3 ia4 ia5 ian1 ian2 ian3 ian4 ian5 iang1 iang2 iang3 iang4 iang5 iao1 iao2 iao3 iao4 iao5 ie1 ie2 ie3 ie4 ie5 in1 in2 in3 in4 in5 ing1 ing2 ing3 ing4 ing5 io1 io2 io3 io4 io5 iong1 iong2 iong3 iong4 iong5 iou1 iou2 iou3 iou4 iou5 ix1 ix2 ix3 ix4 ix5 iy1 iy2 iy3 iy4 iy5 iz1 iz2 iz3 iz4 iz5 o1 o2 o3 o4 o5 ong1 ong2 ong3 ong4 ong5 ou1 ou2 ou3 ou4 ou5 u1 u2 u3 u4 u5 ua1 ua2 ua3 ua4 ua5 uai1 uai2 uai3 uai4 uai5 uan1 uan2 uan3 uan4 uan5 uang1 uang2 uang3 uang4 uang5 uei1 uei2 uei3 uei4 uei5 uen1 uen2 uen3 uen4 uen5 ueng1 ueng2 ueng3 ueng4 ueng5 uo1 uo2 uo3 uo4 uo5 v1 v2 v3 v4 v5 van1 van2 van3 van4 van5 ve1 ve2 ve3 ve4 ve5 vn1 vn2 vn3 vn4 vn5

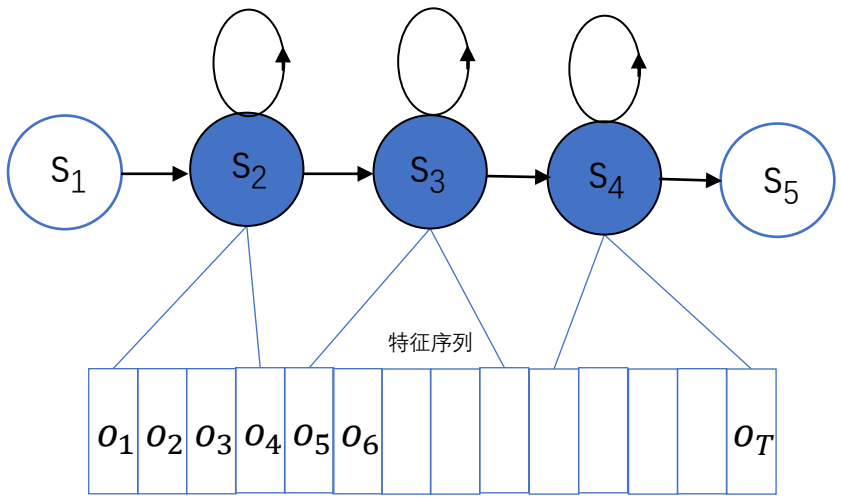
哎哟 _a ai1 _y io1

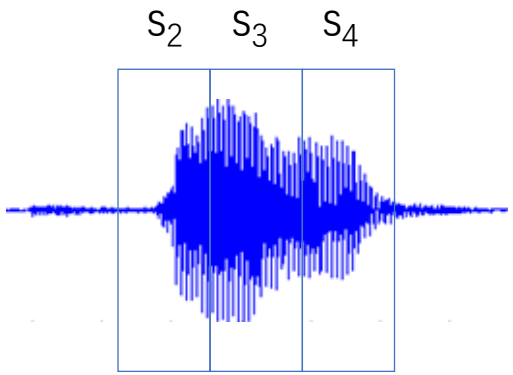
一 _y i1
一丘之貉 _y i1 q iou1 zh ix1 h e2
一丝一毫 _y i1 s iy1 _y i1 h ao2
一日 _y i1 r iz4

汉语建模单元对比

建模单元	模型数目	可训练性	稳定性	应用情况
音节	409个	一般	好	较少
音素(元音/辅音)	32个	好	一般	较少
声韵母	65-67个	较好	好	较普遍
声韵母(声调)	227个	好	较好	很普遍

6.2 发音过程与HMM状态





发音过程的抽象描述：

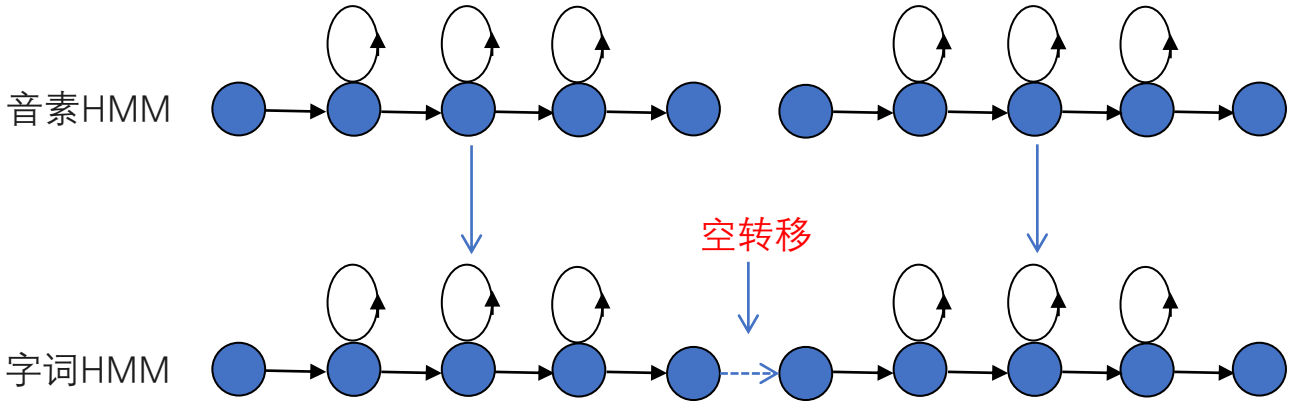
HMM状态起始于S_1，结束于S_5，且只能向自身或向右转移。

S_1模拟发音单元的起始；S_2、S_3、S_4模拟发音单元的发声过程；S_5模拟发音单元的结束。

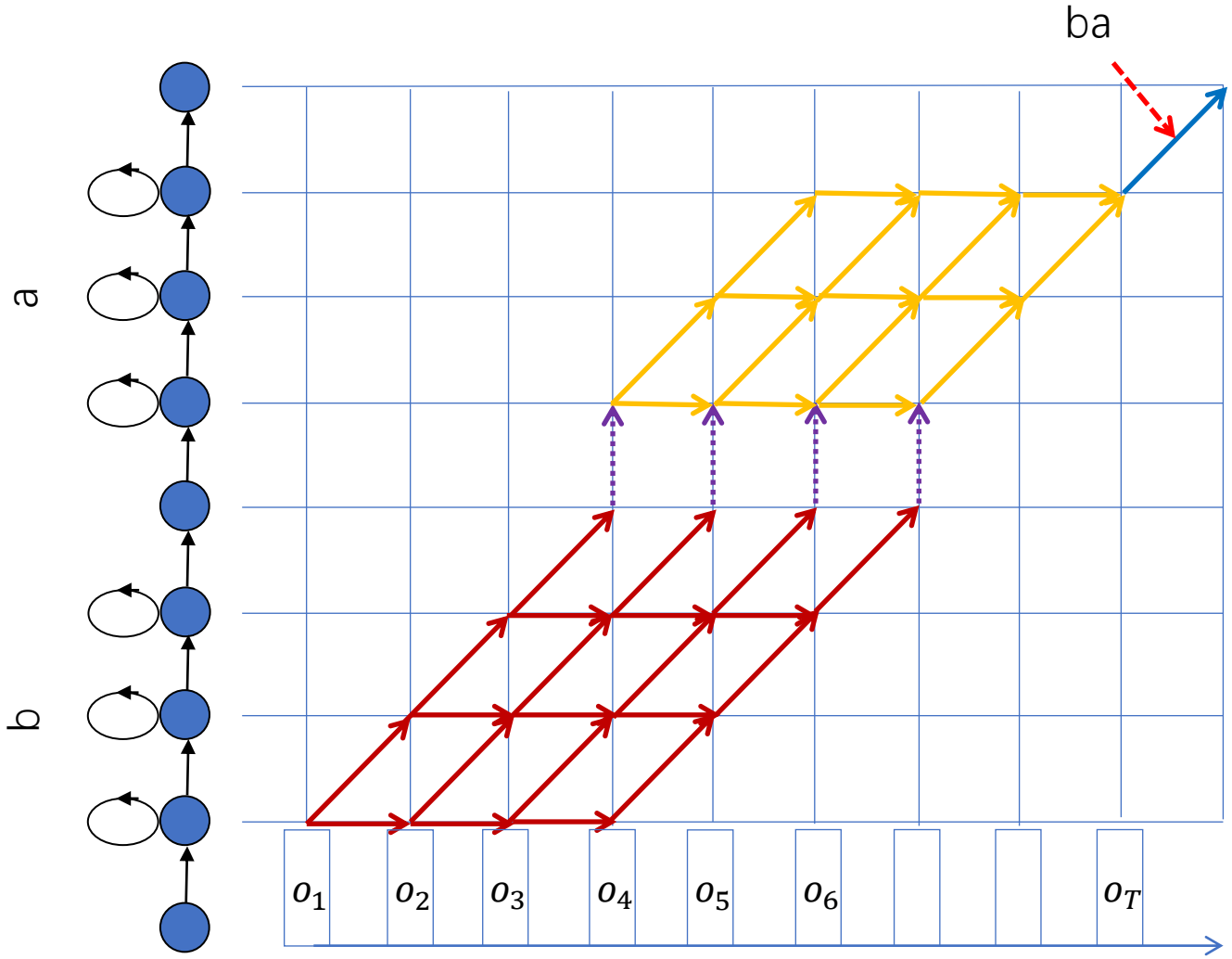
每个状态只能向自身或向右转移。

6.3 串接HMM

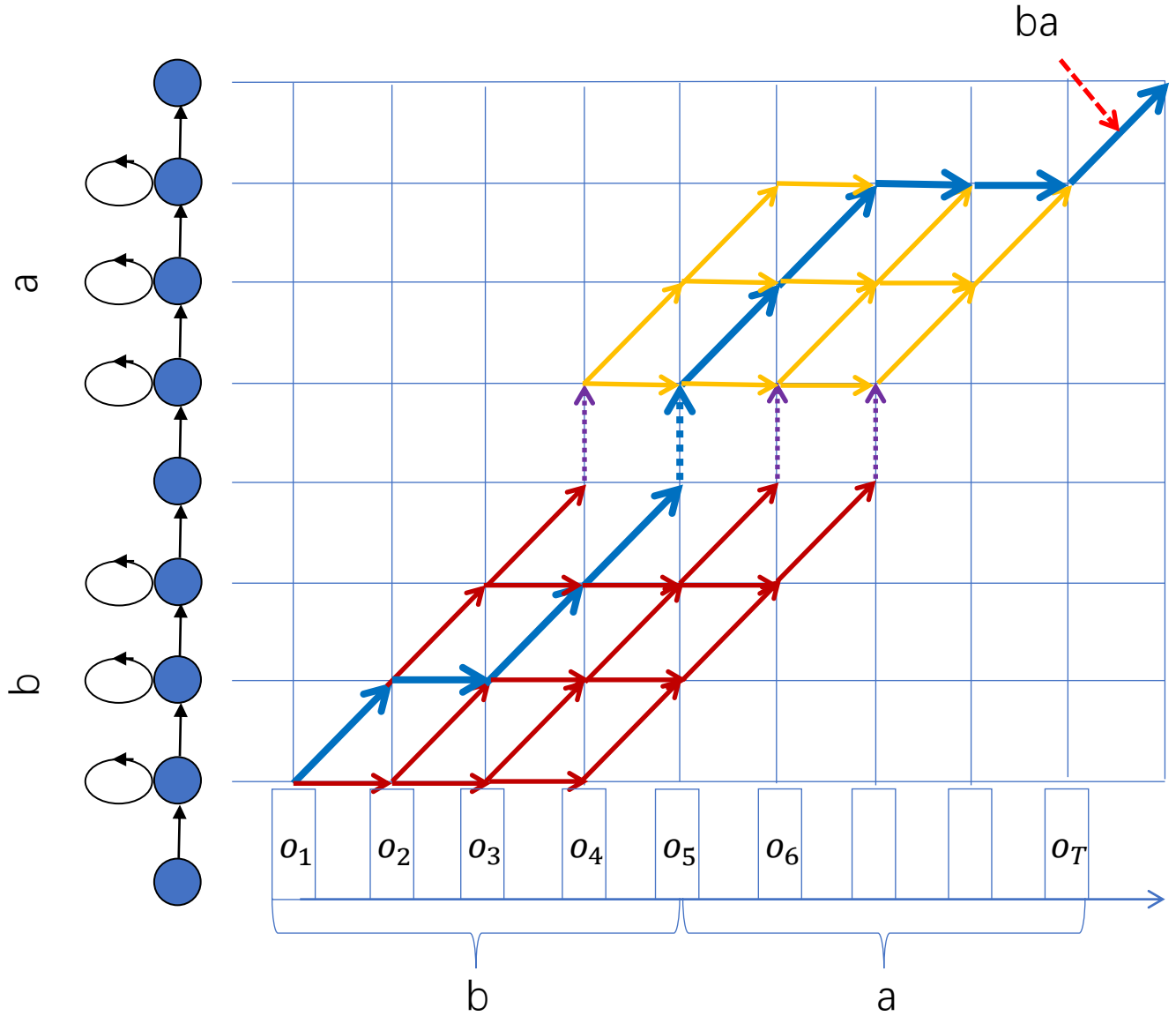
- 在连续语音识别中，字词的HMM都是由子词单元(音素)的HMM连接形成的。一般在连接时，一个子词单元HMM的终止状态和相邻基元HMM的初始状态相连接，这种连接产生的转移弧就是空转移，如下图所示：



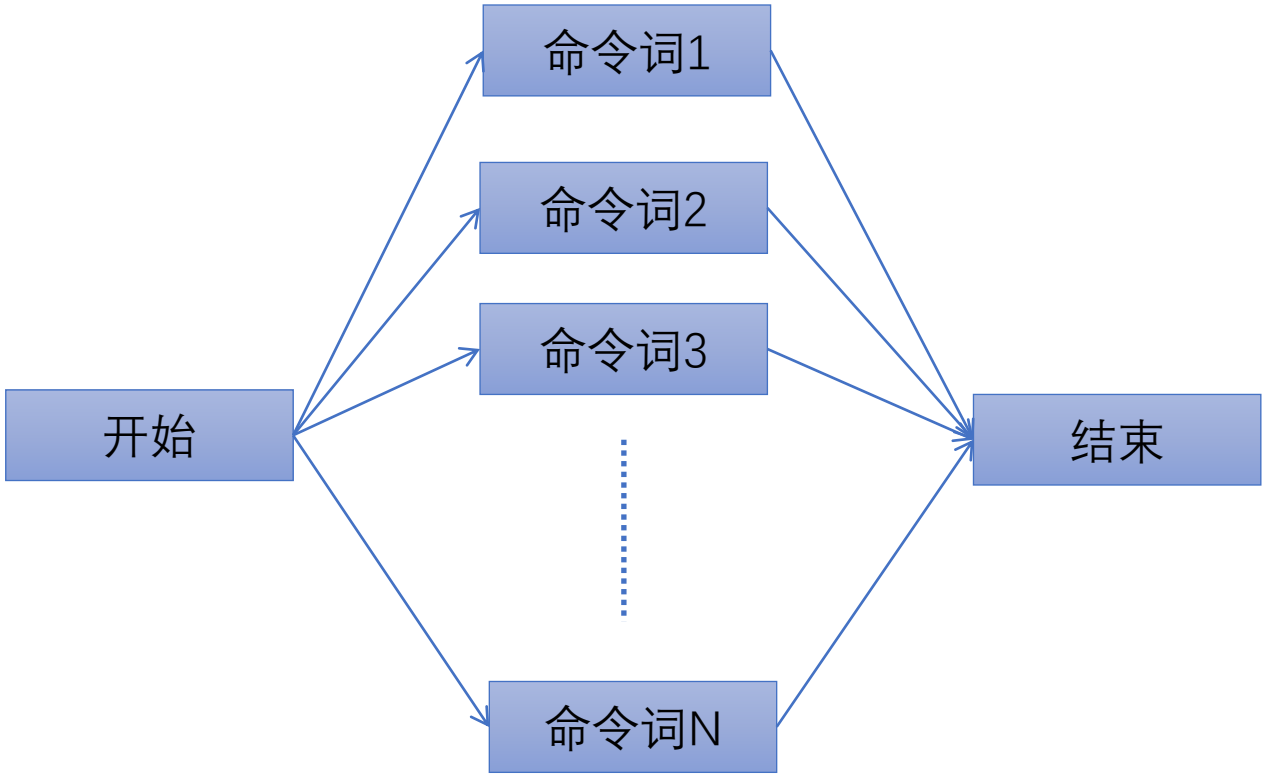
串接HMM的Viterbi识别



串接HMM的Viterbi识别



命令词识别

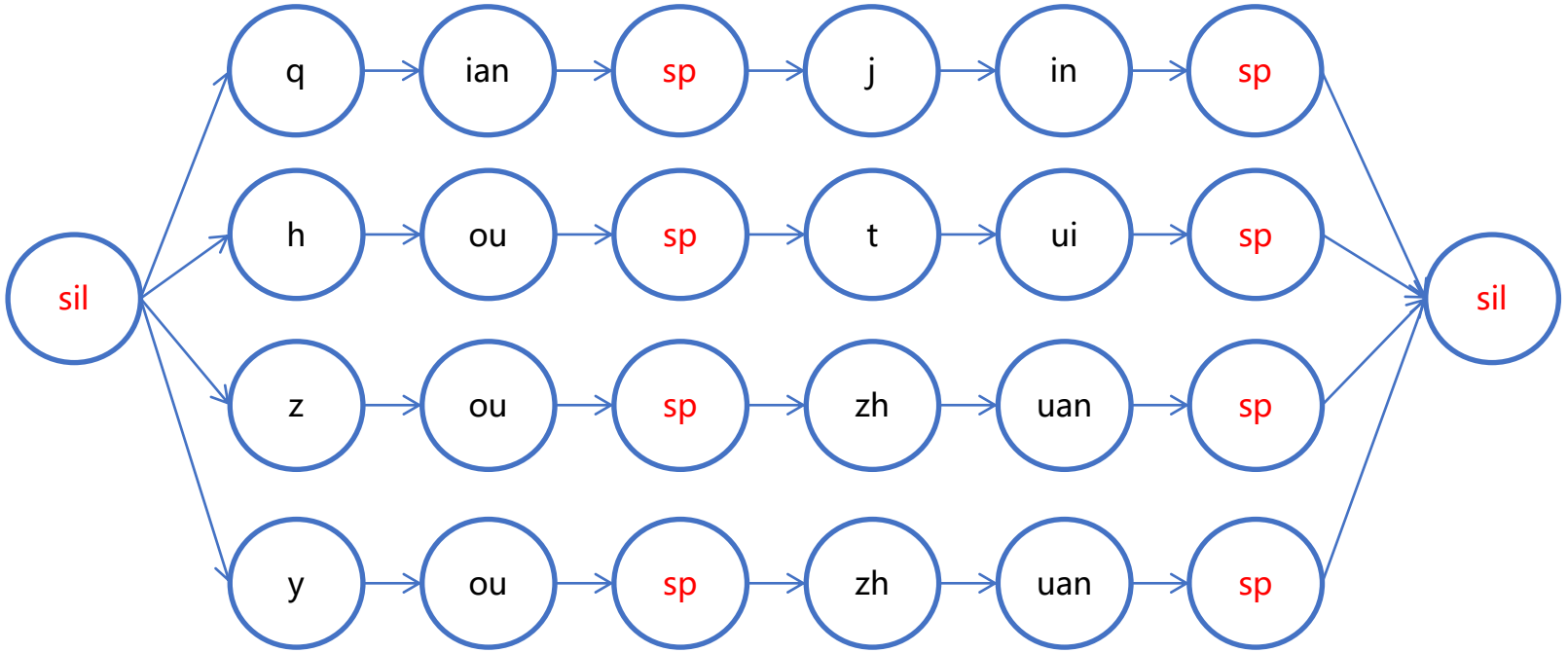


命令词

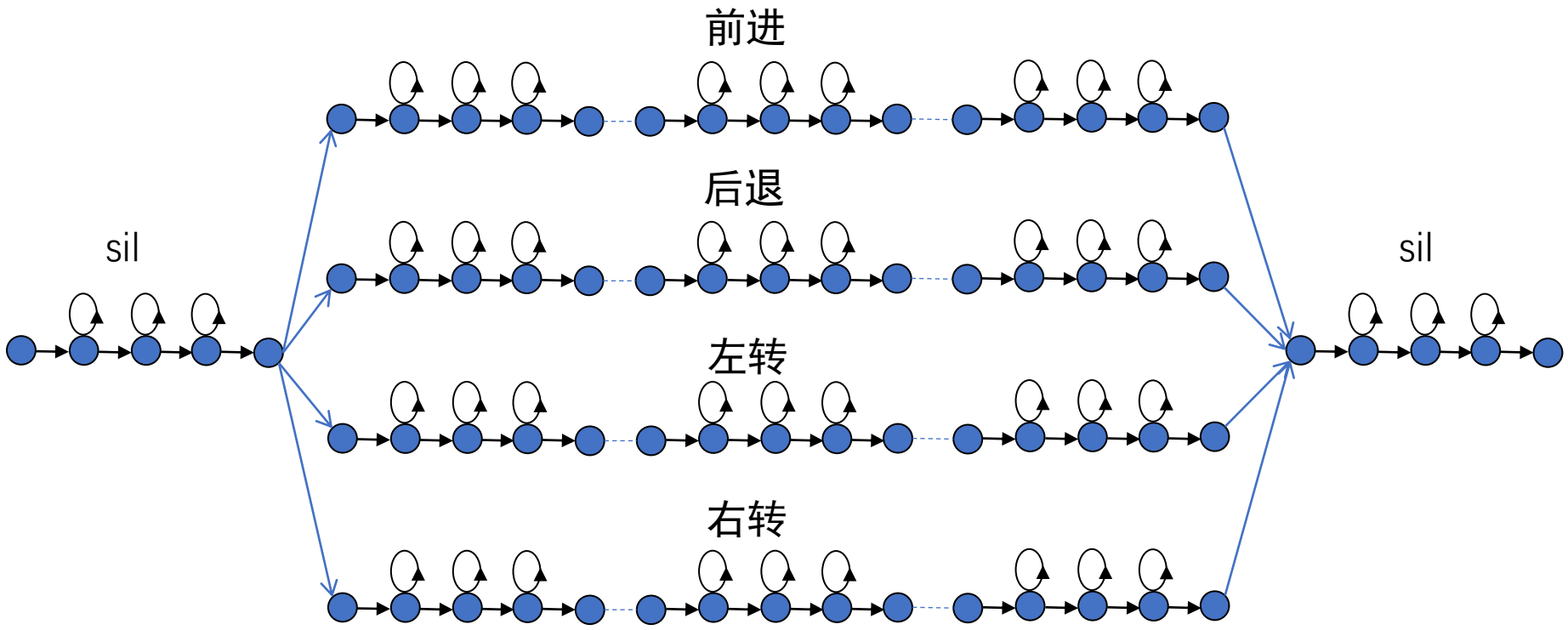
- 前进（qian jin）
- 后退（hou tui）
- 左转（zou zhuan）
- 右转（you zhuan）

识别网络

- 将字、词转化为对应的发音序列
- 增加静音（sil）和停顿（sp）节点，以适应语音前后的静音



对应的HMM网络



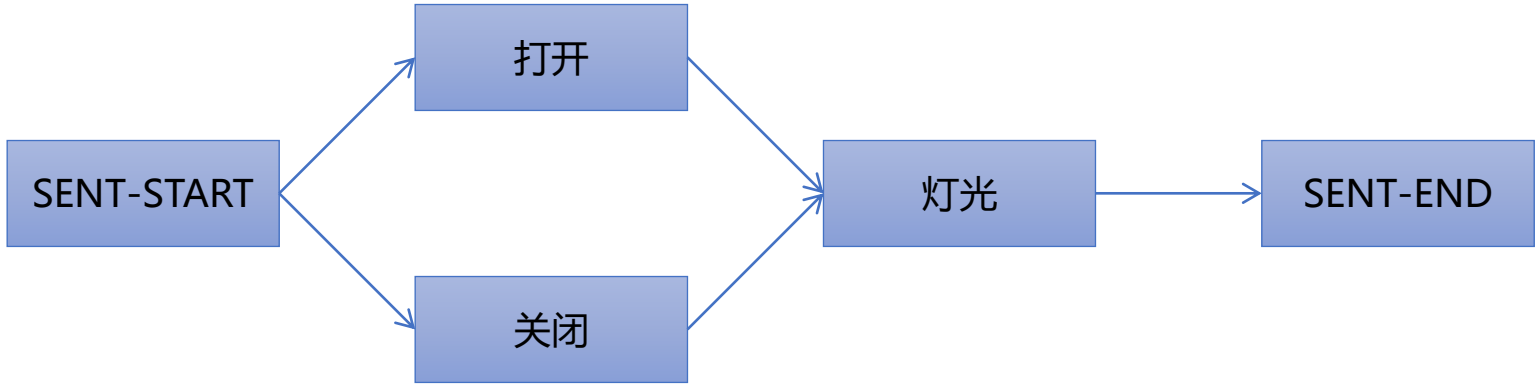
6.4 固定语法的识别

$action = 打开 | 关闭;

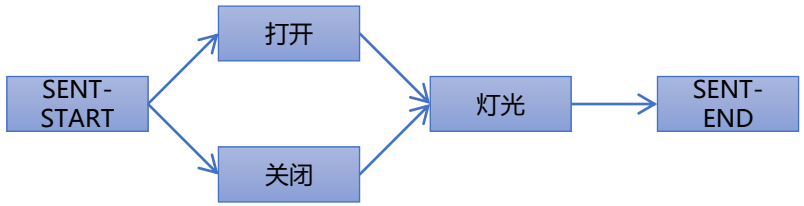
$object = 灯光;

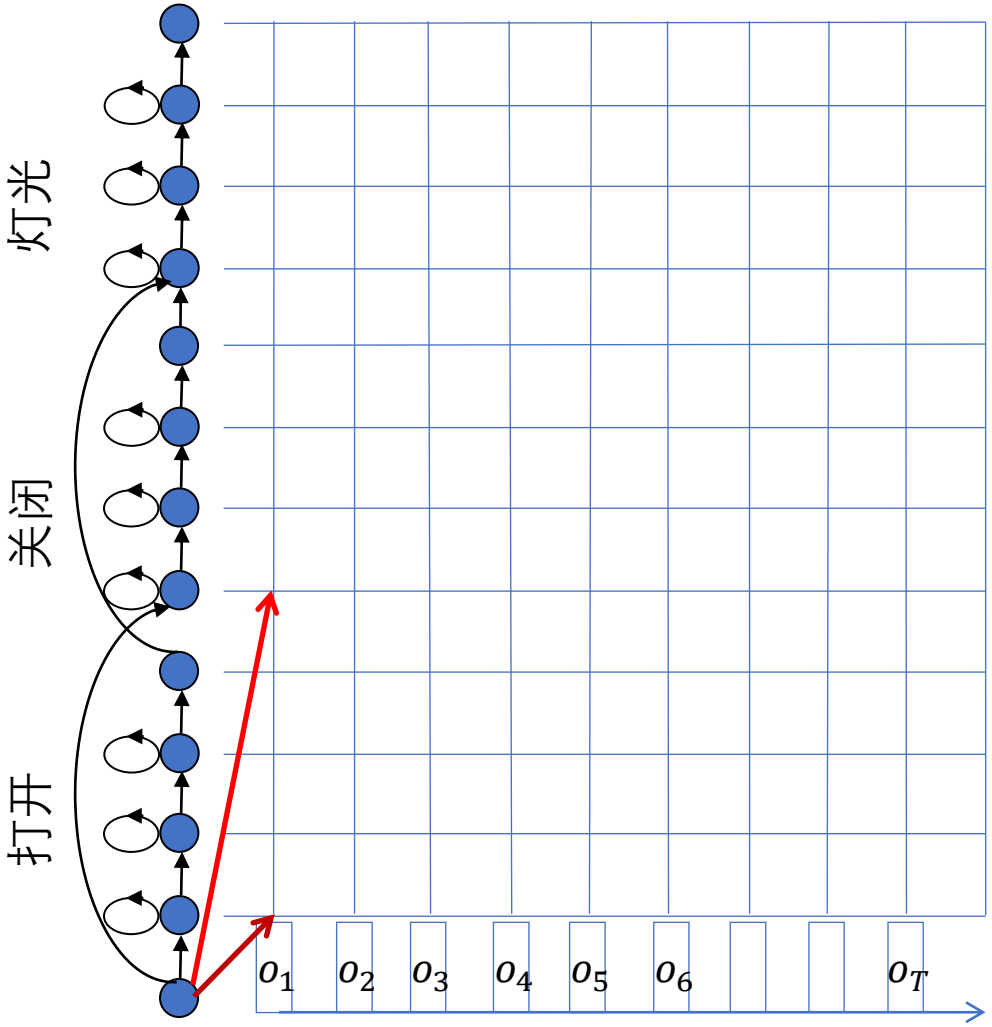
(SENT-START ($action $object) SENT-END)

识别网络

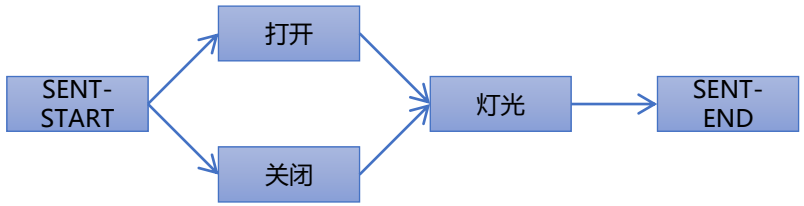


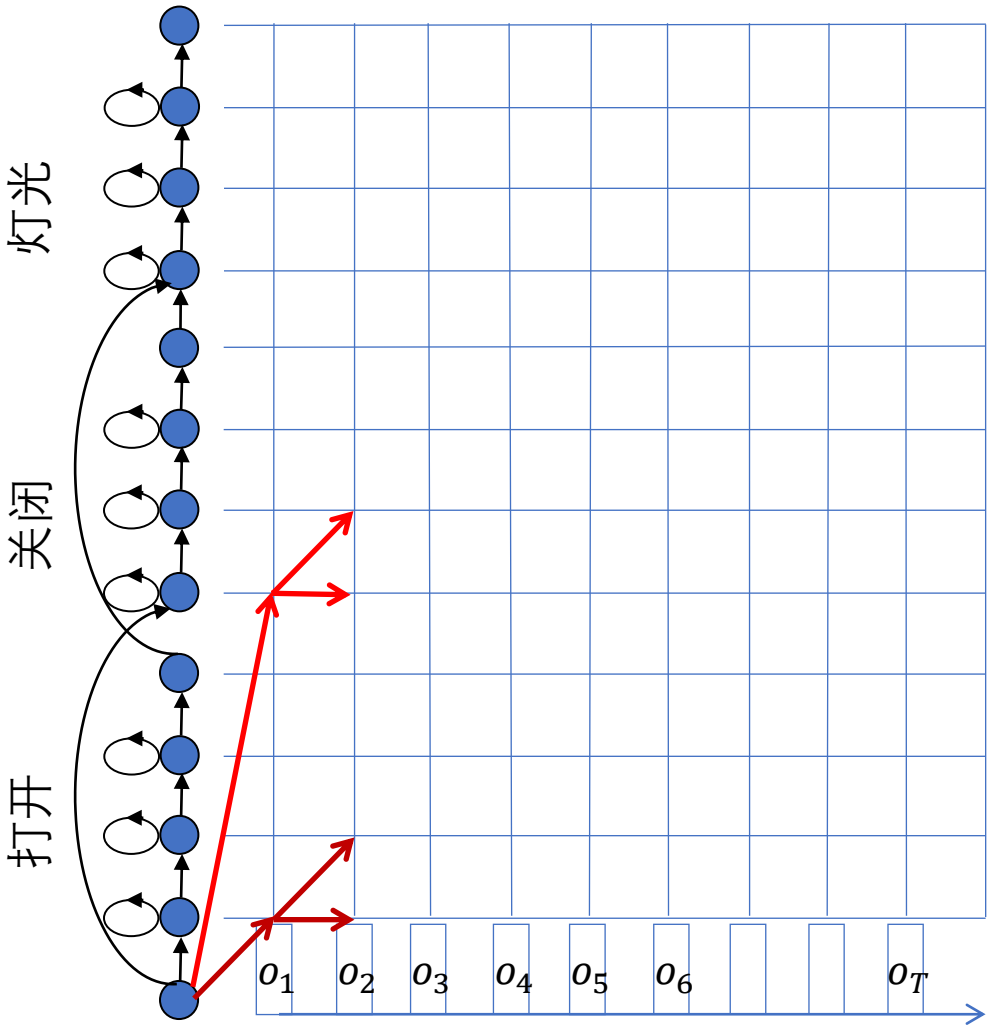
解码过程



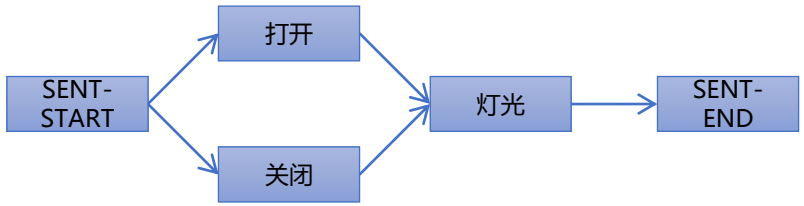


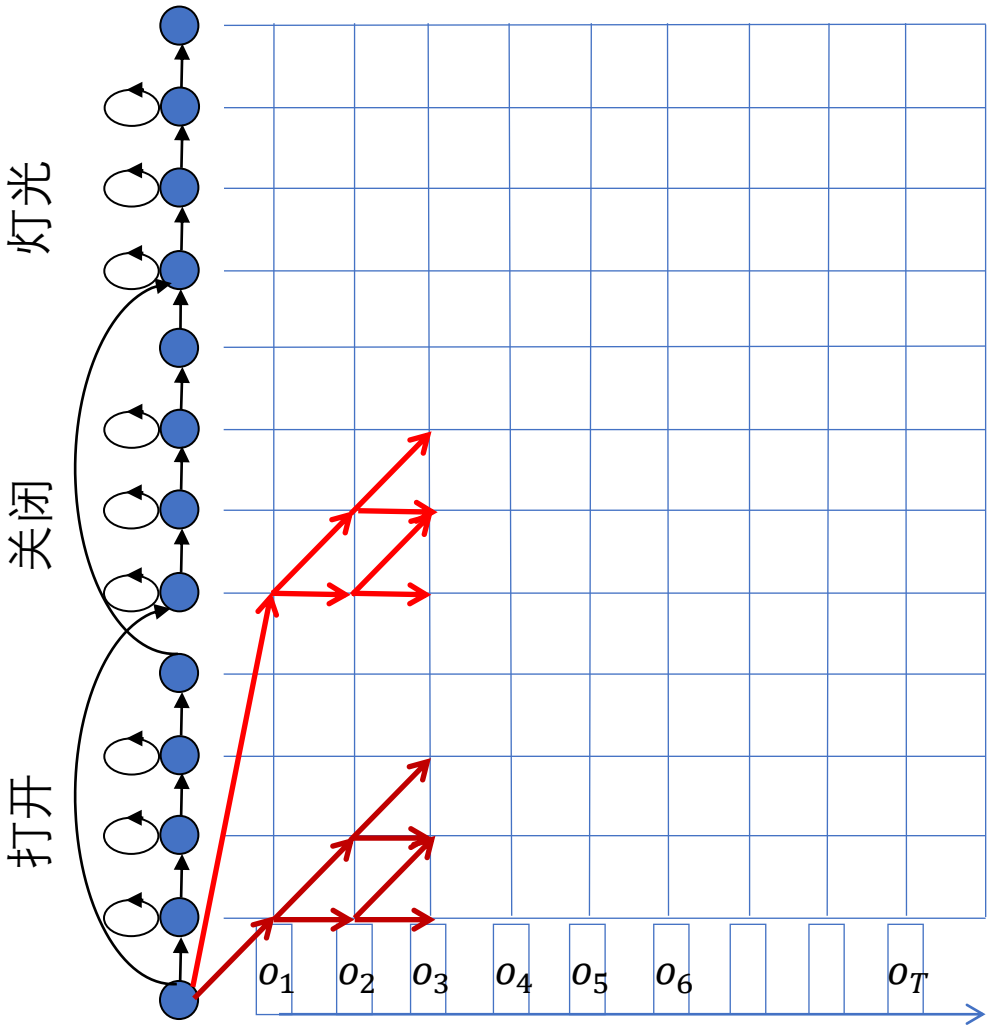
解码过程



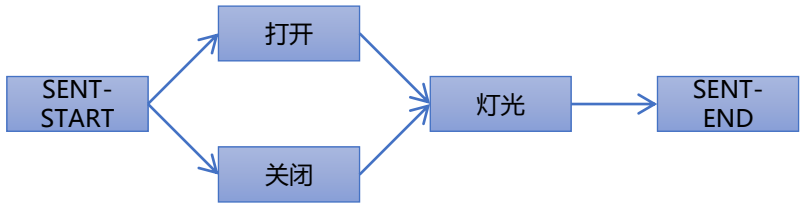


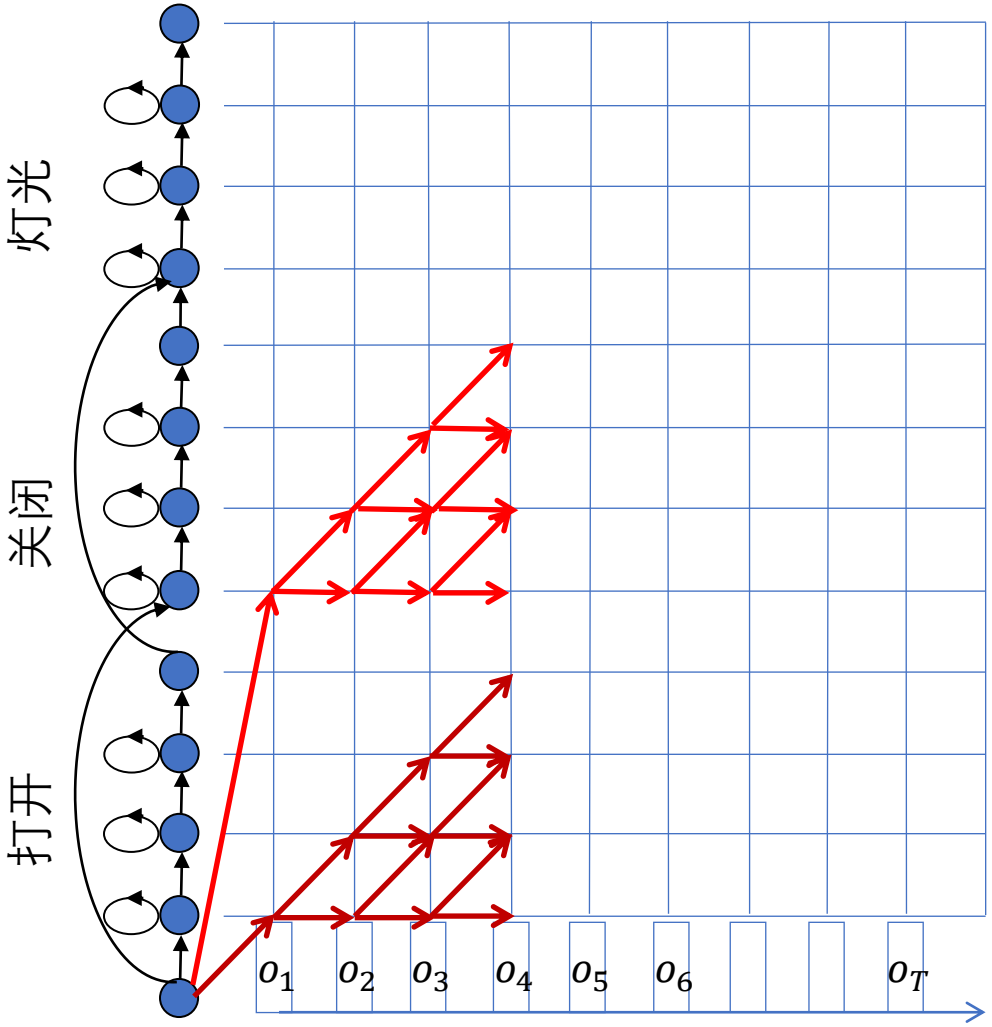
解码过程



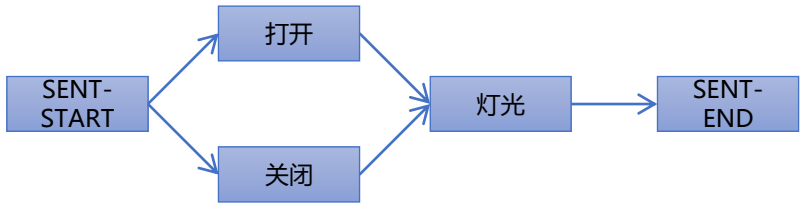


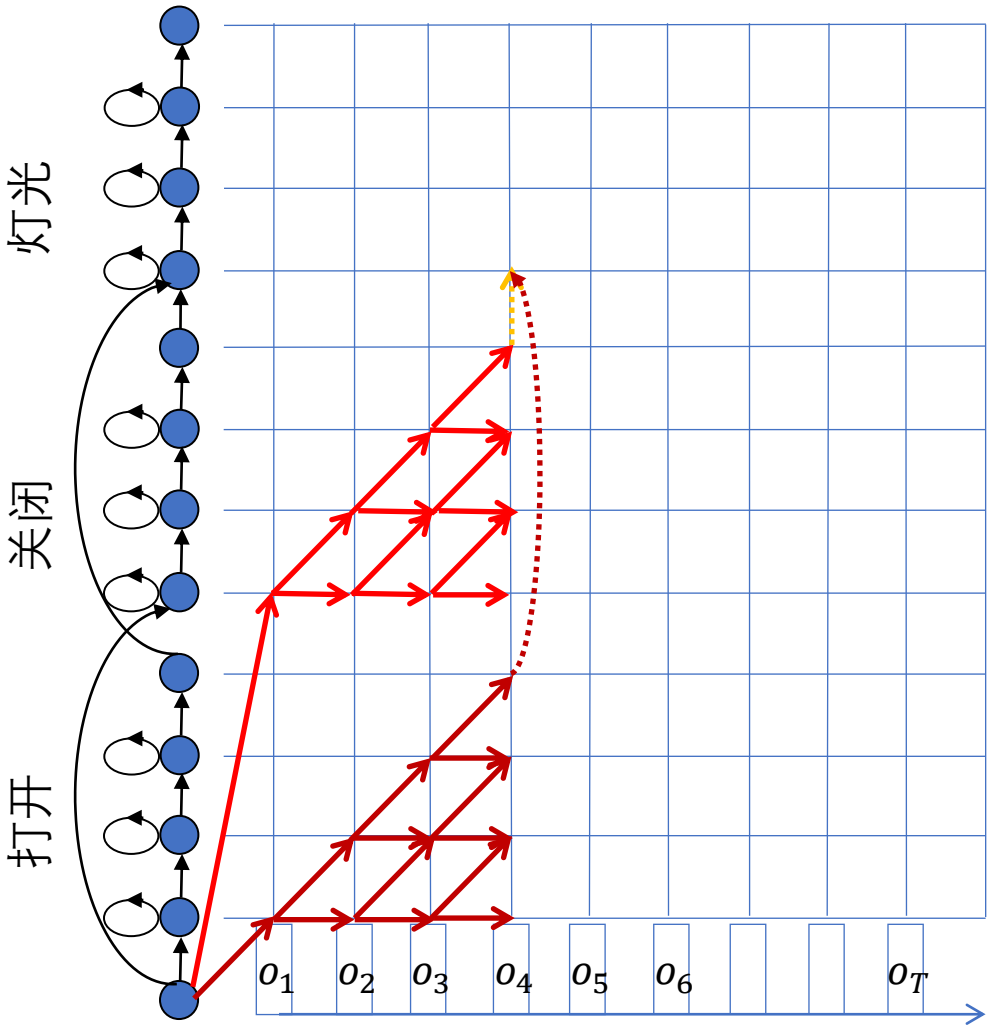
解码过程



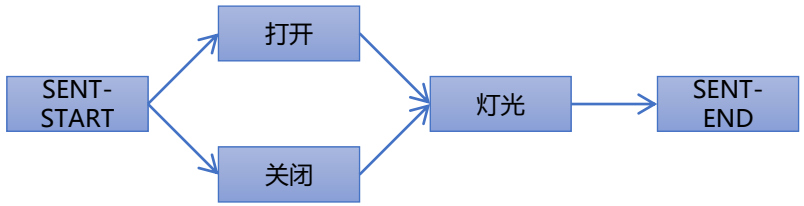


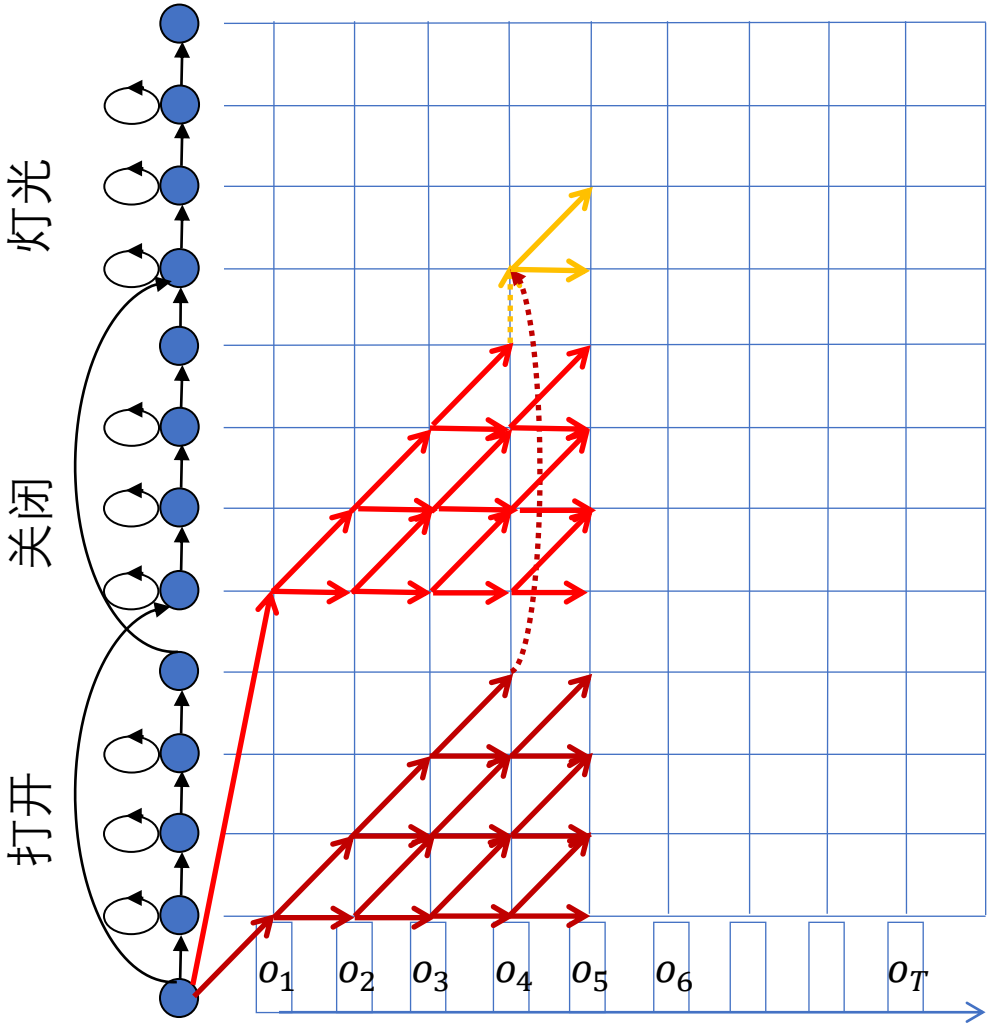
解码过程



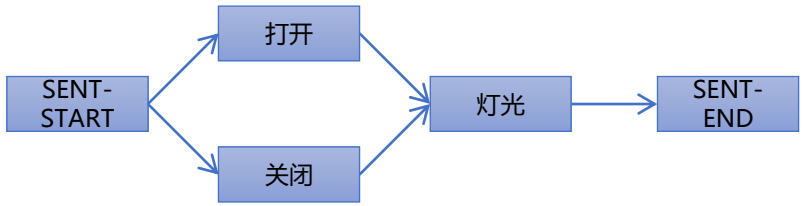


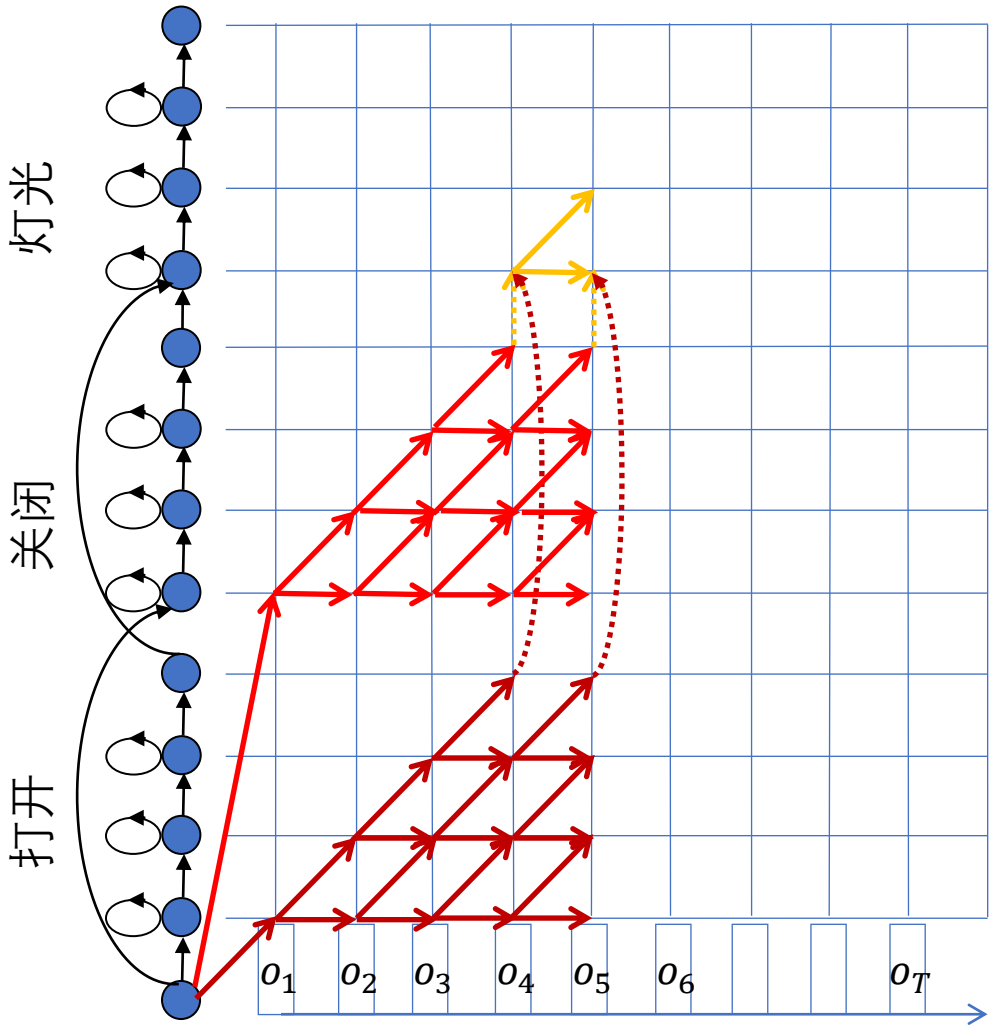
解码过程



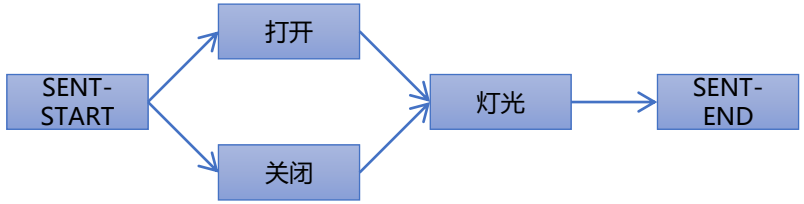


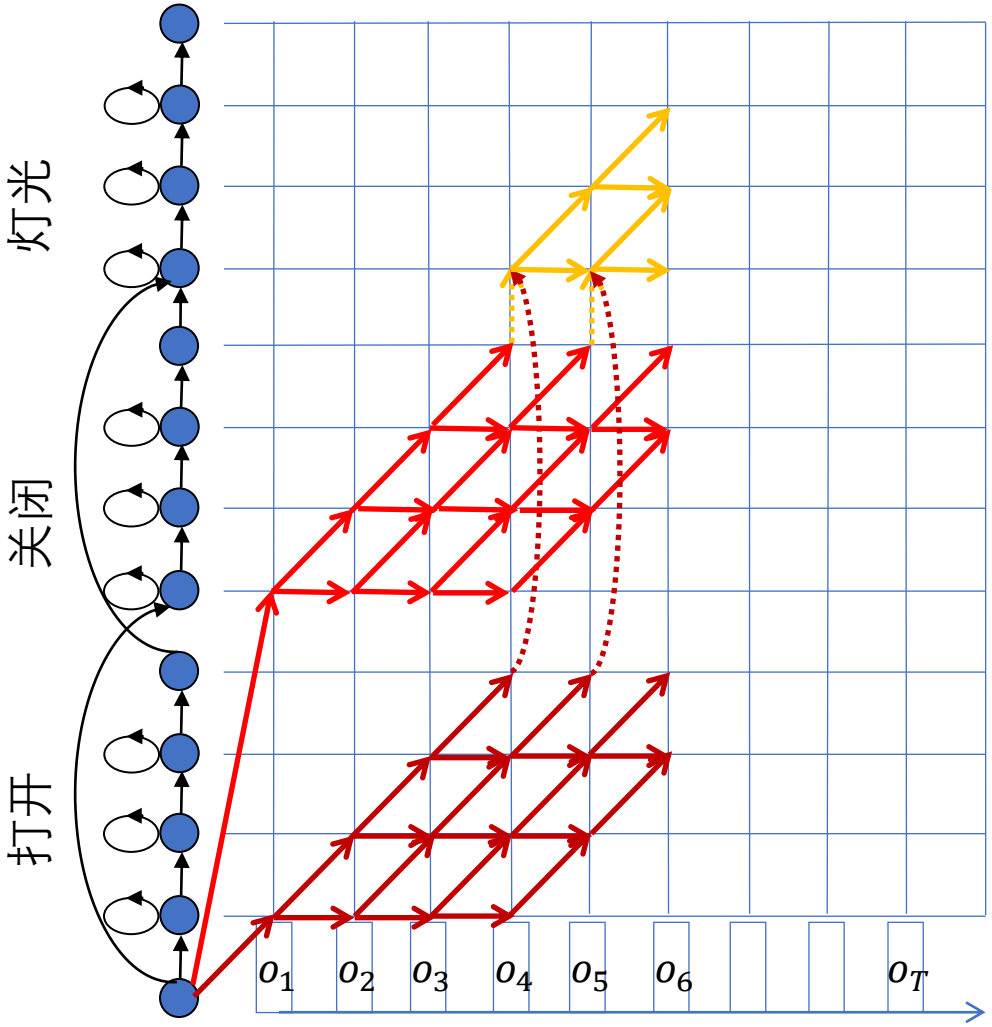
解码过程



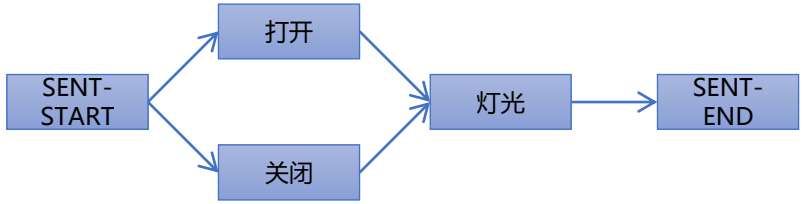


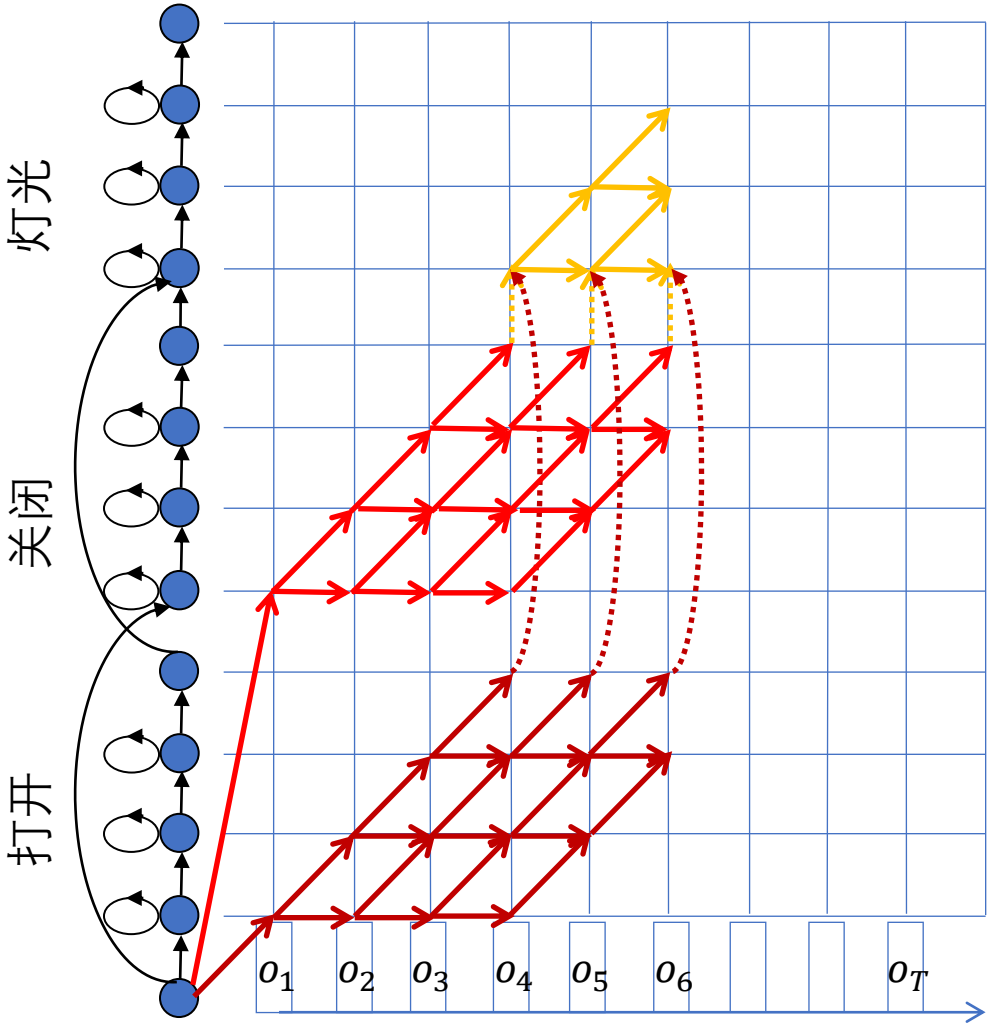
解码过程



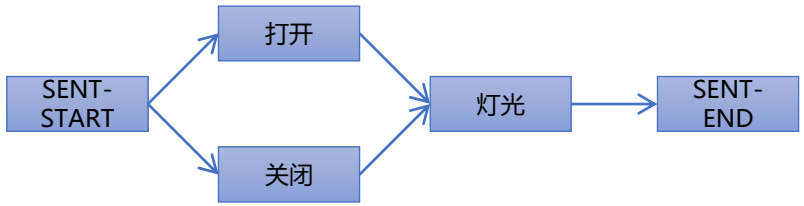


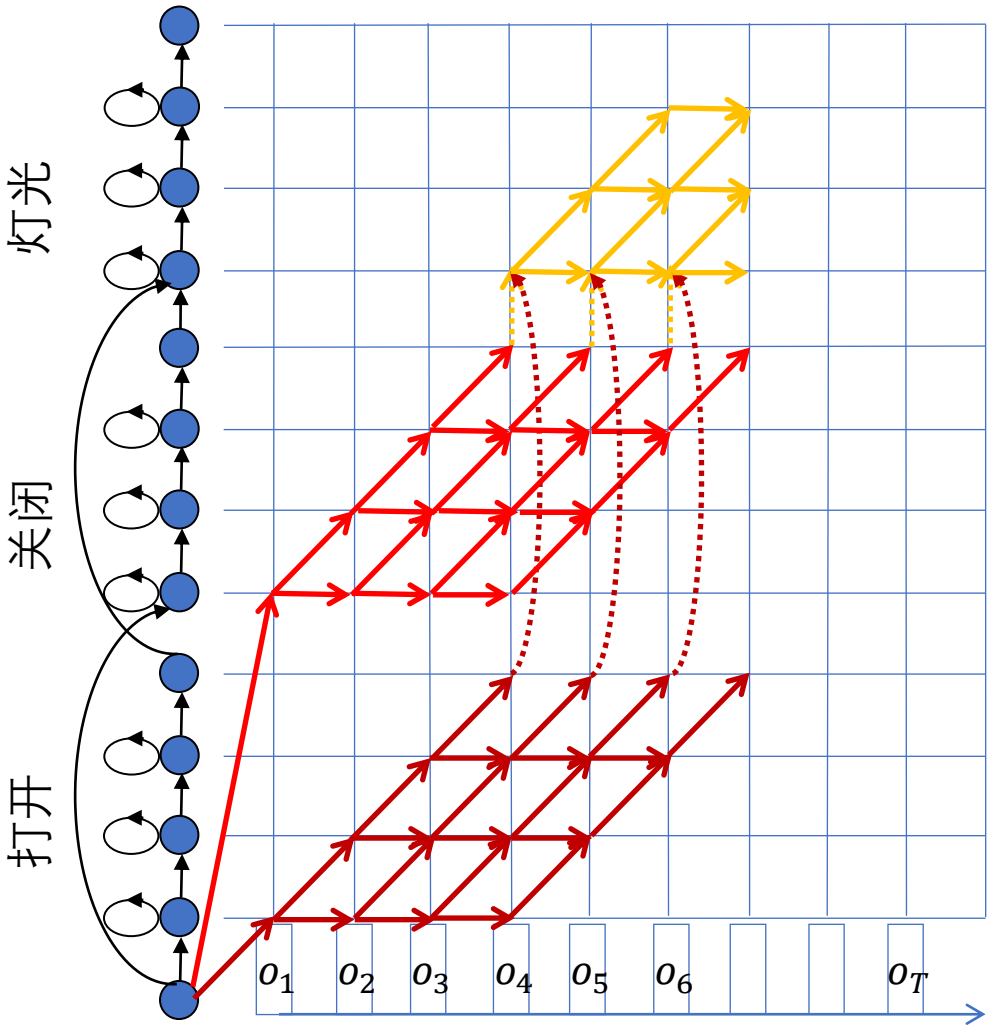
解码过程



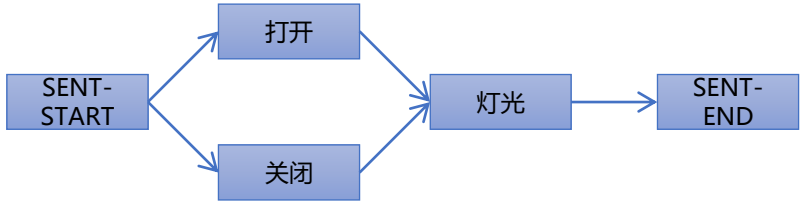


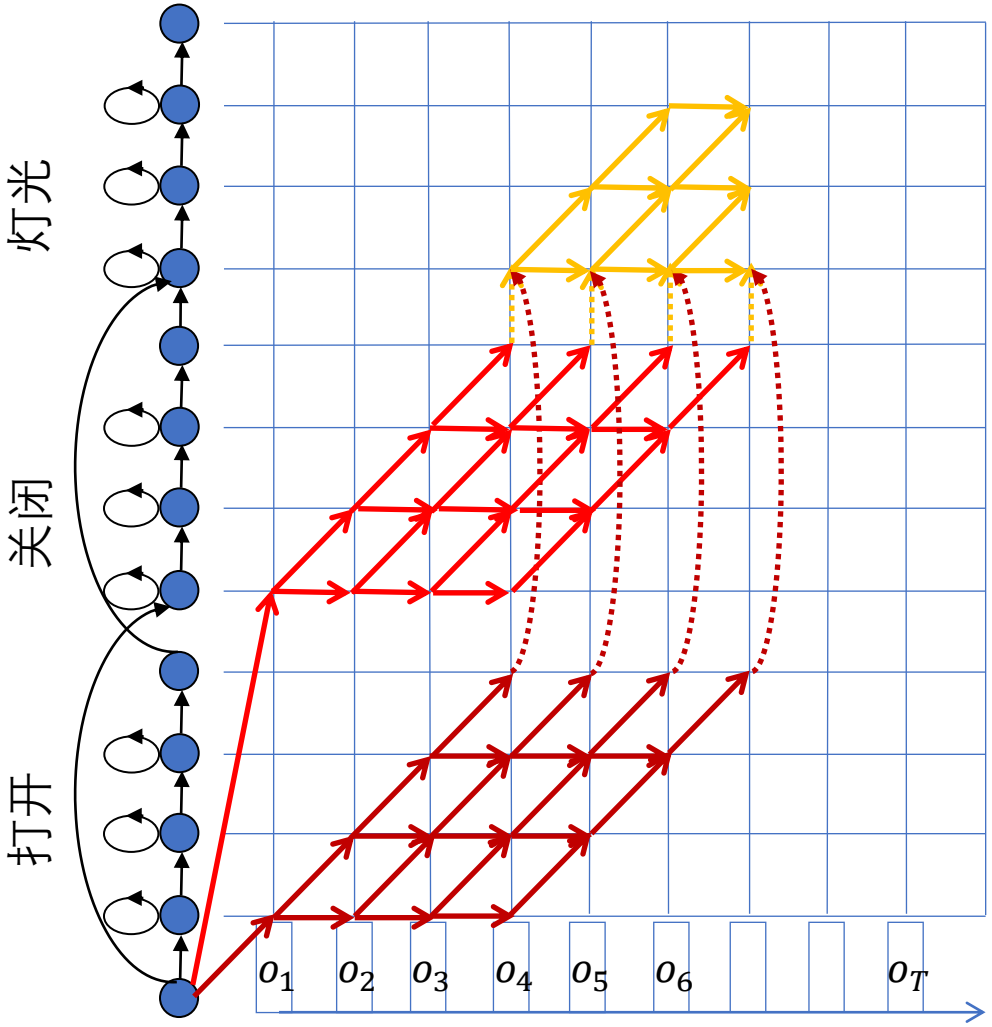
解码过程



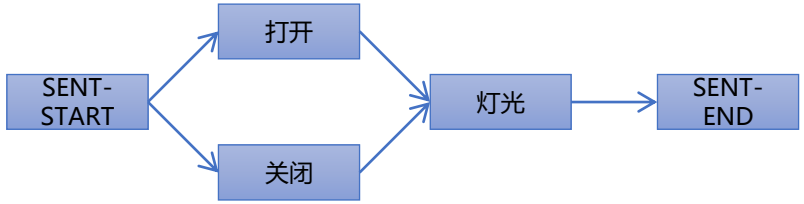


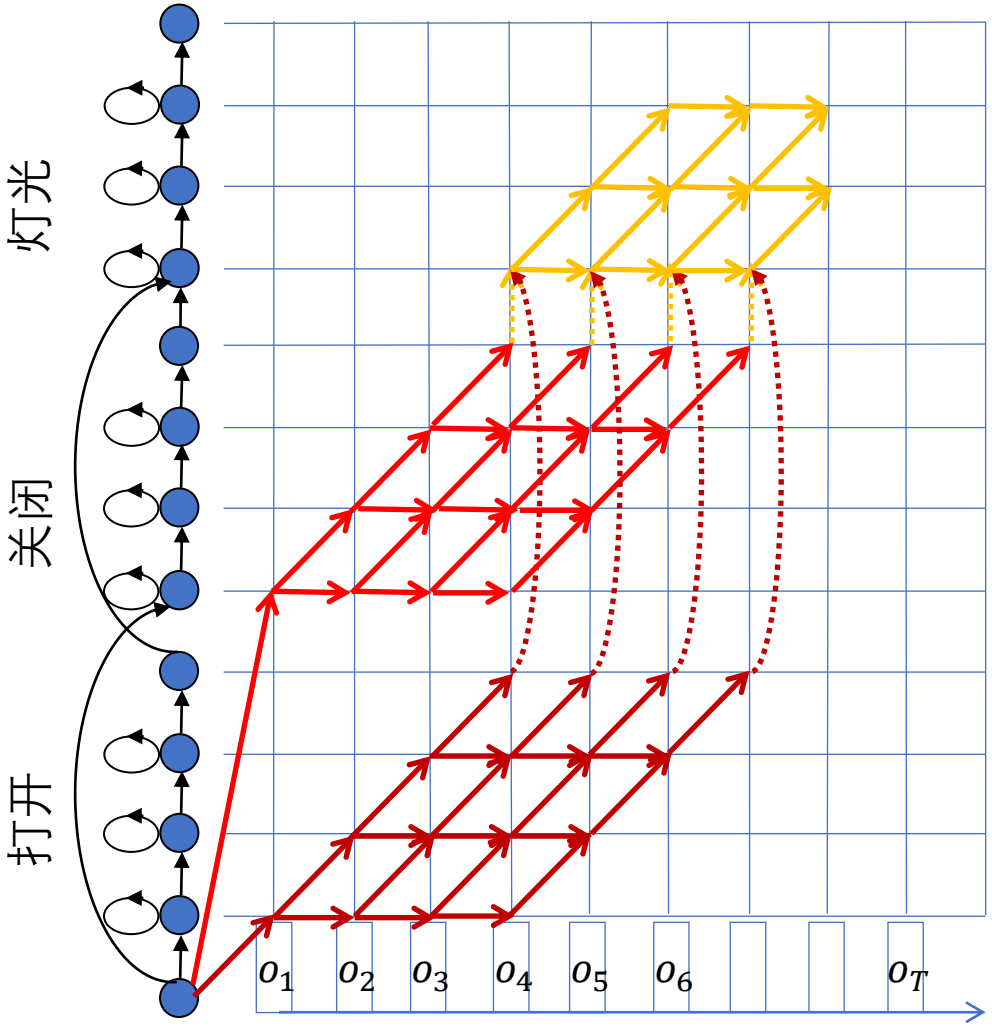
解码过程



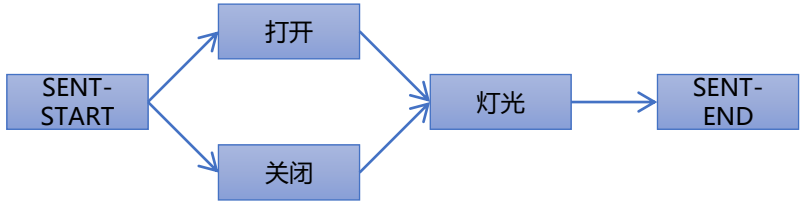


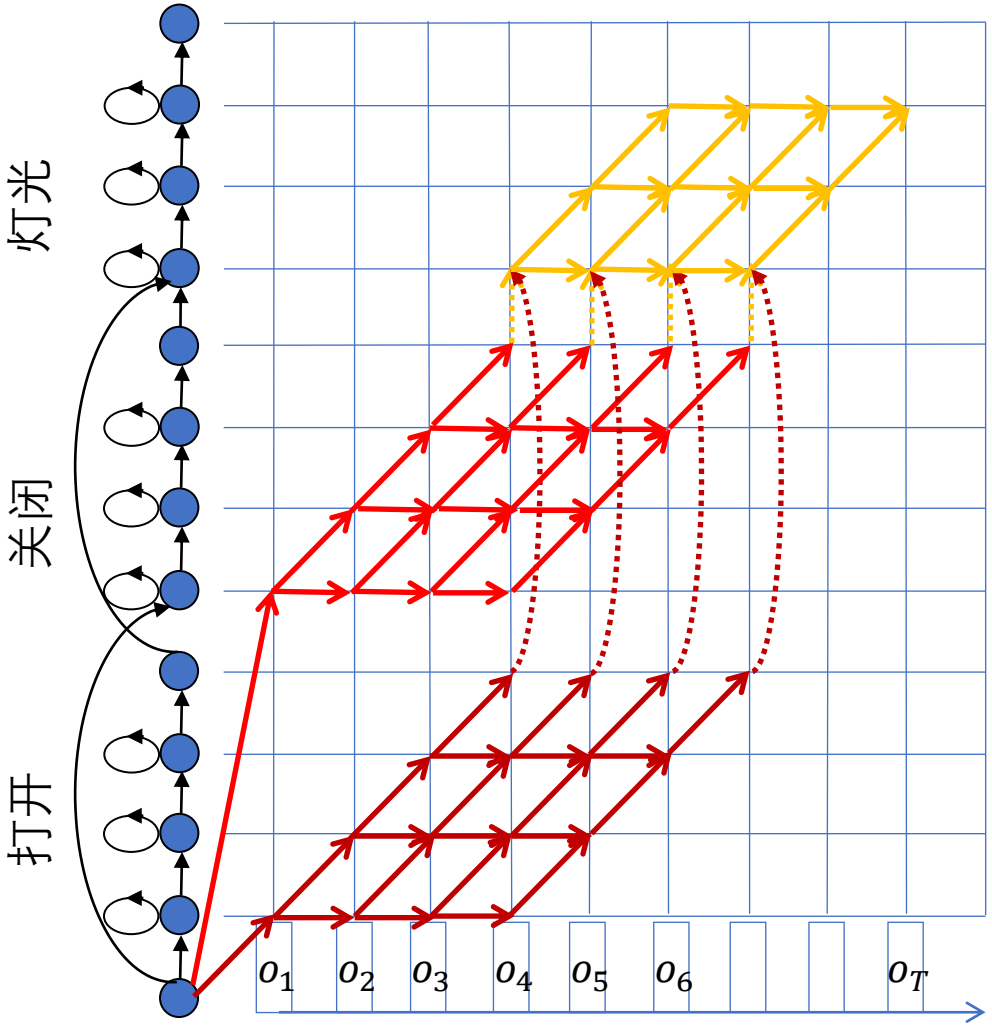
解码过程



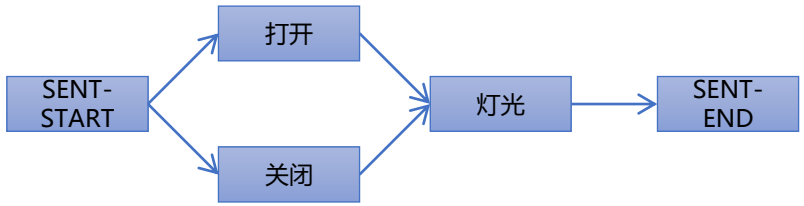


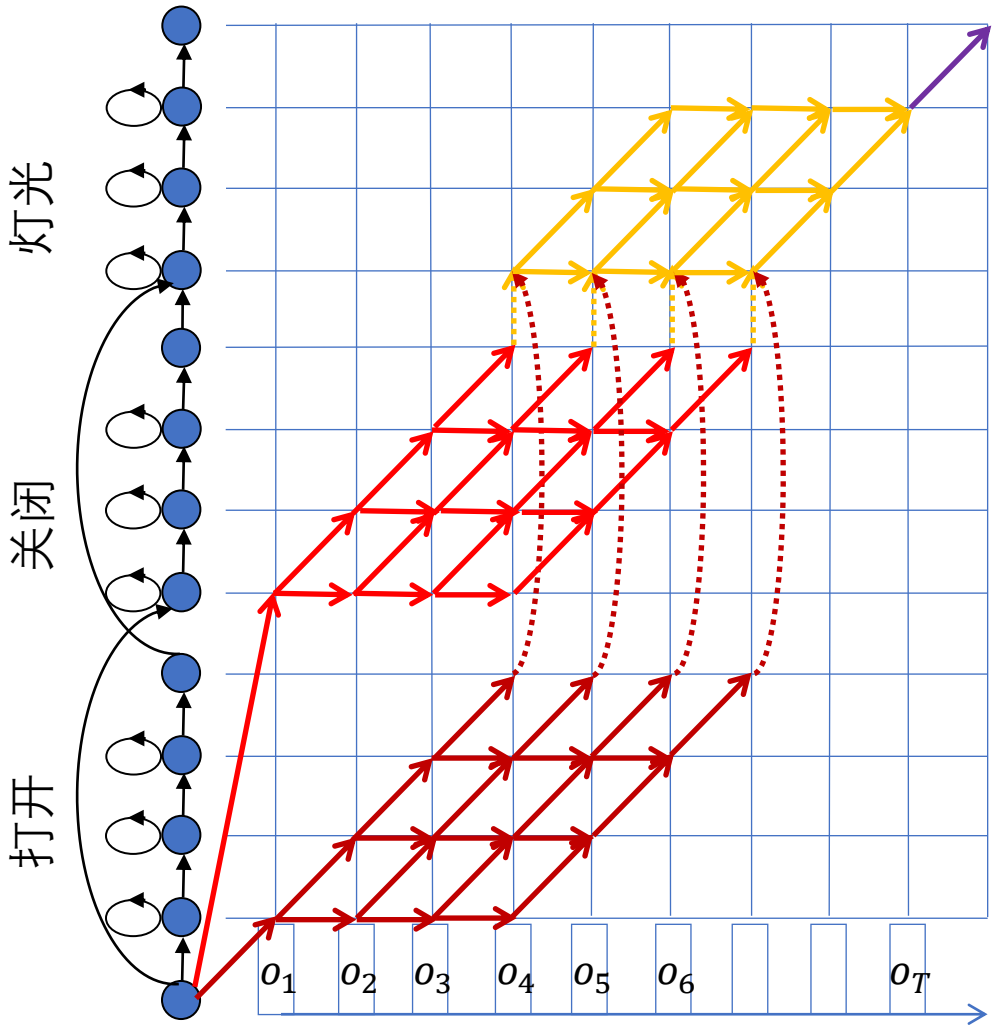
解码过程



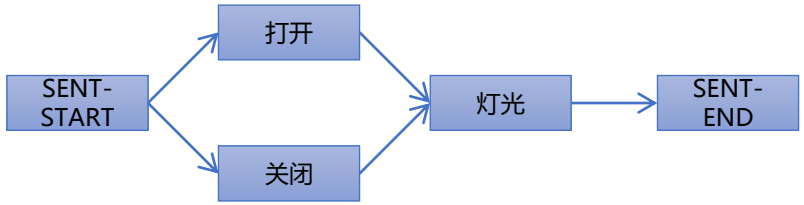


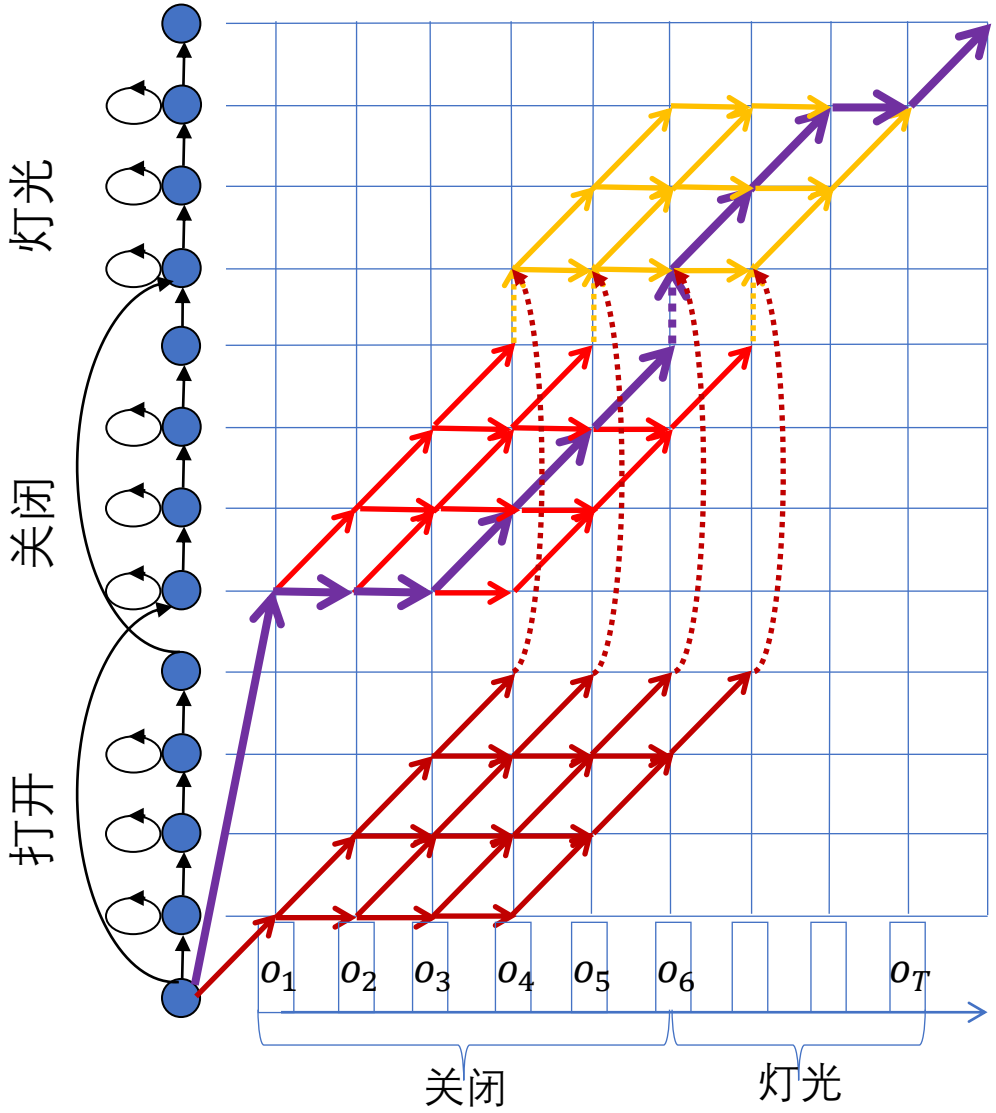
解码过程





解码过程





识别结果：关闭灯光

6.5 随机语法的识别

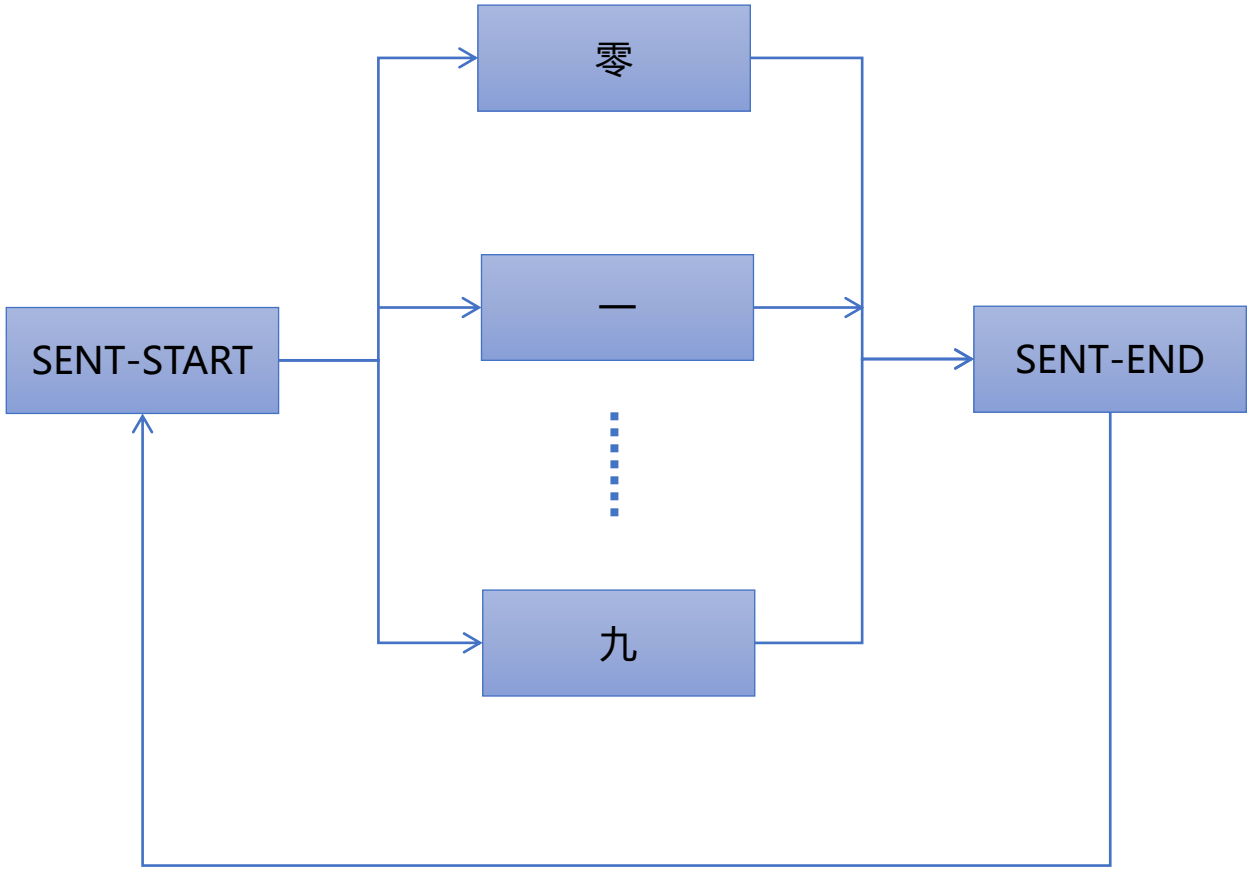
- 随机数字串识别

$digit = 零 | 一 | 二 | 三 | 四 | 五 | 六 | 七 | 八 | 九 ;

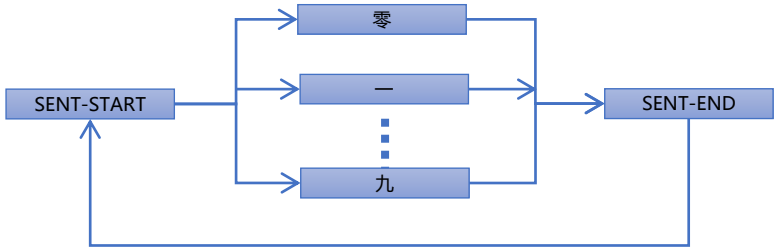
(SENT-START (<$digit >) SENT-END)

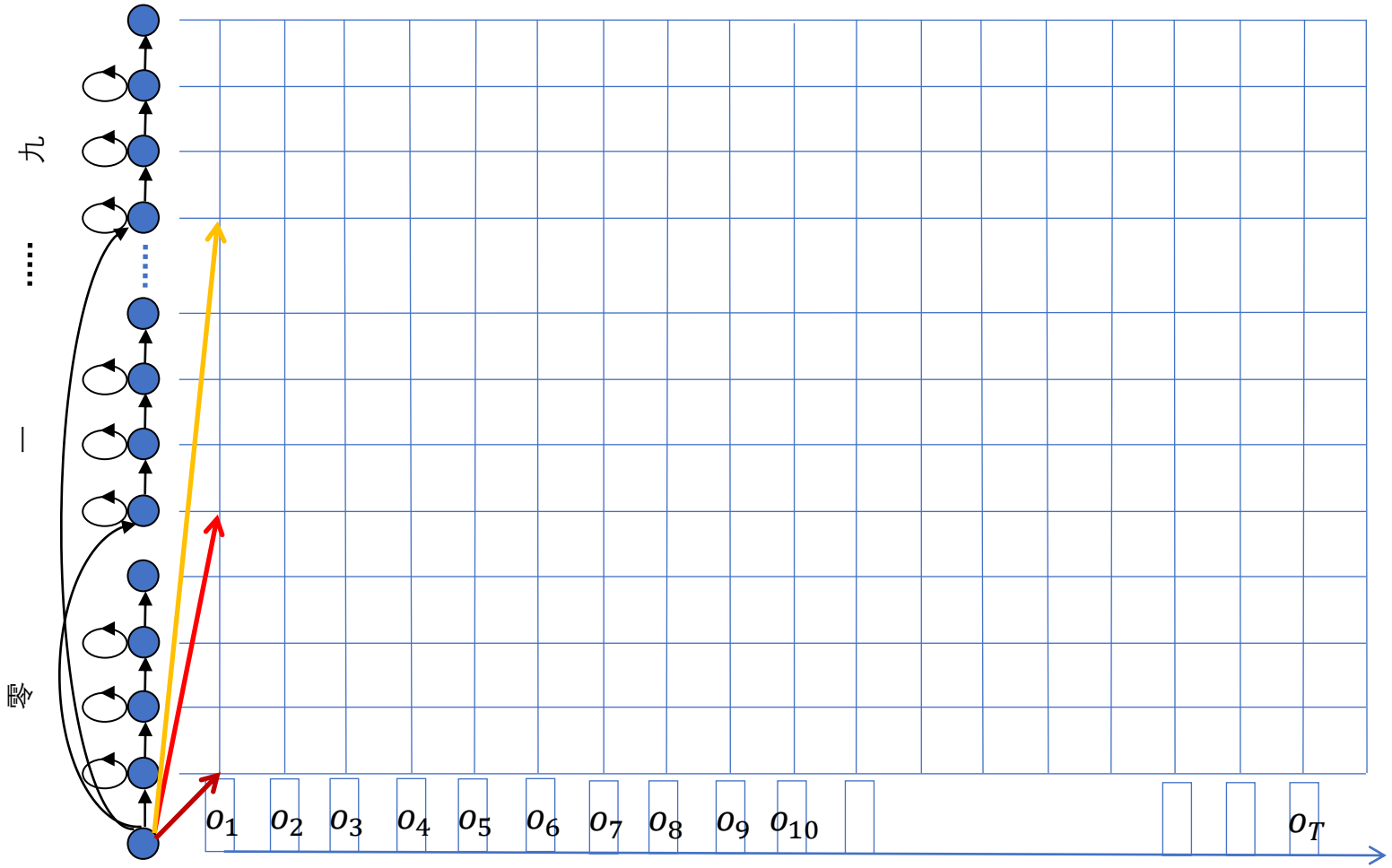
其中<>表示可以识别连续的多个重复的digit单元。

识别网络

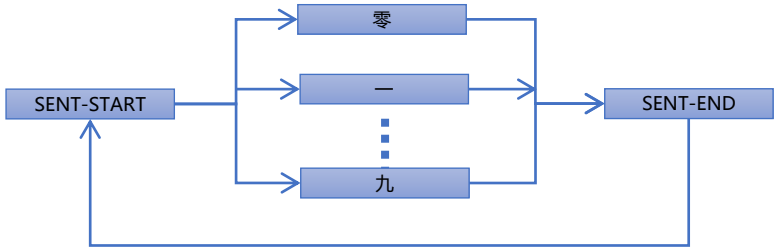


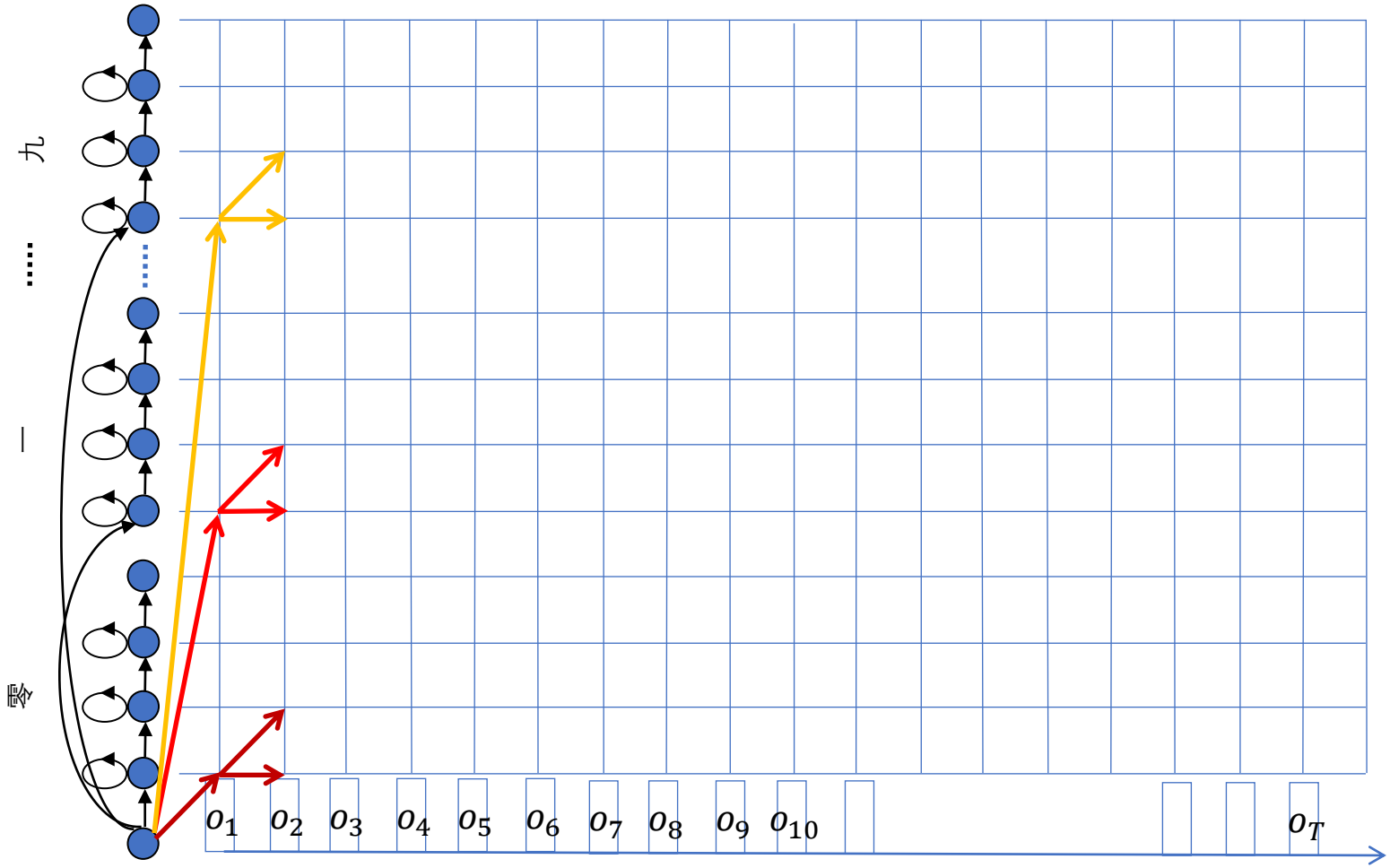
解码过程



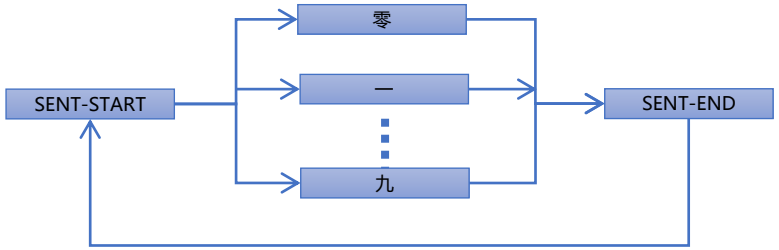


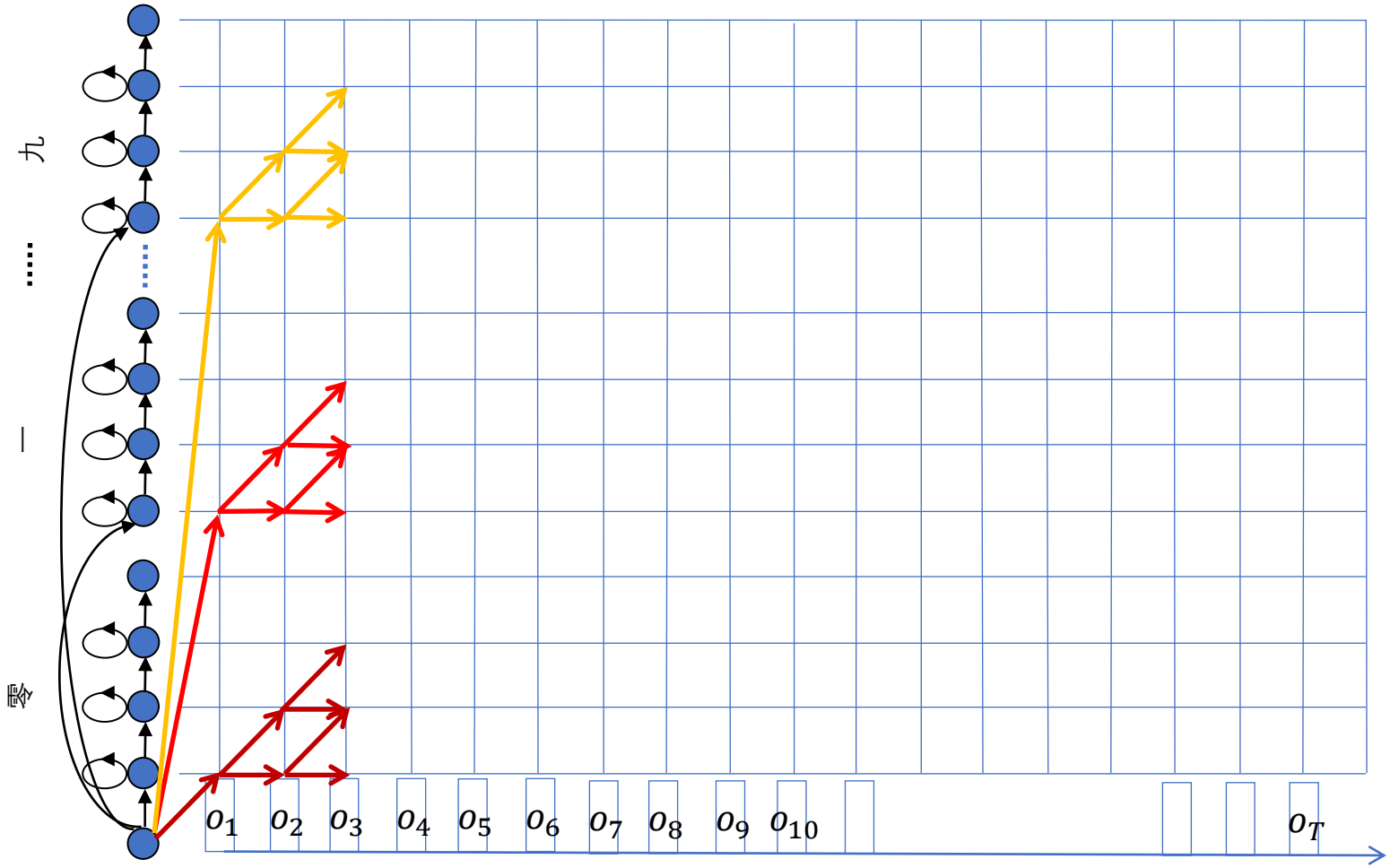
解码过程



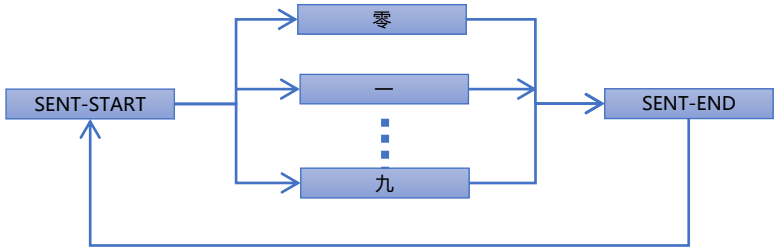


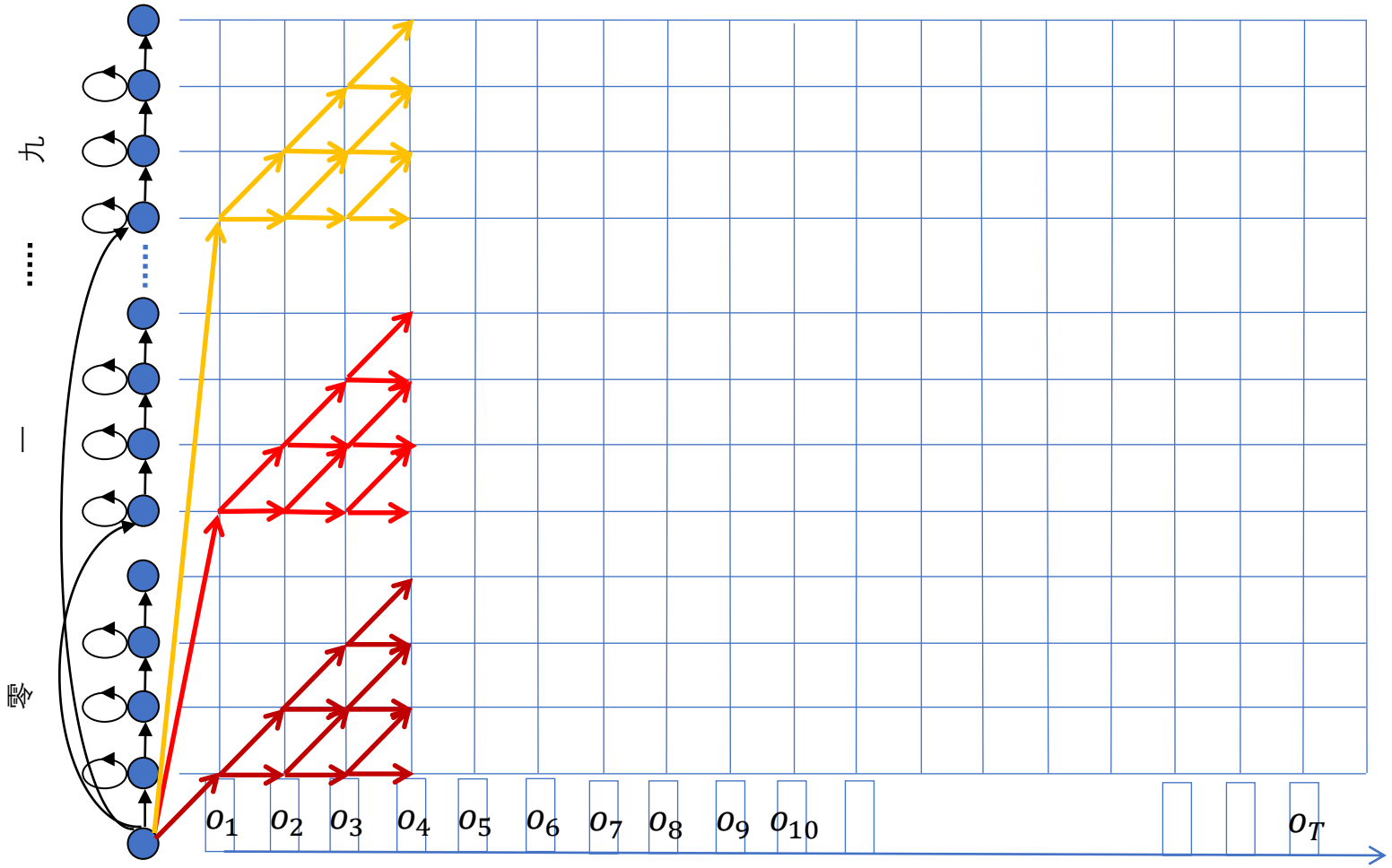
解码过程



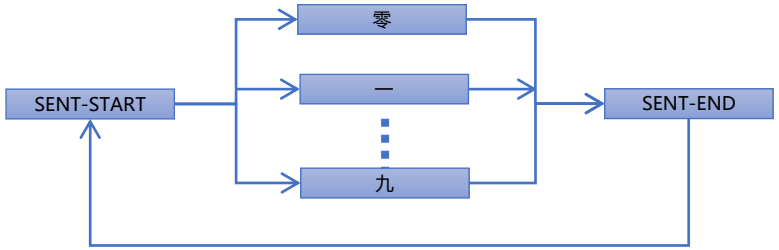


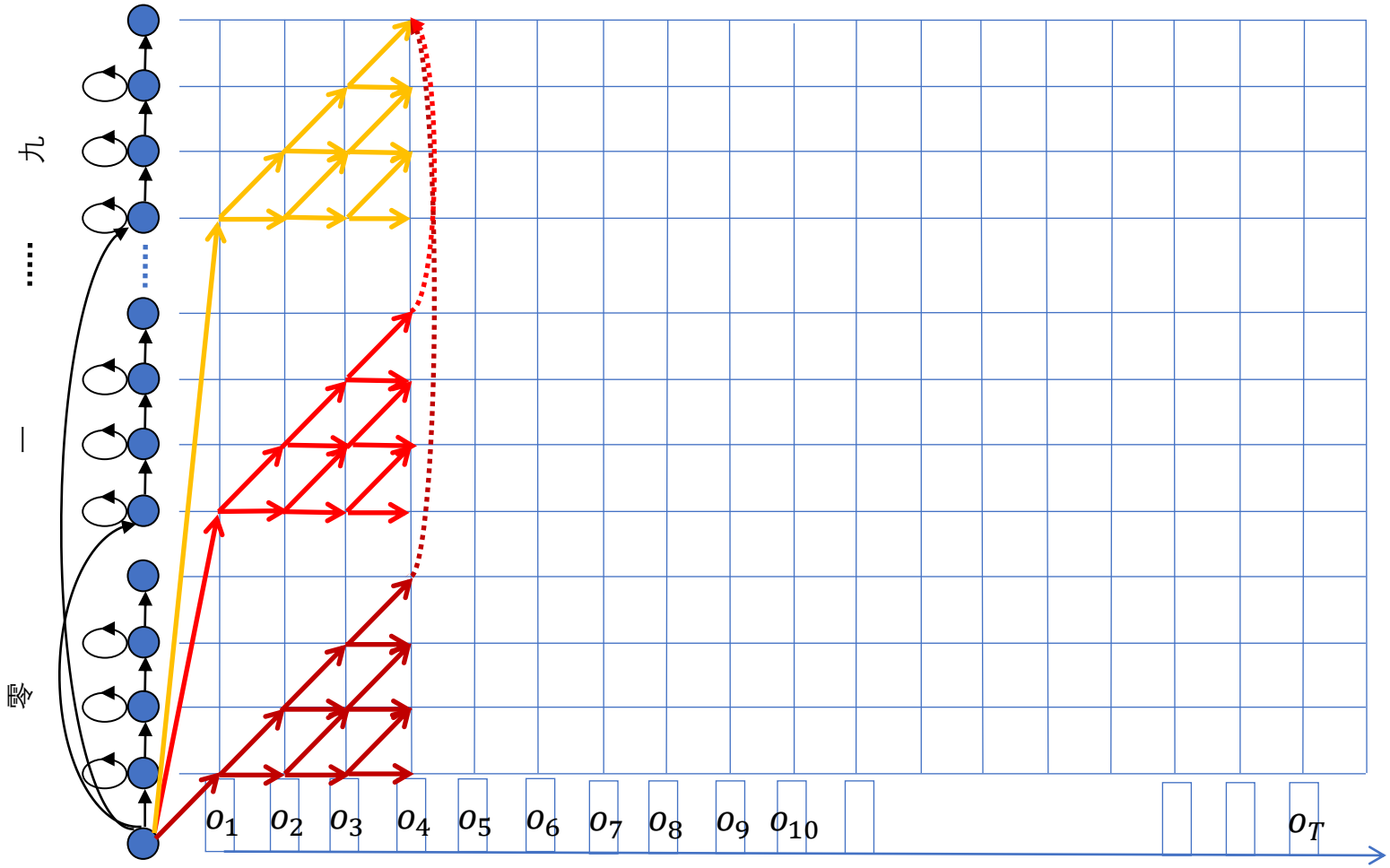
解码过程



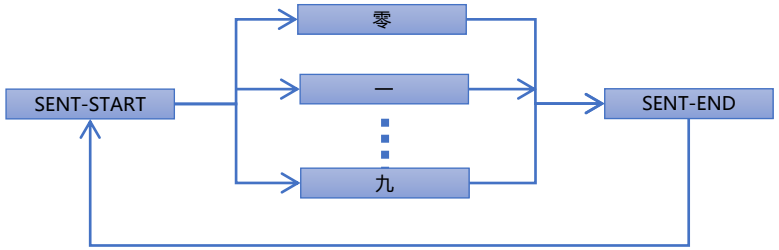


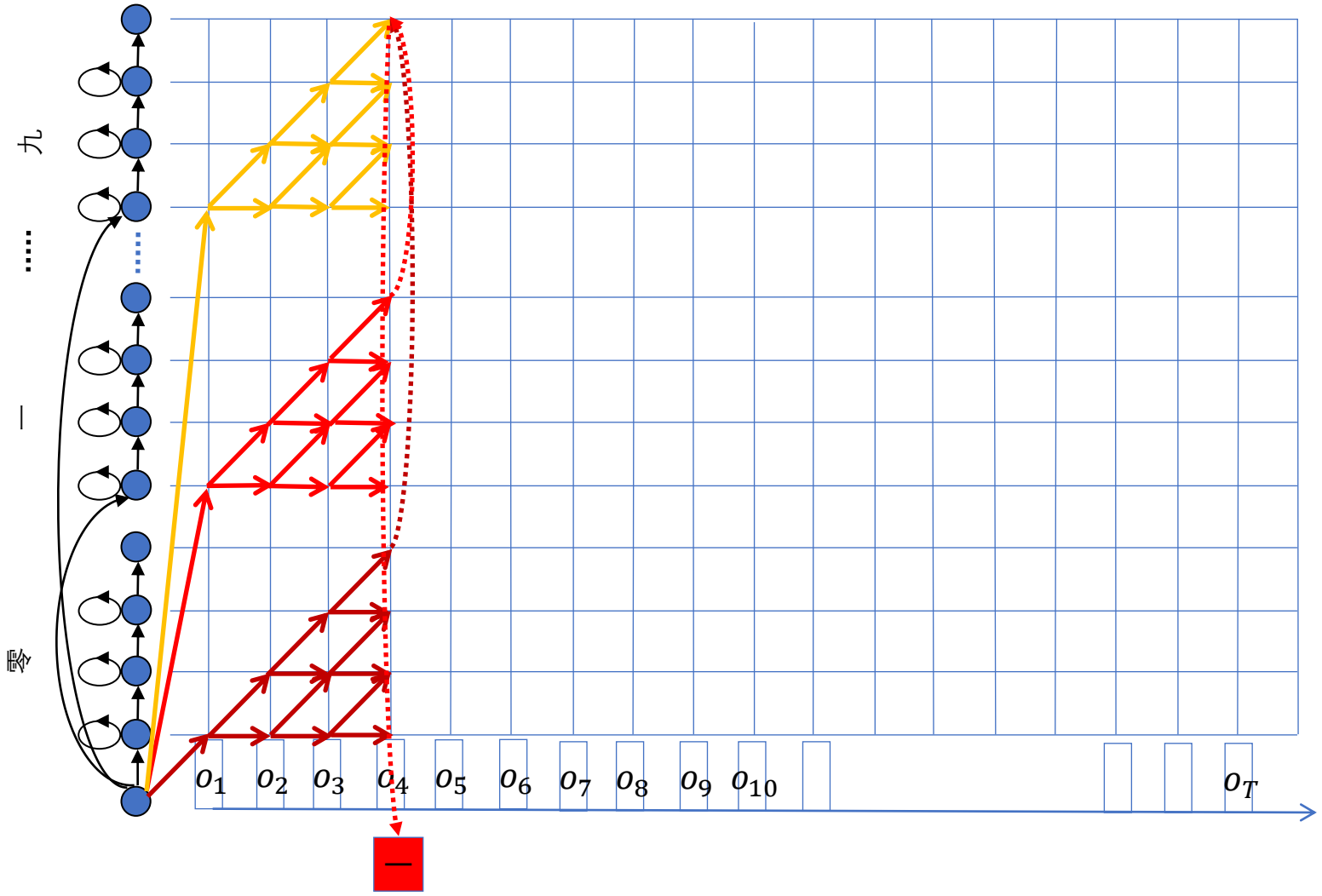
解码过程



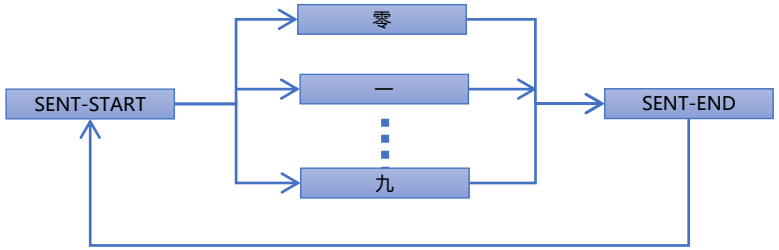


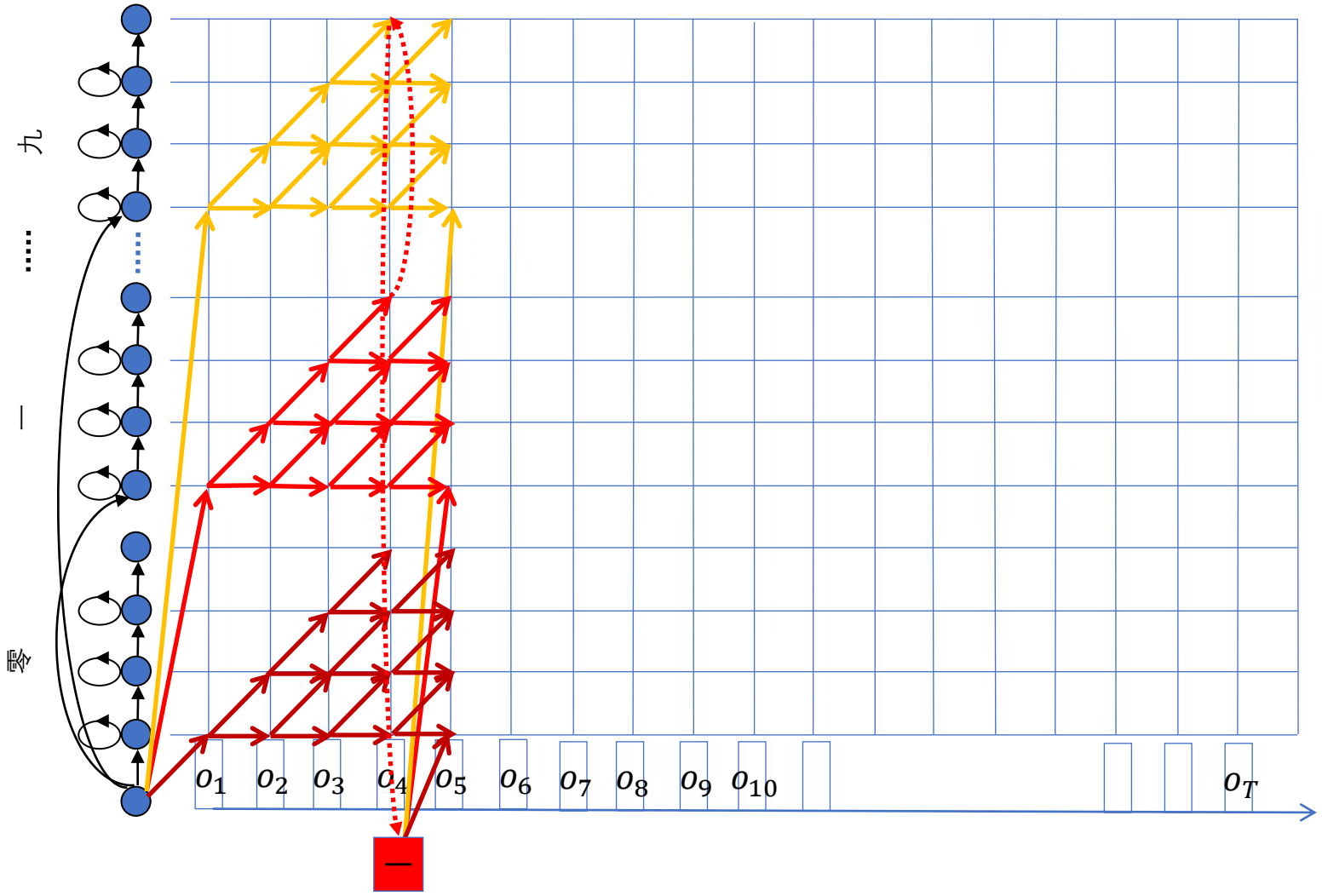
解码过程



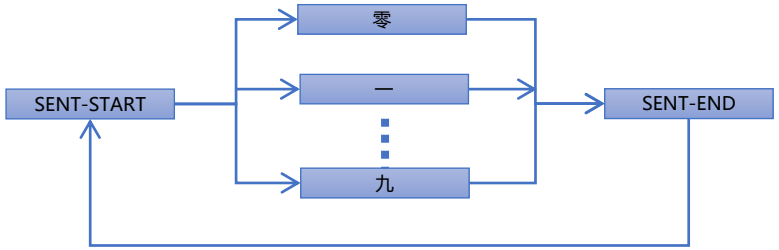


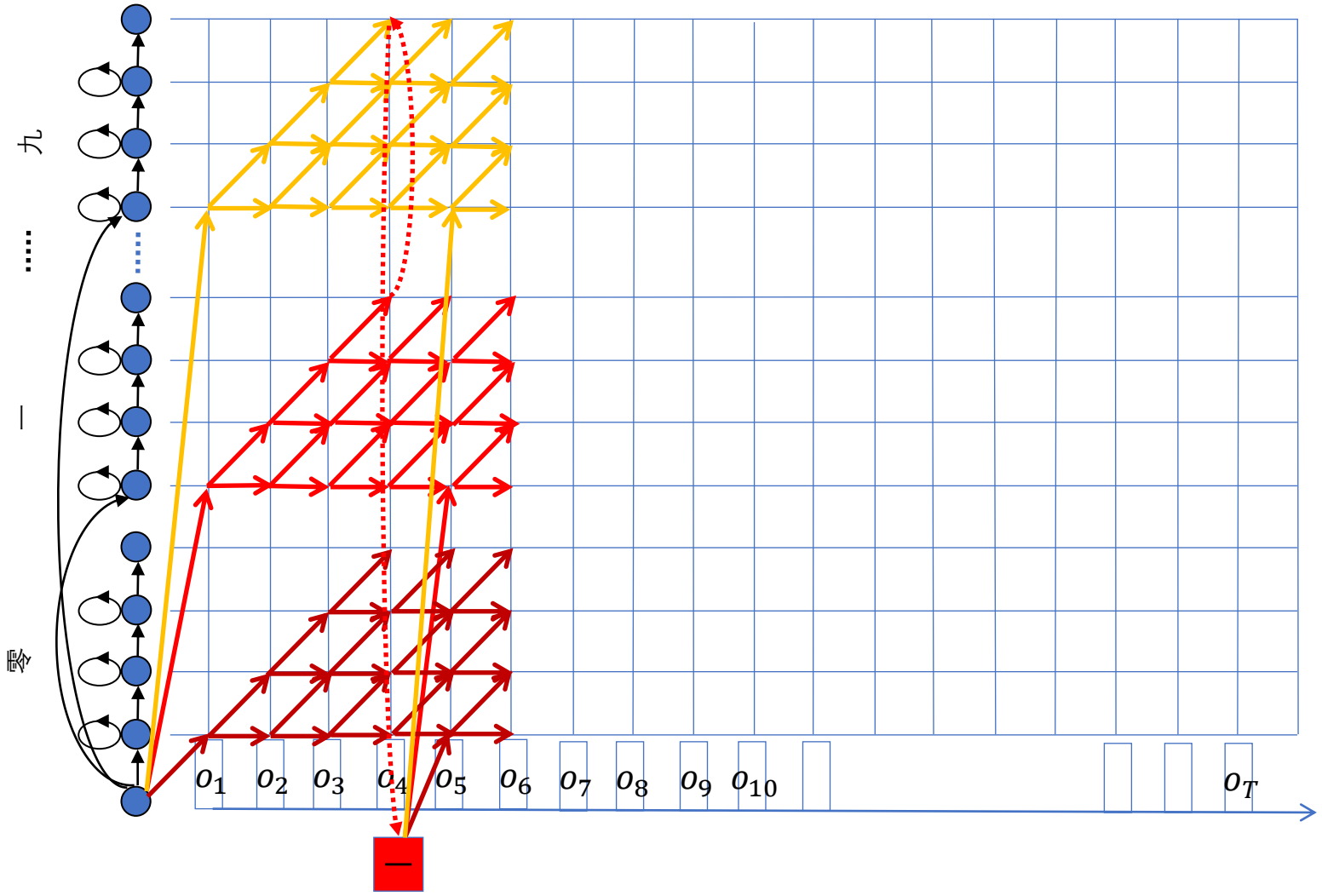
解码过程



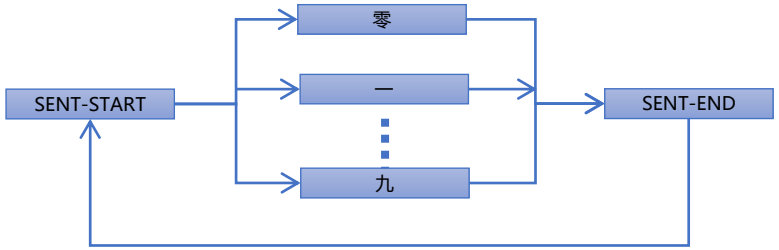


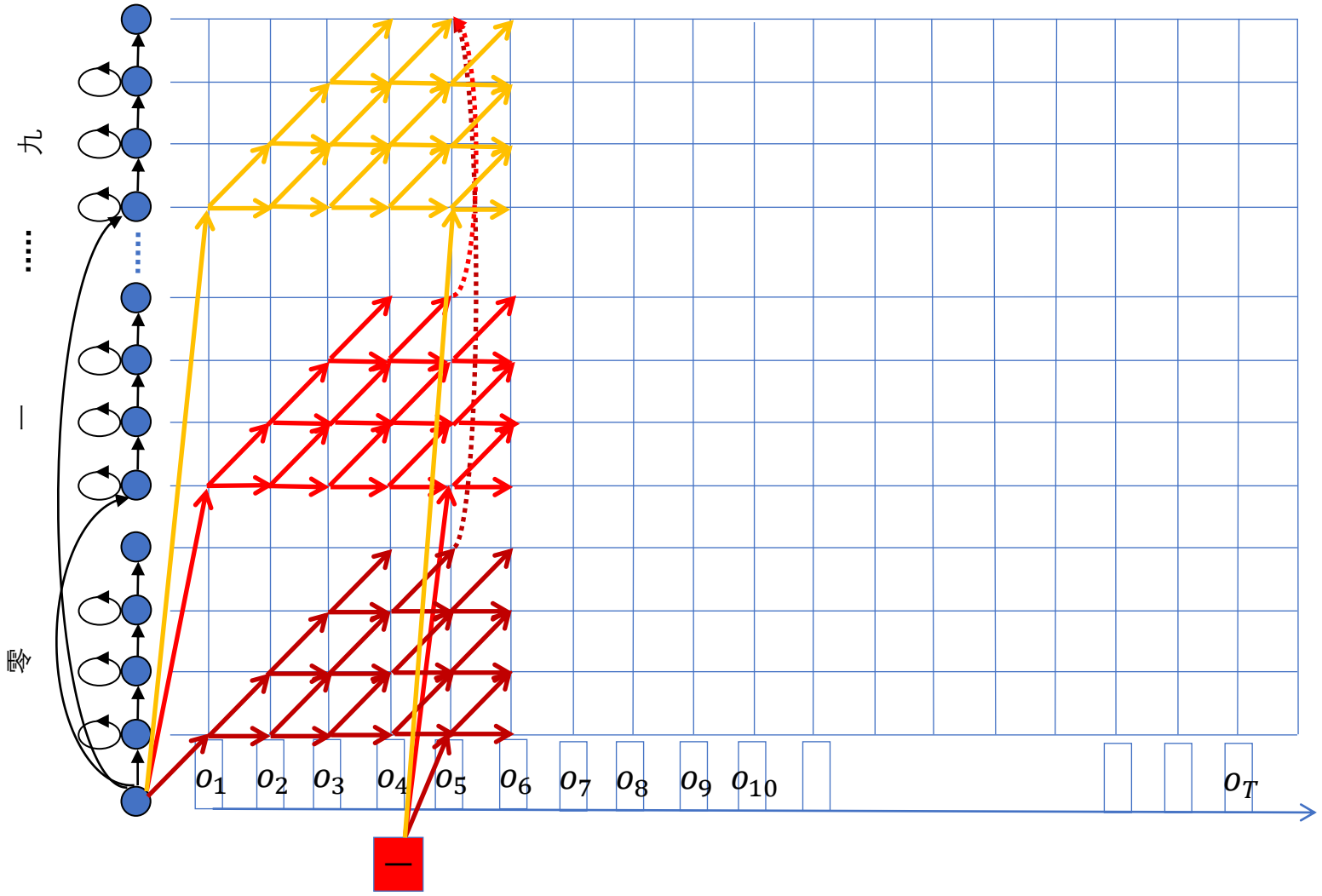
解码过程



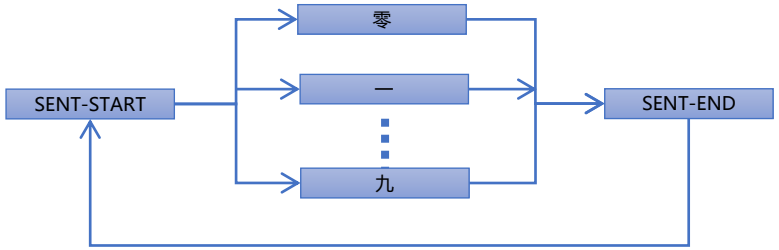


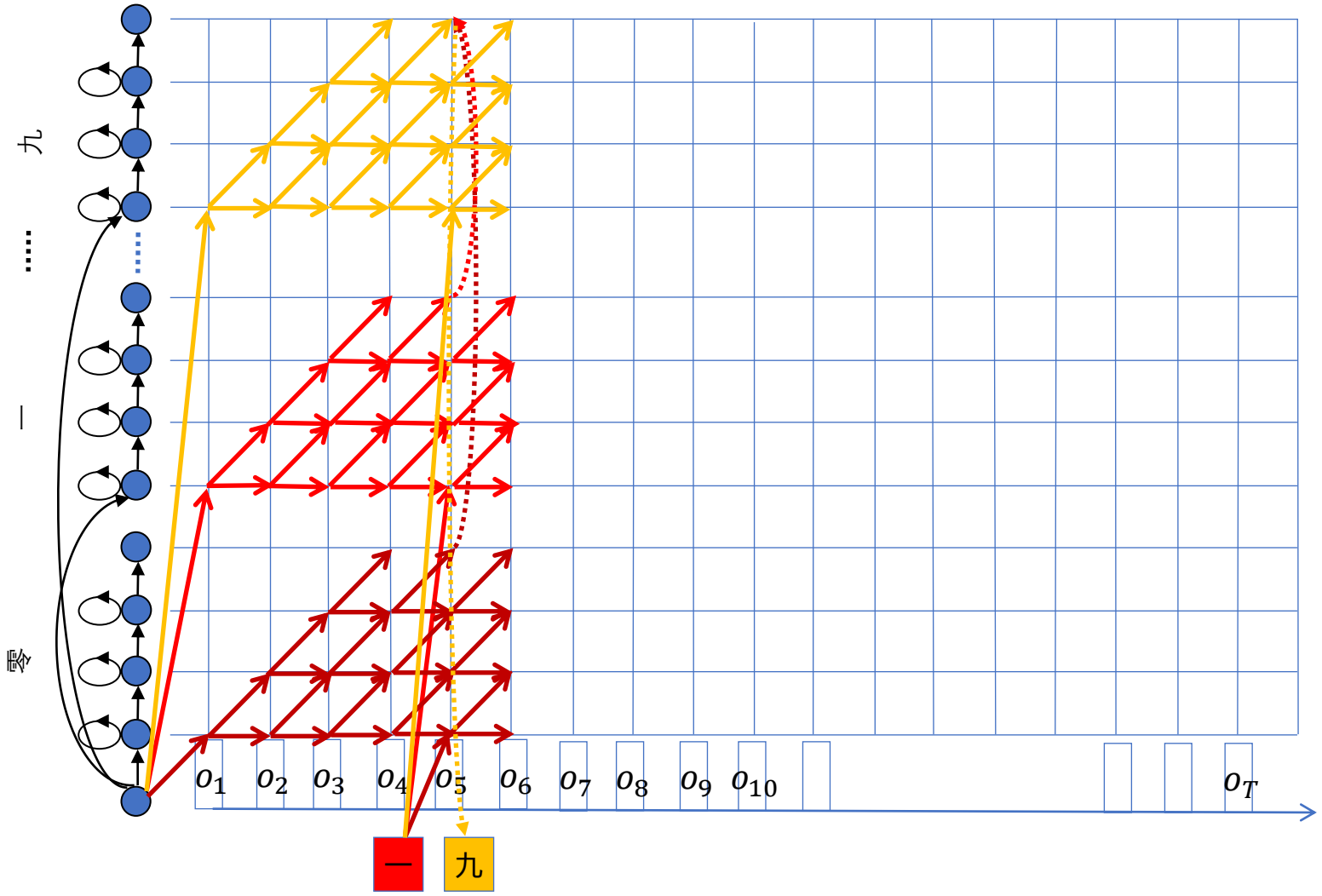
解码过程



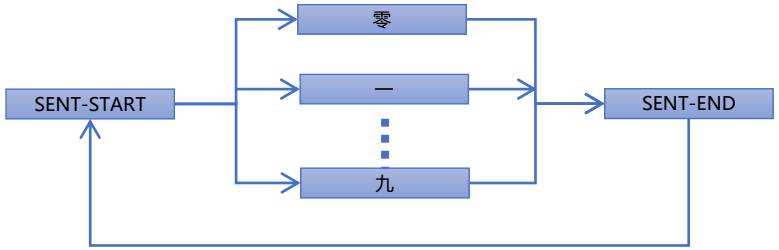


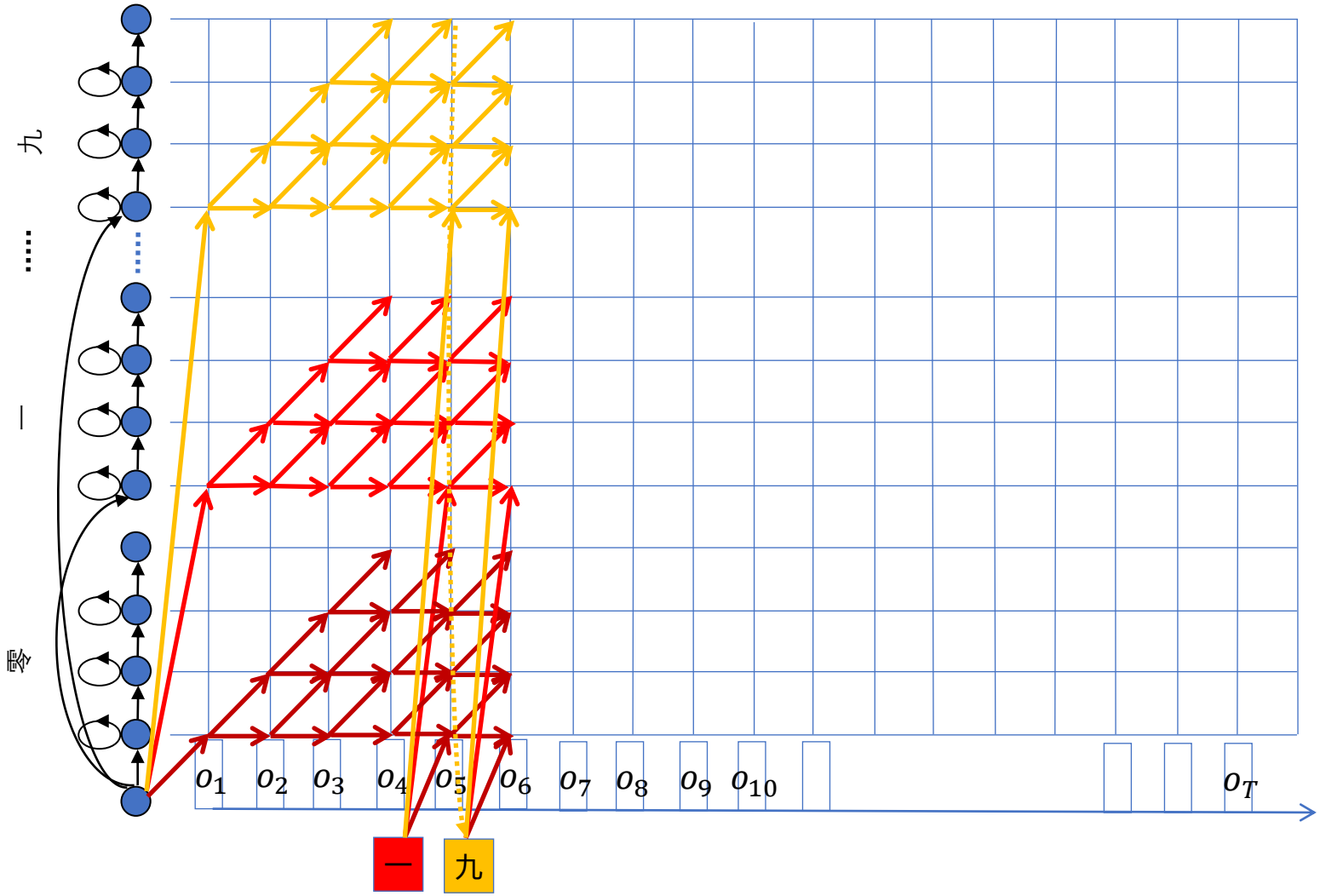
解码过程



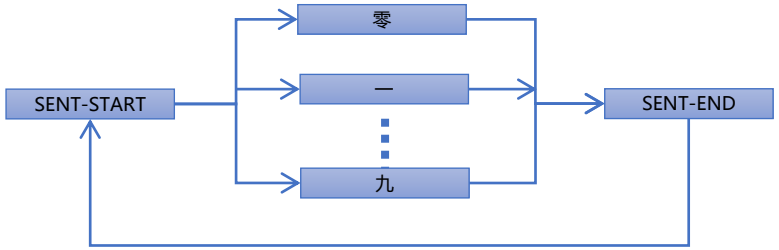


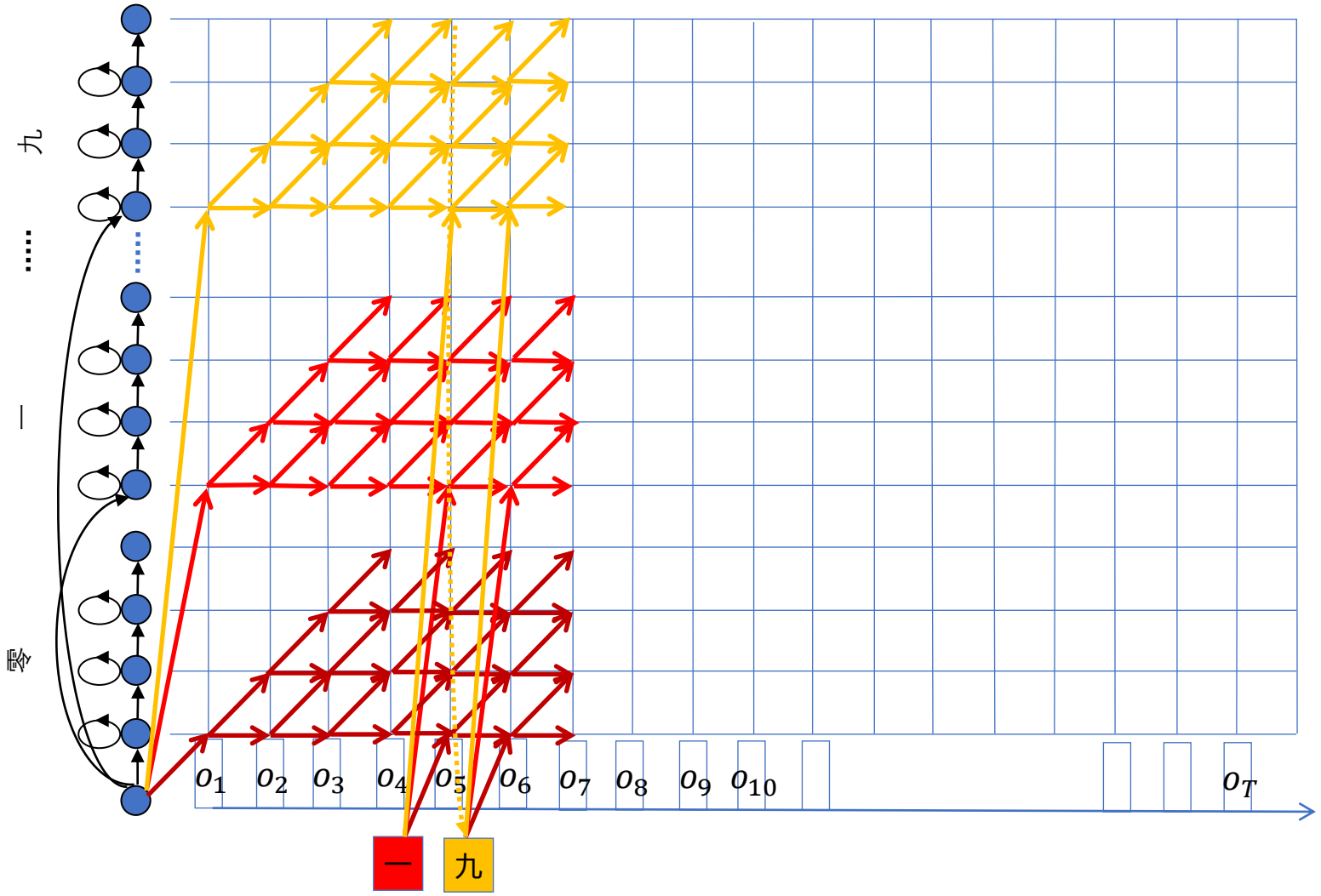
解码过程



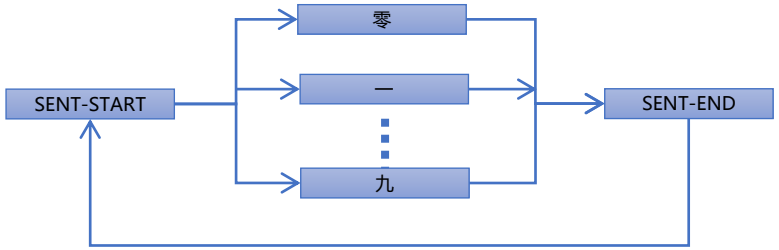


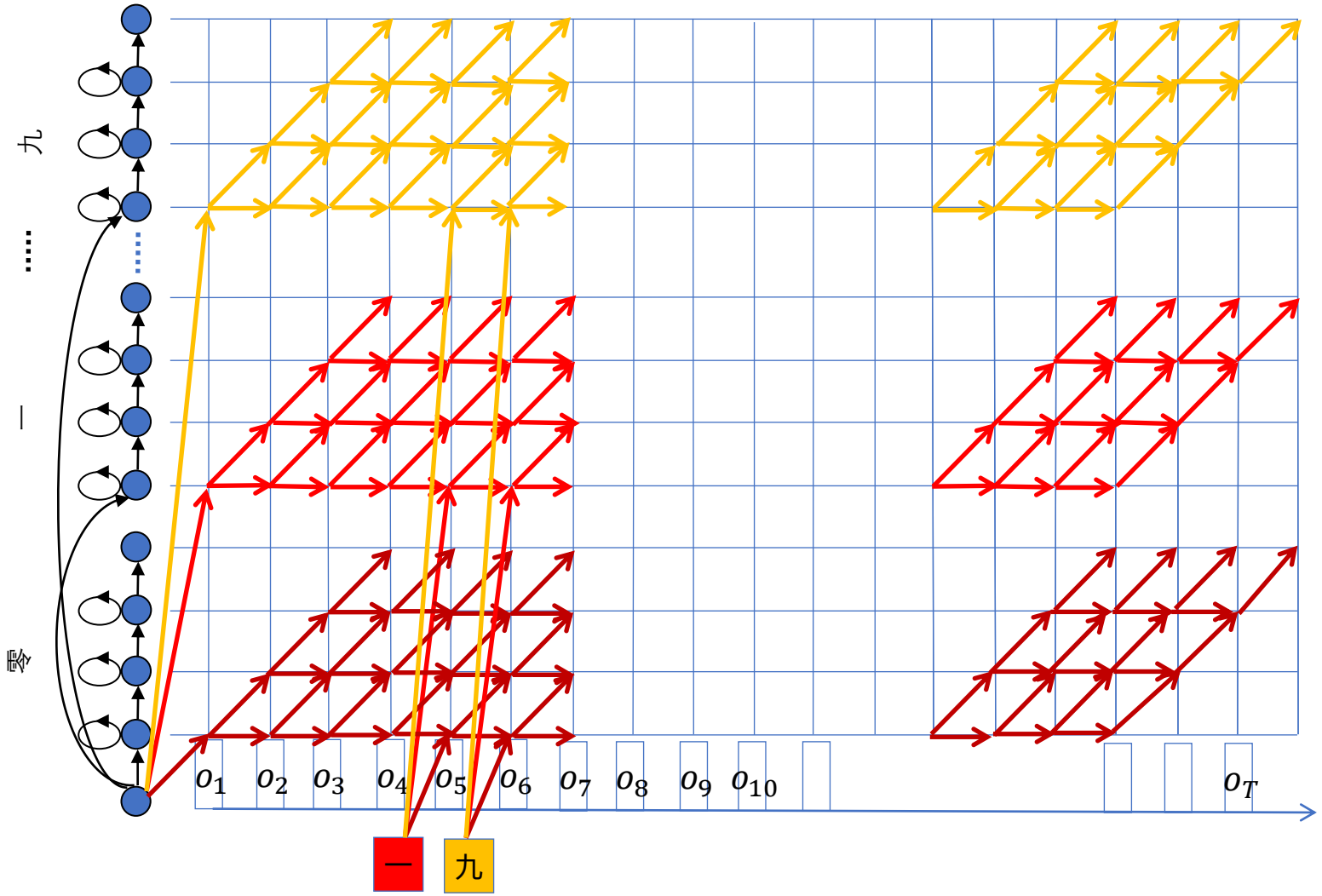
解码过程



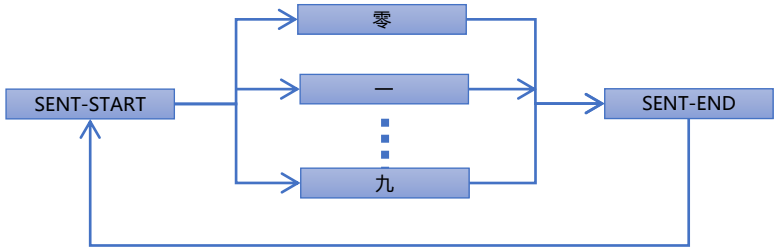


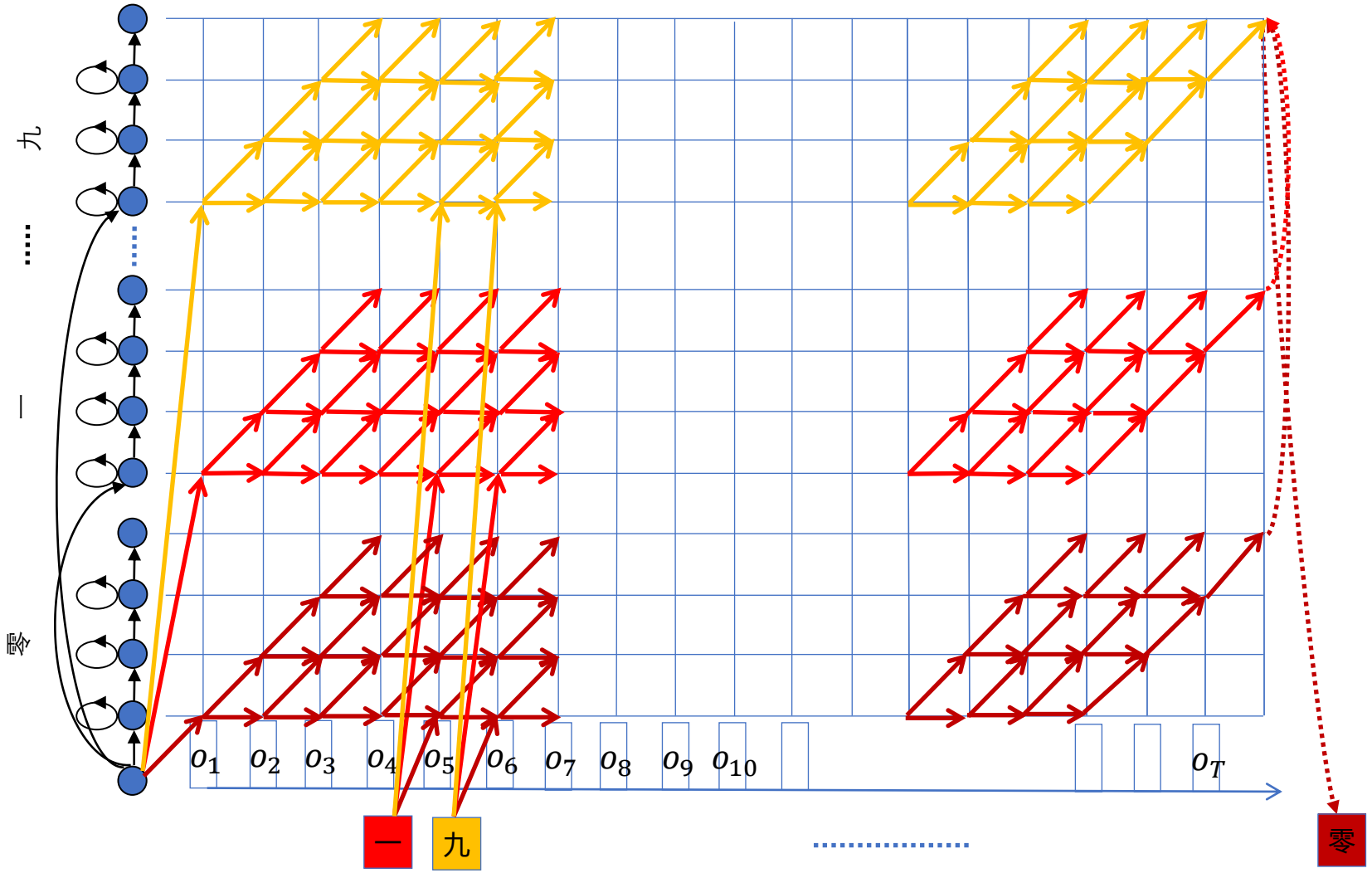
解码过程



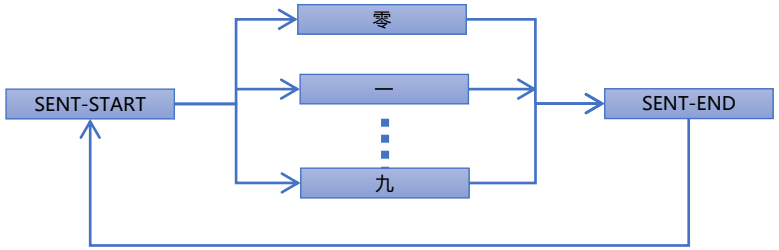


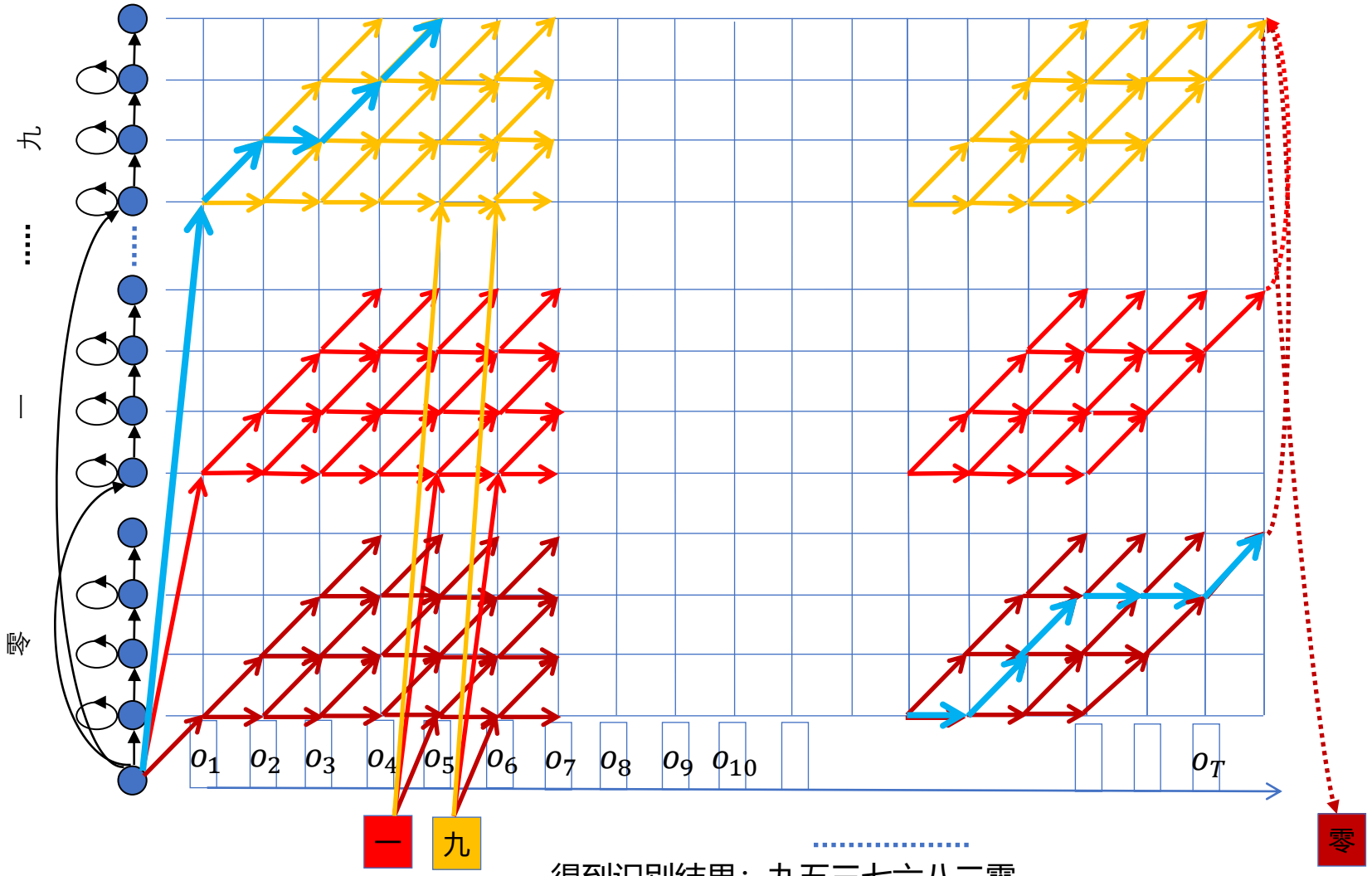
解码过程





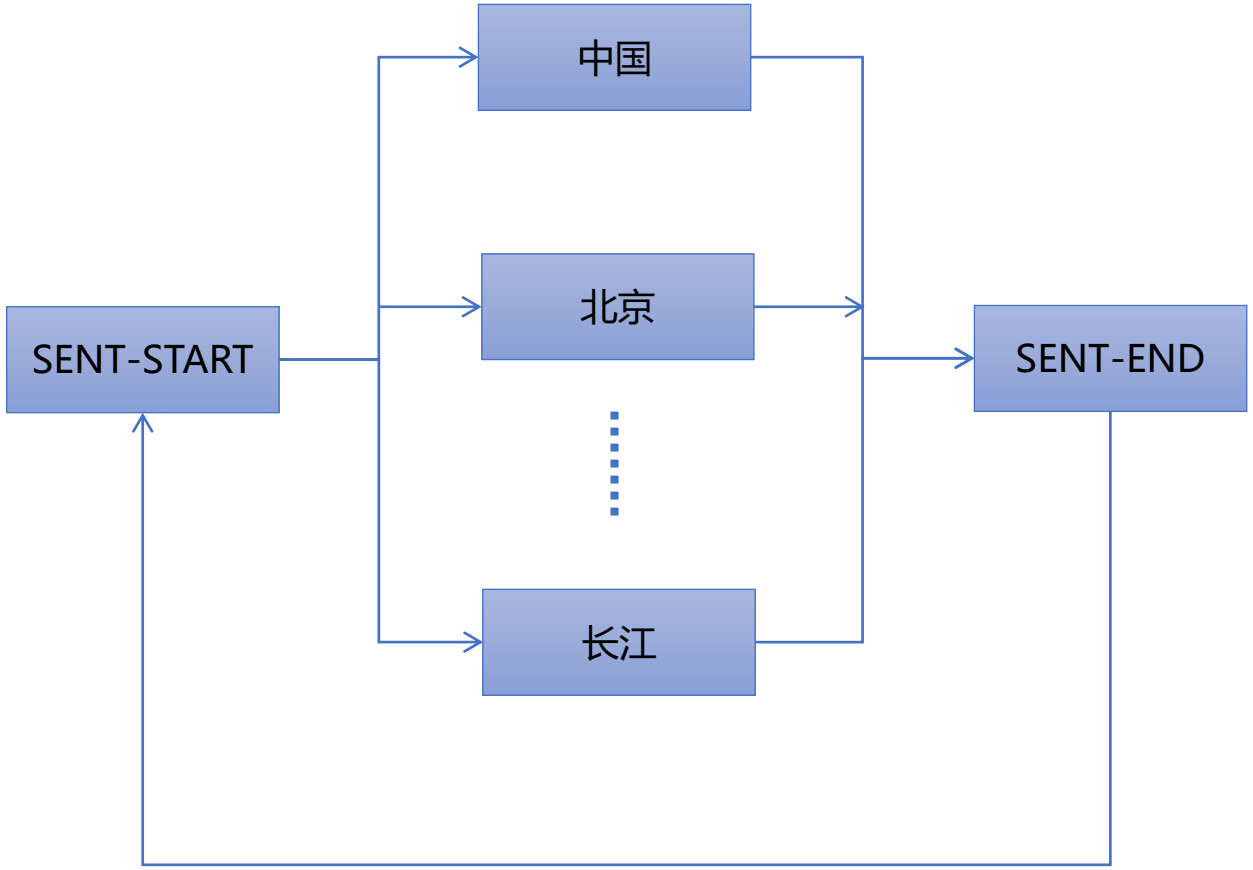
解码过程





得到识别结果：九五三七六八二零

普通话词汇识别网络（随机语法）



包含所有的普通话词汇，有可能达十几万个词，则解码过程将无比庞杂，识别速度会非常慢，识别结果也会不理想，因为各种组合太多，甚至可能出现语法不通顺或奇怪组合的句子，如“北京在中国长江”。

6.6 音素的上下文建模

6.6.1 协同发音

6.6.2 上下文建模

6.6.3 决策树

6.6.4 问题集

6.6.5 三音子模型的训练

6.6.1 协同发音

汉字是我们在日常交流中最易分辨的最小单元，中文以声韵母作为音素，可以表示为：音节（汉字）=声母+带声调的韵母，例如:

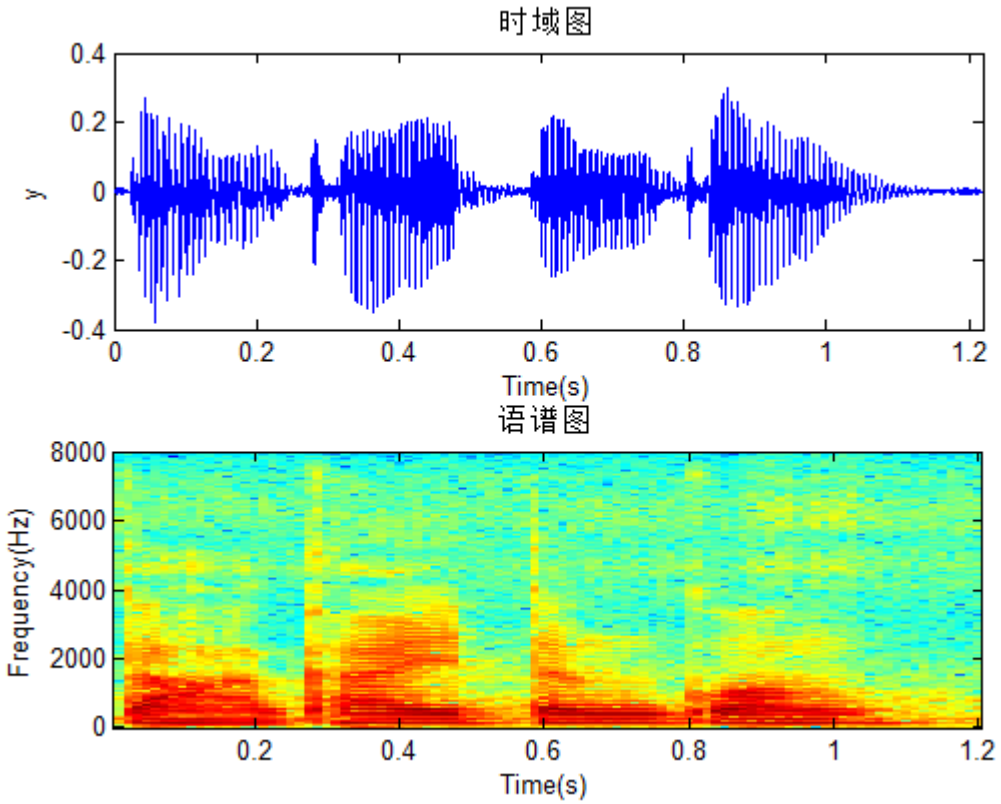
中文：好　好　学　习　天　天　向　上

音节：hao3　hao3　xue2　xi2　tian1　tian1　xiang4　shang4

音素：h ao3　h ao3　x ue2　x i2　t ian1　t ian1　x iang4　sh ang4

其中，音节和音素中的数字代表声调。

6.6.1 协同发音



打 开 灯 光

da kai deng guang

6.6.2 上下文建模

- 每个字都是一个音节，音节内部有发音衔接，音节之间也是有衔接的，体现在语谱图过渡阶段存在交叉，如“打”和“开”之间，“灯”和“光”之间。
- 本质上，音素属于单音子（monophone），没有考虑上下文的协同发音，不适合连续语音流的HMM建模。
- 单音子（monophone）没有考虑上下文的协同发音；
- 为了更好地匹配协同发音现象，需要对音素进行上下文建模。

双音子（diphone）

- 只考虑音节内部声母和韵母之间的关联，如“灯”的拼音deng，本来是分为d和eng两部分，可拆分成两种形式：

  deng d+eng

  deng d-eng

- d+eng表示发音偏向于d，但后面衔接eng，d-eng表示发音偏向于eng，但前面是d。这样把d和eng发音密切关联起来，跟直接分开对比，多考虑了两者之间的发音连读。

双音子组合数目

- 普通话(不带声调)：27(声母)*38(韵母)*2=2052个
- 英语：28(辅音)*20(元音)*2=1120个

三音子（triphone）

如“打开灯光”的拼音：

d a k ai d eng g uang

加上前后静音

sil d a k ai d eng g uang sil

可转化为三音子如下：

sil sil-d+a d-a+k a-k+ai k-ai+d ai-d+eng d-eng+g eng-g+uang g-guang+sil sil

其中sil-d+a 表示d的左边（‘-’表示）是音素sil，右边（‘+’表示）是音素a，d-a+k是由da和kai的声母组合而成，形成跨词三音子。

三音子（triphone）

sil	d	a	k	ai	d	eng	g	uang	sil
sil	sil-d+a	d-a+k	a-k+ai	k-ai+d	ai-d+eng	d-eng+g	eng-g+uang	g-uang+sil	sil

三音子（triphone）

中文:	好	好	学	习	天	天	向	上
音节:	hao3	hao3	xue2	xi2	tian1	tian1	xiang4	shang4
单音子:	sil h ao3	h ao3	x ue2	x i2	t ian1	t ian1	x iang4	sh ang4 sil

三音子: sil sil-h+ao3 h-ao3+h ao3-h+ao3 h-ao3+x …… sh-ang4+sil sil

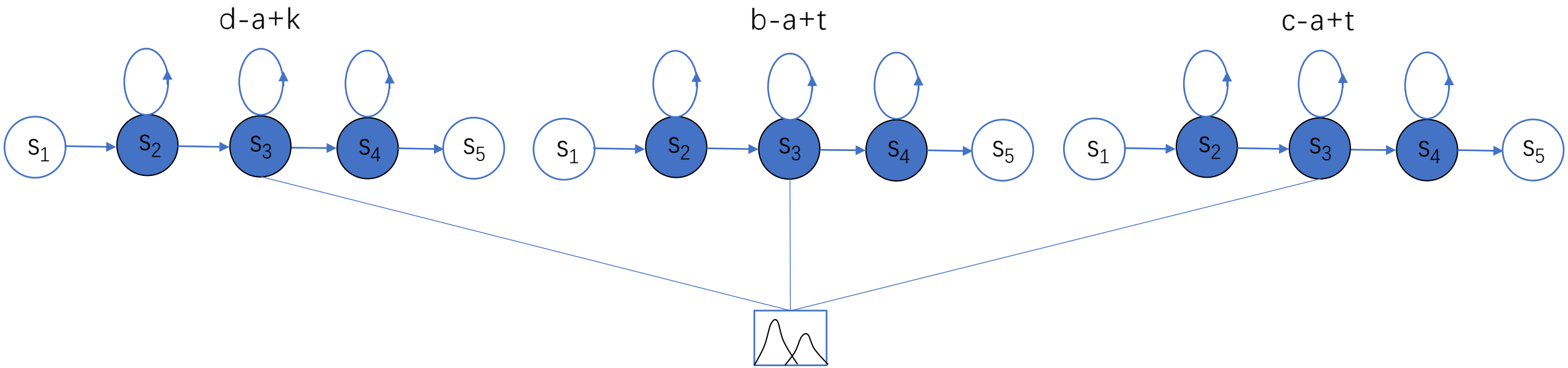
三音子组合数目

- 三音子可实现对音素上下文更精细的建模，但数量巨大！例如
  - 普通话：27(声母)*38(韵母)*29(声母+sil/sp)=29754
  - 英语：28(辅音)*20(元音)*30(辅音+sil/sp)=16800

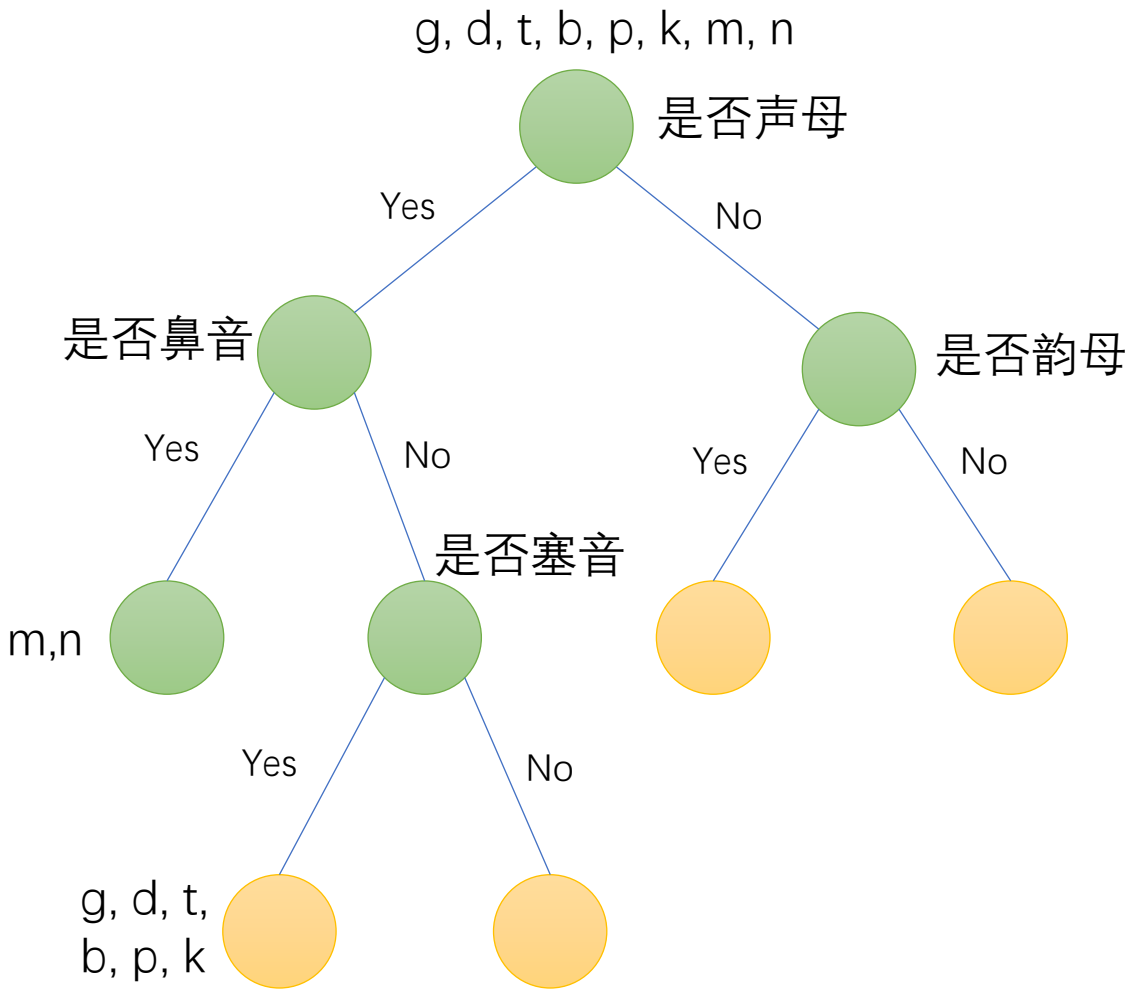
  其中sil和sp分别表示静音和停顿。

- 不考虑音素的先后顺序，则N个音素至少有N^3种组合；
- 如果加上声调，则数目更多！
- 而训练数据往往有限，无法对这么多模型的HMM状态参数进行充分训练。
- 解决办法：状态绑定。

状态绑定



6.6.3 决策树


g, d, t, b, p, k, m, n
是否声母
Yes
No
是否鼻音
是否韵母
Yes
No
Yes
No
是否塞音
m,n
Yes
No
g, d, t,
b, p, k

6.6.4 问题集

- 人工设计
- 自动生成

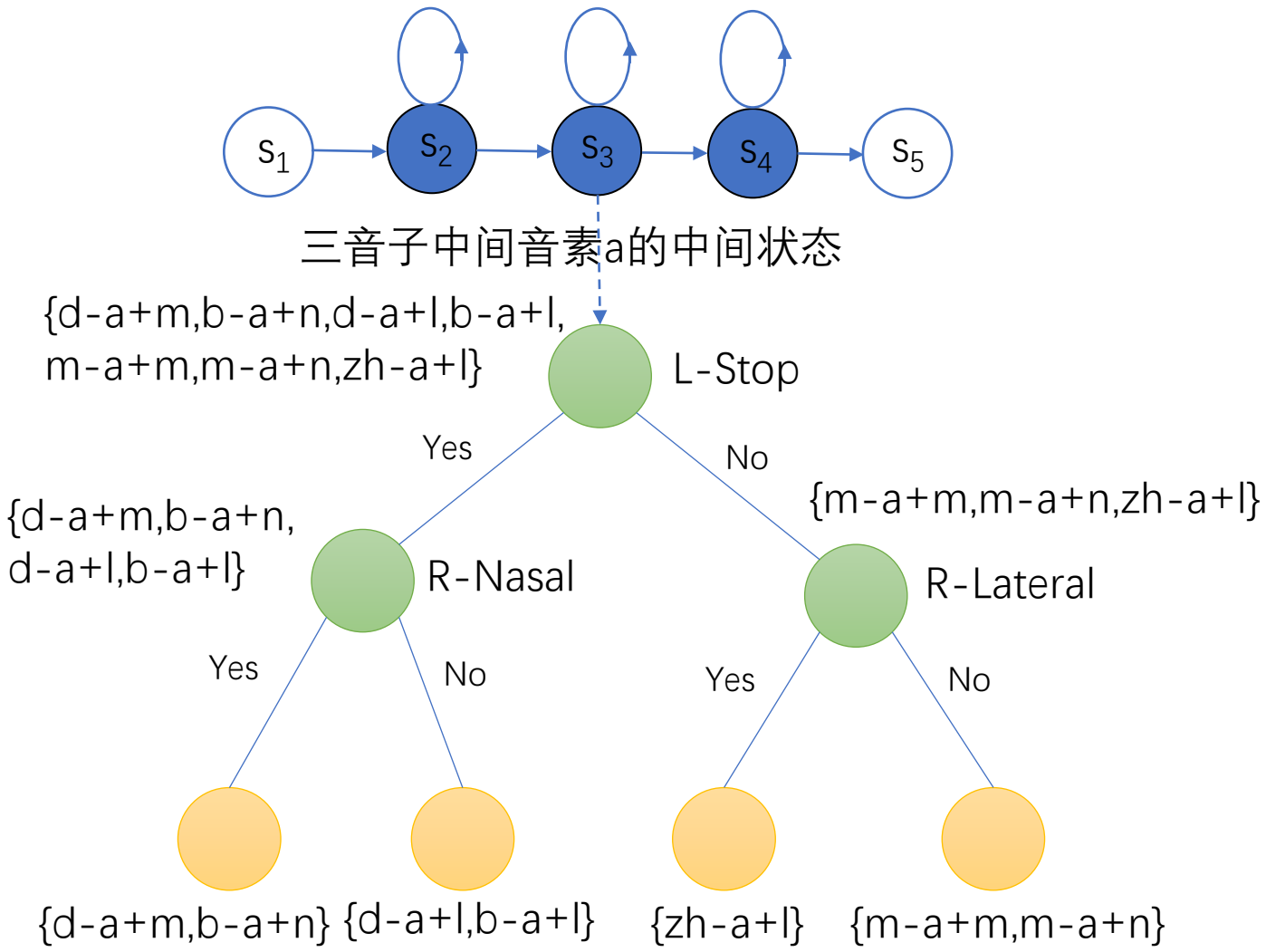
问题集—人工设计

分类	音素列表
声母（Consonant）	b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, r, z, c, s, y, w
塞音（Stop）	g, d, t, b, p, k
送气塞音（Aspirated_Stop）	p, t, k
不送气音（Non_Aspirated_Stop）	g, b, d
擦音（Fricative）	x, f, s, h, sh, r
清擦音（Breach_Fricative）	x, f, s, h, sh
浊擦音（Voiced_Fricative）	r
塞擦音（Affricate）	j, z, zh, q, c, ch
不送气塞擦音（Non_Aspirated_Affricate）	z, zh, j
送气塞擦音（Aspirated_Affricate）	c, ch, q
唇音（Labial）	b, p, m, f
边音（Lateral）	l
鼻音（Nasal）	m, n

问题集—人工设计

表达式	含义
QS "L-Consonant"	左边音素是否为声母
QS "L-Stop"	左边音素是否为塞音
QS "L-Aspirated_Stop"	左边音素是否为送气塞音
QS "L-Lateral"	左边音素是否为边音
QS "L-Nasal"	左边音素是否为鼻音
QS "R-Consonant"	右边音素是否为声母
QS "R-Stop"	右边音素是否为塞音
QS "R-Aspirated_Stop"	右边音素是否为送气塞音
QS "R-Lateral"	右边音素是否为边音
QS "R-Nasal"	右边音素是否为鼻音

决策树


s_1
s_2
s_3
s_4
s_5
三音子中间音素a的中间状态
{d-a+m,b-a+n,d-a+l,b-a+l,
m-a+m,m-a+n,zh-a+l}
L-Stop
Yes
No
{d-a+m,b-a+n,
d-a+l,b-a+l}
R-Nasal
{m-a+m,m-a+n,zh-a+l}
R-Lateral
Yes
No
Yes
No
{d-a+m,b-a+n}
{d-a+l,b-a+l}
{zh-a+l}
{m-a+m,m-a+n}

问题集—自动生成

- 设 $S=\{s_1,s_2,s_3,\cdots,s_N\}$ 为一个分类的状态集， N 为所有的状态数，状态集整体的均值和方差为μ_S和Σ_S。
- 设$X=\{x_1,x_2,s_3,\cdots,x_T\}$为训练数据集，$T$为所有的帧数。$t$时刻处于状态$s_n$的概率为$\gamma_t(s_n)$，$\gamma_t(s_n)$ 可在帧与状态对齐时，由前向后向算法得到，只需计算一次。
- 整个数据集X由状态集S产生的似然概率为：

$$L(S)=\sum_t\left[lnP(x_t|\mu_S,\Sigma_S)\sum_n\gamma_t(s_n)\right]$$

问题集—自动生成

- 针对高斯概率密度函数$P(x_t|\mu_S,\Sigma_S)$，可展开如下：

$$
\begin{aligned}
&lnP(x_t|\mu_S,\Sigma_S)\\
&= ln\left(\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\Sigma_S|^{\frac{1}{2}}}exp\left\{-\frac{1}{2}(x_t-\mu_S){\Sigma_S}^{-1}(x_t-\mu_S)^T\right\}\right)\\
&= -\frac{1}{2}\left(D\ln(2\pi)+\ln(|\Sigma_S|)+(x_t-\mu_S){\Sigma_S}^{-1}(x_t-\mu_S)^T\right)
\end{aligned}
$$

问题集—自动生成

Σ_S可计算如下：

$$\Sigma_S = \frac{\sum_t \{(\sum_n \gamma_t(s_n))(x_t - \mu_S)(x_t - \mu_S)^T\}}{\sum_t \sum_n \gamma_t(s_n)}$$

- 进一步推导得到：

$$\sum_t \left\{ (x_t - \mu_S) {\Sigma_S}^{-1} (x_t - \mu_S)^T \sum_n \gamma_t(s_n) \right\} = D \sum_t \left(\sum_n \gamma_t(s_n) \right)$$

其中D是数据特征的维度。

问题集—自动生成

因此，我们有

$$L(S) = -\frac{1}{2}\left(D\left(1+\ln(2\pi)\right)+\ln(|\Sigma_S|)\right)\sum_t\left(\sum_n \gamma_t(s_n)\right)$$

- 把状态集S分成两类，得到S_L和S_R，分别计算新的均值和方差。

分类后的似然增益：

$$\Delta = L(S_L) + L(S_R) - L(S)$$

如果$\Delta > 0$，说明状态集分类后有似然增益。

问题集—自动生成步骤

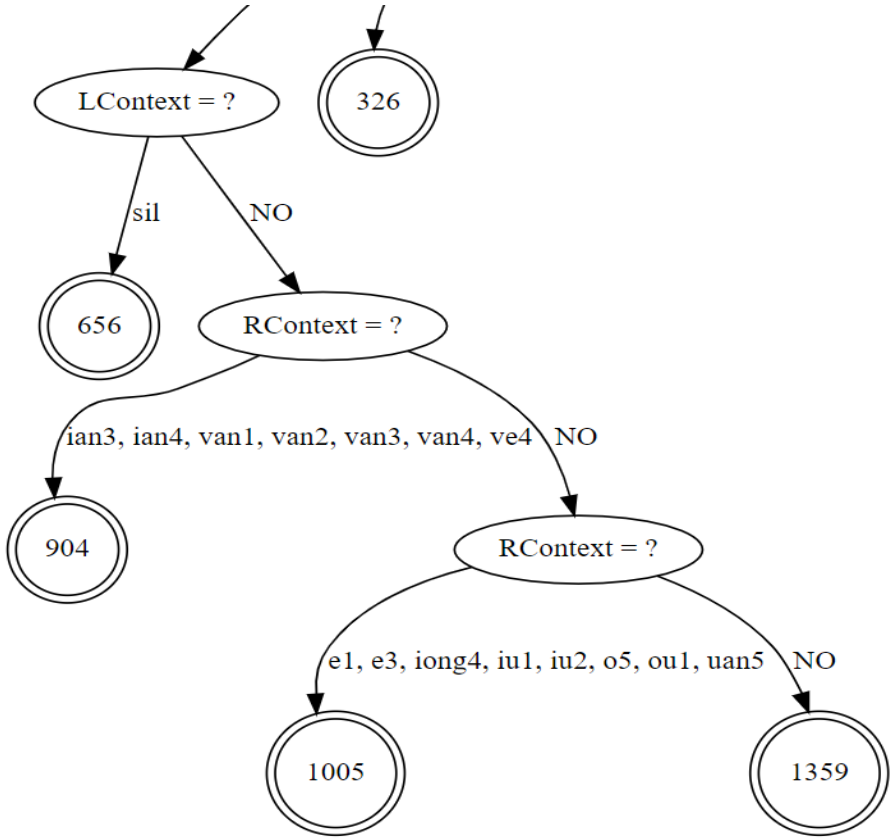
- **计算统计量：** 对于每帧语音x_t，由其Viterbi对齐序列可以得知对应哪个音素的某个状态。对每个单音子做上下文的扩展（如“a”=>“*-a+*”），得到三音子某个状态s_n对应的特征序列，然后计算该三音子状态对应的统计量$\gamma_t(s_n)$。

- **音素状态聚类：** 首先将每个中间音素（如“*-a+*”）作为树的根节点，进行二叉树分裂，根据所有可能的划分，分别利用似然度计算公式计算似然概率，得到能带来最大似然度提升的最优划分。然后不断的递归对节点进行二叉树分裂，就可以得到一颗音素聚类的决策树。一旦某一层两个节点的最优划分带来的似然增益Δ的值都小于设定的阈值，则该层终止分裂。

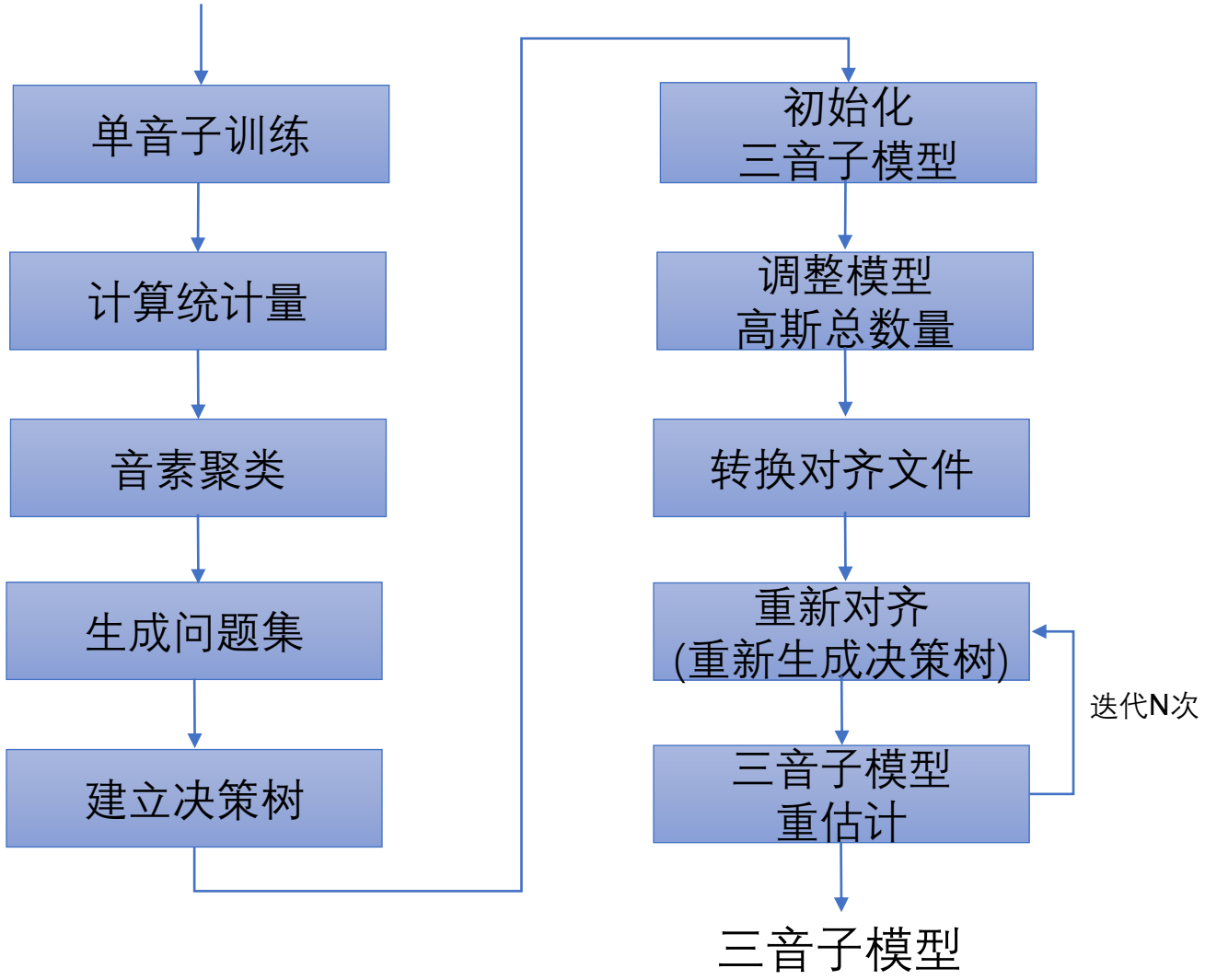
- **终止分裂：** 最终形成一棵聚类树，每个叶子结点都包含不同的音素集合，一个叶子结点就是被聚类后的一个类。

- **生成问题集：** 从根结点开始，向下统计所能到达的所有叶子结点中包含的音素集合，把这个音素集合的统计集合作为一个问题集。

决策树的一部分


LContext = ?
326
sil
NO
656
RContext = ?
ian3, ian4, van1, van2, van3, van4, ve4
NO
904
RContext = ?
e1, e3, iong4, iu1, iu2, o5, ou1, uan5
NO
1005
1359

6.6.5 三音子模型的训练



6.7 本章小结

- 针对HMM在语音识别的具体应用，介绍了不同的建模单元，以及建模单元与HMM之间的关系，特别是发音过程与HMM状态的对应，从而构建了适用的声学模型。
- 基于串接HMM的解码过程相对复杂，尤其是随机语法的识别存在词数未知、词位未知等困难，本章通过图示详细地描述了Viterbi解码步骤。
- 本章还针对协同发音现象，介绍了音素的上下文建模方法，包括双音子和三音子。
- 为了减少三音子模型参数，比较有效的方案是让不同模型状态共享概率密度函数。共享的原则是基于决策树，该决策树每个节点对应一个问题。

Thank you!

Any questions?