POST

Sequence Modelling.

IE

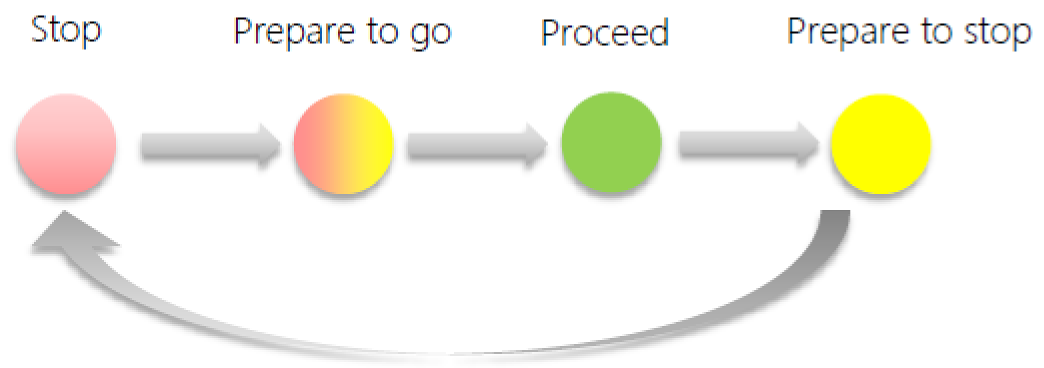
Named Entity Recognition

Hidden Markov Models

Ack: Derek Kane

Hidden Markov Models

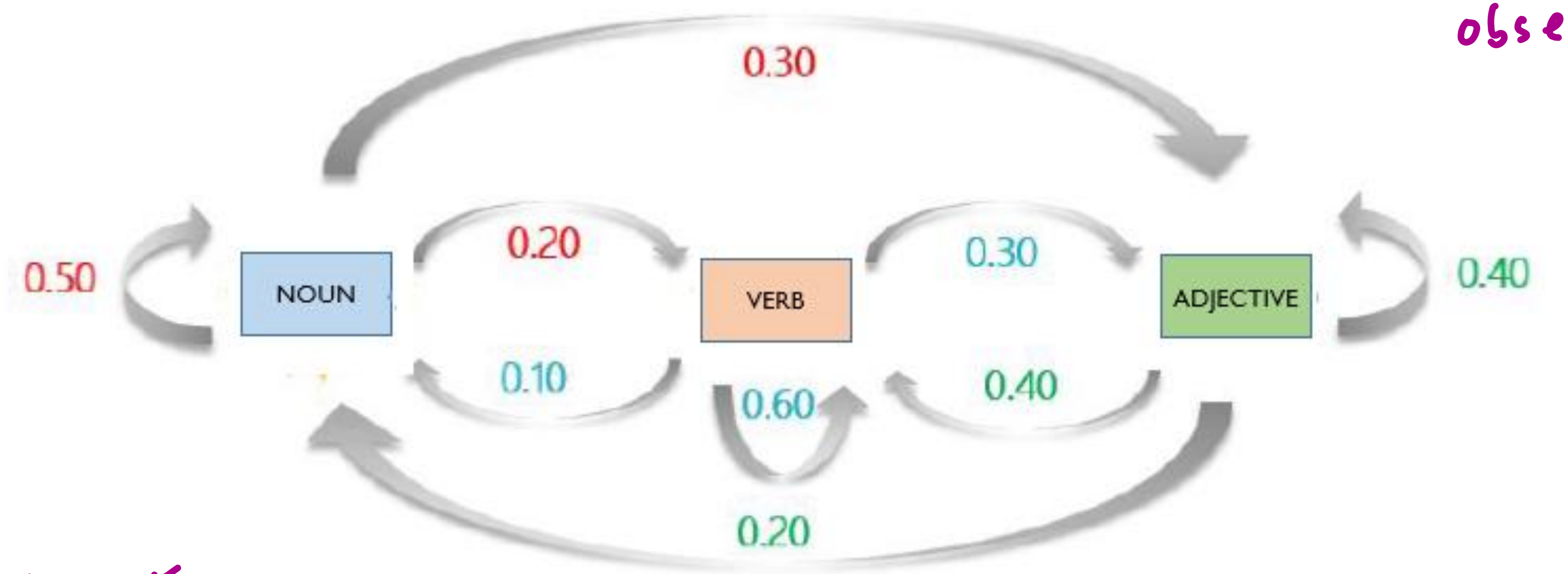
- Foundation: Markov Processes
- Observable Markov Process: Traffic Light example



Hidden Markov Model: an example

hidden state : POS tags
observation : words.

WSD
POST



Ram is happy

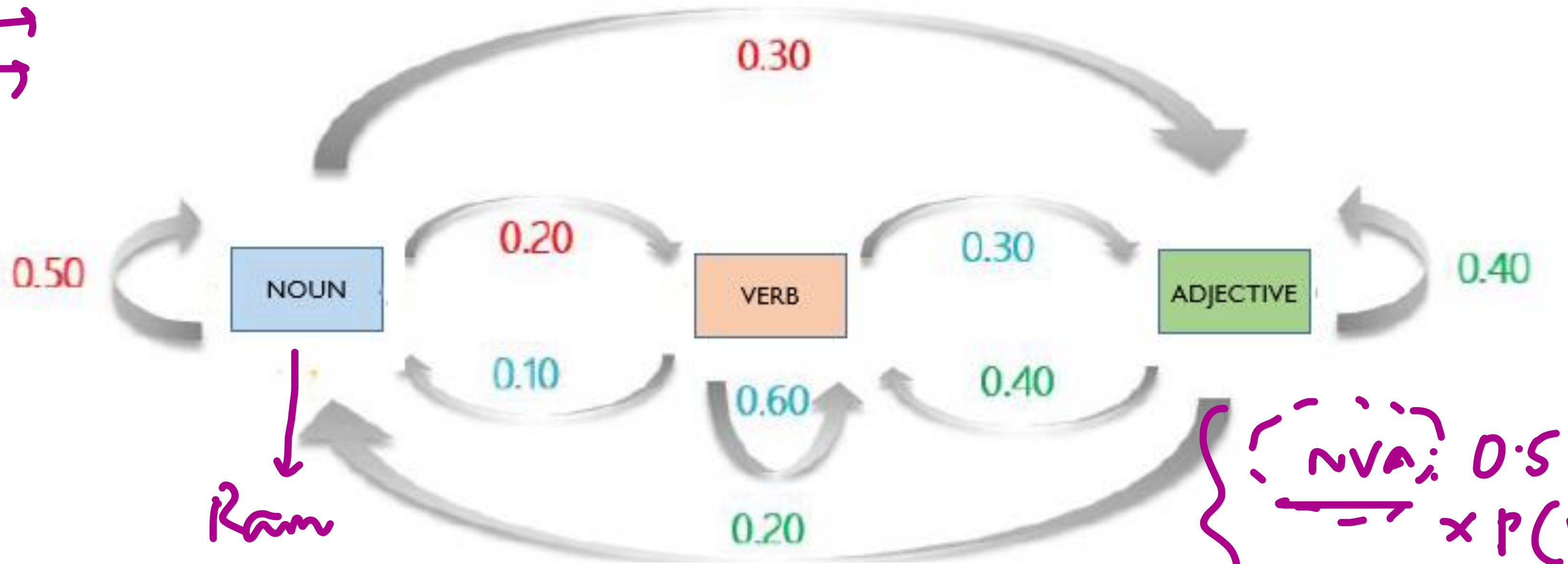
Transition Probabilities

	N	V	A
N	0.50	0.20	0.30
V	0.10	0.60	0.30
A	0.20	0.40	0.40

Hidden Markov Model: example (contd)

NNN →
N ~~N~~ V →
⋮
$\boxed{\text{NVA}}$

AAA



Ram

M states (POS tags)

N types (dist. words)

NVA: $0.5 \times p(\text{Ram}|N) \times p(V|N) \times p(\text{is}|V) \times p(A|V) \times p(\text{hap}$

Transition Probabilities MxM

	N	V	A
N	0.50	0.20	0.30
V	0.10	0.60	0.30
A	0.20	0.40	0.40

Emission Probabilities MxN

	Ram	is	happy
N	0.8	0.1	0.1
V	0.1	0.8	0.1
A	0.1	0.1	0.8

P vector

π

N	V	A
0.5	0.3	0.2

→ $[0.5 \times 0.8 \times 0.2 \times 0.8 \times 0.3 \times 0.8$

Three Problems

HMM_1 HMM_2 HMM_3

- Evaluation: Given a HMM, find probability of an observed sequence

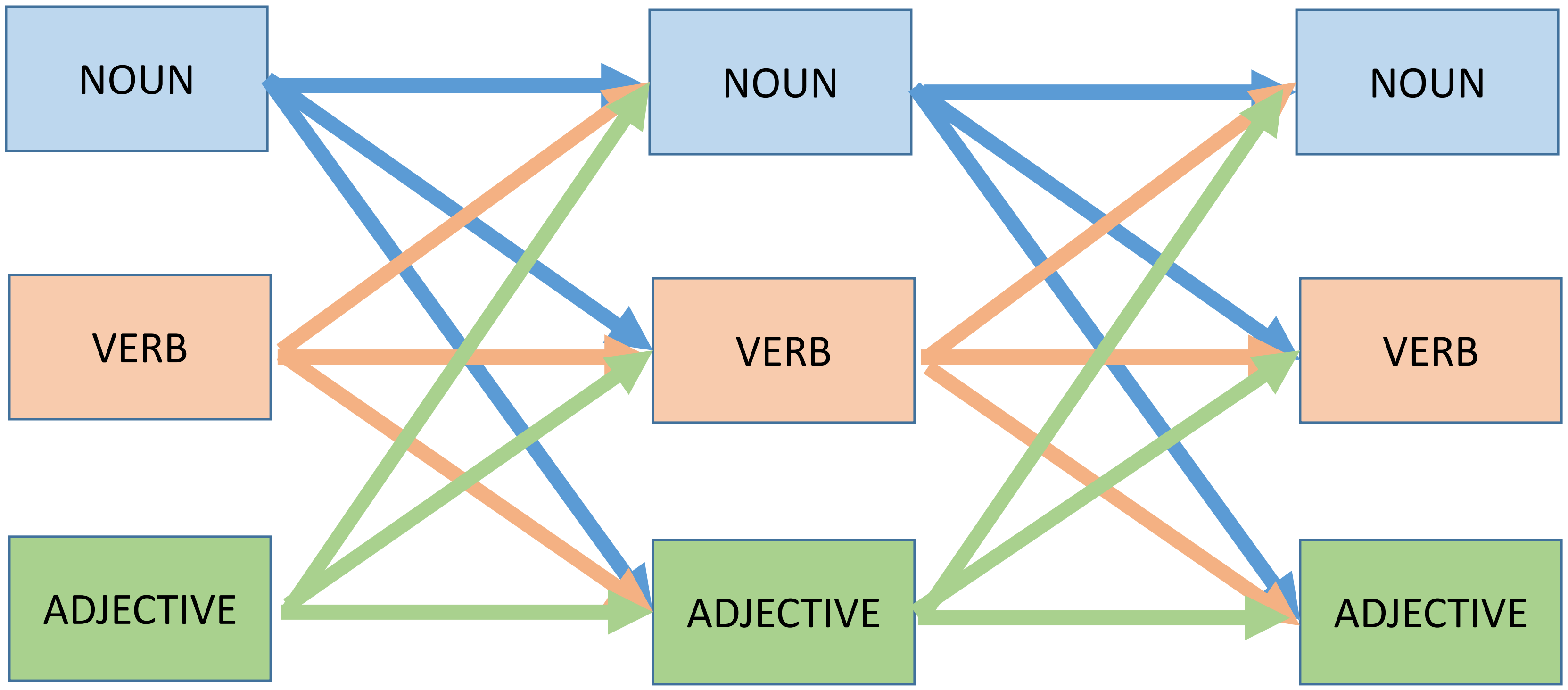
POST

- Decoding: Given a HMM, find most probable sequence of states that could have given rise to an observation.

- Learning: Generate a HMM that best explains a given set of observations

The forward algorithm

[Evaluation problem.]

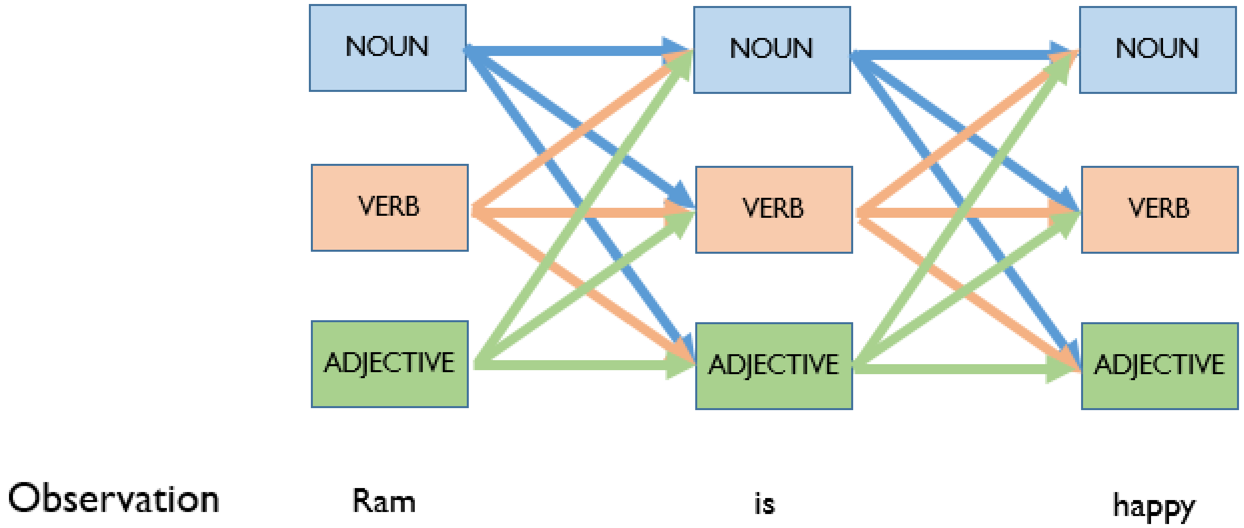


Observation Ram is happy

Goal : We want to find the probability of the observed sequence, given an HMM whose parameters are known.

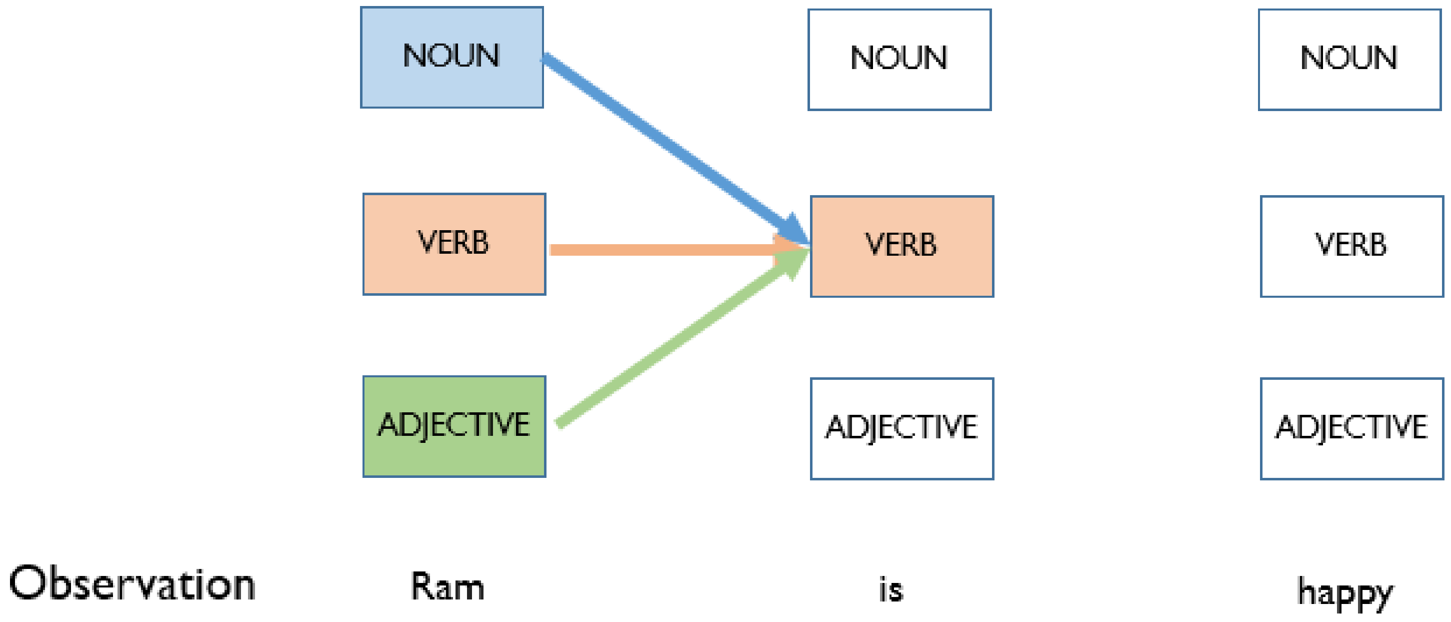
Brute Force Approach

{Dynamic Programming



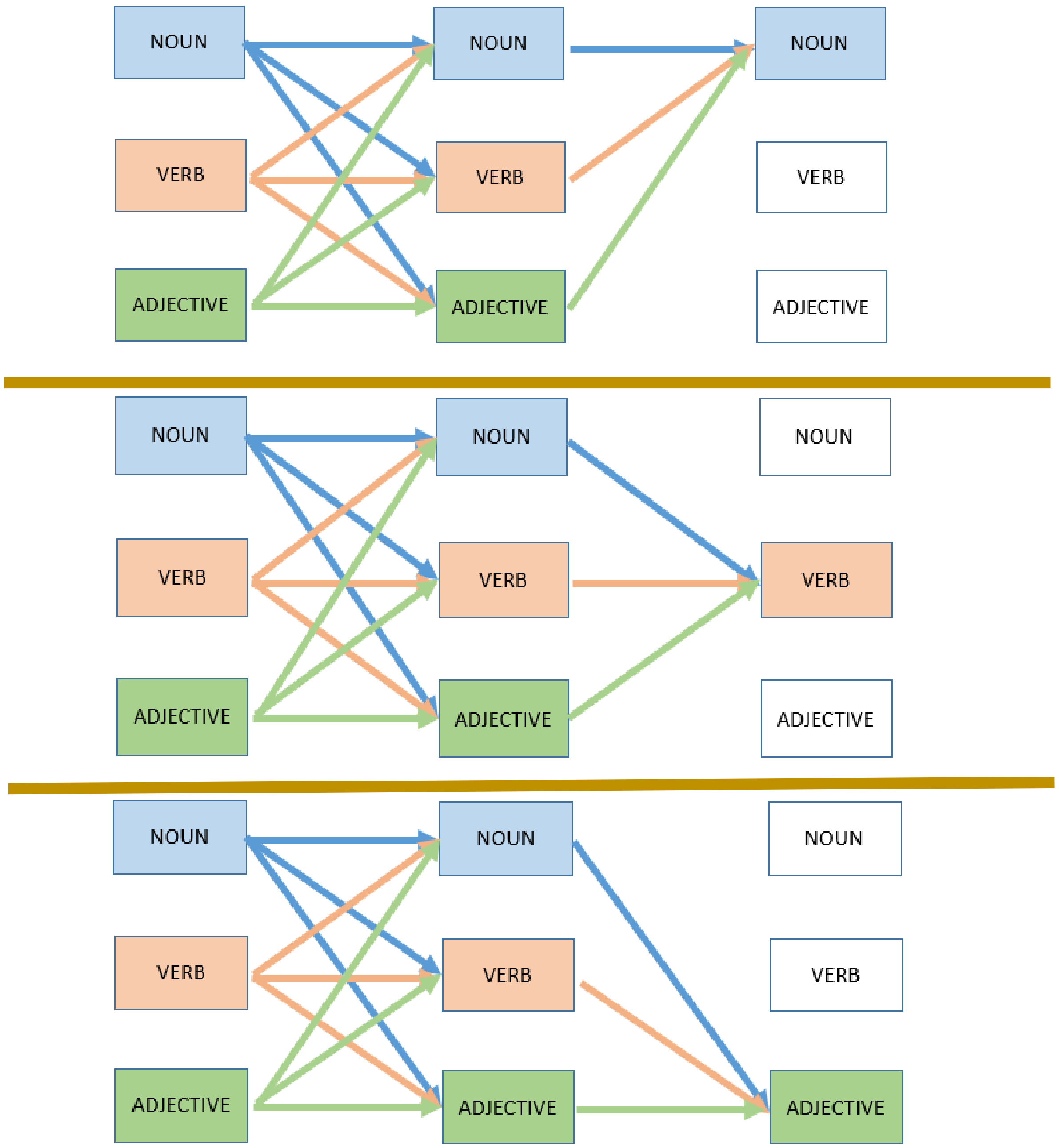
Sum over probabilities of each possible sequence of state transitions that could have given rise to the observed sequence , there are 27 such possibilities in this simple example.

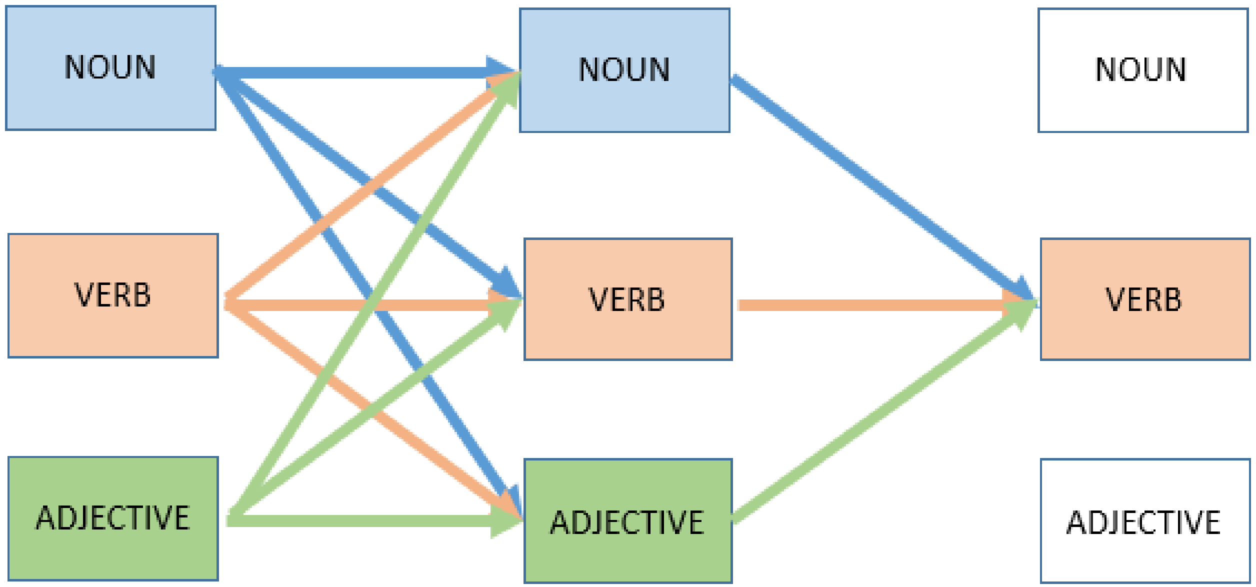
Observation



We can calculate the probability of reaching an intermediate state in the trellis as the sum of all possible paths to that state.

For example, the probability of being at a verb at t = 2 is calculated from the paths.

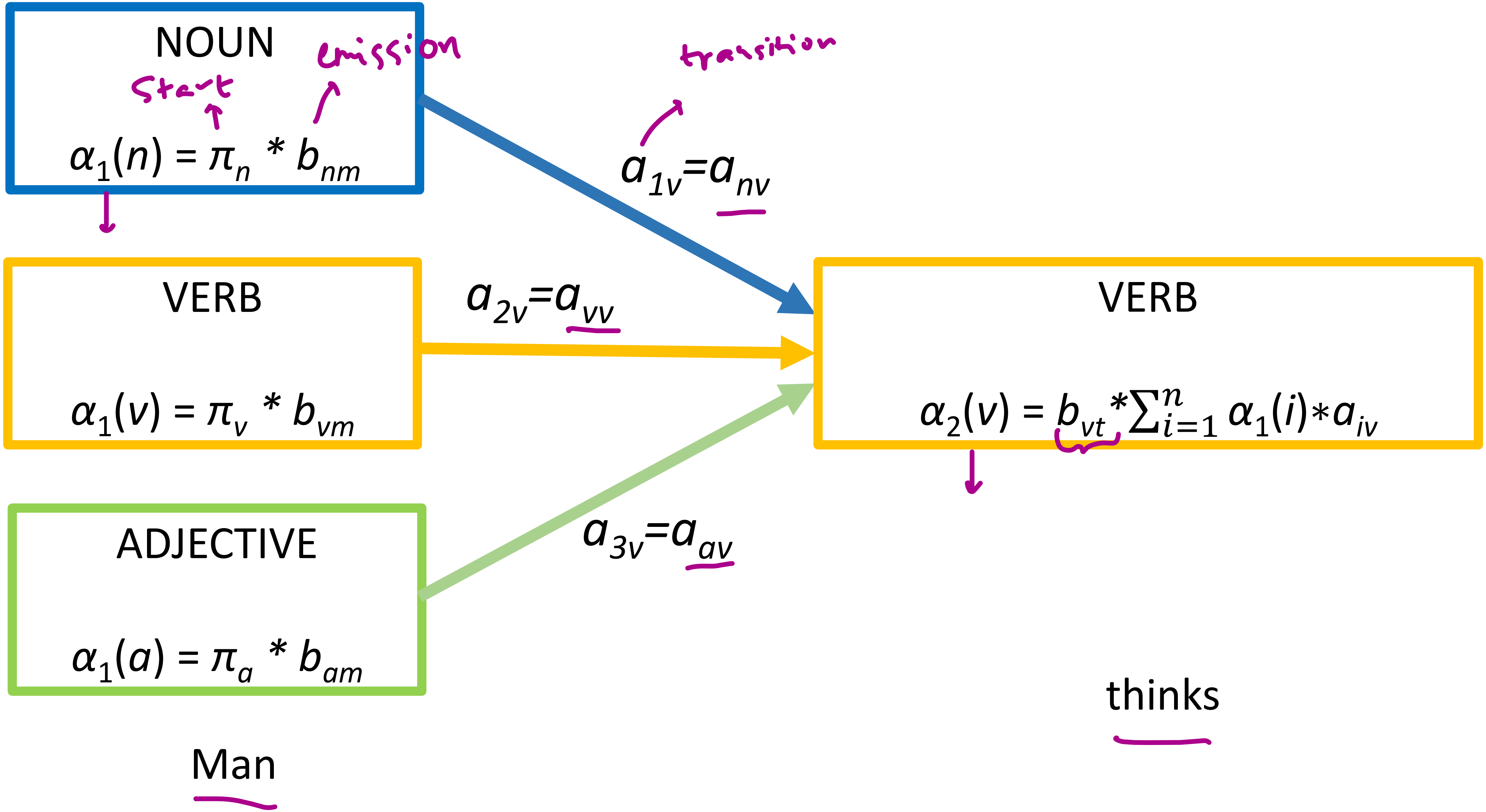

NOUN
VERB
ADJECTIVE
NOUN
VERB
ADJECTIVE
NOUN
VERB
ADJECTIVE
NOUN
VERB
ADJECTIVE
NOUN
VERB
ADJECTIVE
NOUN
VERB
ADJECTIVE
NOUN
VERB
ADJECTIVE
NOUN
VERB
ADJECTIVE
NOUN
VERB
ADJECTIVE



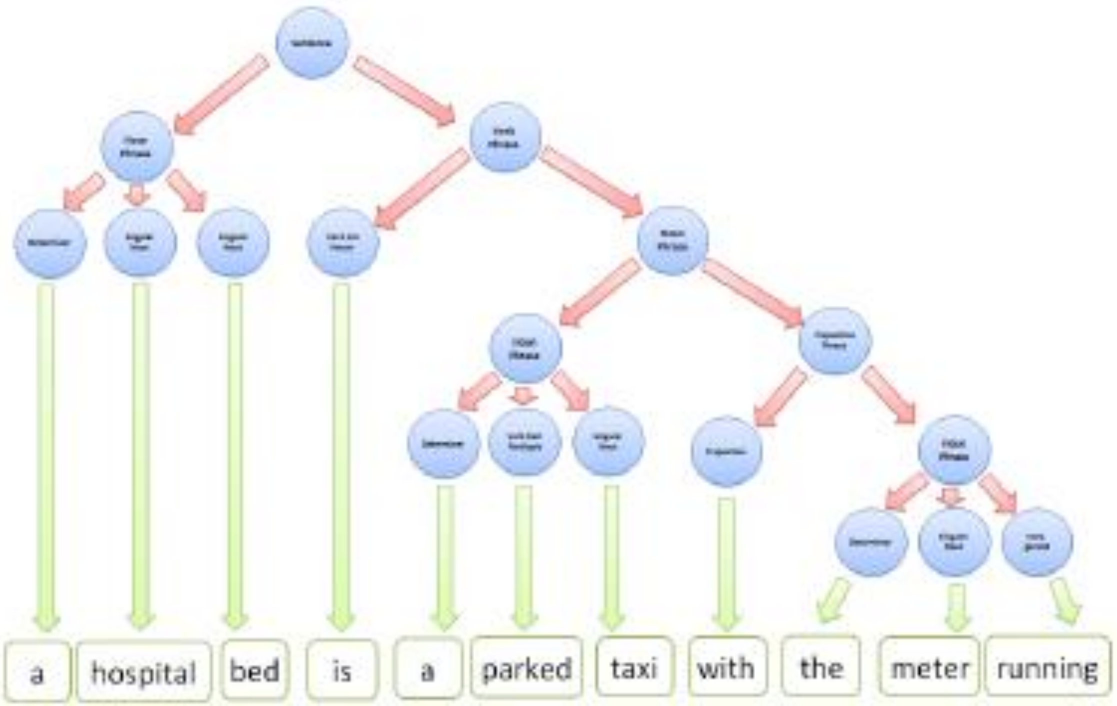
t=0

t=1,
3 routes to
each state

t=2,
9 routes to
each state


NOUN
start
emission
$\alpha_1(n) = \pi_n * b_{nm}$
transition
$a_{1v}=a_{nv}$
VERB
$\alpha_1(v) = \pi_v * b_{vm}$
$a_{2v}=a_{vv}$
VERB
$\alpha_2(v) = b_{vt} * \sum_{i=1}^{n} \alpha_1(i) * a_{iv}$
ADJECTIVE
$\alpha_1(a) = \pi_a * b_{am}$
$a_{3v}=a_{av}$
Man
thinks

Google Translate



Key insight

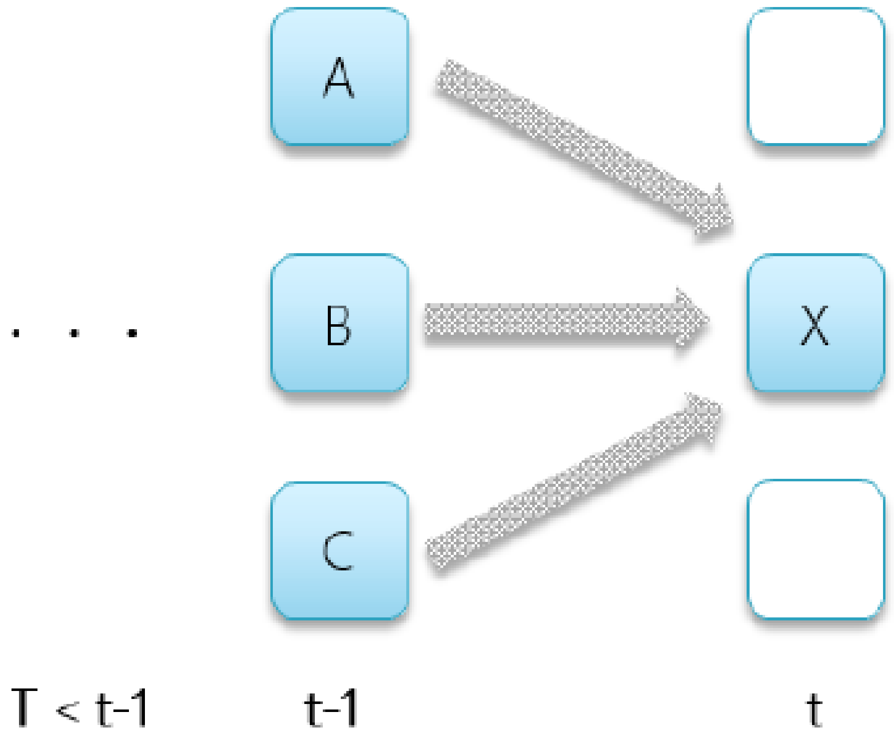
We need to find the most probable path to the state X at time t

Consider the trellis to the right

The most probable path through X must pass through one of A, B or C at time t-1

Therefore the most probable path to X will be one of

sequence of states , … ,A,X
sequence of states , … ,B,X
sequence of states , … ,C,X



Key insight

The most probable path to X will be one of

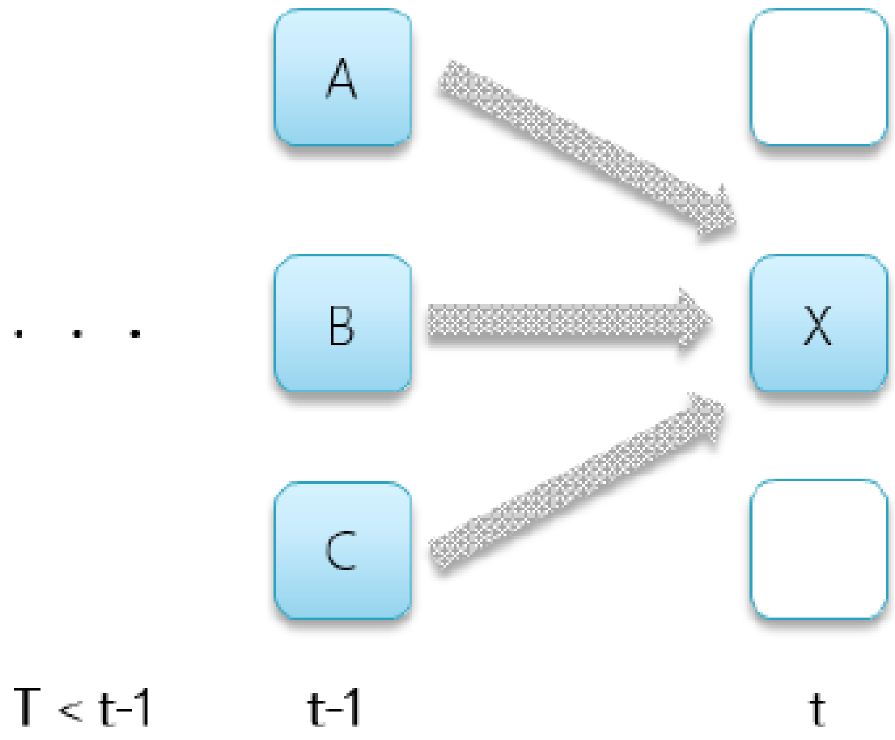
sequence of states , … ,A,X

sequence of states , … ,B,X

sequence of states , … ,C,X

1st order Markov assumption: probability of X occurring after a sequence depends only on the previous state

P(most probable path to A) = P(X|A).P(Obs|X)



Generalizations

$$\Pr(X\ at\ time\ t) = \max_{i=A_iB_iC_i} \Pr\left(i\ at\ time\ (t-1)\right) * \Pr(X|i) * \Pr(obs.\ at\ time\ t\ |\ X)$$

$$\delta_t(i) = \max_j(\delta_{t-1}(j) a_{ij} b_{ik_t})$$

The idea of back pointers

Viterbi Algorithm

NOUN

$\delta_1(n) = \pi_n * b_{nm}$

VERB

$\delta_1(v) = \pi_v * b_{vm}$

ADJECTIVE

$\delta_1(a) = \pi_a * b_{am}$

Man

$a_{1v}=a_{nv}$

$a_{2v}=a_{vv}$

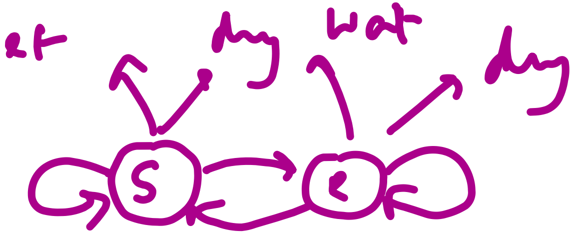
$a_{3v}=a_{av}$

VERB

$\delta_2(v) = max_i(\delta_1(i)* a_{iv})*b_{vt}$

$\varphi_2(v) = argmax_i(\delta_1(i)* a_{iv})$

thinks



Baum Welch Initialization

	~ ~~Sunny~~	✓ ~~Rainy~~
~ ~~Sunny~~	0.50	0.50
✓ ~~Rainy~~	0.40	0.60

Man thinks

	~~Wet~~	~~Dry~~
~ ~~Sunny~~	0.20	0.80
✓ Rainy	0.90	0.10

Complete Data	Incomplete Data
~ ✓ A	
	do not come in POSTs

EM

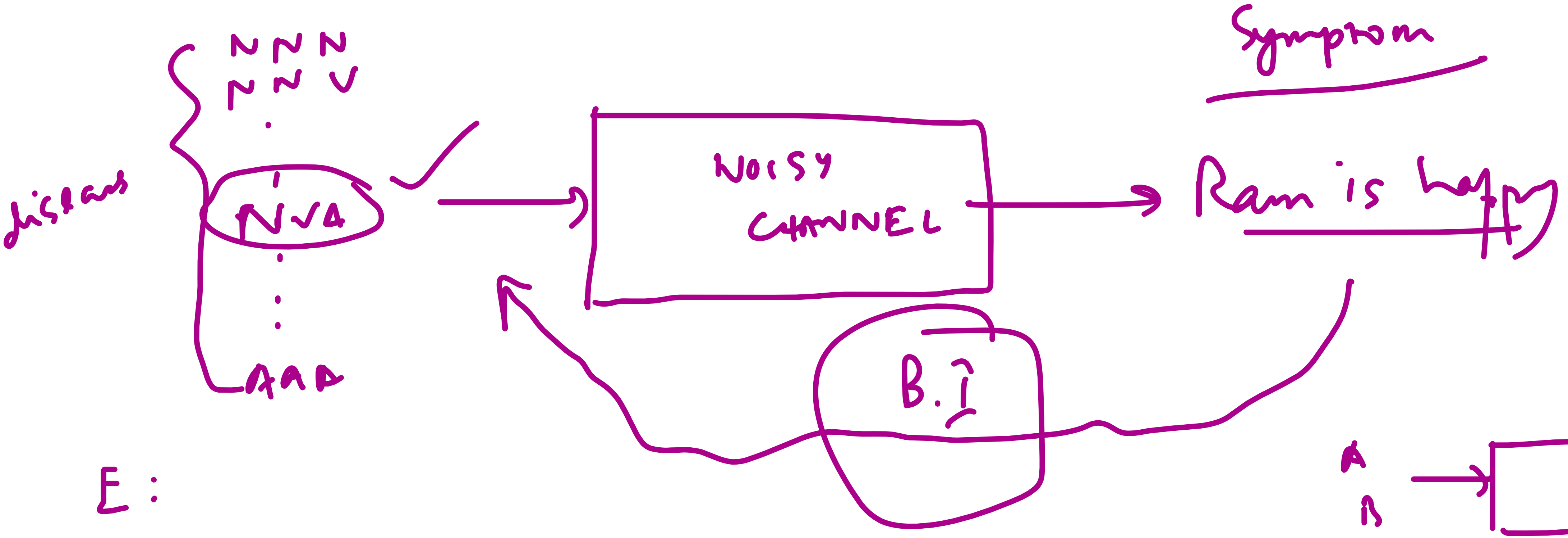
$p(\cancel{Sunny}) = 0.3$ and $p(\cancel{Rainy}) = 0.7$.

Mo:

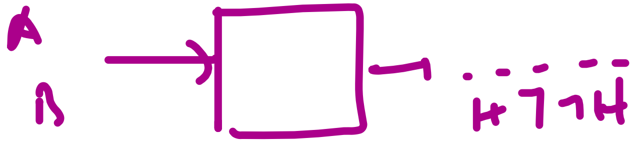
E step → Preparatory

{Forward model}

Ram is happy



E :



P(NNN | Ram is happy)

FM Part 2

Group:

$$P(S|O) = \frac{P(O|S)\,P(S)}{\sum_i P(O|S_i) \cdot P(S_i)}$$

generative process

Preparation

FM Part 1

P(A|O) =

P(B|O) =

Baum Welch

Observations:

WW, WW, WW, WW, WD, DD, DW, WW, WW

W = Wet

D = Dry

If WW came from Sunny → Rainy sequence, the probability would be:

$\boxed{0.3*0.2*0.5*0.9 = 0.027}$

π_S P(W|S) emission P(R|S) state P(W|R) emission

	Sunny	Rainy
Sunny	0.50	0.50
Rainy	0.40	0.60

	Wet	Dry
Sunny	0.20	0.80
Rainy	0.90	0.10

p(Sunny) = 0.3 and p(Rainy) = 0.7.

Homework: complete the E_1 and M_1 steps

Baum Welch

- The revised parameters feed into the next E step, which in turn leads to fresh estimation of parameters
- The process is repeated till parameters converge