CS626: Speech, NLP and the Web

POS Tagging

Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay

Week of 17th August, 2020

Part of Speech Tagging

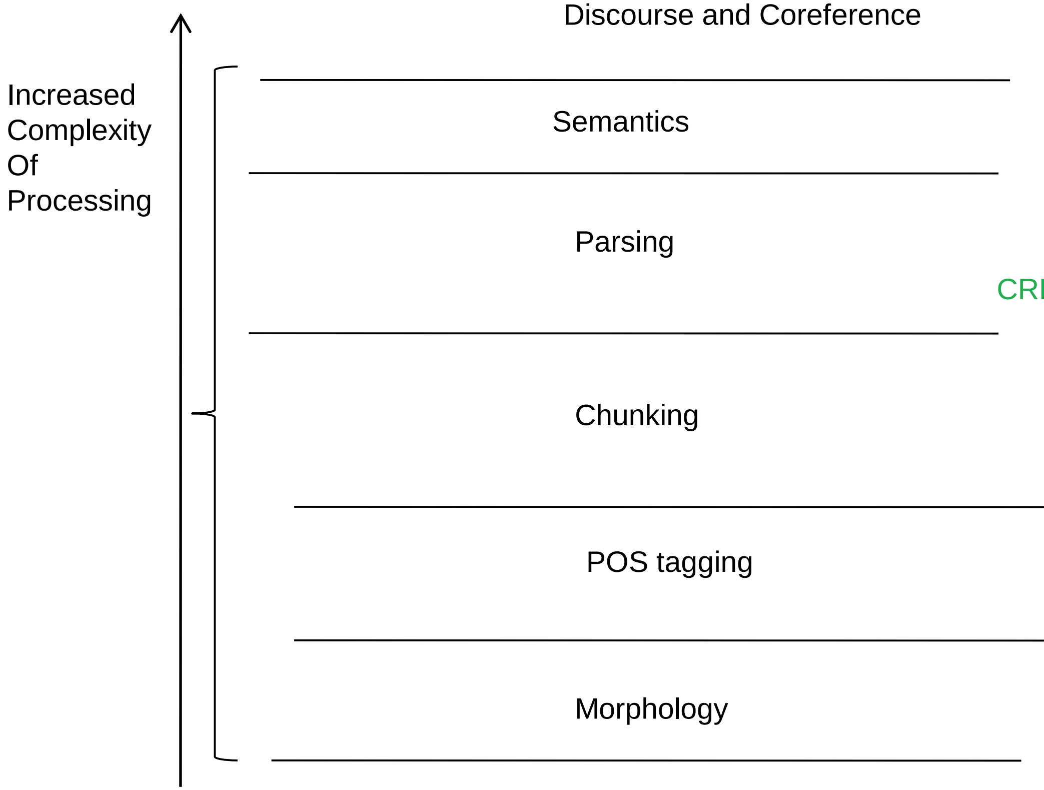
Agenda

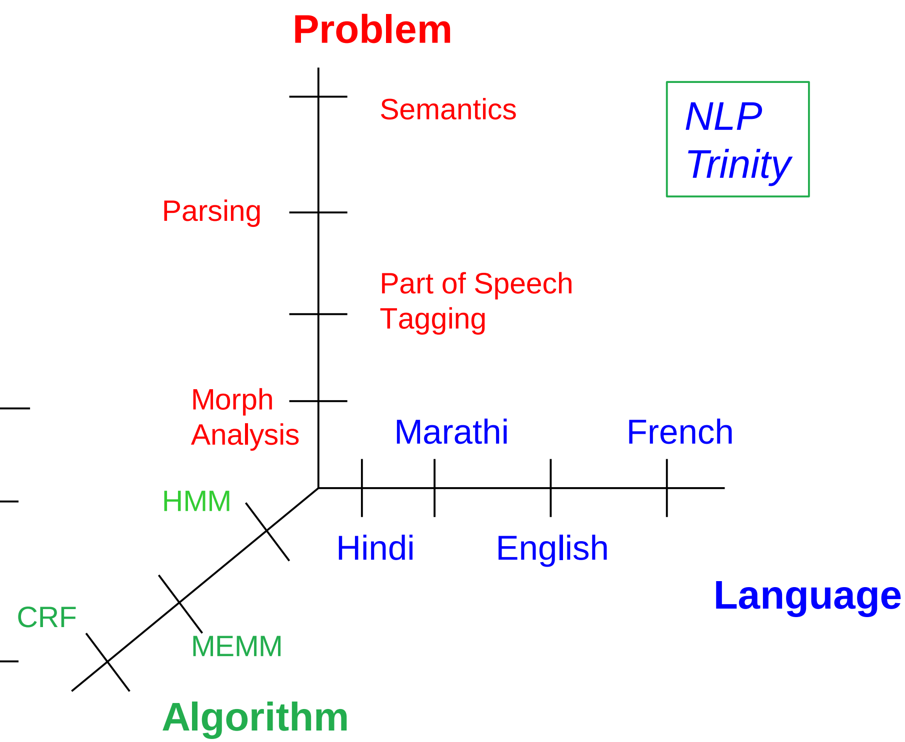
- Rule Based POS Tagging

- Statistical ML based POS Tagging (*Hidden Markov Model, Support Vector Machine*)

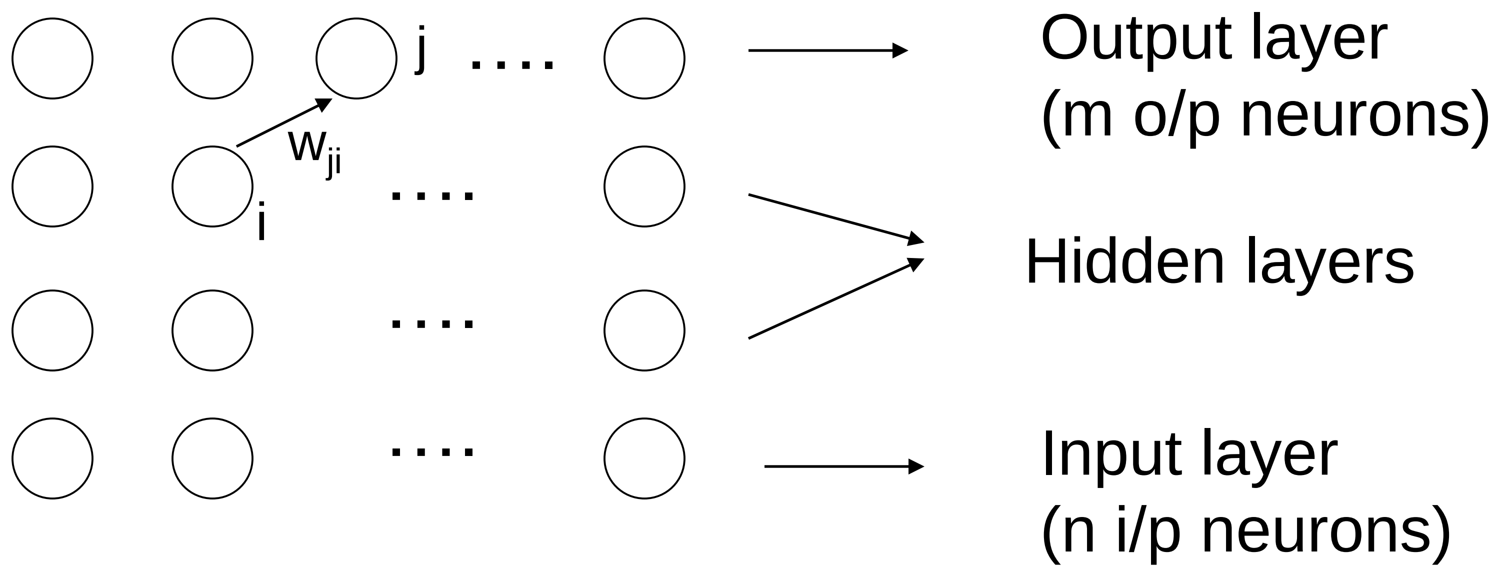
- Neural (Deep Learning) based POS Tagging

NLP: multilayered, multidimensional





Multilayer neural net



- NLP pipeline ←→ NN layers
- Discover bigger structures bottom up, starting from character?
- Words, POS, Parse, Sentence,

Subwords (for “jaauMgaa”, जाऊंगा)

- Characters: “j+aa+u+M+g+aa”
- Morphemes: “jaa”+”uMgaa”
- Syllables: “jaa”+”uM”+”gaa”
- Orthographic syllables: “jaau”+”Mgaa”
- BPE (depends on corpora, statistically frequent patterns): both “jaa” and “uMgaa” are likely

NLP Layer

What a gripping movie was Three_Idiots!

What/WP a/DT gripping/JJ movie/NN was/VBD Three_Idiots/NNP !/!

Parse

```
(ROOT
  (FRAG
    (SBAR
      (WHNP
        (WP What))
        (S
          (NP
            (DT a)
            (JJ gripping)
            (NN movie)
          )
          (VP
            (VBD was)
            (NP
            (NNP Three_idiots)))))
          (. !)
  )
)
```

Universal dependencies

dobj(Three_Idiots-6, What-1)
det(movie-4, a-2)
amod(movie-4, gripping-3)
nsubj(Dangal-6, movie-4)
cop(Dangal-6, was-5)
root(ROOT-0, Three_idiots-6)

Part of Speech Tagging

- Attach to each word a tag from **Tag-Set**
- Standard Tag-set : Penn Treebank (for English).

POS ambiguity instances

best ADJ ADV NP V
better ADJ ADV V DET

close ADV ADJ V N (*running close to the competitor, close escape, close the door, towards the close of the play*)

cut V N VN VD
even ADV DET ADJ V
grant NP N V –
hit V VD VN N
lay ADJ V NP VD
left VD ADJ N VN
like CNJ V ADJ P –
near P ADV ADJ DET
open ADJ V N ADV
past N ADJ DET P
present ADJ ADV V N
read V VN VD NP
right ADJ N DET ADV
second NUM ADV DET N
set VN V VD N –
that CNJ V WH DET

POS Ambiguity

- E.g.
  1. *What a **gripping** movie was Abhiman!*
  2. *He is **gripping** it firm.*

Adjective

Verb

Linguistic fundamentals

- A word can have two roles
  - Grammatical role (Dictionary POS tag)
  - Functional role (Contextual POS tag)

  - E.g. ***Golf** stick*
- POS tag of "*Golf*"
  - Grammatical: Noun
  - Functional: Adjective (+ al)

The “al” rule!

- If a word has different functional POS tag than its grammatical pos then add **“al”** to the functional POS tag

- E.g. ***<u>Golf</u>*** *stick*

Adjective + al

Adjectival

Noun	**+ al**	**= Nominal**
Verb	**+ al**	**= Verbal**
Adjective	**+ al**	**= Adjectival**
Adverb	**+ al**	**= Adverbial**

Dictionary meaning of “Golf”

noun

- a game in which clubs with wooden or metal heads are used to hit a small, white ball into a number of holes, usually 9 or 18, in succession,
- situated at various distances over a course having natural or artificial obstacles, the object being to get the ball into each hole in as few strokes as possible.
- a word used in communications to represent the letter *G.*

Golf stick

verb

(used without object) to play golf.

We golfed the whole day in the weekend

The "al" rule cntd.

- Examples:
  - Nominal
    - *Many don't understand the problem of* ***hungry****.*
  - Adverbial
    - *Come* ***quick.***
  - Verbal ?

adjective, **hun·gri·er, hun·gri·est.**
having a desire, craving, or need for food; feeling hunger. indicating, characteristic of, or characterized by hunger:

He approached the table with a hungry look.

strongly or eagerly desirous.
lacking needful or desirable elements; not fertile; poor:
hungry land.

marked by a scarcity of food:
The depression years were hungry times.

Learning POS Tags

- Question
  - Is one instance of example enough for ML?
  - E.g. common example of “people”

    People → Noun

    POS Ambiguity

  - But it can be verb as well

    People → Verb (to populate)

- Answer
  - We need at least as many instances as number of different labels *#POS tags-1* to make decision.

Disambiguation of POS tag

- If no ambiguity, learn a table of words and its corresponding tags.

- If ambiguity, then look for the contextual information i.e. look-back or look-ahead.

Data for “present”

He gifted me the/a/this/that **present_NN**.

They **present_VB** innovative ideas.

He was **present_JJ** in the class.

Rules for disambiguating “present”

- For Present_NN (look-back)
  - ***If present is preceded by determiner (the/a) or demonstrative (this/that), then POS tag will be noun.***

- Does this rule guarantee 100% precision and 100% recall?
  - False positive:
    - *The **present_ADJ** case is not convincing.*

      Adjective preceded by “the”
  - False negative:
    - ***Present** foretells the future.*

      Noun but not preceded by “the”

Rules for disambiguating “present”

- For Present_NN (look-back and look ahead)
  - ***If present is preceded by determiner (the/a) or demonstrative (this/that) or followed by a verb, then POS tag will be noun*.**
  - E.g.
    - **Present_NN** will tell the future.
    - **Present_NN** fortells the future.
- Does this rule guarantee 100% precision and 100% recall?

Need for ML in POS tagging

- Rules are challenged by new data
- Need a robust system.

- Machine learning based POS tagging:
  - HMM (Accuracy increased by 10-20% against rule based systems)
  - Jelinek’s work inspired from ASR

Noisy Channel Model

W Noisy Channel **T**

$(w_n, w_{n-1}, \ldots, w_1)$ $(t_m, t_{m-1}, \ldots, t_1)$

Sequence *W* is transformed into sequence *T*

$$T^* = \underset{T}{\operatorname{argmax}}(P(T|W))$$

$$W^* = \underset{W}{\operatorname{argmax}}(P(W|T))$$

Mathematics of POS tagging

Argmax computation (1/2)

Best tag sequence
= T*
= argmax P(T|W)
= argmax P(T)P(W|T)   (by Baye's Theorem)

$$
\begin{aligned}
\mathbf{P(T)} &= \mathbf{P(t_0{=}\hat{}\; t_1 t_2 \ldots t_{n+1}{=}.)} \\
&= \mathbf{P(t_0)P(t_1|t_0)P(t_2|t_1t_0)P(t_3|t_2t_1t_0) \ldots} \\
&\qquad \mathbf{P(t_n|t_{n-1}t_{n-2}\ldots t_0)P(t_{n+1}|t_nt_{n-1}\ldots t_0)} \\
&= \mathbf{P(t_0)P(t_1|t_0)P(t_2|t_1) \ldots P(t_n|t_{n-1})P(t_{n+1}|t_n)} \\
&= \prod_{i=0}^{N+1} \mathbf{P(t_i|t_{i-1})}
\end{aligned}
$$

Bigram Assumption

Argmax computation (2/2)

$$P(W|T) = P(w_0|t_0\text{-}t_{n+1})P(w_1|w_0t_0\text{-}t_{n+1})P(w_2|w_1w_0t_0\text{-}t_{n+1}) \ldots P(w_n|w_0\text{-}w_{n-1}t_0\text{-}t_{n+1})P(w_{n+1}|w_0\text{-}w_nt_0\text{-}t_{n+1})$$

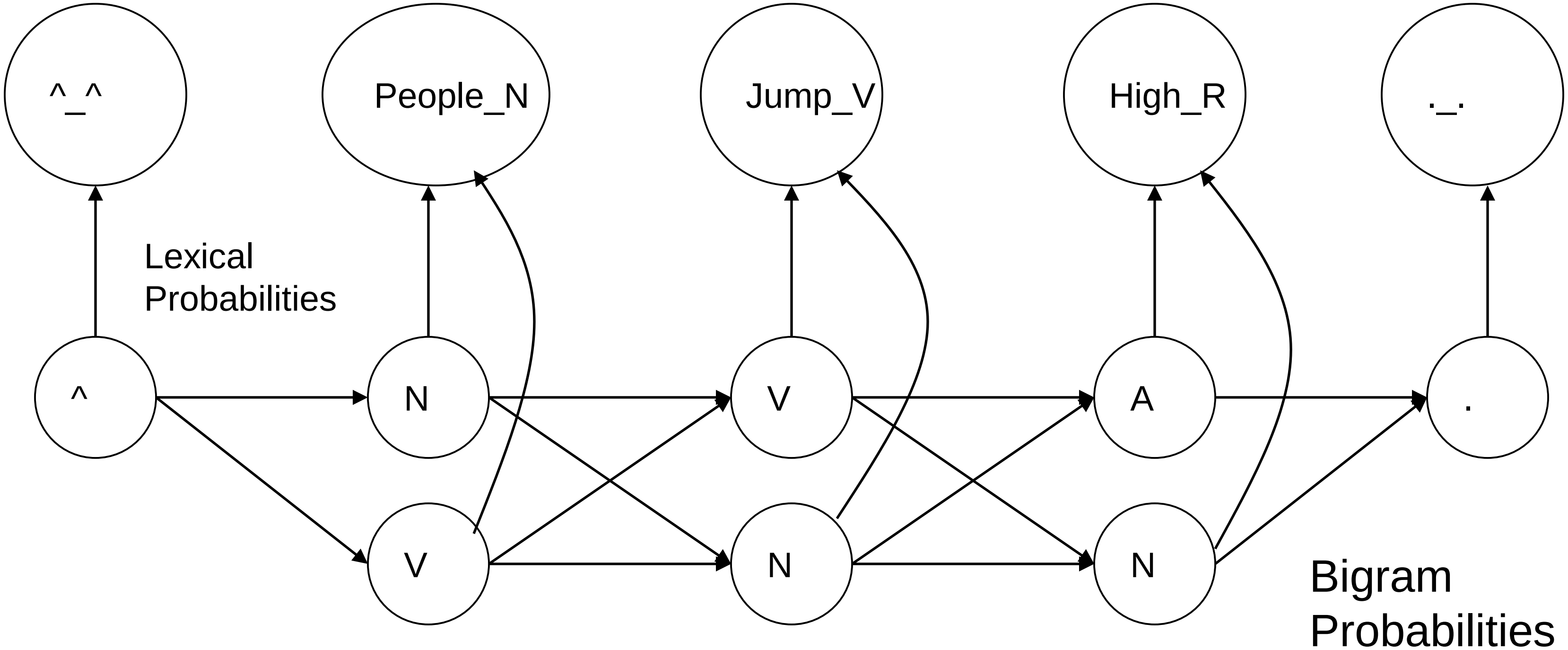
Assumption: A word is determined completely by its tag. This is inspired by speech recognition

$$= P(w_o|t_o)P(w_1|t_1) \ldots P(w_{n+1}|t_{n+1})$$

$$= \prod_{i=0}^{n+1} P(w_i|t_i)$$

$$= \prod_{i=1}^{n+1} P(w_i|t_i) \quad \text{(Lexical Probability Assumption)}$$

Generative Model



This model is called Generative model.
Here words are observed from tags as states.
This is similar to HMM.

Typical POS tag steps

- Implementation of Viterbi – Unigram, Bigram.
- Five Fold Evaluation.
- Per POS Accuracy.
- Confusion Matrix.

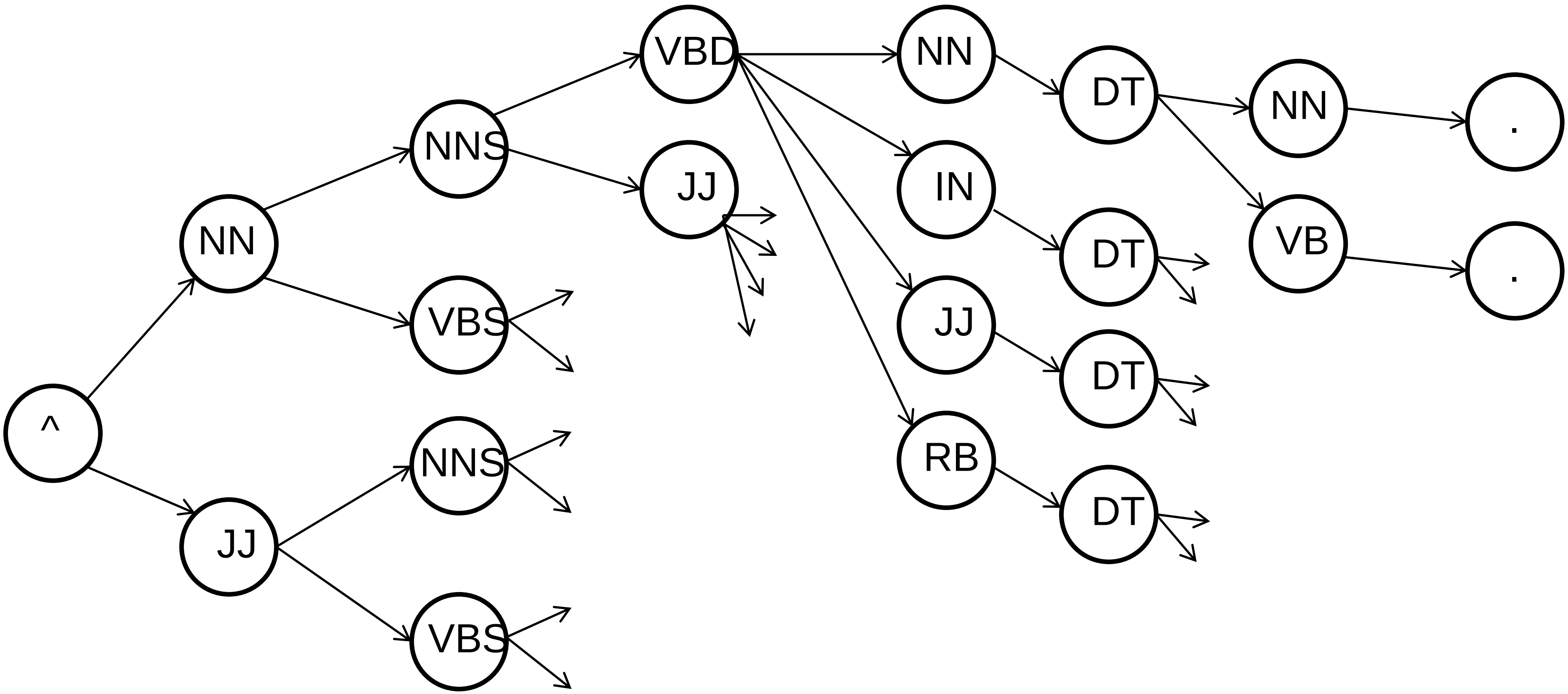
Screen shot of typical Confusion Matrix

	AJ0	AJ0-AV0	AJ0-NN1	AJ0-VVD	AJ0-VVG	AJ0-VVN	AJC	AJS	AT0	AV0	AV0-AJ0	AVP
AJ0	2899	20	32	1	3	3	0	0	18	35	27	1
AJ0-AV0	31	18	2	0	0	0	0	0	0	1	15	0
AJ0-NN1	161	0	116	0	0	0	0	0	0	0	1	0
AJ0-VVD	7	0	0	0	0	0	0	0	0	0	0	0
AJ0-VVG	8	0	0	0	2	0	0	0	1	0	0	0
AJ0-VVN	8	0	0	3	0	2	0	0	1	0	0	0
AJC	2	0	0	0	0	0	69	0	0	11	0	0
AJS	6	0	0	0	0	0	0	38	0	2	0	0
AT0	192	0	0	0	0	0	0	0	7000	13	0	0
AV0	120	8	2	0	0	0	15	2	24	2444	29	11
AV0-AJ0	10	7	0	0	0	0	0	0	0	16	33	0
AVP	24	0	0	0	0	0	0	0	1	11	0	737

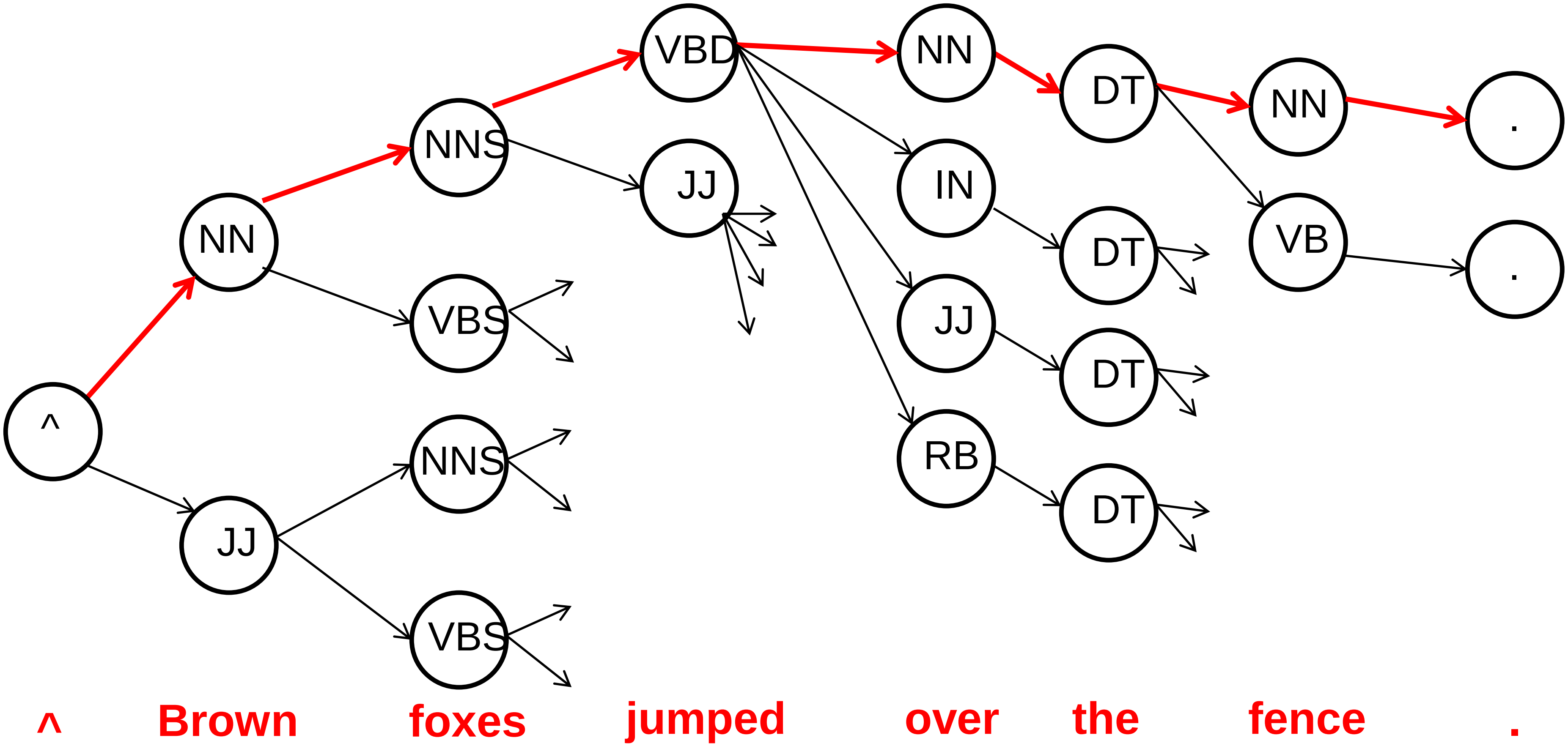
Computation of POS tags

DECODING

W:	^	Brown	foxes	jumped	over	the	fence	.
T:	^	JJ	NNS	VBD	NN	DT	NN	.
		NN	VBS	JJ	IN		VB	
					JJ			
					RB			



^ Brown foxes jumped over the fence .



Probability of a path (e.g. Top most path) = *P*(T) * *P*(W|T)

P(^) . P(NN|^) . P(NNS|NN) . P(VBD|NNS) . P(NN|VBD) . P(DT|NN) . P(NN|DT) . P(.|NN) . P(.)

*

P(^|^) . P(brown|NN) . P(foxes|NNS) . P(jumped|VBD) . P(over|NN) . P(the|DT) . P(fence|NN) . P(.|.)

Questions?

- Where do tags come from?
  - Tag set
- How to get probability values i.e. P(.)?
  - Annotated corpora

After modeling of the problem, emphasis should be on the corpus

Computing P(.) values

Let us suppose annotated corpus has the following sentence

I	**have**	**a**	**brown**	**bag**	**.**
PRN	**VB**	**DT**	**JJ**	**NN**	**.**

$$P(NN \mid JJ) = \frac{Number\_of\_times\_JJ\_followed\_by\_NN}{Number\_of\_times\_JJ\_appeared}$$

$$P(Brown \mid JJ) = \frac{Number\_of\_times\_Brown\_tagged\_as\_JJ}{Number\_of\_times\_JJ\_appeared}$$

Why Ratios?

- This way of computing parameter probabilities: is this <u>correct</u>?
- What does “correct” mean?
- Is this principled?
- We are using Maximum Likelihood Estimate (**<u>MLE</u>**)
- Assumption: underlying distribution is multinomial

Explanation with coin tossing

- A coin is tossed 100 times, Head appears 40 times
- P(H)= 0.4
- Why?
- Because of maximum likelihood

N tosses, K Heads, parameter *P(H)=p*

- Construct Maximum Likelihood Expression
- Take log likelihood and take derivative
- Equate to 0 and Get *p*

$$L = p^K (1-p)^{N-K}$$

$$\Rightarrow LL = \log(L) = K \log p + (N-K) \log(1-p)$$

$$\Rightarrow \frac{d(LL)}{dp} = \frac{K}{p} - \frac{N-K}{1-p}$$

$$\Rightarrow \frac{d(LL)}{dp} = 0 \quad gives \quad p = \frac{K}{N}$$

Exercise

- Following the process for finding the probability of Head from N tosses of coin yielding K Heads, prove that the transition probabilities can be found from MLE
- Most important: get the likelihood expression
- Use chapter 2 of the book
  - Pushpak Bhattacharyya: Machine translation, CRC Press, Taylor & Francis Group, Boca Raton, USA, 2015, ISBN: 978-1-4398-9718-8

Next question?

- How to decode efficiently?
- E.g.
  - T: Tags
  - W: Words
  - Two special symbol: '^' and '.'

  Find out number of paths in the tree given word sequence.

Exponential *w.r.t.* number of wor

Number of path = Number of leaves in the tree.

$$O(T^n)$$

How to avoid it?

We do not need exponential work!

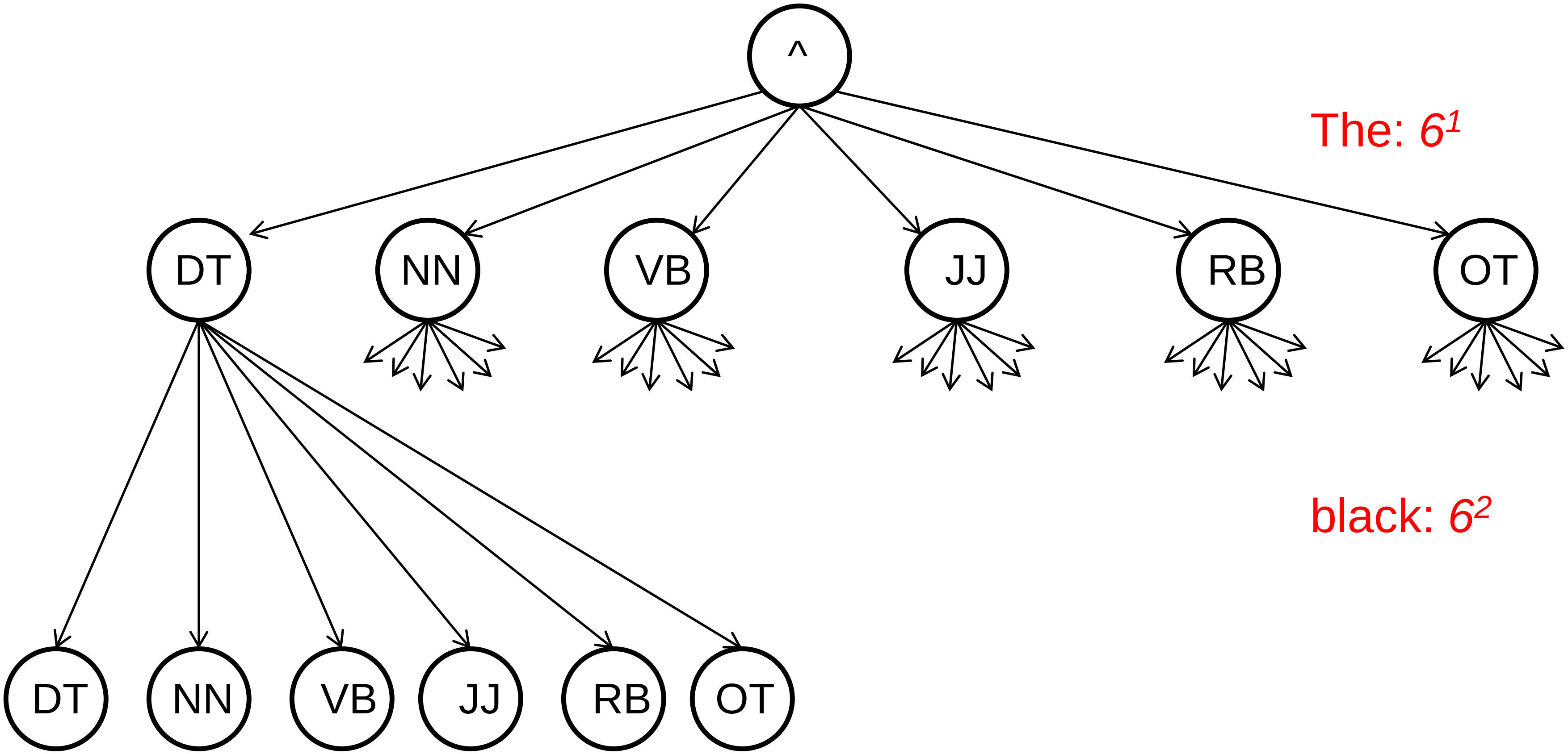
- Suppose our tags are
  - DT, NN, VB, JJ, RB and OT
- E.g.

DT- determiner
NN- Noun
VB- Verb
JJ- Adjective
RB- Adverb
OT- others

^	**The**	**black**	**dog**	**barks**	**.**
^	DT NN VB JJ RB OT	DT NN VB JJ RB OT	DT NN VB JJ RB OT	DT NN VB JJ RB OT	.

Possible tags

So, 6^4 possible path



The: 6^1

black: 6^2

dog: 6^3

barks: 6^4

.: 6^4

Total 6^4 paths

- Now consider the paths that end in NN after seeing input “The black”

^	The	black	
^	DT	NN	***P*(T).*P*(W|T) =** *P*(DT|^) . *P*(NN|DT) . *P*(The|DT) . *P*(Black|NN)
^	NN	NN	***P*(T).*P*(W|T) =** *P*(NN|^) . *P*(NN|NN) . *P*(The|NN) . *P*(Black|NN)
^	VB	NN	***P*(T).*P*(W|T) =** *P*(VB|^) . *P*(NN|VB) . *P*(The|VB) . *P*(Black|NN)
^	JJ	NN	***P*(T).*P*(W|T) =** *P*(JJ|^) . *P*(NN|JJ) . *P*(The|JJ) . *P*(Black|NN)
^	RB	NN	***P*(T).*P*(W|T) =** *P*(RB|^) . *P*(NN|RB) . *P*(The|RB) . *P*(Black|NN)
^	OT	NN	***P*(T).*P*(W|T) =** *P*(OT|^) . *P*(NN|OT) . *P*(The|OT) . *P*(Black|NN)

$$Complexity = W_n * T$$

For each tag, only path with ***highest probability*** value are ***retained***, others are simply ***discarded***.

Machine Translation v/s POS tagging!

- Similarity
  - POS
    - Every word in a sentence has one corresponding tag.
  - MT
    - Every word in a sentence has one (or more) corresponding translated word.
- Difference
  - Order: Order of translated word may change.
  - Fertility: One word corresponds to many. Many to one also possible.

Complexity

- POS and HMM
  - Linear time complexity
- MT and Bean search
  - Exponential time complexity
  - Permutation of words produces exponential searc space
  - However, for related languages, MT is like POS tagging

Properties of related languages

1. Order preserving

2. Fertility ~ 1

3. Morphology preserving

Hindi	*Jaaunga*
Bengali	*Jaabo*
English	*Will go*

Hindi & Bengali ↑

Hindi & English ↓

Properties of related languages

4. Syncretism: Suffix features should be similarly loaded

Hindi	Main *jaaunga*	Hum *jaayenge*
Bengali	Ami *jaabo*	Aamra *jaabo*

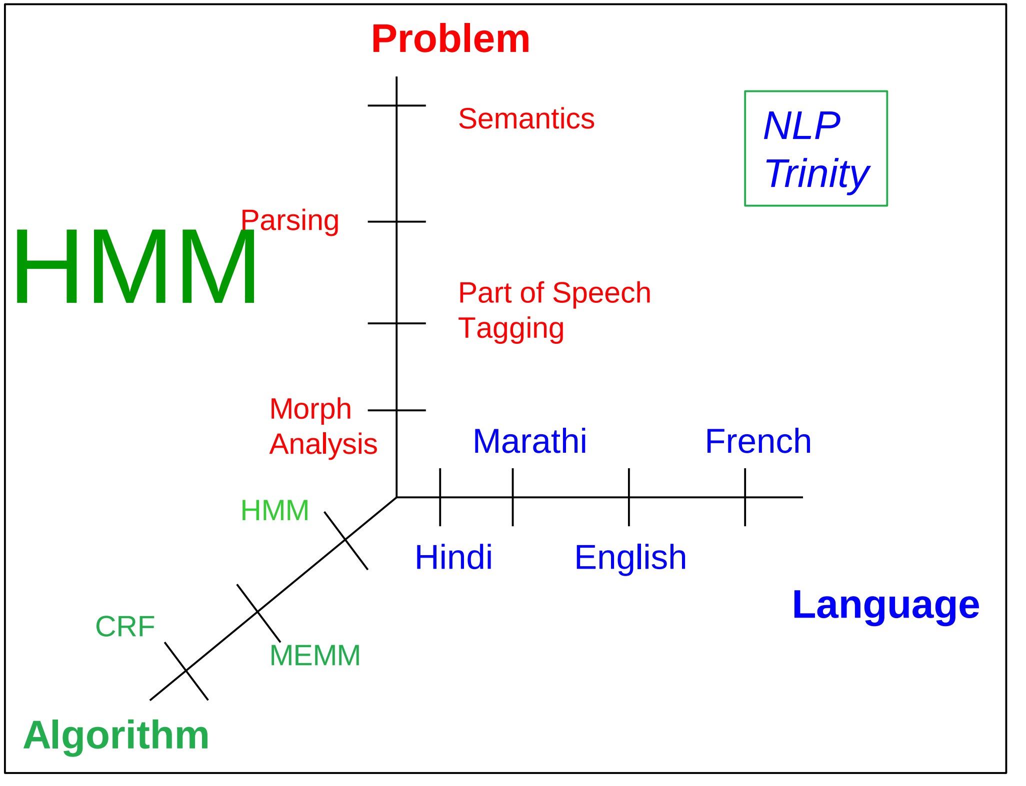
Hindi & Bengali ↓

5. Idiomaticity: Literal translation should be high

Hindi	*Aap Kaise Ho?*
Bengali	*Aapni Kemon Achen?*
English	*How do you do?*

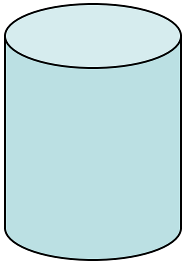
Hindi & Bengali ↑

Hindi & English ↓


Problem
Semantics
NLP
Trinity
Parsing
HMM
Part of Speech
Tagging
Morph
Analysis
Marathi
French
HMM
Hindi
English
Language
CRF
MEMM
Algorithm

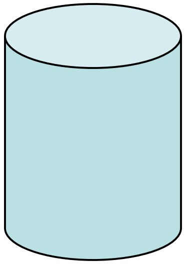
A Motivating Example

Colored Ball choosing



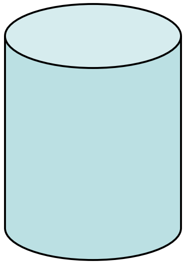
Urn 1

# of Red = 30
# of Green = 50
# of Blue = 20



Urn 2

# of Red = 10
# of Green = 40
# of Blue = 50



Urn 3

# of Red =60
# of Green =10
# of Blue = 30

Example (contd.)

Given :

	U_1	U_2	U_3
U_1	0.1	0.4	0.5
U_2	0.6	0.2	0.2
U_3	0.3	0.4	0.3

Transition probability table

and

	R	G	B
U_1	0.3	0.5	0.2
U_2	0.1	0.4	0.5
U_3	0.6	0.1	0.3

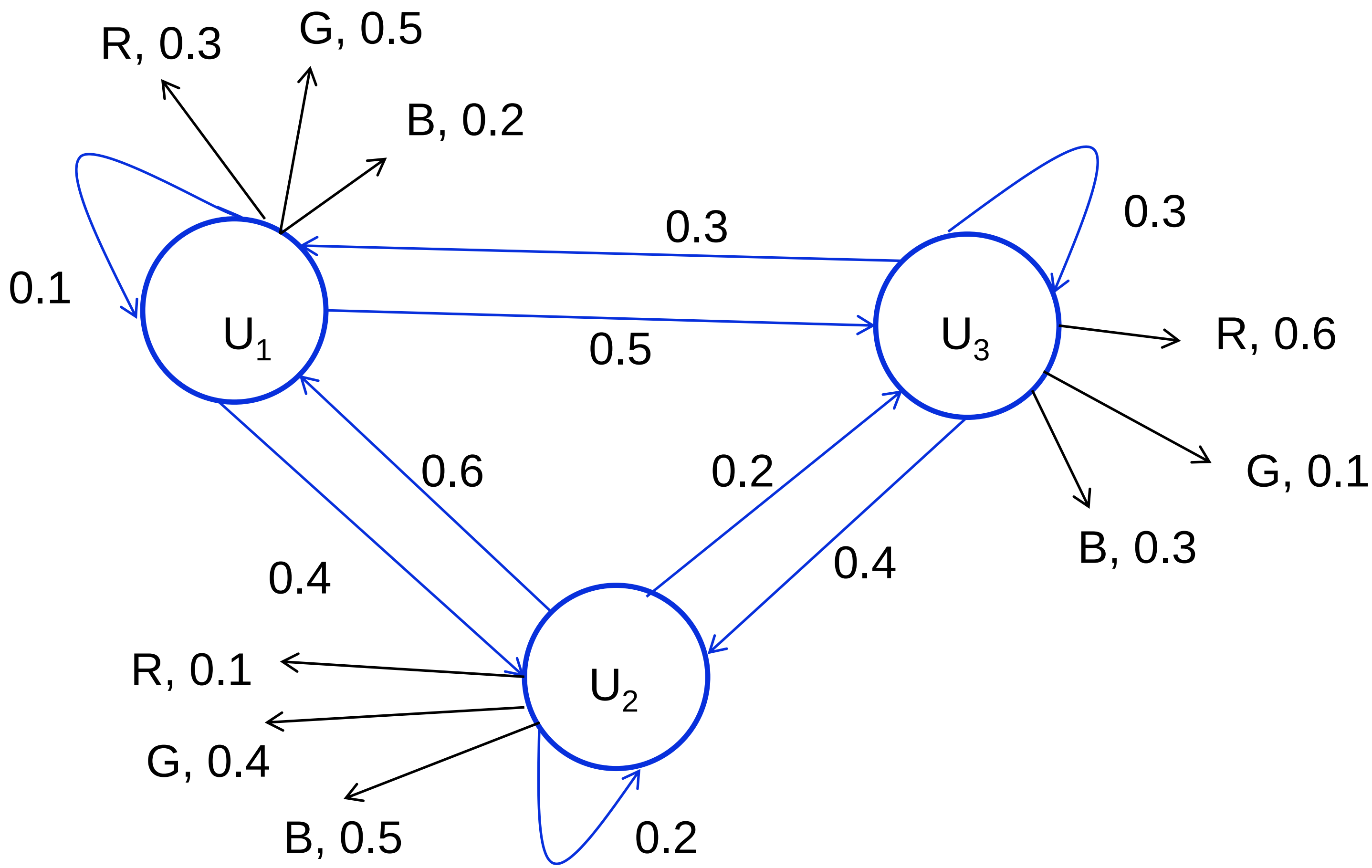
Emission probability table

Observation : RRGGBRGR

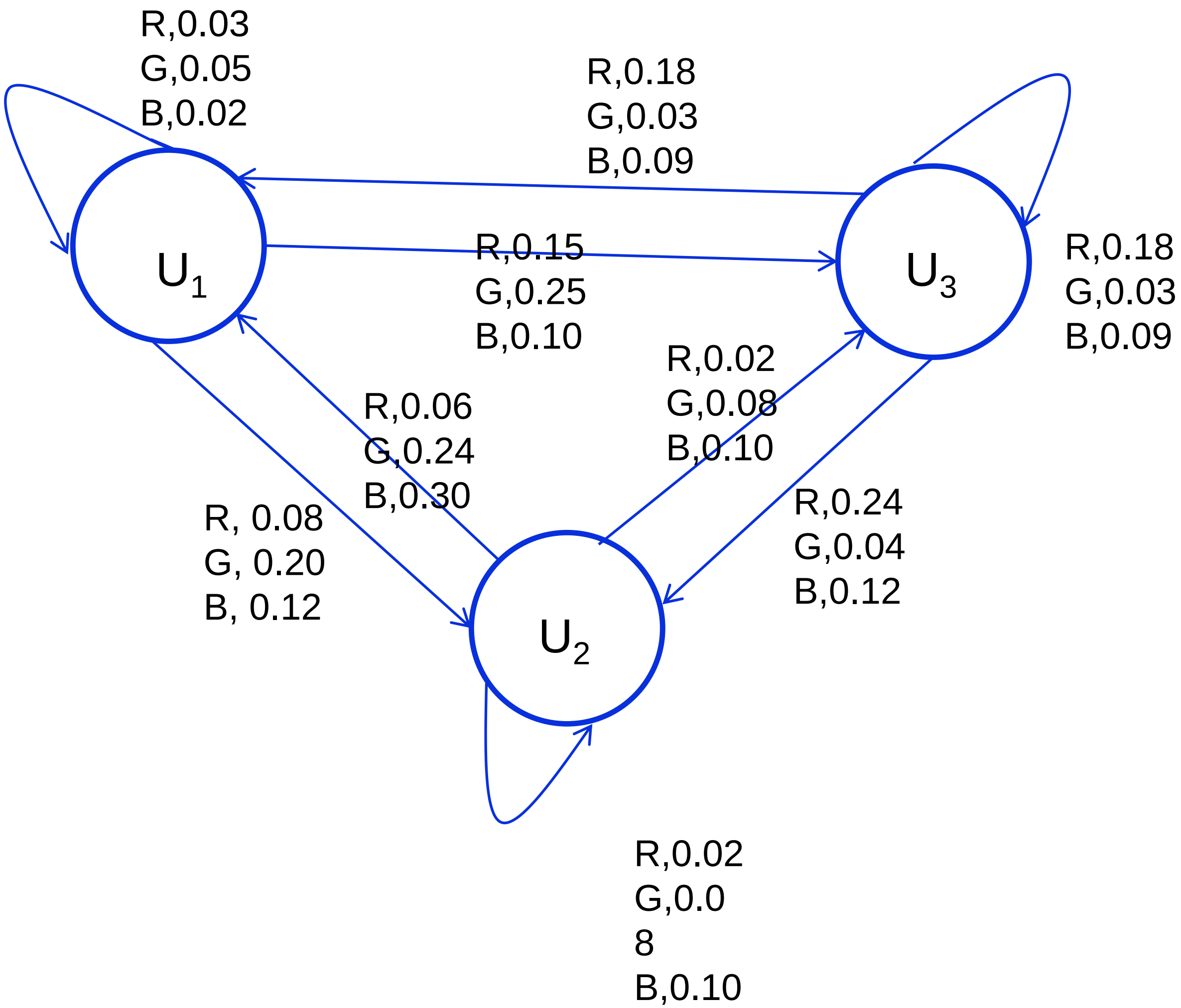
State Sequence : ??

Not so Easily Computable.

Diagrammatic representation (1/2)



Diagrammatic representation (2/2)



Classic problems with respect to HMM

1. Given the observation sequence, find the possible state sequences- Viterbi
2. Given the observation sequence, find its probability- forward/backward algorithm
3. Given the observation sequence find the HMM prameters.- Baum-Welch algorithm

Illustration of Viterbi

- The “start” and “end” are important in a sequence.
- Subtrees get eliminated due to the Markov Assumption.

<u>POS Tagset</u>

- N(noun), V(verb), O(other)   [simplified]
- ^ (start), . (end)   [start & end states]

Illustration of Viterbi

Lexicon

people: N, V

laugh: N, V

.

.

.

Corpora for Training

_______^ $w_{11\_}t_{11}$ $w_{12\_}t_{12}$ $w_{13\_}t_{13}$ ……………….$w_{1k\_1\_}t_{1k\_1}$.

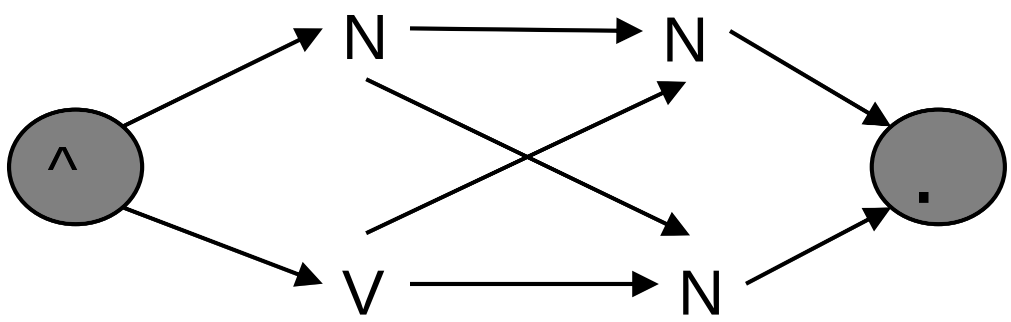
^ $w_{21\_}t_{21}$ $w_{22\_}t_{22}$ $w_{23\_}t_{23}$ ……………….$w_{2k\_2\_}t_{2k\_2}$.

.

.

^ $w_{n1\_}t_{n1}$ $w_{n2\_}t_{n2}$ $w_{n3\_}t_{n3}$ ……………….$w_{nk\_n\_}t_{nk\_n}$.

Inference



Partial sequence graph

	^	N	V	O	.
^	0	0.6	0.2	0.2	0
N	0	0.1	0.4	0.3	0.2
V	0	0.3	0.1	0.3	0.3
O	0	0.3	0.2	0.3	0.2
.	1	0	0	0	0

This transition table will change from language to language due to language divergences.

Lexical Probability Table

	Є	people	laugh	…	…
^	1	0	0	…	0
N	0	$1x10^{-3}$	$1x10^{-5}$	…	…
V	0	$1x10^{-6}$	$1x10^{-3}$	…	…
O	0	0	0	…	…
.	1	0	0	0	0

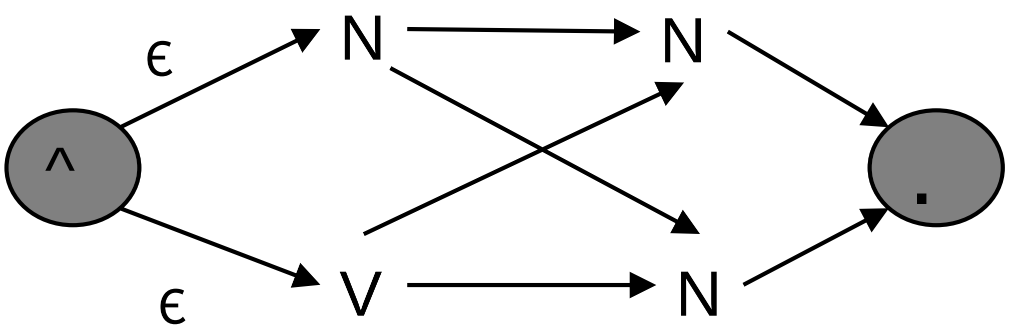
Size of this table = # pos tags in tagset X vocabulary size

vocabulary size = # unique words in corpus

Inference

New Sentence:

^ people laugh .



p(^ N N . | ^ people laugh .)

$= (0.6 \times 0.1) \times (0.1 \times 1 \times 10^{-3}) \times (0.2 \times 1 \times 10^{-5})$

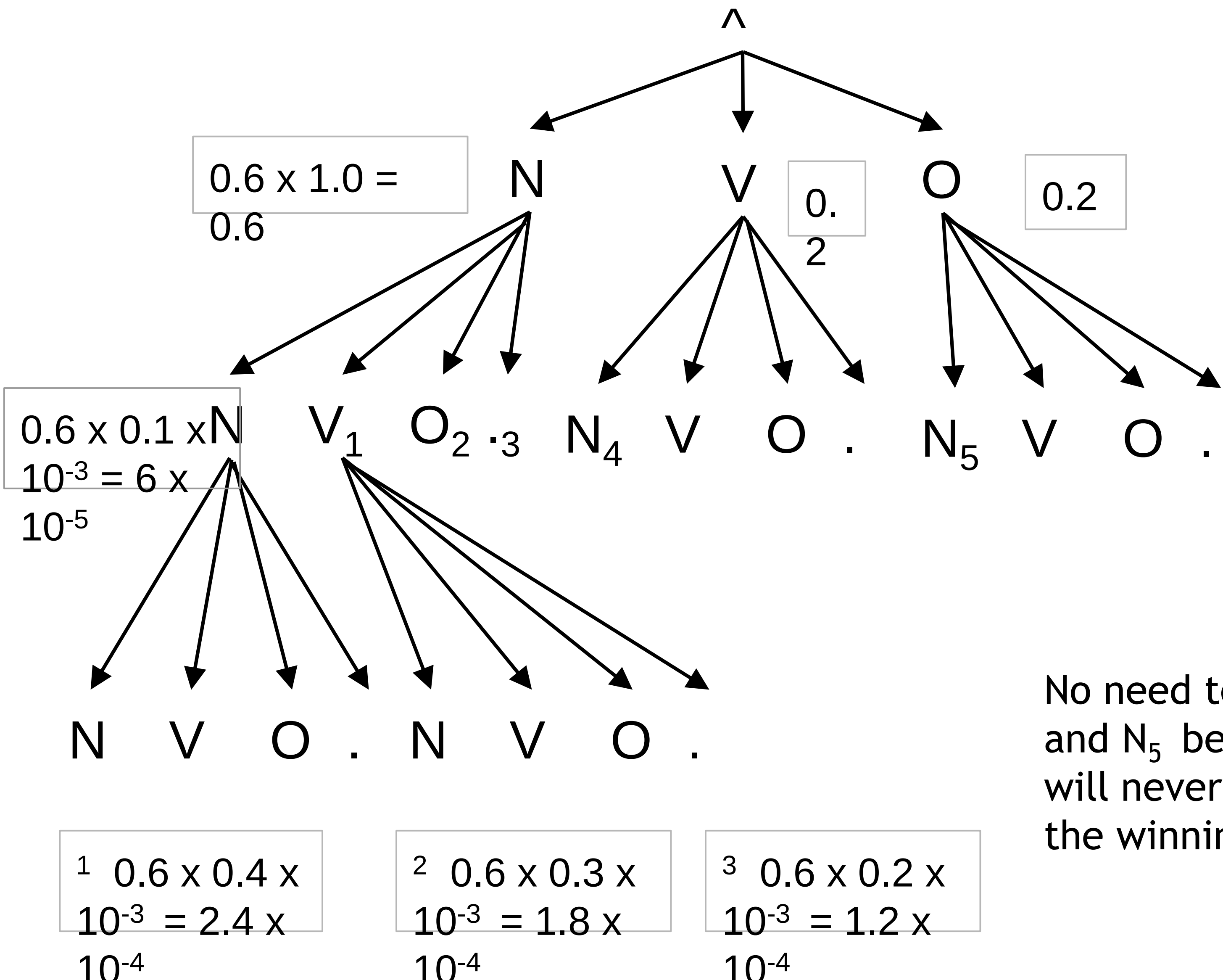
Computational Complexity

- If we have to get the probability of each sequence and then find maximum among them, we would run into exponential number of computations.
- If $|s|$ = #states (tags + ^ + .)

  and $|o|$ = length of sentence (words + ^ + .)

  Then, #sequences = $s^{|o|-2}$
- But, a large number of partial computations can be reused using Dynamic Programming.

Dynamic Programming



No need to expand N_4 and N_5 because they will never be a part of the winning sequence.

Computational Complexity

- Retain only those N / V / O nodes which ends in the highest sequence probability.
- Now, complexity reduces from $|s|^{|o|}$ to $|s|.|o|$
- Here, we followed the Markov assumption of order 1.

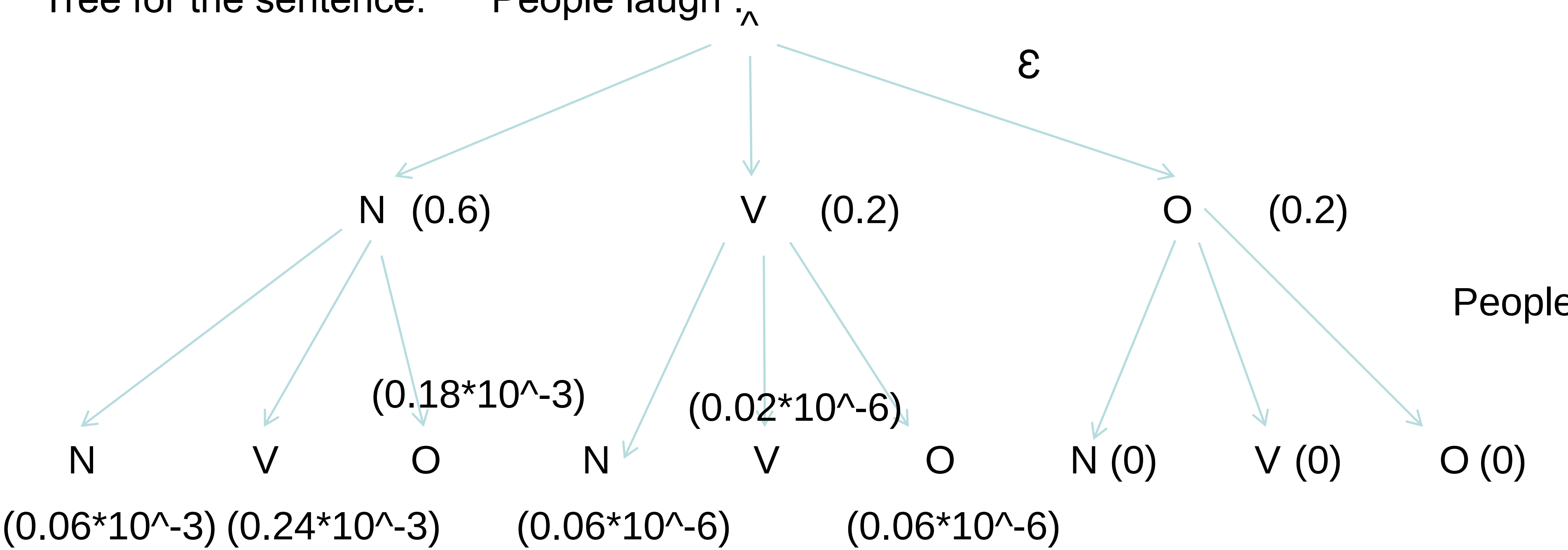
Points to ponder wrt HMM and Viterbi

Viterbi Algorithm

- Start with the start state.
- Keep advancing sequences that are “maximum” amongst all those ending in the same state

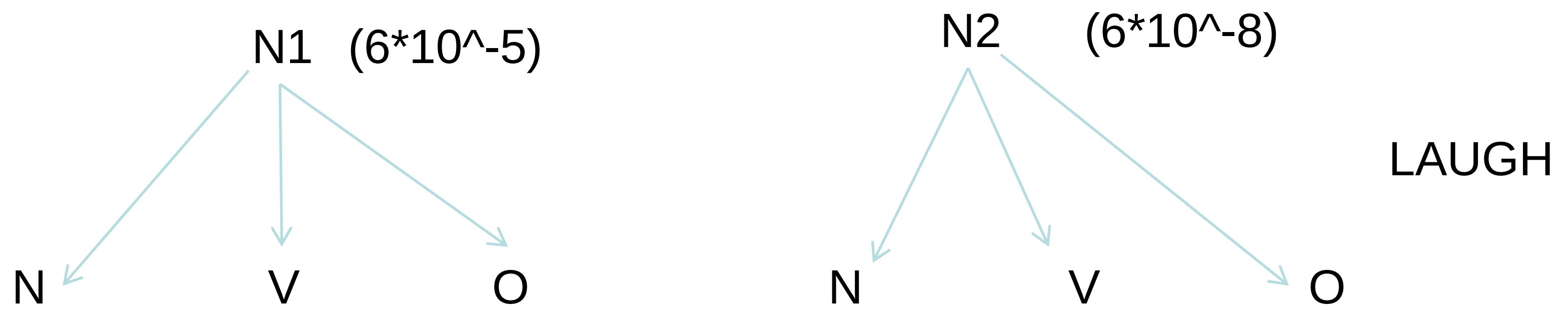
Viterbi Algorithm

Tree for the sentence: “^ People laugh .”



Claim: We do not need to draw all the subtrees in the algorithm

Viterbi phenomenon (Markov process)



Next step all the probabilities will be multiplied by identical probability (lexical and transition). So children of N2 will have probability less than the children of N1.

What does P(A|B) mean?

- P(A|B)= P(B|A)

  If P(A)=P(B)

- P(A|B) means??
  - Causality?? B causes A??
  - Sequentiality?? A follows B?

Back to the Urn Example

- Here :
  - $S = \{U1, U2, U3\}$
  - $V = \{ R,G,B\}$
- For observation:
  - $O = \{o_1 \ldots o_n\}$
- And State sequence
  - $Q = \{q_1 \ldots q_n\}$
- π is $\pi_i = P(q_1 = U_i)$

A =

	U_1	U_2	U_3
U_1	0.1	0.4	0.5
U_2	0.6	0.2	0.2
U_3	0.3	0.4	0.3

B=

	R	G	B
U_1	0.3	0.5	0.2
U_2	0.1	0.4	0.5
U_3	0.6	0.1	0.3

Observations and states

	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8
OBS:	R	R	G	G	B	R	G	R
State:	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8

$S_i = U_1/U_2/U_3$; A particular state

S: State sequence

O: Observation sequence

S* = “best” possible state (urn) sequence

Goal: Maximize P(S*|O) by choosing “best” S

Goal

- Maximize P(S|O) where S is the State Sequence and O is the Observation Sequence

$$S^* = \arg\max_S (P(S \mid O))$$

False Start

	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8
OBS:	R	R	G	G	B	R	G	R
State:	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8

$$P(S \mid O) = P(S_{1-8} \mid O_{1-8})$$

$$P(S \mid O) = P(S_1 \mid O).P(S_2 \mid S_1, O).P(S_3 \mid S_{1-2}, O)\ldots P(S_8 \mid S_{1-7}, O)$$

By Markov Assumption (a state depends only on the previous state)

$$P(S \mid O) = P(S_1 \mid O).P(S_2 \mid S_1, O).P(S_3 \mid S_2, O)\ldots P(S_8 \mid S_7, O)$$

Baye's Theorem

$$P(A \mid B) = P(A).P(B \mid A) / P(B)$$

P(A) -: Prior

P(B|A) -: Likelihood

$$\arg\max_S P(S \mid O) = \arg\max_S P(S).P(O \mid S)$$

State Transitions Probability

$$P(S) = P(S_{1-8})$$

$$P(S) = P(S_1).P(S_2 \mid S_1).P(S_3 \mid S_{1-2}).P(S_4 \mid S_{1-3})\ldots P(S_8 \mid S_{1-7})$$

By Markov Assumption (k=1)

$$P(S) = P(S_1).P(S_2 \mid S_1).P(S_3 \mid S_2).P(S_4 \mid S_3)\ldots P(S_8 \mid S_7)$$

Observation Sequence probability

$$P(O|S) = P(O_1|S_{1-8}).P(O_2|O_1, S_{1-8}).P(O_3|O_{1-2}, S_{1-8})...P(O_8|O_{1-7}, S_{1-8})$$

Assumption that ball drawn depends only on the Urn chosen

$$P(O|S) = P(O_1|S_1).P(O_2|S_2).P(O_3|S_3)...P(O_8|S_8)$$

$$P(S|O) = P(S).P(O|S)$$

$$P(S|O) = P(S_1).P(S_2|S_1).P(S_3|S_2).P(S_4|S_3)...P(S_8|S_7).$$

$$P(O_1|S_1).P(O_2|S_2).P(O_3|S_3)...P(O_8|S_8)$$

Grouping terms

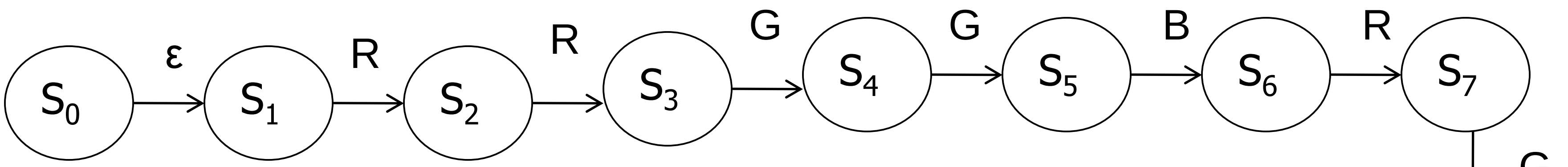
	O_0	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	
Obs:	ε	R	R	G	G	B	R	G	R	
State:	S_0	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9

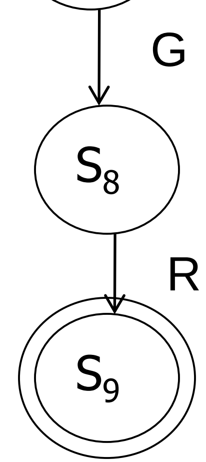
$$
\begin{aligned}
& P(S).P(O|S) \\
= \; & [P(O_0|S_0).P(S_1|S_0)]. \\
& [P(O_1|S_1).\quad P(S_2|S_1)]. \\
& [P(O_2|S_2).\quad P(S_3|S_2)]. \\
& [P(O_3|S_3).P(S_4|S_3)]. \\
& [P(O_4|S_4).P(S_5|S_4)]. \\
& [P(O_5|S_5).P(S_6|S_5)]. \\
& [P(O_6|S_6).P(S_7|S_6)]. \\
& [P(O_7|S_7).P(S_8|S_7)]. \\
& [P(O_8|S_8).P(S_9|S_8)].
\end{aligned}
$$

- We introduce the states S_0 and S_9 as initial and final states respectively.
- After S_8 the next state is S_9 with probability 1, i.e., $P(S_9|S_8)=1$
- O_0 is ε-transition

Introducing useful notation

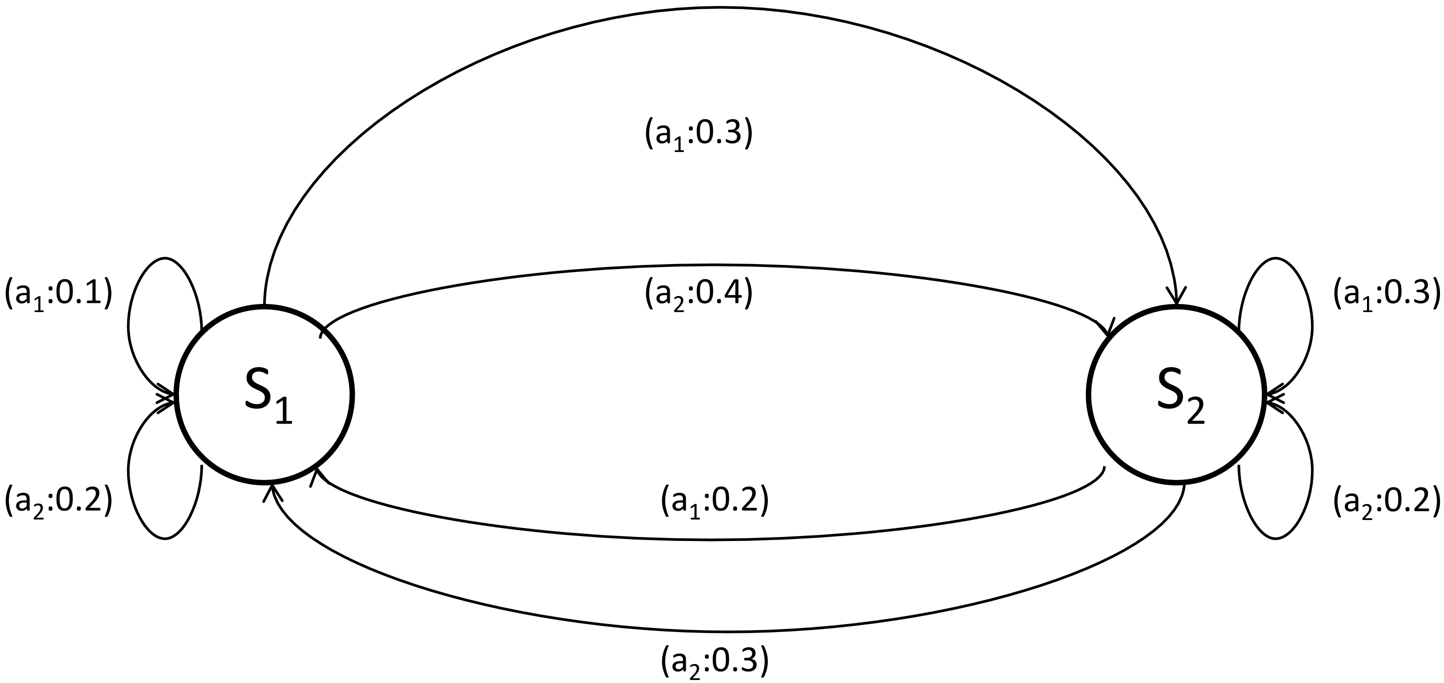
	O_0	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	
Obs:	ε	R	R	G	G	B	R	G	R	
State:	S_0	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9





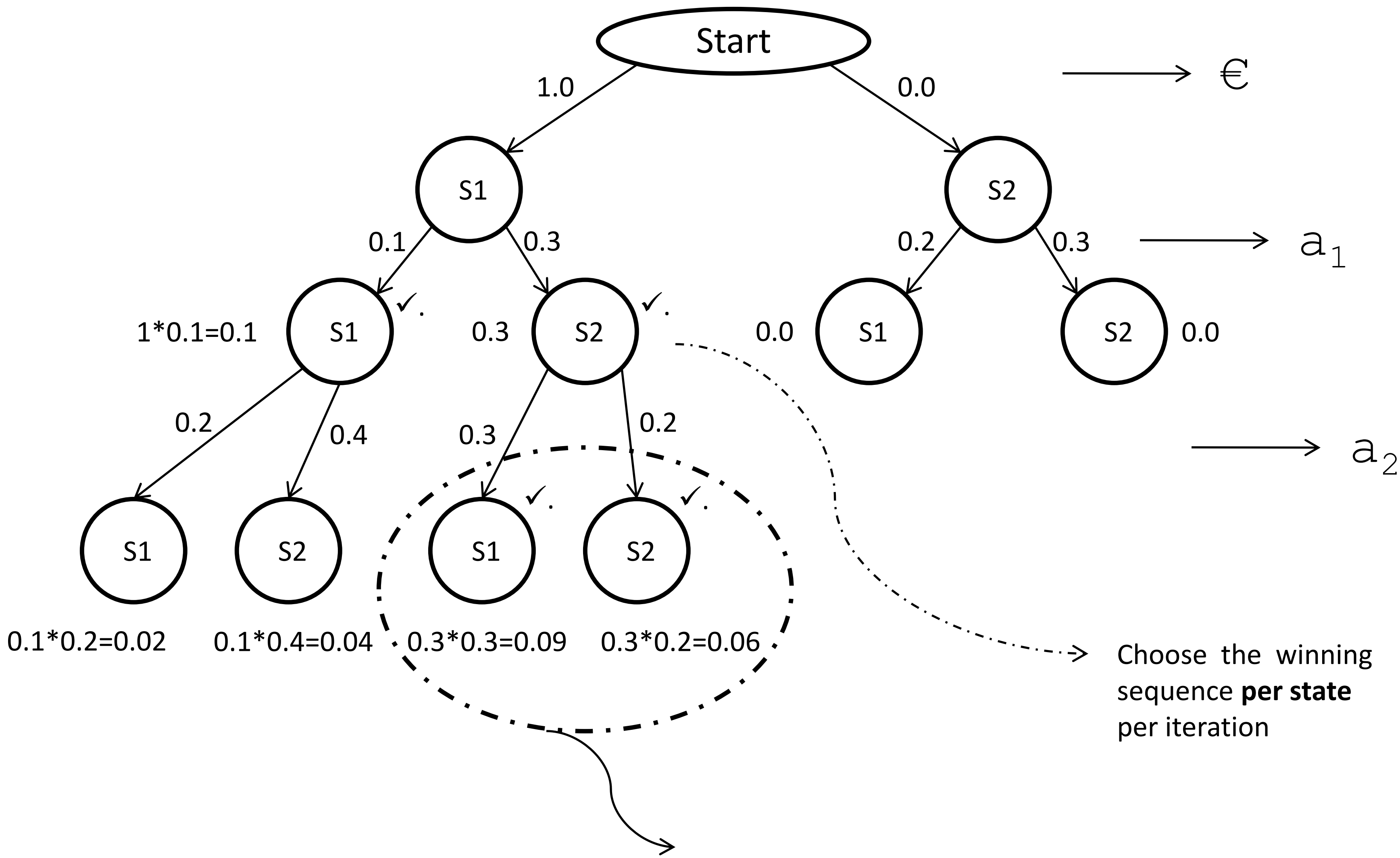
$$P(O_k|S_k).P(S_{k+1}|S_k)=P(S_k \xrightarrow{O_k} S_{k+1})$$

Probabilistic FSM

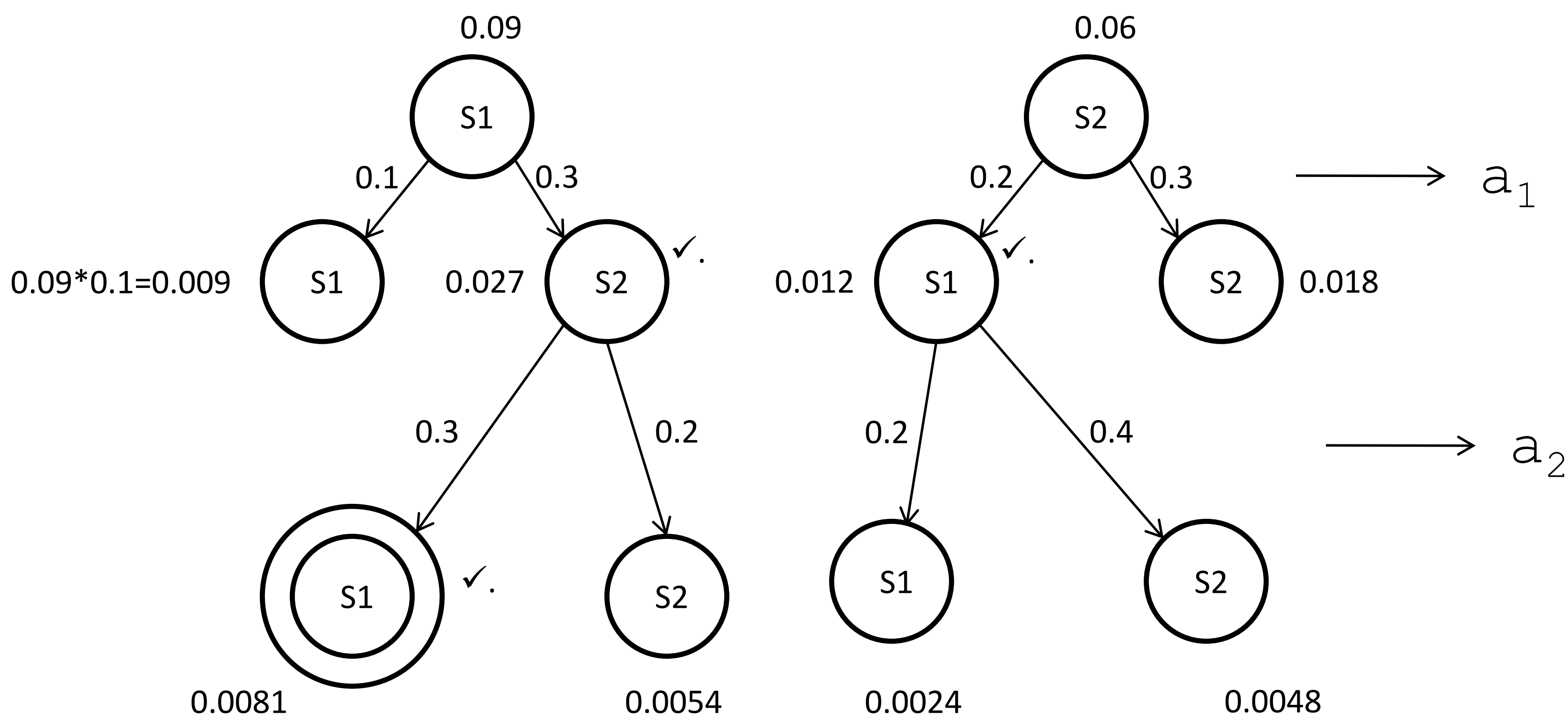


The question here is:
"what is the most likely state sequence given the output sequence seen"

Developing the tree



Tree structure contd…



The problem being addressed by this tree is $S^* = \arg\max_{s} P(S \mid a_1 - a_2 - a_1 - a_2, \mu)$

a1-a2-a1-a2 is the output sequence and μ the model or the machine

Path found: (working backward)

$$S_1 \underbrace{\quad}_{a_1} S_2 \underbrace{\quad}_{a_2} S_1 \underbrace{\quad}_{a_1} S_2 \underbrace{\quad}_{a_2} S_1$$

Problem statement: Find the best possible sequence

$$S^* = \arg\max_{S} P(S \mid O, \mu)$$

$$where,\ S \rightarrow \text{State Seq},\ O \rightarrow \text{Output Seq},\ \mu \rightarrow \text{Model or Machine}$$

$$\text{Model or Machine} = \{S_0, S, A, T\}$$

Start symbol State collection Alphabet set Transitions

T is defined as $P(S_i \xrightarrow{a_k} S_j) \quad \forall i, j, k$

Tabular representation of the tree

Latest symbol observed / Ending state	€	a_1	a_2	a_1	a_2
S_1	1.0	(1.0*0.1,0.0*0.2) =(**0.1**,0.0)	(0.02, **0.09**)	(0.009, **0.012**)	(0.0024, **0.0081**)
S_2	0.0	(1.0*0.3,0.0*0.3) =(**0.3**,0.0)	(0.04,**0.06**)	(**0.027**,0.018)	(0.0048,0.0054)

Note: Every cell records the winning probability ending in that state

Final winner

The bold faced values in each cell shows the sequence probability ending in that state. Going backward from final winner sequence which ends in state S_2 (indicated By the 2nd tuple), we recover the sequence.

Algorithm

(following James Alan, Natural Language Understanding (2nd edition), Benjamin Cummins (pub.), 1995

Given:

1. The HMM, which means:
   a. Start State: S_1
   b. Alphabet: $A = \{a_1, a_2, \ldots a_p\}$
   c. Set of States: $S = \{S_1, S_2, \ldots S_n\}$
   d. Transition probability $P(S_i \xrightarrow{a_k} S_j) \quad \forall i, j, k$
      which is equal to $P(S_j, a_k \mid S_i)$
2. The output string $a_1 a_2 \ldots a_T$

To find:

The most likely sequence of states $C_1 C_2 \ldots C_T$ which produces the given output sequence, *i.e.*, $C_1 C_2 \ldots C_T = \arg\max_C [P(C \mid a_1, a_2, \ldots a_T, \mu)]$

1. Initialization

SEQSCORE(1,1)=1.0

BACKPTR(1,1)=0

For(i=2 to N) do

    SEQSCORE(i,1)=0.0

[expressing the fact that first state is S_1]

2. Iteration

For(t=2 to T) do

  For(i=1 to N) do

    SEQSCORE(i,t) = $\text{Max}_{(j=1,N)}$

$$[SEQSCORE(j,(t-1)) * P(Sj \xrightarrow{a_k} Si)]$$

BACKPTR(I,t) = index *j* that gives the MAX above

3. Seq. Identification

C(T) = i that maximizes SEQSCORE(i,T)

For i from (T-1) to 1 do

C(i) = BACKPTR[C(i+1),(i+1)]

Optimizations possible:

1. BACKPTR can be 1*T
2. SEQSCORE can be T*2

Homework:- Compare this with A*, Beam Search [Homework]

Reason for this comparison:

Both of them work for finding and recovering sequence

Reading List

- https://www.nltk.org/book/ch05.html
- TnT (http://www.aclweb.org/anthology-new/A/A00/A00-1031.pdf)
- Hindi POS Tagger built by IIT Bombay (http://www.cse.iitb.ac.in/pb/papers/ACL-2006-Hindi-POS-Tagging.pdf)

Assignment Discussion

Build a POS Tagger (due date: 5th September, 2020)

- Using
  - HMM
  - SVM
  - Bi-LSTM

- Training corpora (Brown Copus)
  - http://www.nltk.org/nltk_data/

Project Discussion

Different Areas of NLP

Questions from Mohith

1. You mentioned about "ellipsis" during the lecture. In NLP, does "ellipsis" refer to the computational problem that arises due to
skipping a few words in a sentence, or the whole act of skipping words is itself termed as ellipsis?

   - Ans: whole act; “where do you live?”- “Delhi” and “your friend?$_{\text{ellipsis}}$”- “Mumbai”

2. You mentioned about shallow/deep parsing. Was this in the context of dependency/constituency parsing? That is, can I say that dependency parser does shallow parsing and constituency parser does deep parsing, or some other similar relations exist between them?

- - both constituency and dependency are “deep” parsing tasks; pos tagging, chunking [(the_DT blue_JJ sky_NN)$_{\text{chunk}}$ was (vast_JJ and_CONJ deep_JJ)$_{\text{chunk}}$]

Questions from Mohith

3. In the Learning POS Tags slide, it is mentioned, "We need at least as many instances as number of different labels #POS tags-1 to make decision". That means, corresponding to every tag(except one) we are giving one example to the learning algorithm. If the algorithm encounters a previously unseen example, give it the last tag. I just wanted to know if my understanding is correct here.

-Ans: Correct; give the remaining tag if none of the tags from the training data is applicable

Mohith

Besides my doubts, I have also solved the first homework question that you had provided in the class. Could you please let me know if my answers are correct?

- Question 1: Example of Verbal, Answer: "Could you please google this topic?". Here "google" is a noun, but in this context it is being used as a verb. Therefore "google" would be tagged as a "Verbal“

-Ans: No, dictionary definition of “google” includes *verb* also

I haven't got an example for the second question of finding false positive/negative for rules of "present"

-Ans: The rule is “***If present is preceded by determiner (the/a) or demonstrative (this/that) or followed by a verb, then POS tag will be noun***.” Still fails for “The present situation is comfortable”