CS626 : Natural Language Processing, Speech and the Web
(Lecture 4,5 – HMM, POS tagging)

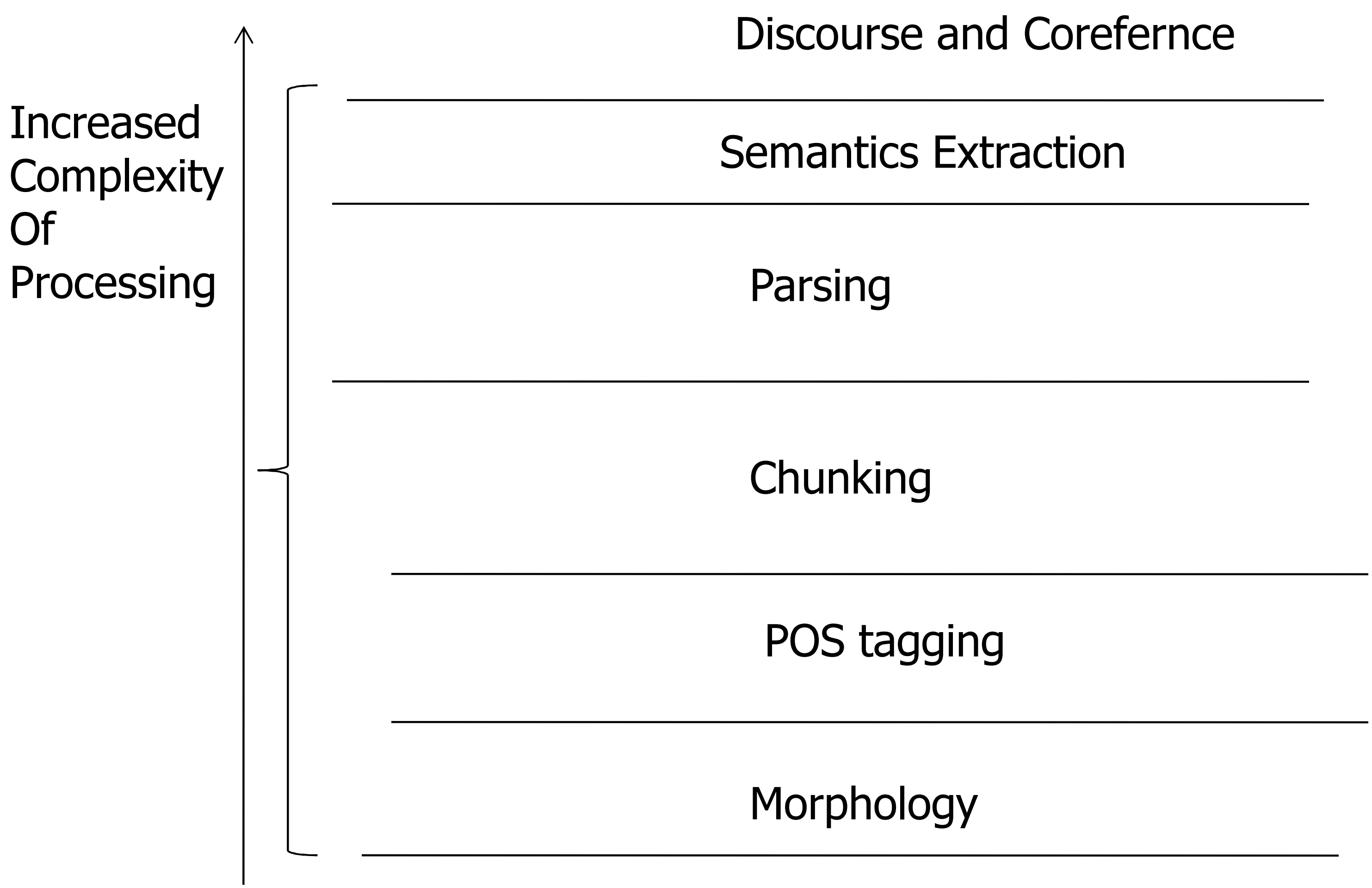
Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

30th July and 2nd August, 2012

POS tagging: Definition

- Tagging is the assignment of a singlepart-of-speech tag to each word (and punctuation marker) in a corpus.
  - "_" The_DT guys_NNS that_WDT make_VBP traditional_JJ hardware_NN are_VBP really_RB being_VBG obsoleted_VBN by_IN microprocessor-based_JJ machines_NNS ,_, "_" said_VBD Mr._NNP Benton_NNP ._.

Where does POS tagging fit in



Behaviour of “That”

- That
  - *That man is known by the company he keeps.* (Demonstrative)
  - *Man that is known by the company he keeps, gets a good job*. (Pronoun)
  - *That man is known by the company he keeps, is a proverb*. (Complementation)
- Chaotic systems: Systems where a small perturbation in input causes a large change in output

Argmax computation (1/2)

Best tag sequence

= T*

= argmax P(T|W)

= argmax P(T)P(W|T) (by Baye's Theorem)

$$P(T) = P(t_0{=}\hat{}\ t_1 t_2 \ldots t_{n+1}{=}.)$$

$$= P(t_0)P(t_1|t_0)P(t_2|t_1t_0)P(t_3|t_2t_1t_0) \ldots P(t_n|t_{n-1}t_{n-2}\ldots t_0)P(t_{n+1}|t_nt_{n-1}\ldots t_0)$$

$$= P(t_0)P(t_1|t_0)P(t_2|t_1) \ldots P(t_n|t_{n-1})P(t_{n+1}|t_n)$$

$$= \prod_{i=0}^{N+1} P(t_i|t_{i-1})$$ Bigram Assumption

Argmax computation (2/2)

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

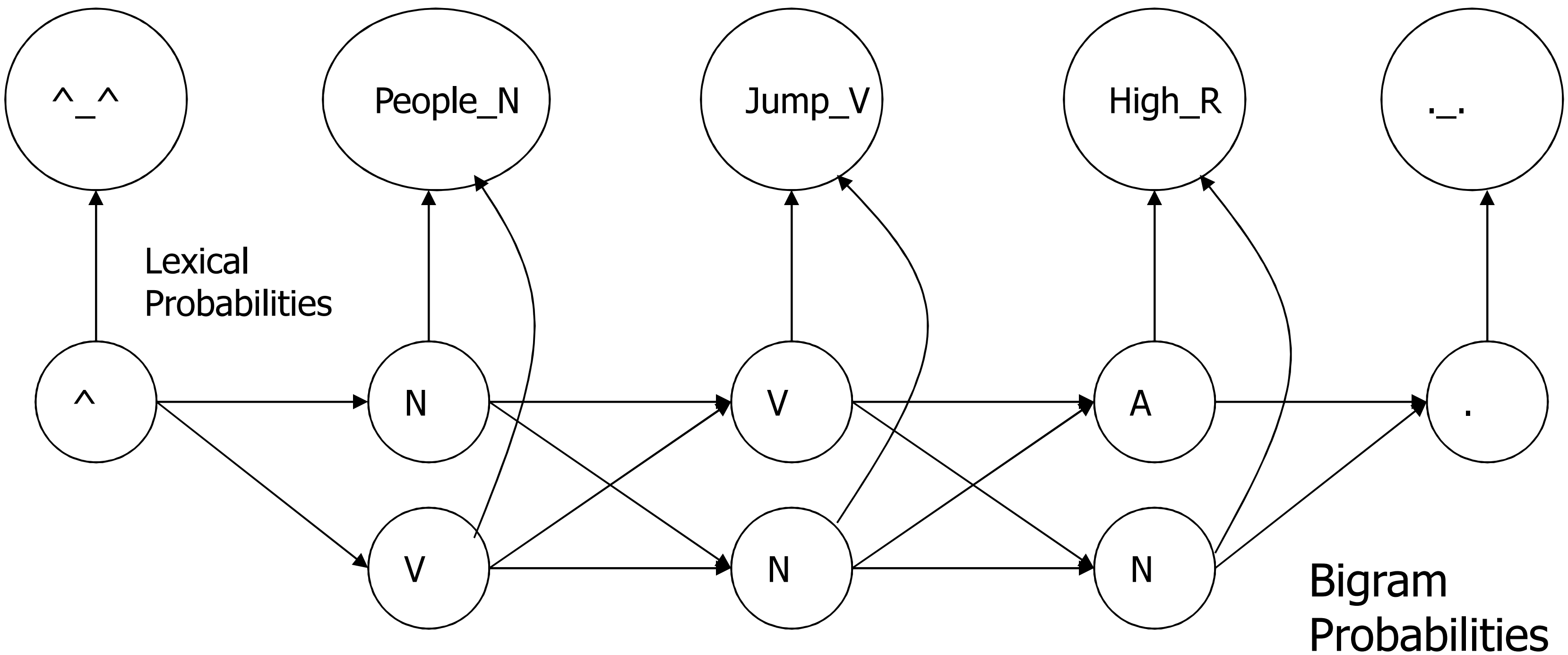
Assumption: A word is determined completely by its tag. This is inspired by speech recognition

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

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

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

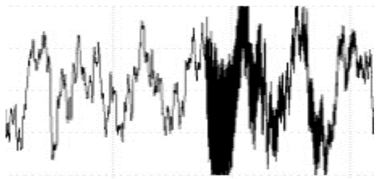
Generative Model



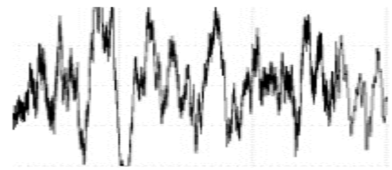
This model is called Generative model.
Here words are observed from tags as states.
This is similar to HMM.

Inspiration from Automatic Speech Recognition

- Isolated Word Recognition (IWR)



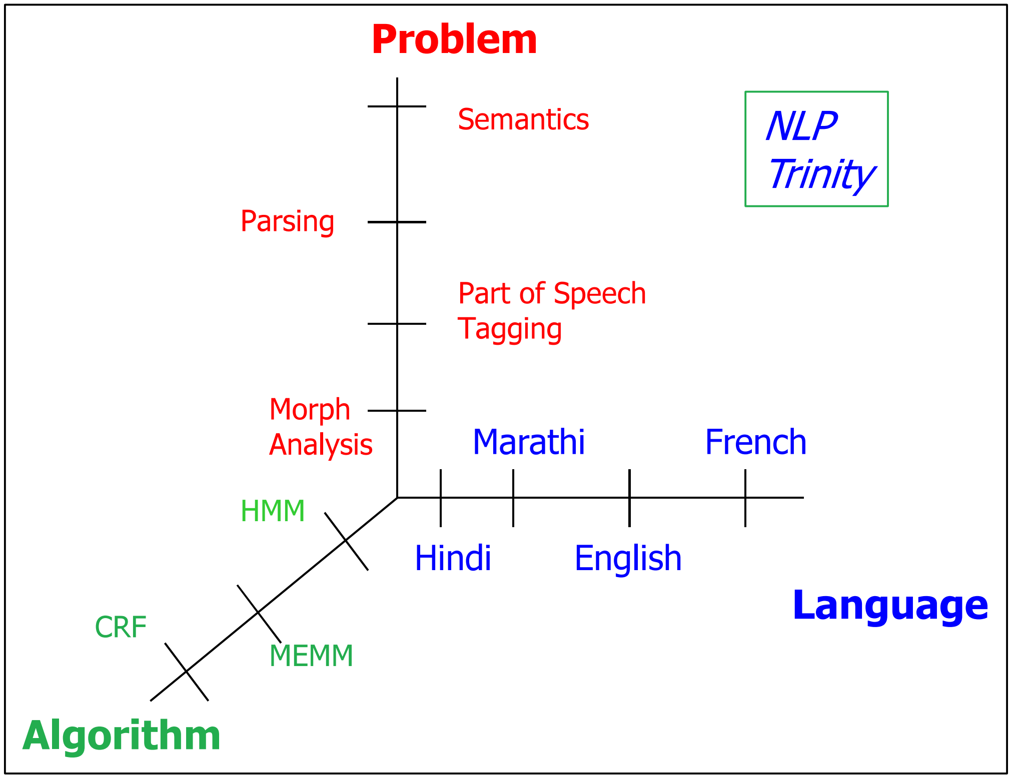
apple



dog

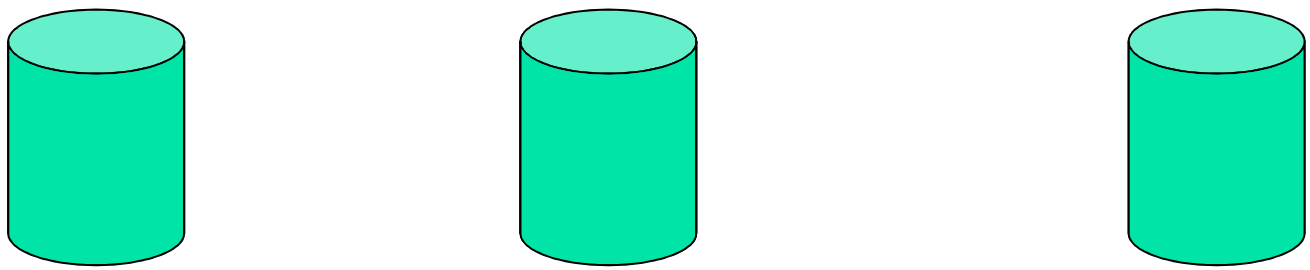
- $w^* = \text{argmax}_w (P(w|s))$
  - w=word, s=speech signal
- $P(w|s) = P(w) \,.\, P(s|w)$
  - P(w) – word model (how probable is a word) – learnt from any corpus
  - P(s|w) – translation model (how a word is spoken) – learnt from annotated speech corpus
- Brittle, britle, brite
  - P(w) will be extremely low (~0) for the words britle and brite

HMM


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

Probability of transition to another Urn after picking a ball:

	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

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

and

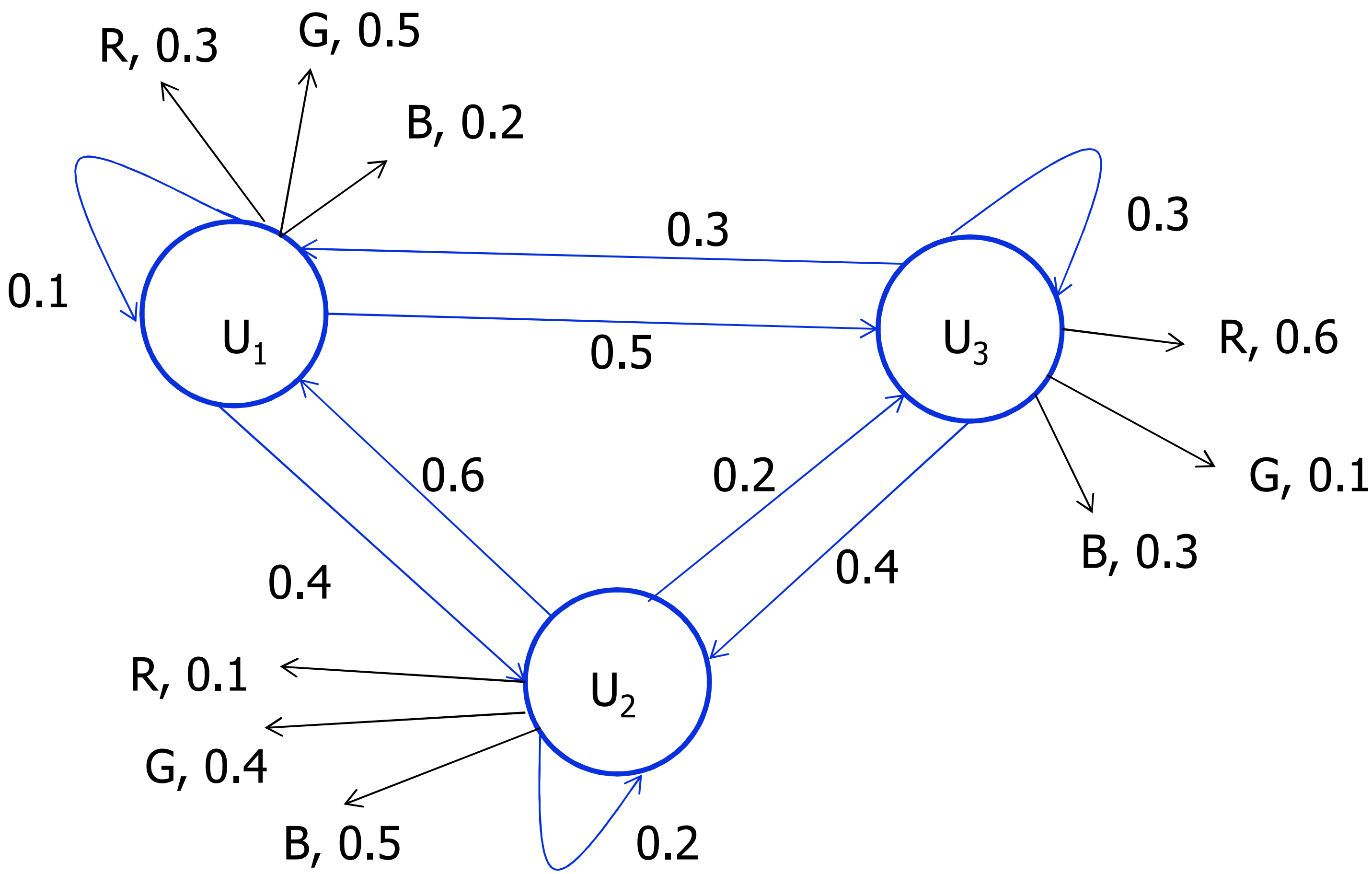
	R	G	B
U_1	0.3	0.5	0.2
U_2	0.1	0.4	0.5
U_3	0.6	0.1	0.3

Observation : RRGGBRGR

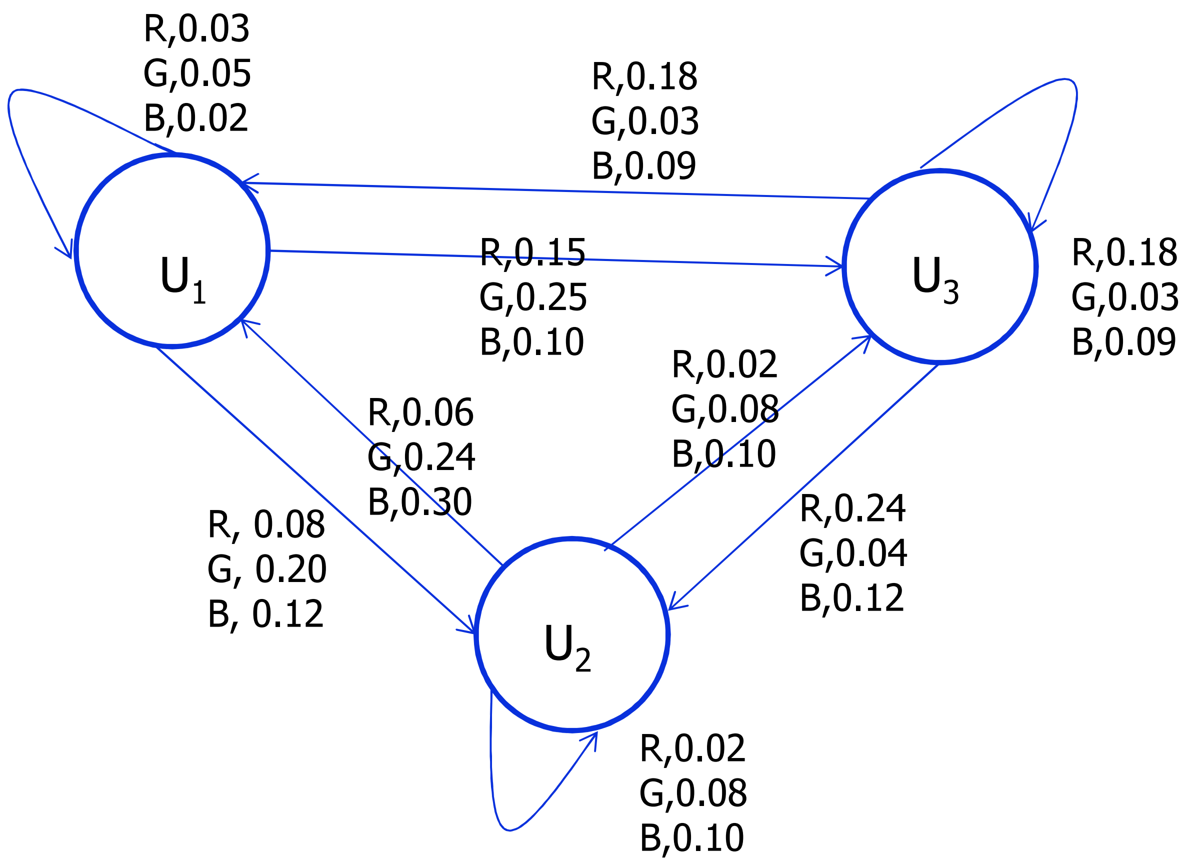
State Sequence : ??

Not so Easily Computable.

Diagrammatic representation (1/2)



Diagrammatic representation (2/2)



Example (contd.)

- Here :
  - S = {U1, U2, U3}
  - V = { R,G,B}
- For observation:
  - O ={o_1… o_n}
- And State sequence
  - Q ={q_1… 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

$$P(A|B) = P(A).P(B|A) / P(B)$$

P(A) -: Prior
P(B|A) -: Likelihood

$$\operatorname{argmax}_S P(S|O) = \operatorname{argmax}_S P(S).P(O|S)$$

State Transitions Probability

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

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

By Markov Assumption (k=1)

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

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

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

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

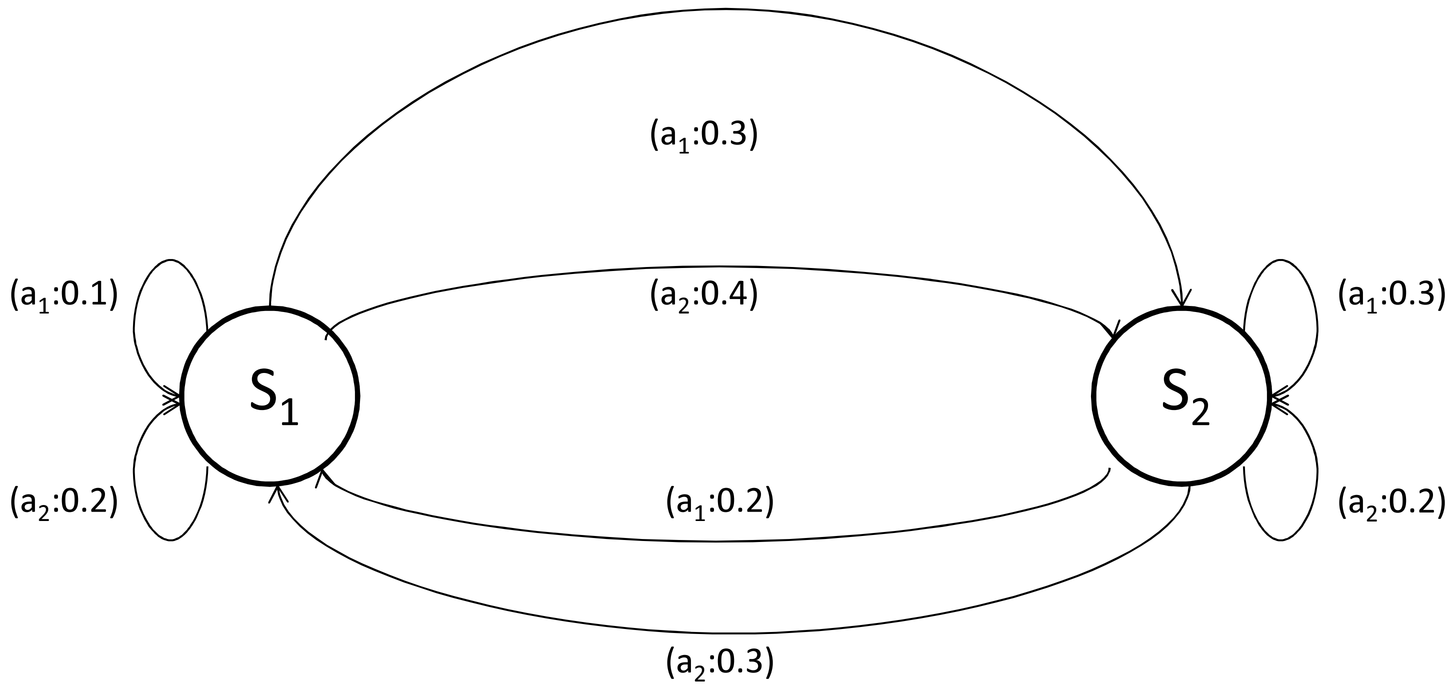
Introducing useful notation

	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

$$
\begin{tikzcd}
S_0 \arrow[r, "\varepsilon"] & S_1 \arrow[r, "R"] & S_2 \arrow[r, "R"] & S_3 \arrow[r, "G"] & S_4 \arrow[r, "G"] & S_5 \arrow[r, "B"] & S_6 \arrow[r, "R"] & S_7 \arrow[d, "G"] \\
& & & & & & & S_8 \arrow[d, "R"] \\
& & & & & & & S_9
\end{tikzcd}
$$

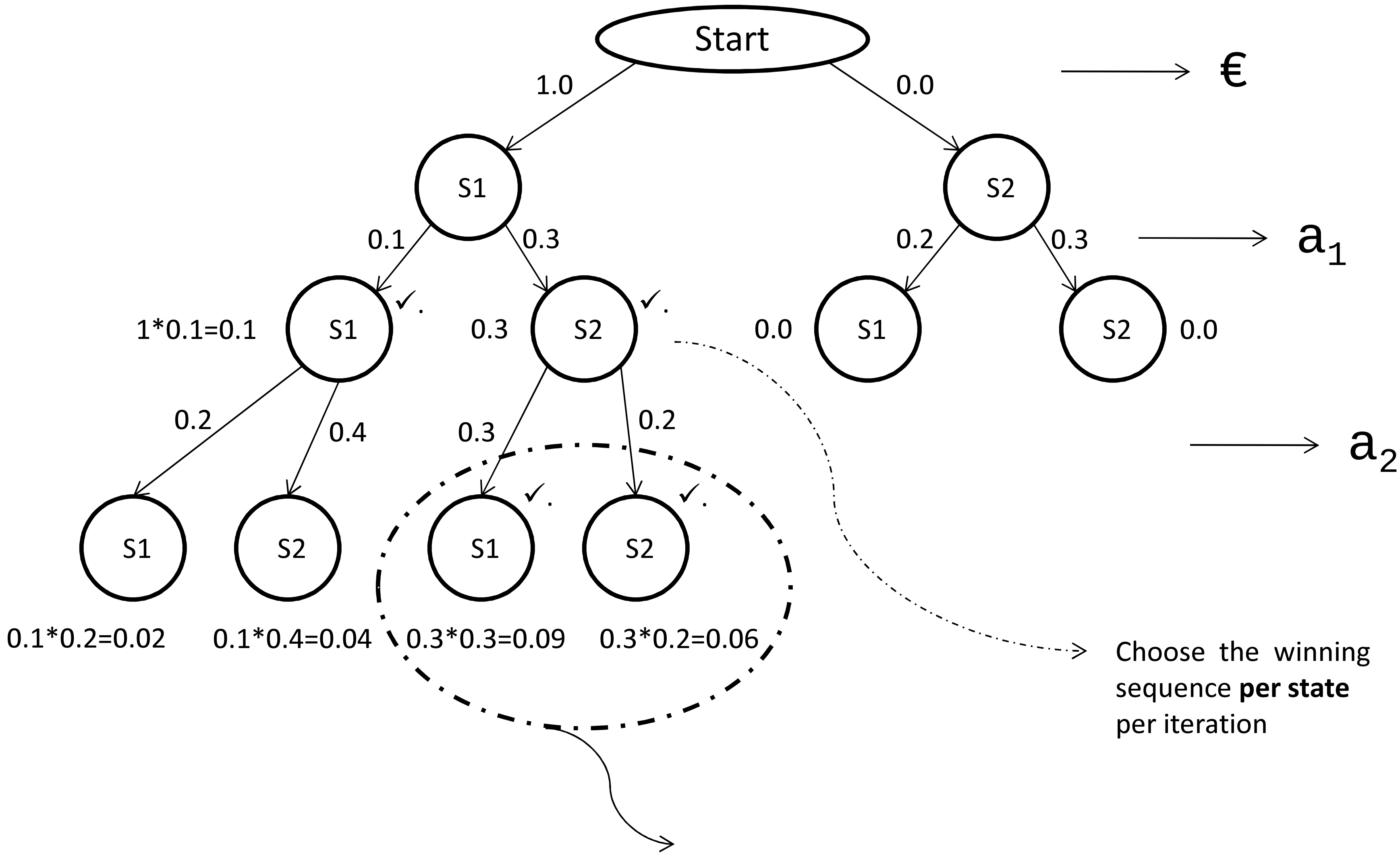
$$P(O_k|S_k).P(S_{k+1}|S_k)=P(S_k \xrightarrow{O_k} S_{k+1})$$

Probabilistic FSM

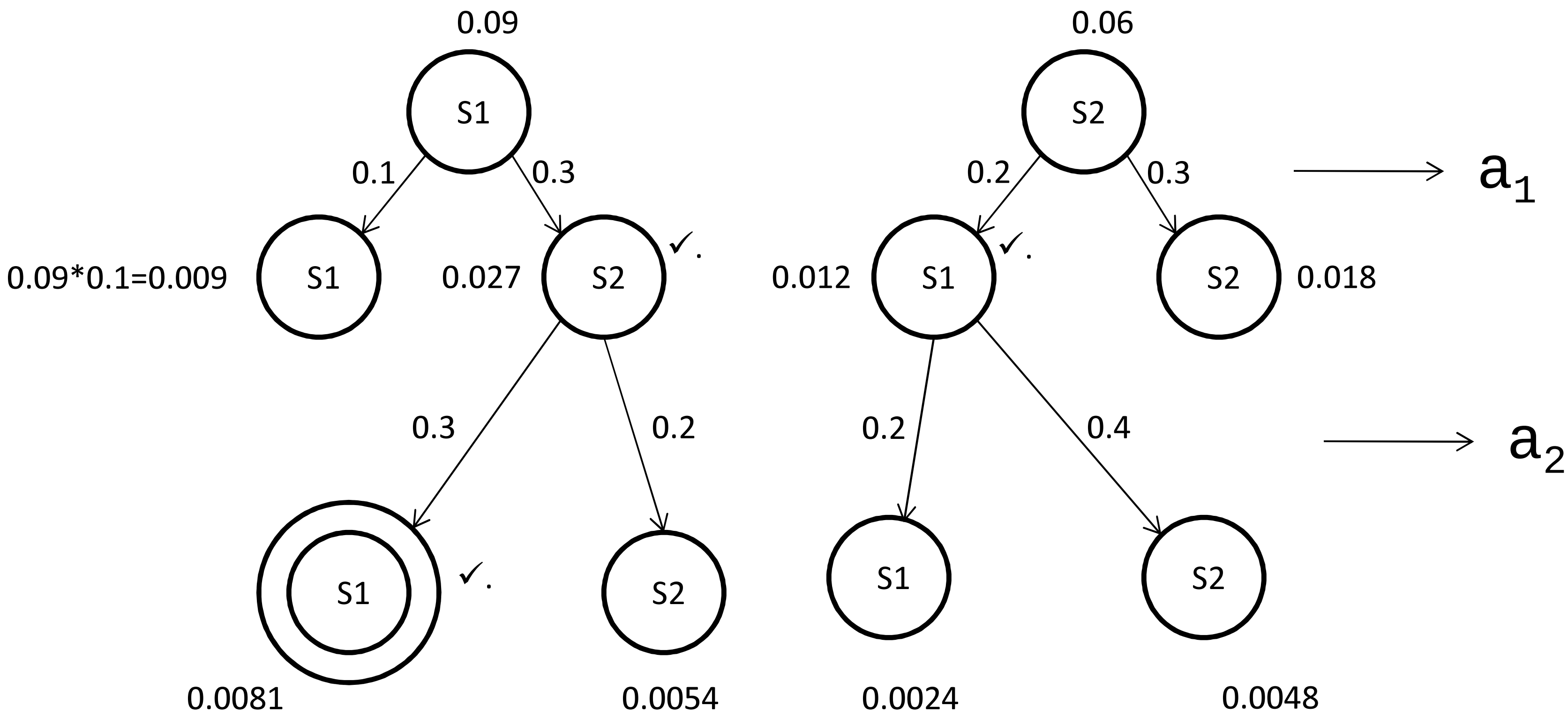


The question here is:
"what is the most likely state sequence given the output sequence seen"

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)

$$S_1 \qquad S_2 \qquad S_1 \qquad S_2 \qquad S_1$$
$$a_1 \qquad a_2 \qquad a_1 \qquad a_2$$

Problem statement: Find the best possible sequence

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

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

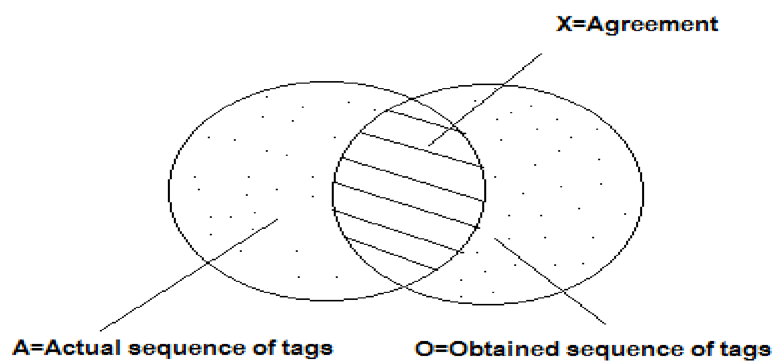
$$\text{Model or Machine} = \{S_0, S, A, T\}$$

Start symbol   State collection   Alphabet set   Transitions

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

Evaluation of POS Tagging

- $\hat{}=w_0 \quad w_1 \quad w_2 \quad w_3 \quad \ldots \quad w_n \quad w_{n+1}=\$$
- $\hat{}=T_0 \quad T_1 \quad T_2 \quad T_3 \quad \ldots \quad T_n \quad T_{n+1}=\$$
- Gold Standard - 80-20 rule: 5 fold cross validation
  - Divide data into 5 folds, 4 folds for training, 1 fold for testing



- Precision P = $\frac{|X|}{|O|}$ Recall R = $\frac{|X|}{|A|}$
- F-measure F = $\frac{2PR}{P+R}$

POS: Tagset

Penn tagset (1/2)

CC	Coord Conjuncn	*and,but,or*	NN	Noun, sing. or mass	*dog*
CD	Cardinal number	*one,two*	NNS	Noun, plural	*dogs*
DT	Determiner	*the,some*	NNP	Proper noun, sing.	*Edinburgh*
EX	Existential there	*there*	NNPS	Proper noun, plural	*Orkneys*
FW	Foreign Word	*mon dieu*	PDT	Predeterminer	*all, both*
IN	Preposition	*of,in,by*	POS	Possessive ending	*'s*
JJ	Adjective	*big*	PP	Personal pronoun	*I,you,she*
JJR	Adj., comparative	*bigger*	PP$	Possessive pronoun	*my,one's*
JJS	Adj., superlative	*biggest*	RB	Adverb	*quickly*
LS	List item marker	*1,One*	RBR	Adverb, comparative	*faster*
MD	Modal	*can,should*	RBS	Adverb, superlative	*fastest*

Penn tagset (2/2)

RP	Particle	*up,off*	WP$	Possessive-Wh	*whose*
SYM	Symbol	+,%,&	WRB	Wh-adverb	*how,where*
TO	"to"	*to*	$	Dollar sign	$
UH	Interjection	*oh, oops*	#	Pound sign	#
VB	verb, base form	*eat*	“	Left quote	‘ “
VBD	verb, past tense	*ate*	”	Right quote	’ ”
VBG	verb, gerund	*eating*	(	Left paren	(
VBN	verb, past part	*eaten*	)	Right paren	)
VBP	Verb, non-3sg, pres	*eat*	,	Comma	,
VBZ	Verb, 3sg, pres	*eats*	.	Sent-final punct	. ! ?
WDT	Wh-determiner	*which,that*	:	Mid-sent punct.	: ; — ...
WP	Wh-pronoun	*what,who*			

Indian Language Tagset: Noun

Sl. No	Category			Label	Annotation Convention**	Examples
	Top level	Subtype (level 1)	Subtype (level 2)			
1	**Noun**			**N**	**N**	ladakaa, raajaa, kitaaba
1.1		Common		NN	N__NN	kitaaba, kalama, cashmaa
1.2		Proper		NNP	N__NNP	Mohan, ravi, rashmi
1.4		Nloc		NST	N__NST	Uupara, niice, aage,

Indian Language Tagset: Pronoun

2	**Pronoun**			**PR**	**PR**	Yaha, vaha, jo
2.1		Personal		PRP	PR__PRP	Vaha, main, tuma, ve
2.2		Reflexive		PRF	PR__PRF	Apanaa, swayam, khuda
2.3		Relative		PRL	PR__PRL	Jo, jis, jab, jahaaM,
2.4		Reciprocal		PRC	PR__PRC	Paraspara, aapasa
2.5		Wh-word		PRQ	PR__PRQ	Kauna, kab, kahaaM
		Indefinite		PRI	PR__PRI	Koii, kis

Indian Language Tagset: Quantifier

10.1		General		QTF	QT__QTF	thoRaa, bahuta, kucha
10.2		Cardinals		QTC	QT__QTC	eka, do, tiina,
10.3		Ordinals		QTO	QT__QTO	pahalaa, duusaraa

Indian Language Tagset: Demonstrative

3	**Demonstrative**			**DM**	**DM**	Vaha, jo, yaha,	
3.1		Deictic		DMD	DM__DMD	Vaha, yaha	
3.2		Relative		DMR	DM__DMR	jo, jis	
3.3		Wh-word		DMQ	DM__DMQ	kis, kaun	
		Indefinite		DMI	DM__DMI	KoI, kis	

Indian Language Tagset: Verb, Adjective, Adverb

4	**Verb**			**V**	**V**	giraa, gayaa, sonaa, haMstaa, hai, rahaa
4.1		Main		VM	V__VM	giraa, gayaa, sonaa, haMstaa,
4.2		Auxiliary		VAUX	V__VAUX	hai, rahaa, huaa,
5	**Adjective**			**JJ**	**JJ**	sundara, acchaa, baRaa
6	**Adverb**			**RB**	**RB**	jaldii, teza

Indian Language Tagset: Postposition, conjunction

7	**Postposition**			**PSP**	**PSP**	ne, ko, se, mein
8	**Conjunction**			**CC**	**CC**	aur, agar, tathaa, kyonki
8.1		Co-ordinator		CCD	CC__CCD	aur, balki, parantu
8.2		Subordinator		CCS	CC__CCS	Agar, kyonki, to, ki

Indian Language Tagset: Particle

9	**Particles**			**RP**	**RP**	to, bhii, hii
9.1		Default		RPD	RP__RPD	to, bhii, hii
9.3		Interjectio n		INJ	RP__INJ	are, he, o
9.4		Intensifier		INTF	RP__INTF	bahuta, behada
9.5		Negation		NEG	RP__NEG	nahiin, mata, binaa

Indian Language Tagset: Residuals

11	**Residuals**			**RD**	**RD**		
11.1		Foreign word		RDF	RD__RDF		A word written in script other than the script of the original text
11.2		Symbol		SYM	RD__SYM	$, &, *, (,)	For symbols such as $, & etc
11.3		Punctuation		PUNC	RD__PUNC	., : ;	Only for punctuations
11.4		Unknown		UNK	RD__UNK		
11.5		Echowords		ECH	RD__ECH	(Paanii-) vaanii, (khaanaa-) vaanaa	

Challenge of POS tagging

Example from Indian Language

Tagging of *jo*, *vaha*, *kaun* and their inflected forms in Hindi and their equivalents in multiple languages

DEM and PRON labels

- ***Jo_DEM*** *ladakaa kal aayaa thaa, vaha cricket acchhaa khel letaa hai*
- ***Jo_PRON*** *kal aayaa thaa, vaha cricket acchhaa khel letaa hai*

Disambiguation rule-1

- ***If***
  - ***Jo is followed by noun***
- ***Then***
  - ***DEM***
- ***Else***
  - ***…***

False Negative

- When there is arbitrary amount of text between the *jo* and the noun
- *Jo_??? **bhaagtaa huaa, haftaa huaa, rotaa huaa, chennai academy a koching lenevaalaa** ladakaa kal aayaa thaa, vaha cricket acchhaa khel letaa hai*

False Positive

- *Jo_DEM* (wrong!) ***duniyadarii samajhkar chaltaa hai, …***
- *Jo_DEM/PRON? manushya manushyoM ke biich ristoM naatoM ko samajhkar chaltaa hai, …* (ambiguous)

False Positive for Bengali

- *Je_DEM* (wrong!) ***bhaalobaasaa paay,*** *sei bhaalobaasaa dite paare*

  (one who gets love can give love)
- *Je_DEM* (right!) *bhaalobaasa tumi kalpanaa korchho, taa e jagat e sambhab nay*

  (*the love that you imagine exits, is impossible in this world)*

Will fail

- In the similar situation for
  - *Jis, jin, vaha, us, un*
- All these forms add to corpus count

Disambiguation rule-2

- ***If***
  - ***Jo is oblique (attached with ne, ko, se etc. attached)***
- ***Then***
  - ***It is PRON***
- ***Else***
  - ***<other tests>***

Will fail (false positive)

- In case of languages that demand agreement between *jo*-form and the noun it qualifies
- *E.g.* Sanskrit
- *Yasya_PRON* (wrong!) *baalakasya aananam drshtyaa…* (*jis ladake kaa muha dekhkar*)
- *Yasya_PRON* (wrong!) *kamaniyasya baalakasya aananam drshtyaa…*

Will also fail for

- Rules that depend on the whether the noun following *jo/vaha/kaun or its form* is oblique or not
- Because the case marker can be far from the noun
- *<vaha or its form> ladakii jise piliya kii bimaarii ho gayiii thii* ***ko*** *…*
- **Needs discussions across languages**

DEM vs. PRON cannot be disambiguated

IN GENERAL

At the level of the POS tagger

i.e.

Cannot assume parsing

Cannot assume semantics

POS Tags

- NN – Noun; e.g. *Dog_NN*
- VM – Main Verb; e.g. *Run_VM*
- VAUX – Auxiliary Verb; e.g. *Is_VAUX*
- JJ – Adjective; e.g. *Red_JJ*
- PRP – Pronoun; e.g. *You_PRP*
- NNP – Proper Noun; e.g. *John_NNP*
- etc.

POS Tag Ambiguity

- In English : I bank$_1$ on the bank$_2$ on the river bank$_3$ for my transactions.
  - Bank$_1$ is verb, the other two banks are noun

- In Hindi :
  - ”Khaanaa” : can be noun (food) or verb (to eat)

For Hindi

- *Rama achhaa gaata hai*. (hai is VAUX : Auxiliary verb); *Ram sings well*
- *Rama achha ladakaa hai*. (hai is VCOP : Copula verb); *Ram is a good boy*

Morphology: syncretism

Languages that are poor in Morphology (Chinese, English) have Role Ambiguity or **Syncretism** (fusion of originally different inflected forms resulting in a reduction in the use of inflections)

Eg: *You/They/He/I will* *come* *tomorrow*

Here, just by looking at the verb '*come*' its syntactic features aren't apparent i.e.

Gender, Number, Person, Tense, Aspect, Modality (GNPTAM)

-Aspect tells us how the event occurred; whether it is completed, continuous, or habitual. Eg: *John came, John will be coming*

- Modality indicates possibility or obligation. Eg: *John can arrive / John must arrive*

Contrast this with the Hindi Translation of *'I will <u>come</u> tomorrow'*

मैं Main (I) *कल kal(tomorrow)* *<u>आऊंगा aaunga (will come)</u>*

<u>आऊंगा aaunga</u> – GNPTAM: Male, Singular, First, Future

आओगे (Aaoge) – has number ambiguity, but still contains more information than '*come*' in English

Books etc.

- Main Text(s):
  - Natural Language Understanding: James Allan
  - Speech and NLP: Jurafsky and Martin
  - Foundations of Statistical NLP: Manning and Schutze
- Other References:
  - NLP a Paninian Perspective: Bharati, Cahitanya and Sangal
  - Statistical NLP: Charniak
- Journals
  - Computational Linguistics, Natural Language Engineering, AI, AI Magazine, IEEE SMC
- Conferences
  - ACL, EACL, COLING, MT Summit, EMNLP, IJCNLP, HLT, ICON, SIGIR, WWW, ICML, ECML