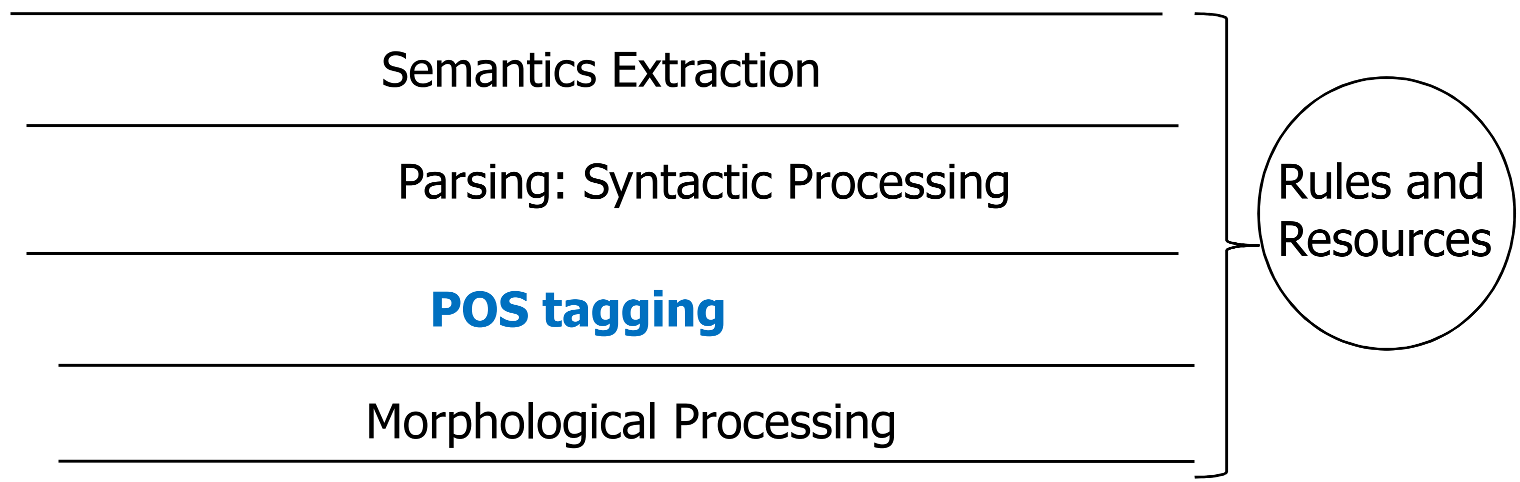
CS344: Introduction to Artificial Intelligence
(associated lab: CS386)

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

Lecture 12, 13: Sequence Labeling using HMM- POS tagging; Baum Welch

1st Feb, 2011

POS tagging sits between Morphology and Parsing



Morph→POS->Parse

- Because of this sequence, at the ***level*** of POS tagging the only information available is the word, its constituents, its properties and its neighbouring words and their properties

W_0 W_1 W_2 … $\mathbf{W_i}$ W_{i+1} W_{n-1} W_n

Word of interest

Cannot assume parsing and semantic processing

- Parsing identifies long distance dependencies
- Needs POS tagging which must finish earlier
- Semantic processing needs parsing and POS tagging

Example

- *Vaha ladakaa so rahaa hai*
- *(that boy is sleeping)*
- *Vaha cricket khel rahaa hai*
- *(he plays cricket)*
- The fact that "*vaha*" is demonstrative in the first sentence and pronoun in the second sentence, needs deeper levels of information

“vaha cricket” is not that simple!

- *Vaha cricket jisme bhrastaachaar ho, hame nahii chaahiye*
- *(that cricket which has corruption in it is not acceptable to us)*
- Here *"vaha"* is demonstrative
- Needs deeper level of processing

Syntactic processing also cannot be assumed

- *raam kaa yaha baar baar shyaam kaa ghar binaa bataaye JAANAA*
  mujhe bilkul pasand nahii haai
- *(I do not at all like the fact that Ram goes to Shyam's house repeatedly*
  without informing (anybody))
- "Ram-GENITIVE this again and again Shyam-GENITIVE house any not saying GOING I-dative at all like not VCOP"
- *JAANAA* can be VINF (verb infinitive) or VN (verb nominal, i.e., gerundial)

Syntactic processing also cannot be assumed (cntd.)

- *raam kaa yaha baar baar shyaam kaa ghar binaa bataaye JAANAA*
  mujhe bilkul pasand nahii haai
- The correct clue for disambiguation here is 'raam kaa', and this word group is far apart
- One needs to determine the structure of intervening constituents
- This needs parsing which in turn needs correct tags
- Thus there is a circularity which can be broken only by retaining ONE of VINF and VN.

Fundamental principle of POS tagset design

- IN THE TAGSET DO NOT HAVE TAGS THAT ARE POTENTIAL COMPETITORS AND TIE BETWEEN WHICH CAN BE BROKEN ONLY BY NLP PROCESSES COMING AFTER THE PARTICULAR TAGGING TASK.

Computation of POS tags

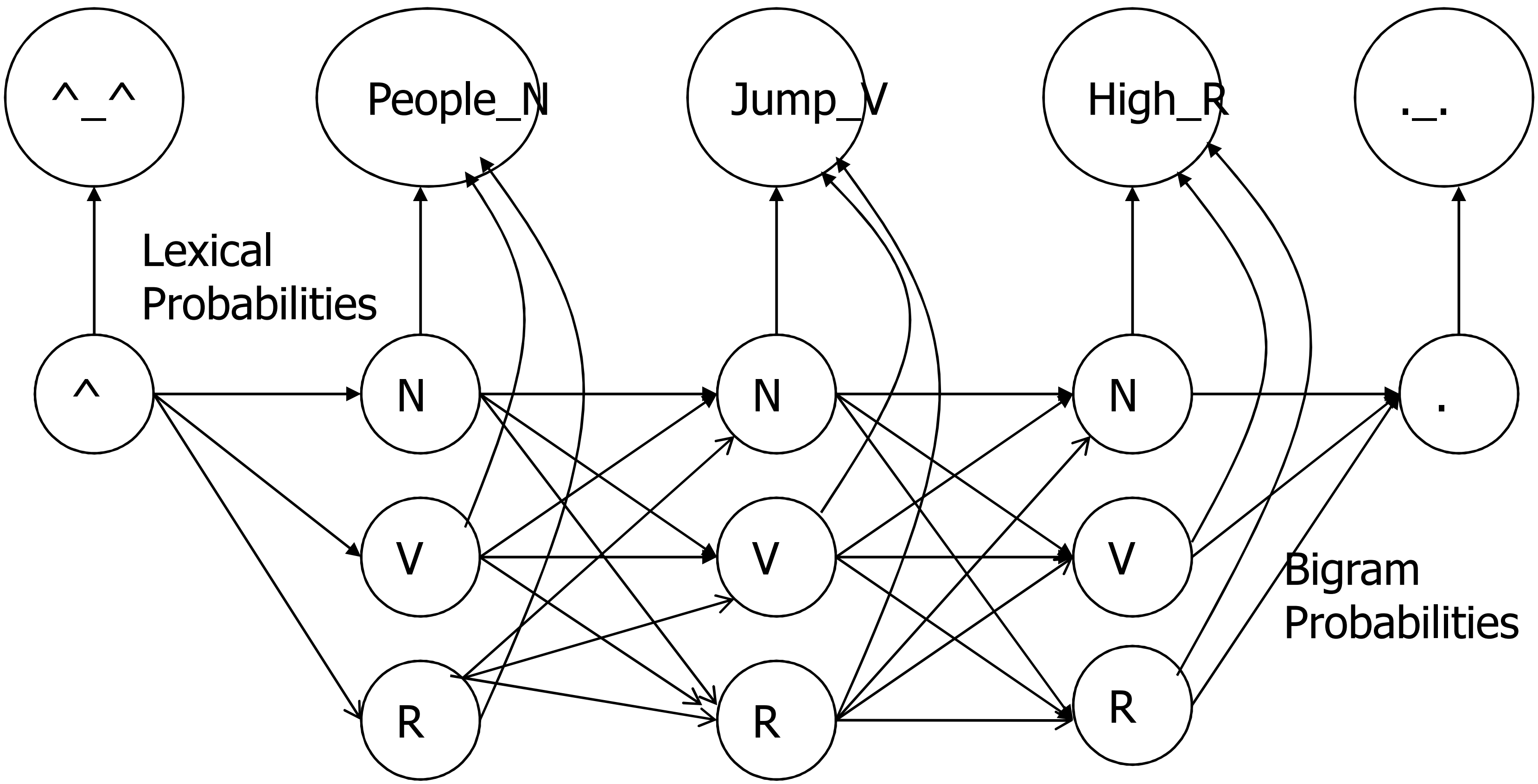
Process

- List all possible tag for each word in sentence.
- Choose best suitable tag sequence.

Example

- ”People jump high”.
- People : Noun/Verb
- jump : Noun/Verb
- high : Noun/Adjective
- We can start with probabilities.

Generative Model



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

Example of Calculation from Actual Data

- Corpus
  - *^ Ram got many NLP books. He found them all very interesting.*
- Pos Tagged
  - *^ N V A N N . N V N A R A .*

Recording numbers (bigram assumption)

	^	**N**	**V**	**A**	**R**	**.**
^	0	2	0	0	0	0
N	0	1	2	1	0	1
V	0	1	0	1	0	0
A	0	1	0	0	1	1
R	0	0	0	1	0	0
.	1	0	0	0	0	0

^ Ram got many NLP books. He found them all very interesting.

Pos Tagged

^ N V A N N . N V N A R A .

Probabilities

	^	N	V	A	R	.
^	0	1	0	0	0	0
N	0	1/5	2/5	1/5	0	1/5
V	0	1/2	0	1/2	0	0
A	0	1/3	0	0	1/3	1/3
R	0	0	0	1	0	0
.	1	0	0	0	0	0

^ Ram got many NLP books. He found them all very interesting.

Pos Tagged

^ N V A N N . N V N A R A .

To find

- $T^* = argmax\ (P(T)\ P(W/T))$
- $P(T).P(W/T) = \prod_{i=1 \rightarrow n+1} P(\ t_i\ /\ t_{i-1}\ ).P(w_i\ /t_i)$
- $P(\ t_i\ /\ t_{i-1}\ )$: *Bigram probability*
- $P(w_i\ /t_i)$: *Lexical probability*

Note: $P(w_i/t_i)=1$ for $i=0$ (^, sentence beginner)) and $i=(n+1)$ (., fullstop)

Bigram probabilities

	N	*V*	*A*	*R*
N	0.15	0.7	0.05	0.1
V	0.6	0.2	0.1	0.1
A	0.5	0.2	0.3	0
R	0.1	0.3	0.5	0.1

Lexical Probability

-

	People	jump	high
N	10^{-5}	$0.4x10^{-3}$	10^{-7}
V	10^{-7}	10^{-2}	10^{-7}
A	0	0	10^{-1}
R	0	0	0

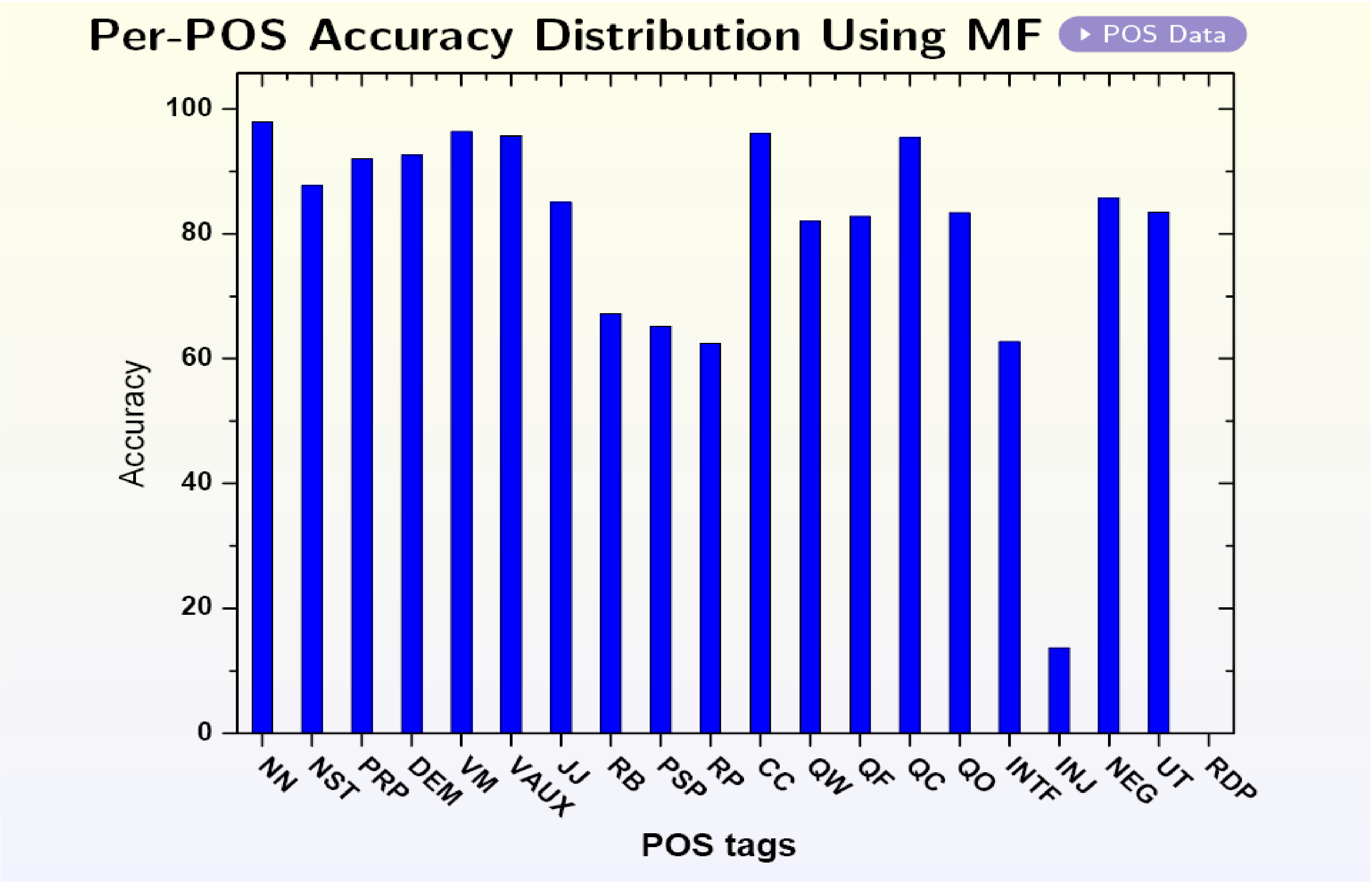
values in cell are P(col-heading/row-heading)

Some notable text corpora of English

- American National Corpus
- Bank of English
- British National Corpus
- Corpus Juris Secundum
- Corpus of Contemporary American English (COCA) 400+ million words, 1990-present. Freely searchable online.
- Brown Corpus, forming part of the "Brown Family" of corpora, together with LOB, Frown and F-LOB.
- International Corpus of English
- Oxford English Corpus
- Scottish Corpus of Texts & Speech

Accuracy measurement in POS tagging

Standard Bar chart: Per Part of Speech Accuracy



Standard Data: Confusion Matrix

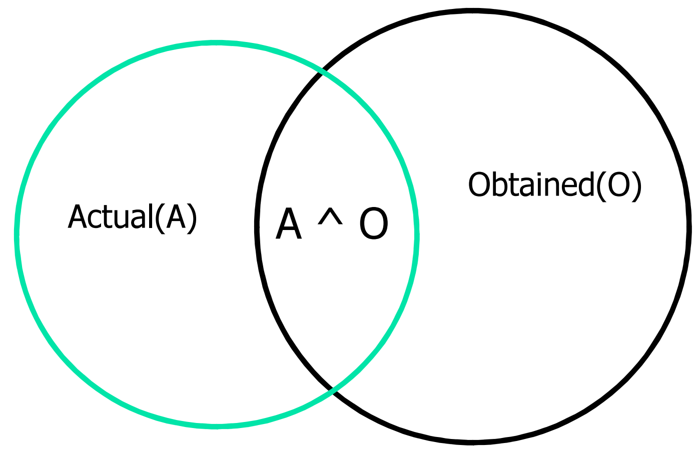
POS Data

	NN	NST	PRP	DEM	VM	VAUX
NN	49988	18	92	2	167	4
NST	33	507	9	0	3	0
PRP	145	3	8071	312	8	5
DEM	3	0	231	3002	2	1
VM	225	1	4	9	17078	347
VAUX	10	0	1	1	257	6025

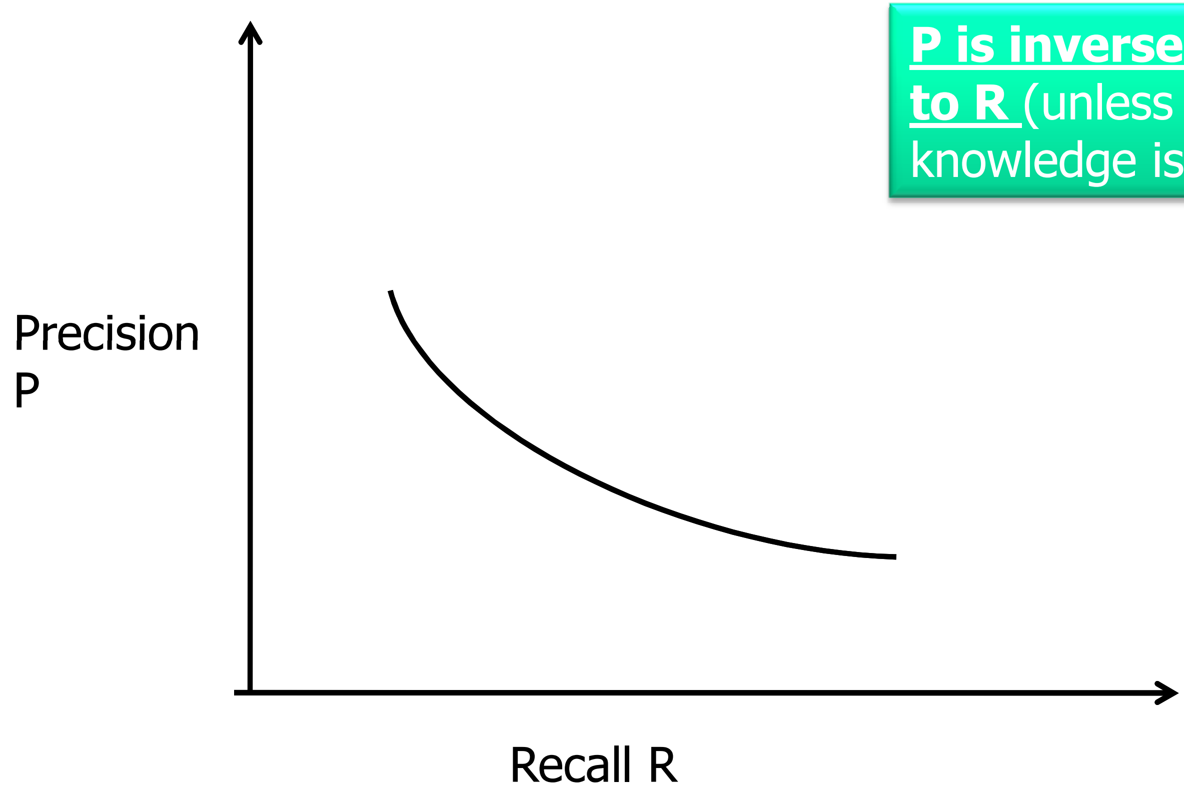
Table: POS Confusion Matrix with MF

How to check quality of tagging (P, R, F)

- Three parameters
  - Precision P = |A ^ O|/|O|
  - Recall R = |A ^ O| / |A|
  - F-score = 2PR/(P+R)
    - Harmonic mean



Relation between P & R

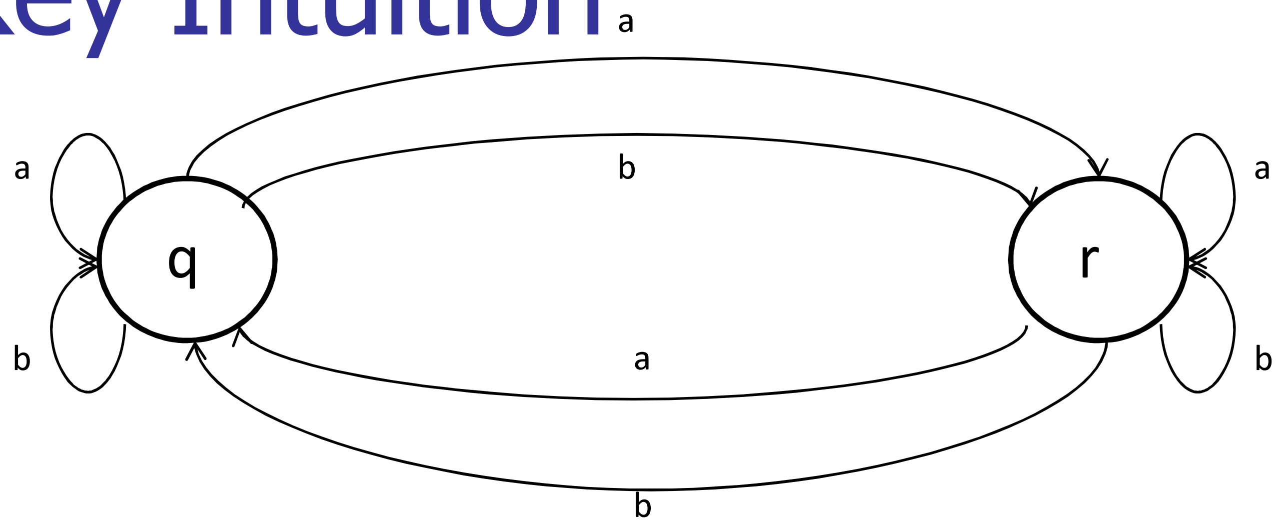


P is inversely related to R (unless additional knowledge is given)

HMM Training

Baum Welch or Forward Backward Algorithm

Key Intuition



Given: Training sequence

Initialization: Probability values

Compute: Pr (state seq | training seq)

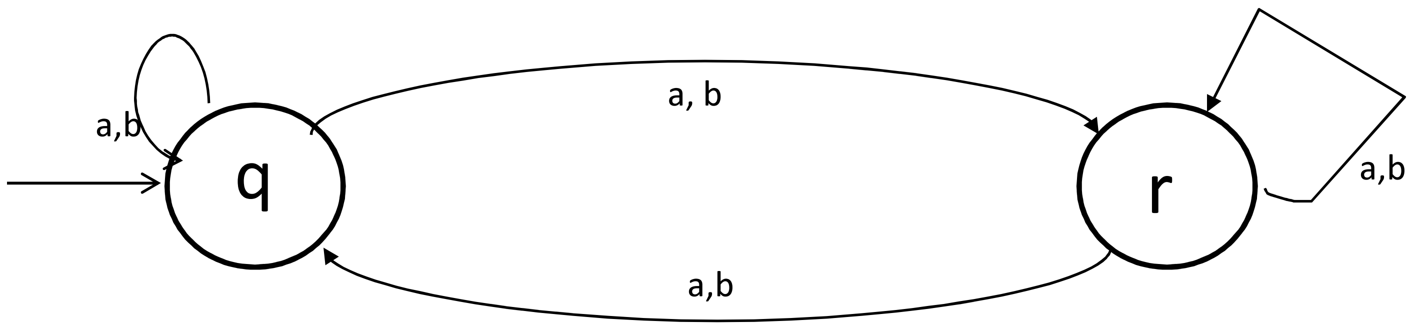
get expected count of transition

compute rule probabilities

Approach: Initialize the probabilities and recompute them…

EM like approach

Baum-Welch algorithm: counts



String = abb aaa bbb aaa

Sequence of states with respect to input symbols

o/p seq

State seq

$$\rightarrow q \xrightarrow{a} r \xrightarrow{b} q \xrightarrow{b} q \xrightarrow{a} r \xrightarrow{a} q \xrightarrow{a} r \xrightarrow{b} q \xrightarrow{b} q \xrightarrow{b} q \xrightarrow{a} r \xrightarrow{a} q \xrightarrow{a} r$$

Calculating probabilities from table

$$P(q \xrightarrow{a} r) = 5/8$$

$$P(q \xrightarrow{b} r) = 3/8$$

$$P(s^i \xrightarrow{w_k} s^j) = \frac{c(s^i \xrightarrow{w_k} s^j)}{\sum_{l=1}^{T} \sum_{m=1}^{A} c(s^i \xrightarrow{w_m} s^l)}$$

T=#states

A=#alphabet symbols

Table of counts

Src	**Dest**	**O/P**	**Count**
q	r	a	5
q	q	b	3
r	q	a	3
r	q	b	2

Now if we have a non-deterministic transitions then multiple state seq possible for the given o/p seq (ref. to previous slide's feature). Our aim is to find expected count through this.

Interplay Between Two Equations

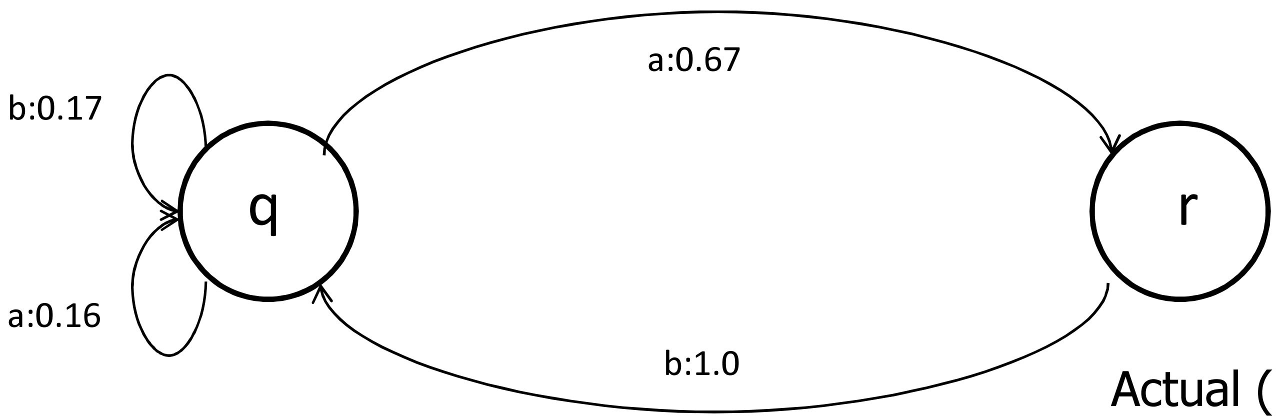
$$P(s^i \xrightarrow{W_k} s^j) = \frac{c(s^i \xrightarrow{W_k} s^j)}{\sum_{l=0}^{T} \sum_{m=0}^{A} c(s^i \xrightarrow{W_m} s^l)}$$

$$C(s^i \xrightarrow{W_k} s^j) =$$

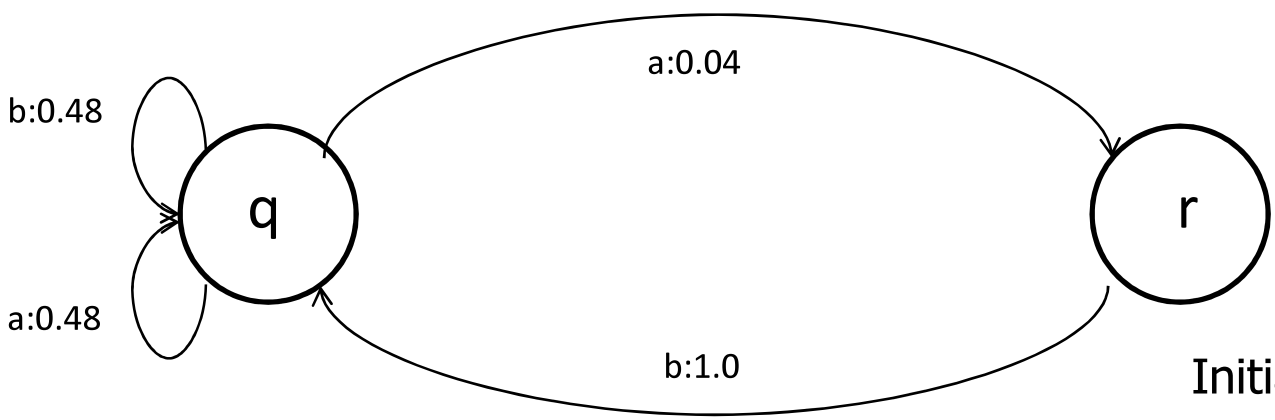
$$\sum_{s_{0,n+1}} P(S_{0,n+1} \mid W_{0,n}) \times n(s^i \xrightarrow{W_k} s^j, S_{0,n+1}, w_{0,n})$$

No. of times the transitions $s^i \xrightarrow{w_k} s^j$ occurs in the string

Illustration



Actual (Desired) HMM



Initial guess

One run of Baum-Welch algorithm: *string ababb*

$\in \to a$	$a \to b$	$b \to a$	$a \to b$	$b \to b$	$b \to \in$	P(path)	$q \xrightarrow{a} r$	$r \xrightarrow{b} q$	$q \xrightarrow{a} q$	$q \xrightarrow{b} q$
q	r	q	r	q	q	0.00077	0.00154	0.00154	0	0.00077
q	r	q	q	q	q	0.00442	0.00442	0.00442	0.00442	0.00884
q	q	q	r	q	q	0.00442	0.00442	0.00442	0.00442	0.00884
q	q	q	q	q	q	0.02548	0.0	0.000	0.05096	0.07644
Rounded Total →						0.035	0.01	0.01	0.06	0.095
New Probabilities (P) → State sequences							0.06 =(0.01/(0.01+0.06+0.095)	1.0	0.36	0.581

* ε is considered as starting and ending symbol of the input sequence string. Through multiple iterations the probability values will converge.

Computational part (1/2)

$$C(s^i \xrightarrow{w_k} s^j) = \sum_{s_{0,n+1}} [P(S_{0,n+1} \mid W_{0,n}) \times n(s^i \xrightarrow{w_k} s^j, S_{0,n+1}, W_{0,n})]$$

$$= \frac{1}{P(W_{0,n})} \sum_{s_{0,n+1}} [P(S_{0,n+1}, W_{0,n}) \times n(s^i \xrightarrow{w_k} s^j, S_{0,n+1}, W_{0,n})]$$

$$= \frac{1}{P(W_{0,n})} \sum_{t=0,n} \sum_{s_{0,n+1}} [P(S_t = s^i, W_t = w_k, S_{t+1} = s^j, S_{0,n+1}, W_{0,n})]$$

$$= \frac{1}{P(W_{0,n})} \sum_{t=0,n} [P(S_t = s^i, W_t = w_k, S_{t+1} = s^j, W_{0,n})]$$

$$S0 \xrightarrow{w_0} S1 \xrightarrow{w_1} S1 \xrightarrow{w_2} \ldots\ Si \xrightarrow{w_k} Sj \ldots \rightarrow Sn\text{-}1 \xrightarrow{w_{n-1}} Sn \xrightarrow{w_n} Sn+1$$

Computational part (2/2)

$$\sum_{t=0}^{n} P(S_t = s^i, S_{t+1} = s^j, W_t = w_k, W_{0,n})$$

$$= \sum_{t=0}^{n} P(W_{0,t-1}, S_t = s^i, S_{t+1} = s^j, W_t = w_k, W_{t+1,n})$$

$$= \sum_{t=0}^{n} P(W_{0,t-1}, S_t = s^i) P(S_{t+1} = s^j, W_t = w_k \mid W_{0,t-1}, S_t = s^i) P(W_{t+1,n} \mid S_{t+1} = s^j)$$

$$= \sum_{t=0}^{n} F(t-1, i) P(S_{t+1} = s^j, W_t = w_k \mid S_t = s^i) B(t+1, j)$$

$$= \sum_{t=0}^{n} F(t-1, i) P(S_{t+1} = s^j, W_t = w_k \mid S_t = s^i) B(t+1, j)$$

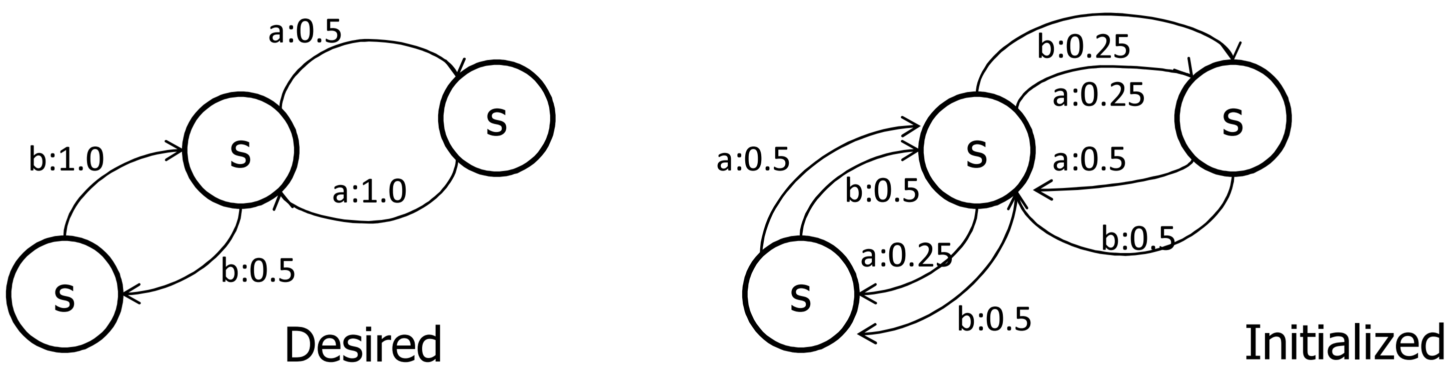
$$= \sum_{t=0}^{n} F(t-1, i) P(s^i \xrightarrow{w_k} s^j) B(t+1, j)$$

$$S0 \xrightarrow{w_0} S1 \xrightarrow{w_1} S1 \xrightarrow{w_2} \ldots Si \xrightarrow{w_k} Sj \ldots \rightarrow Sn\text{-}1 \xrightarrow{w_{n-1}} Sn \xrightarrow{w_n} Sn{+}1$$

Discussions

1. Symmetry breaking:

Example: Symmetry breaking leads to no change in initial values



2 Struck in Local maxima

3. Label bias problem

Probabilities have to sum to 1.

Values can rise at the cost of fall of values for others.