Language and Computation

Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay

10th Jan, 2023

Ode to *Scientists and Engineers*

Scientists ask WHY
Engineers ask WHY NOT
Scientists wonder at WHAT-IS
Engineers wonder WHAT-COULD-BE
World couldn’t do without either.

Scientists STUDY
Engineers MAKE
And ever the twain shall meet.

What is “Language”

Oxford English Dictionary

1. the principal method of human ***communication***, consisting of words used in a structured and conventional way and conveyed by speech, writing, or gesture.

"a study of the way children learn Language"

2. a system of communication used by a particular country or community.

"the book was translated into twenty-five languages"

General point: Properties of Human Languages

(George Yule, “Study of Language”, 1998)

- **Displacement** (Indicators that change with time and place: I saw him yesterday at the market; I will see him tomorrow in the school)
- **Arbitrariness** (name → Meaning; water, chair)
- **Productivity/creativity** (potentially infinite no. of sentences)
- **Cultural Transmission** (child acquires parent’s language)
- **Discreteness** (sound and meaning units separated)
- **Duality** (Surface structure, deep structure)

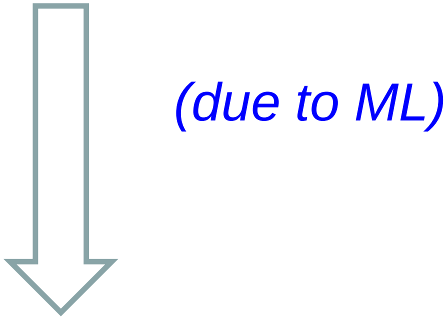
What is “Linguistics”

- Scientific study of language, its underlying and governing rules

- **Descriptive**: describe the language objects, language phenomena AS THEY ARE
- **Prescriptive**: prescribe what is allowed and not allowed, e.g. disallow double negative- “*I did not see nobody in the hall*”; control language behaviour

Natural Language Processing (NLP)

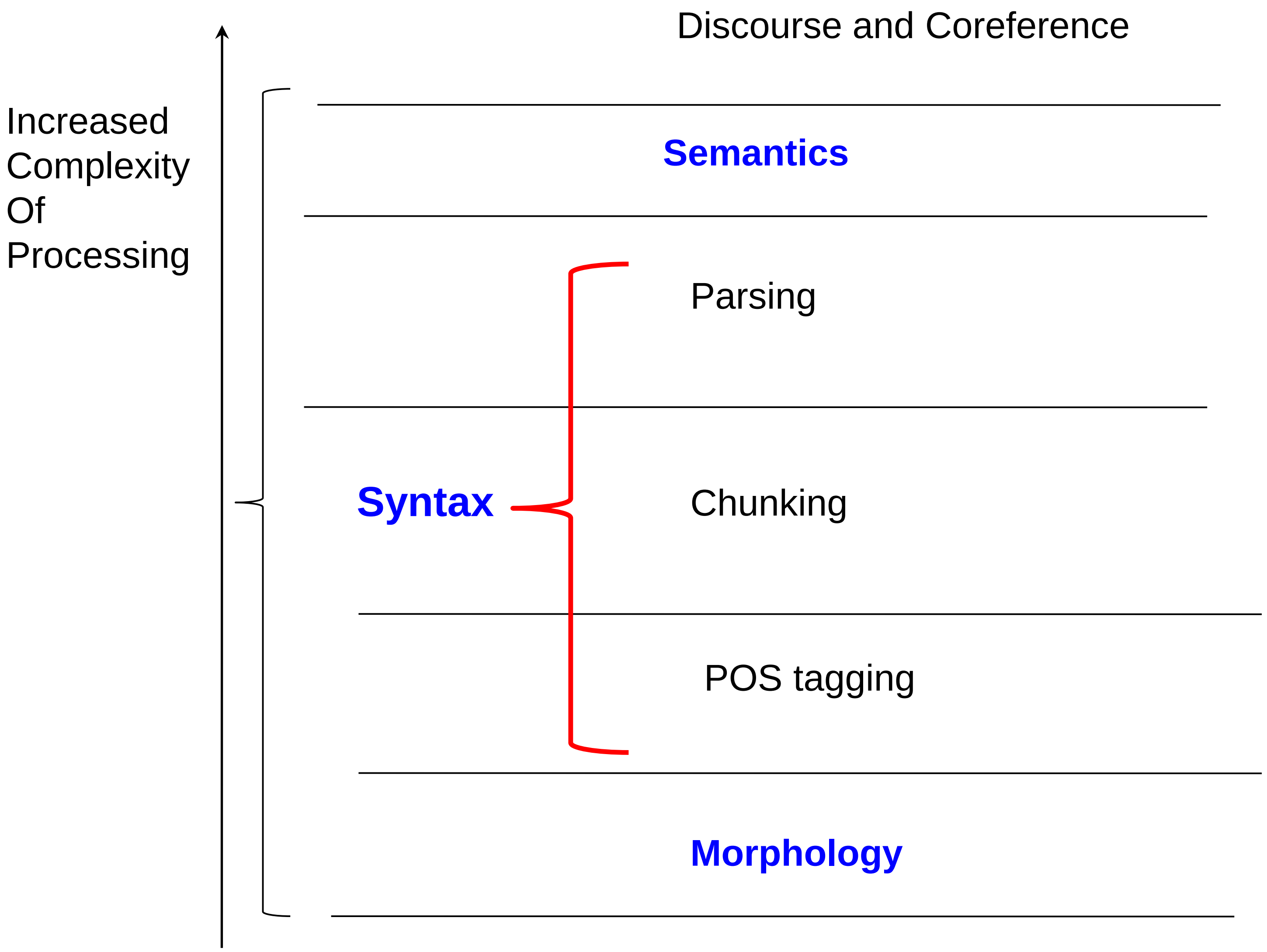
NLP= Language + Computation



(due to ML)

= Linguistics + Probability

NLP Layers



Discourse and Coreference

Increased
Complexity
Of
Processing

Semantics

Parsing

Syntax

Chunking

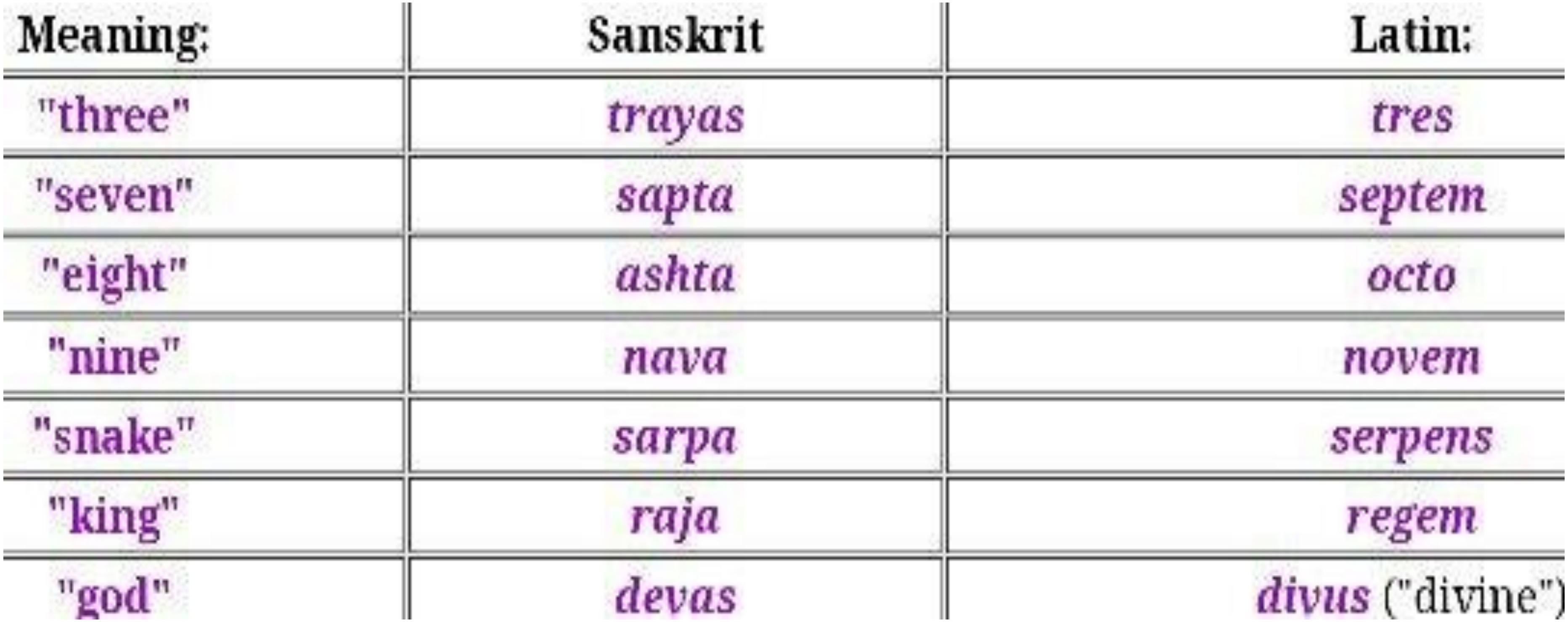
POS tagging

Morphology

Language Typology

Classification according to structural features

Proto-Language *(Wikipedia)*



Meaning:	Sanskrit	Latin:
"three"	*trayas*	*tres*
"seven"	*sapta*	*septem*
"eight"	*ashta*	*octo*
"nine"	*nava*	*novem*
"snake"	*sarpa*	*serpens*
"king"	*raja*	*regem*
"god"	*devas*	*divus* ("divine")

One of the indications that languages descended from a single source

Word order based

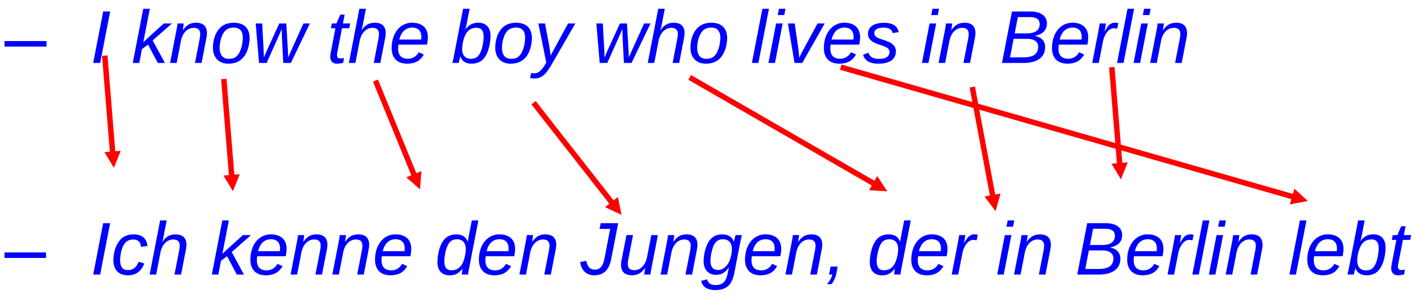
- Object–subject–verb (OSV)
- Object–verb–subject (OVS)
- Subject–verb–object (SVO): English
- Subject–object–verb (SOV): Most Indian Languages
- Verb–subject–object (VSO)
- Verb–object–subject (VOS)

Dominant Word Order Distribution Across Languages (Wikipedia)

Type	Languages	%	Families	%
SOV (Hindi)	2,275	43,3%	239	65.3%
SVO (English)	2,117	40.3%	55	15%
VSO (tagalog in phillipines)	503	9.5%	27	7.4%
VOS (Malagasy in Madagaskar)	174	3.3%	15	4.1%
NODOM (Sanskrit)	124	2.3%	26	7.1%
OVS (Korean and Japanese, many times)	40	0.7%	3	0.8%
OSV (Warao in Venezuela)	19	0.3%	1	0.3%

Some interesting cases

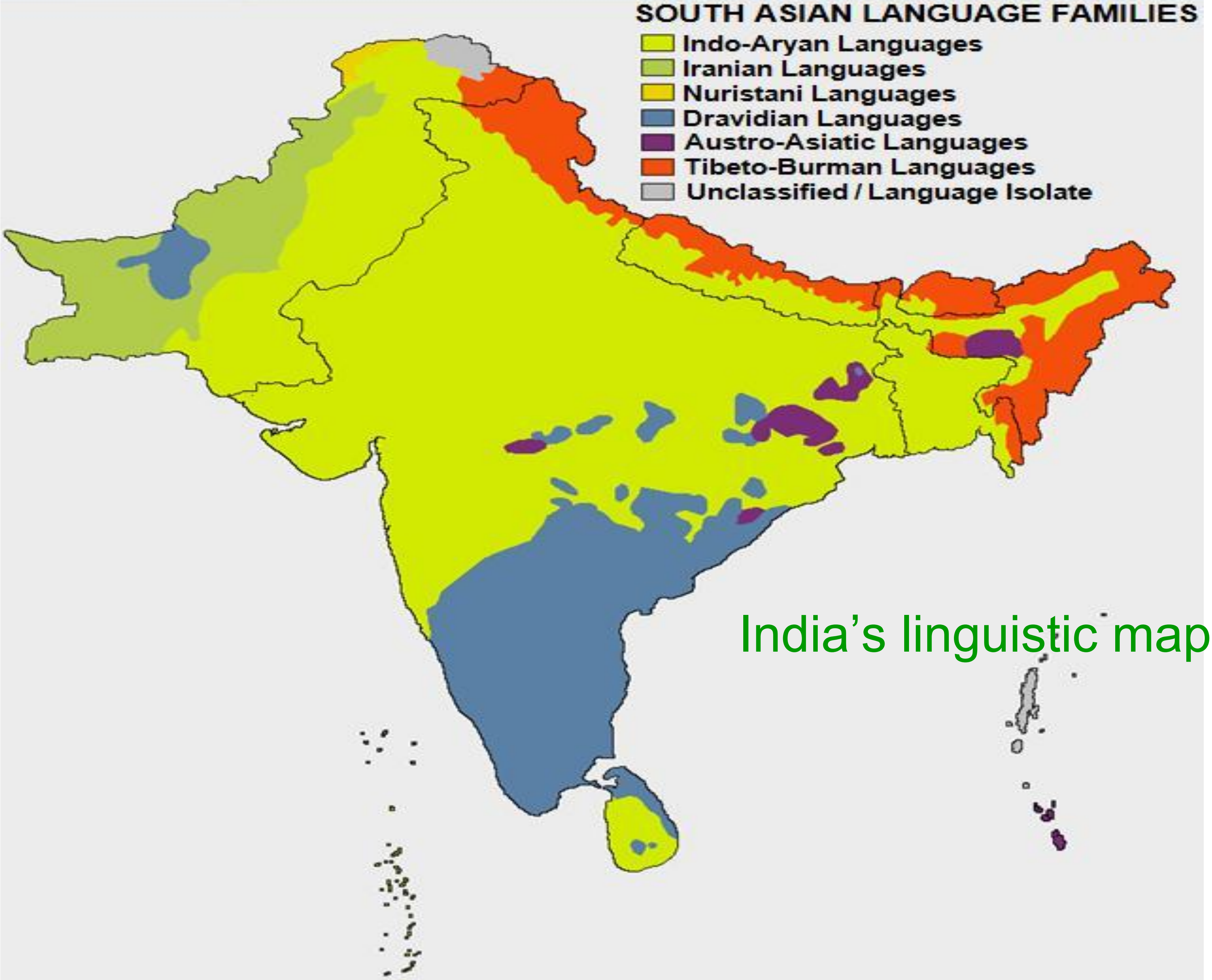
- German word order depends on the position of the main verb (MV)



  – *I know the boy who lives in Berlin*

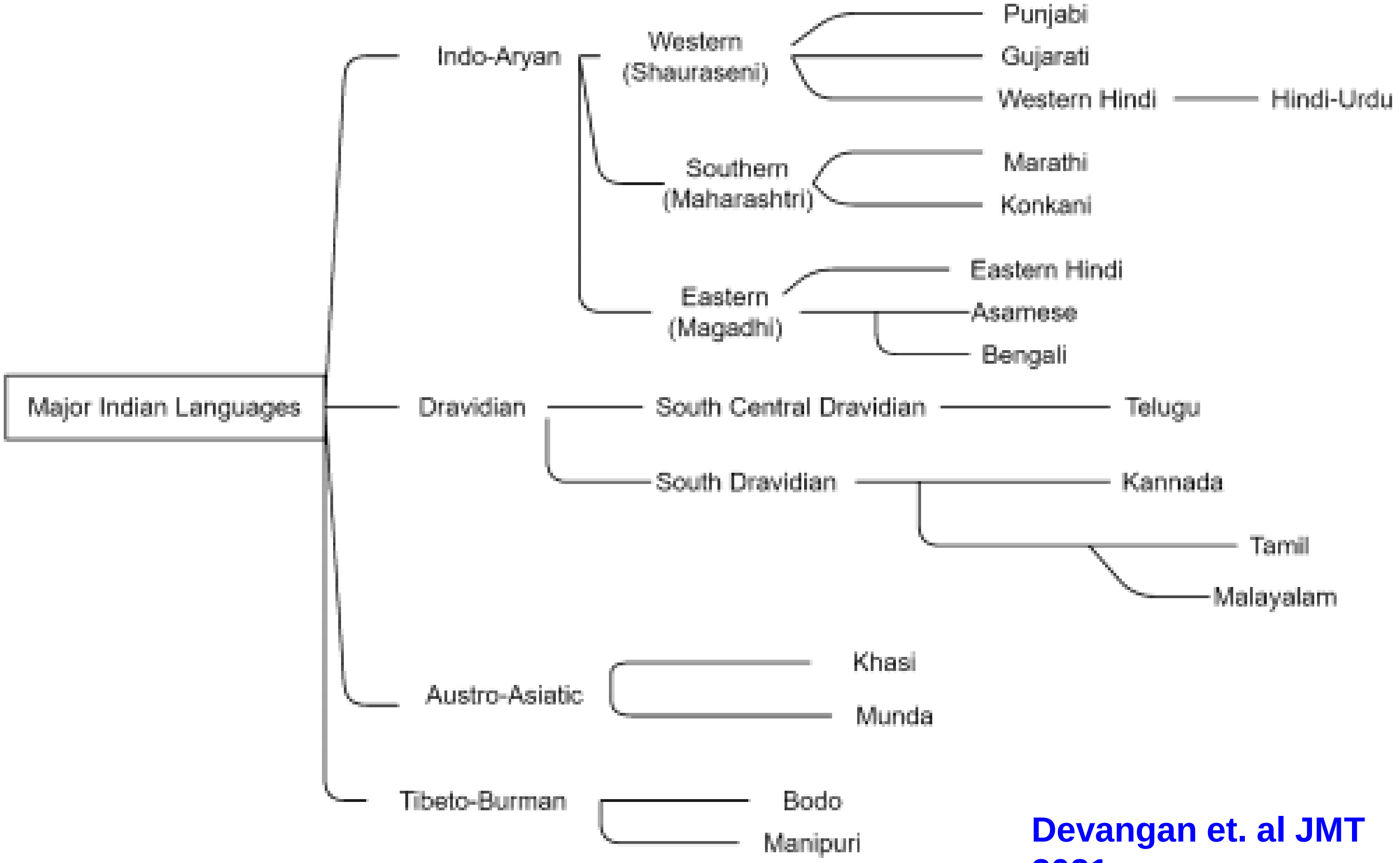
  – *Ich kenne den Jungen, der in Berlin lebt*

- Hindi:
  – *Mei us ladke ko jaantaa hu jo barlin me rahataa haai*
  – *Mei us ladke ko jo barlin me rahataa haai jaantaa hu*



India’s linguistic map

74 S. Dewangan et al.



Devangan et. al JMT 2021

Fig. 1 Tree diagram to illustrate the language closeness of major Indian languages

Main Challenge of NLP:
AMBIGUITY

An interesting whatsapp conversation (English and Bengali)

Lady A: Yesterday you told me about shop that sells artificial jewellery

<bn>ki naam jeno?</bn> (what did you say was the name?)

Lady B: nykaa

Lady A (offended): What do you mean Madam? Is this the way to talk?

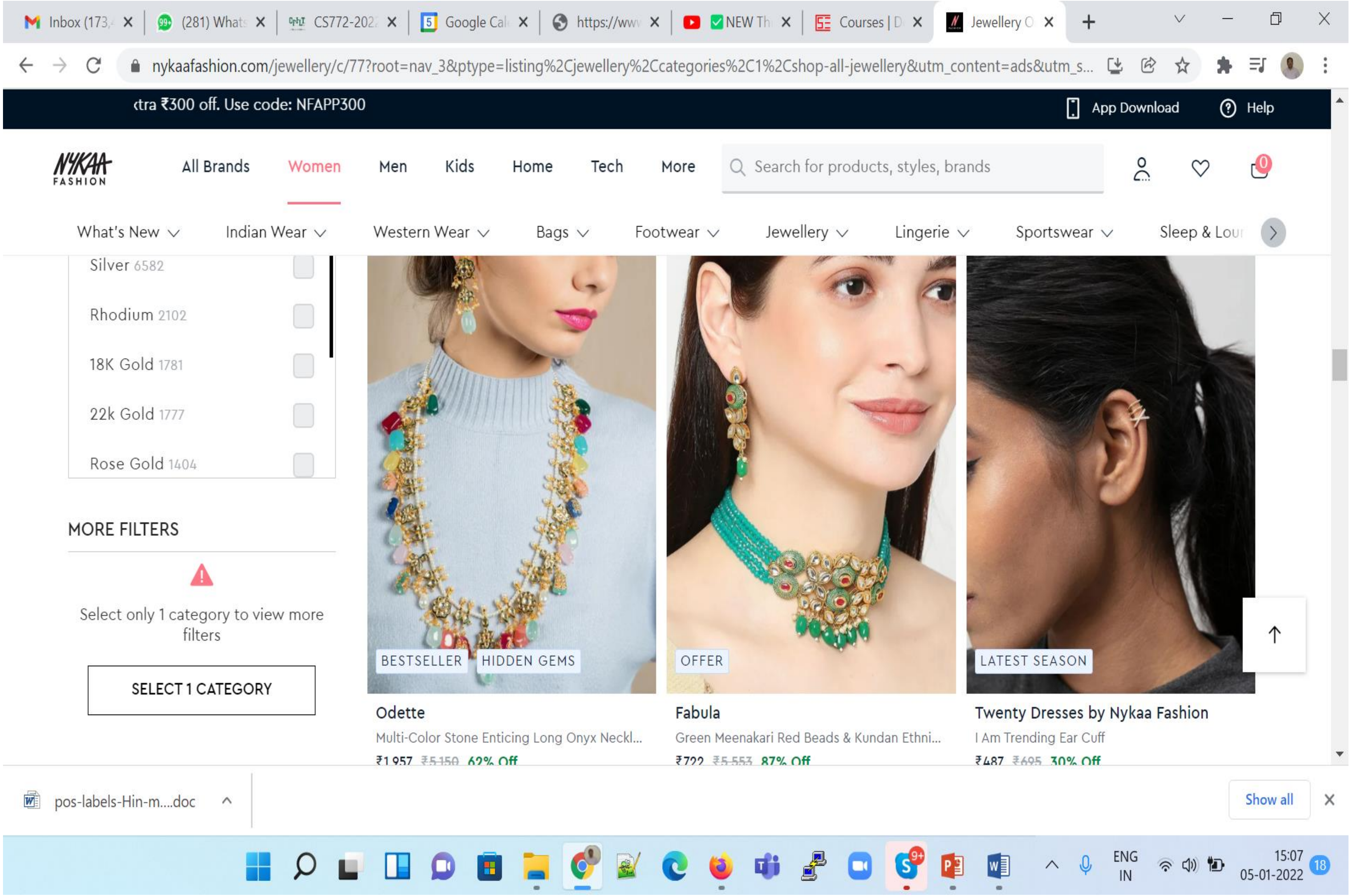
Lady B: <bn> kena ki holo?</bn> (why what happened?)

Lady A did not reply; she was angry!!!

Root cause of the problem: Ambiguity!

- NE-non NE ambiguity (proper noun-common noun)
- Aggravated by code mixing
- “Nykaa”: name of the shop
- Sounds similar to “ন্যাকা” (nyaakaa), meaning somebody “who feigns ignorance/innocence” in a derogatory sense
- An offensive word

NYKAA Fashion

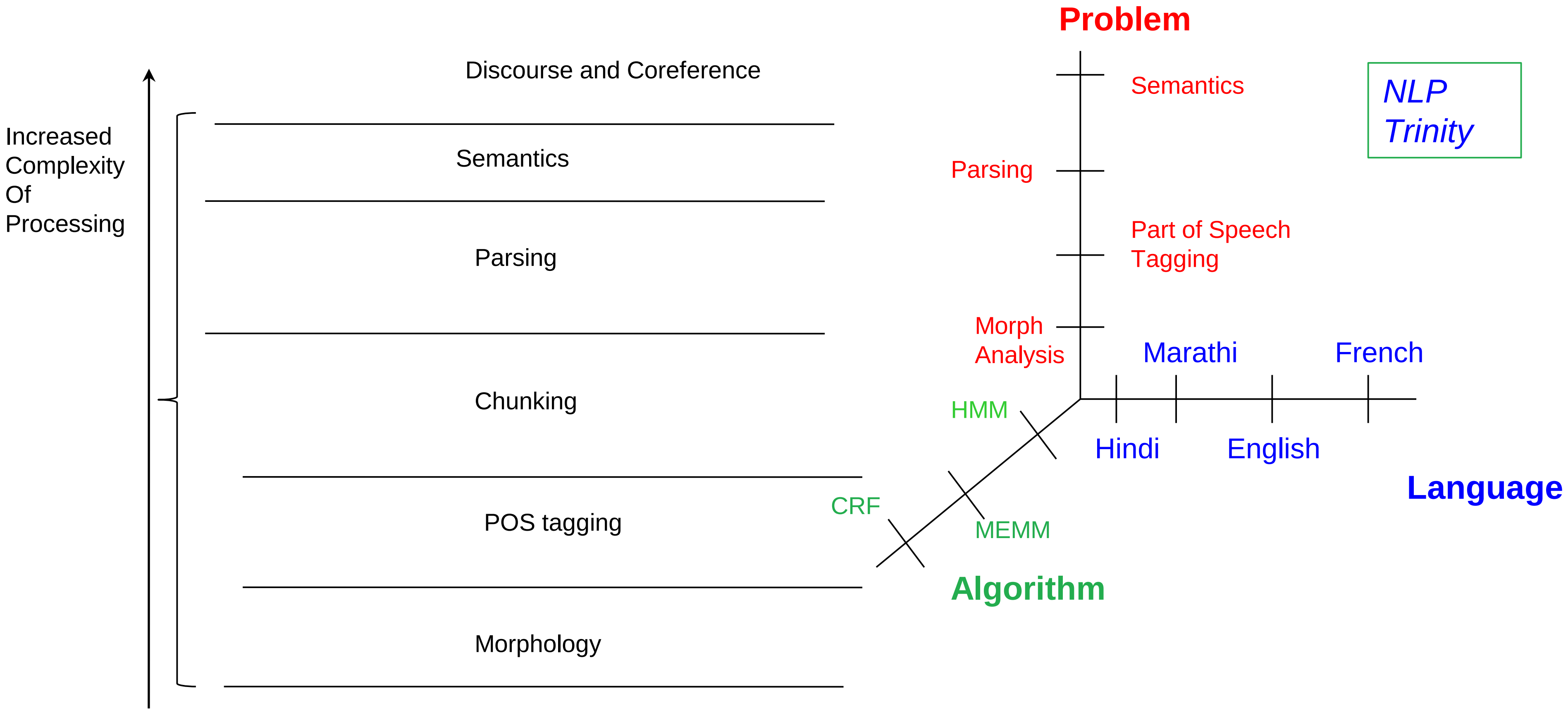


Odette
Multi-Color Stone Enticing Long Onyx Neckl...
₹1 957 ~~₹5 150~~ 62% Off

Fabula
Green Meenakari Red Beads & Kundan Ethni...
₹722 ~~₹5 553~~ 87% Off

Twenty Dresses by Nykaa Fashion
I Am Trending Ear Cuff
₹487 ~~₹695~~ 30% Off

Ambiguity at every layer, for every language, for every mode



Increased Complexity Of Processing

Discourse and Coreference

Semantics

Parsing

Chunking

POS tagging

Morphology

Problem

Semantics

Parsing

Part of Speech Tagging

Morph Analysis

NLP Trinity

Marathi

French

Hindi

English

Language

HMM

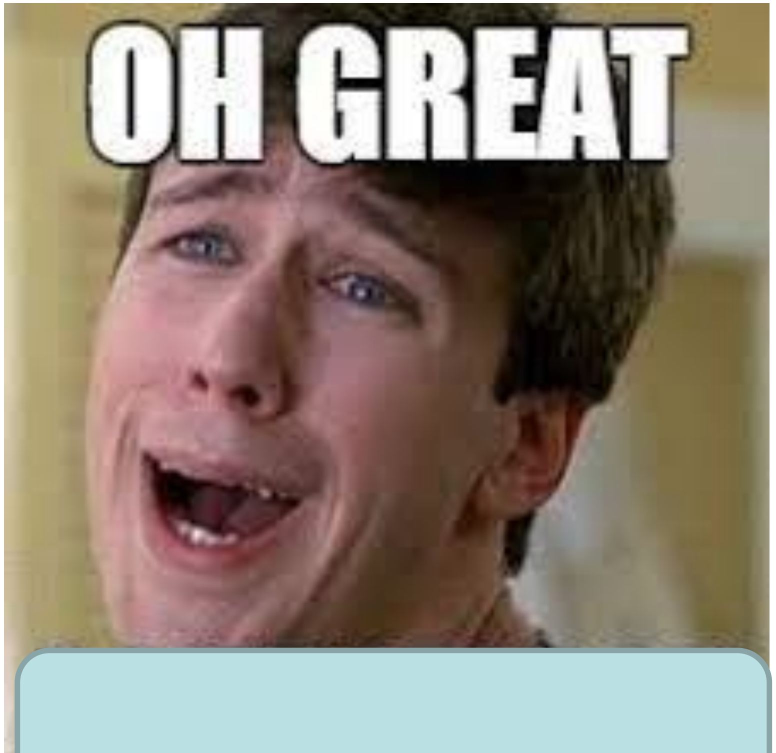
MEMM

CRF

Algorithm

Role of Multimodality

- Signals from other modes
- E.g., Sarcasm



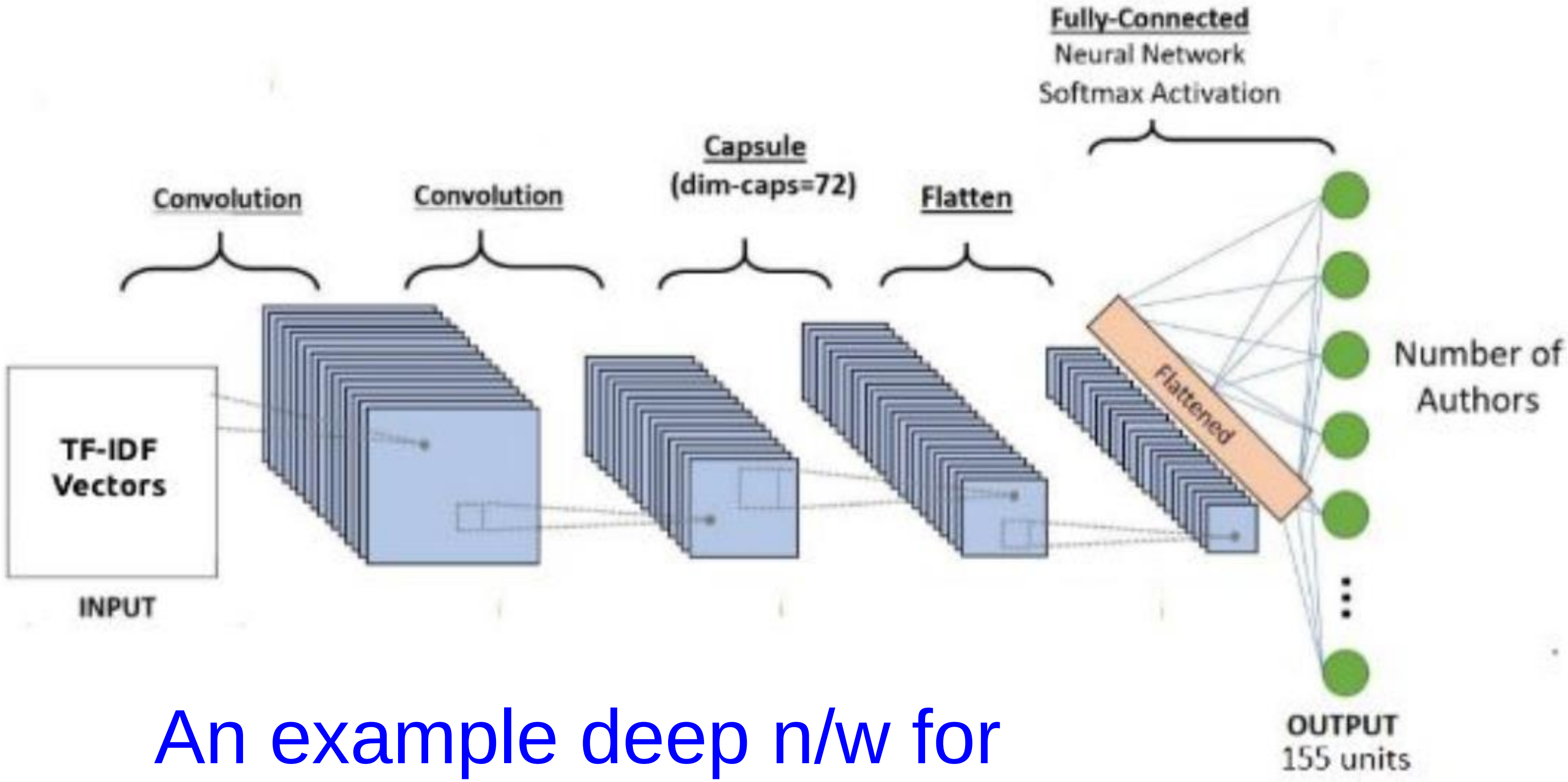
Frequent Observation: Data + Classifier > Human decision maker !!

Case for ML-NLP

LEARN from Data with Probability Based Scoring

- With LOTs of data, learn with
  - High precision (small possibility of error of commission)
  - High recall (small possibility of error of omission)
- But depends on human engineered features, i.e., capturing essential properties

Modern Modus Operandi: End to End DL-NLP



An example deep n/w for author identification

Problem Knowledge and Deep Learning

- Large number of parameter in DL-NLP: Why?
- Fixing large number parameter values need large amounts of data (text for NLP).
- If we know underlying distribution then we can make predictions.

IMP: The number of needed parameters can be reduced by using knowledge.

NLP is Important

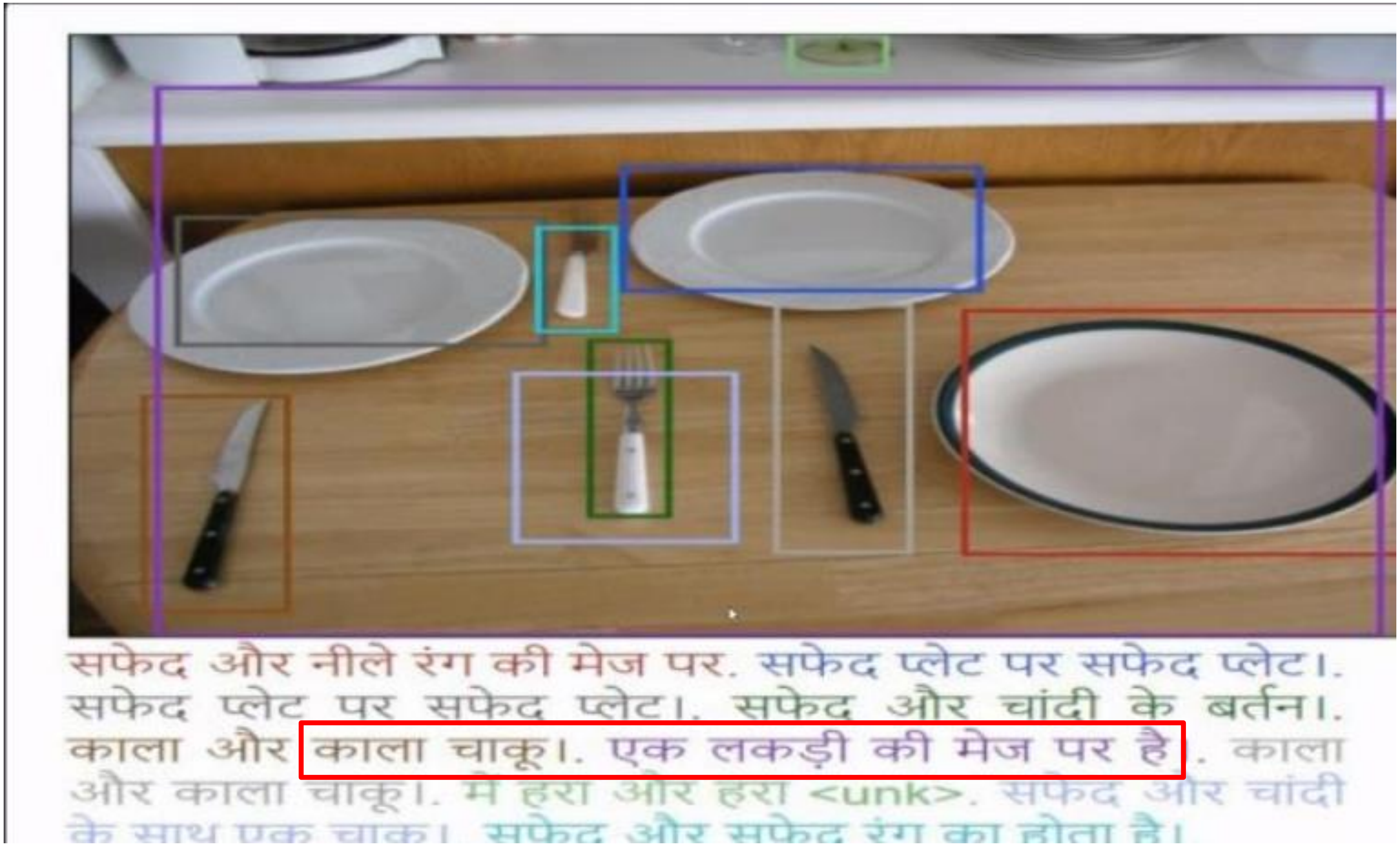
Cutting edge applications

Large Applications to reduce the problem of scale

- (A) Machine Translation (demo)
- (B) Information Extraction
- (C) Sentiment and Emotion Analysis

- Complexity and applicability increases by requirement and introduction of Multilinguality, Multimodality

Dense Image Captioning

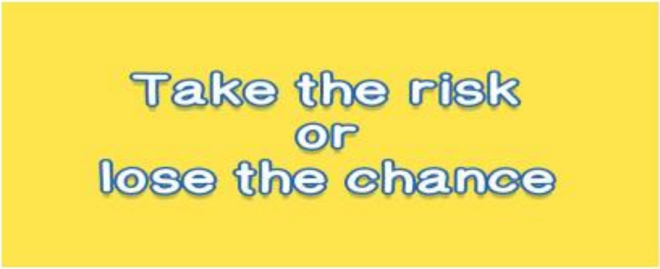


सफेद और नीले रंग की मेज पर. सफेद प्लेट पर सफेद प्लेट।. सफेद प्लेट पर सफेद प्लेट।. सफेद और चांदी के बर्तन।. काला और काला चाकू।. एक लकड़ी की मेज पर है. काला और काला चाकू।. मे हरा और हरा <unk>. सफेद और चांदी के साथ एक चाकू।. सफेद और सफेद रंग का होता है।

OCR-MT-TTS

- Input image:
- English transcription: Take the risk or loose the chance
- Hindi Translation: जोखिम लें या मौका गंवा दें।
- Hindi speech

🔊



Take the risk
or
lose the chance

Demoes

https://chat.openai.com/chat#

https://www.cfilt.iitb.ac.in/ssmt/speech2speech

https://www.cfilt.iitb.ac.in/mtsystem/translate

ML based MT- Czeck-English data

[nesu]	“I carry”
[ponese]	“He will carry”
[nese]	“He carries”
[nesou]	“They carry”
[yedu]	“I drive”
[plavou]	“They swim”

To translate …

- I will carry.
- They drive.
- He swims.
- They will drive.

EM for word alignment from sentence alignment: example

English

(1) three rabbits

a b

(2) rabbits of Grenoble

b c d

French

(1) trois lapins

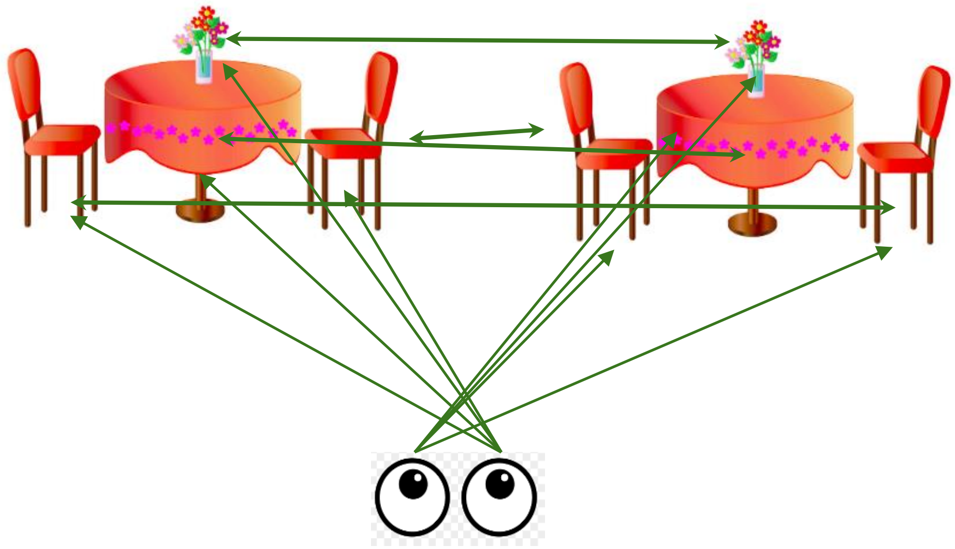
w x

(2) lapins de Grenoble

x y z

How to build part alignment from whole alignment

- Two images are in alignment: images on the two retina
- Need to find alignment of parts of it



Initial Probabilities:
each cell denotes *t(a←→w), t(a←→x) etc.*

	a	b	c	d
w	1/4	1/4	1/4	1/4
x	1/4	1/4	1/4	1/4
y	1/4	1/4	1/4	1/4
z	1/4	1/4	1/4	1/4

Example of expected count

C[w←→a; (a b)←→(w x)]

t(w←→a)

= *------------------------- X #(a in ‘a b’) X #(w in ‘w x’)*

t(w←→a)+t(w←→b)

1/4

= *----------------- X 1 X 1= 1/2*

1/4+1/4

“counts”

a b ←→ ***w x***	a	b	c	d
w	1/2	1/2	0	0
x	1/2	1/2	0	0
y	0	0	0	0
z	0	0	0	0

b c d ←→ ***x y z***	a	b	c	d
w	0	0	0	0
x	0	1/3	1/3	1/3
y	0	1/3	1/3	1/3
z	0	1/3	1/3	1/3

Revised probability: example

$$t_{revised}(a \leftarrow\rightarrow w)$$

$$= \frac{1/2}{(1/2+1/2+0+0)_{(a\ b) \leftarrow\rightarrow (w\ x)} + (0+0+0+0)_{(b\ c\ d) \leftarrow\rightarrow (x\ y\ z)}}$$

Revised probabilities table

	a	b	c	d
w	1/2	1/2	0	0
x	1/4	5/12	1/6	1/6
y	0	1/3	1/3	1/3
z	0	1/3	1/3	1/3

“revised counts”

a b ←→ ***w x***	a	b	c	d
w	1/2	3/8	0	0
x	1/2	5/8	0	0
y	0	0	0	0
z	0	0	0	0

b c d ←→ ***x y z***	a	b	c	d
w	0	0	0	0
x	0	5/9	1/3	1/3
y	0	2/9	1/3	1/3
z	0	2/9	1/3	1/3

Re-Revised probabilities table

	a	b	c	d
w	1/2	1/2	0	0
x	3/16	**85/144**	1/9	1/9
y	0	1/3	1/3	1/3
z	0	1/3	1/3	1/3

Continue until convergence; notice that (b,x) binding gets progressively stronger; b=rabbits, x=lapins

Appendix

Derivation of EM based Alignment Expressions

$$V_E = \text{vocalbulary of language } L_1 \text{ (Say English)}$$

$$V_F = \text{vocabulary of language } L_2 \text{ (Say Hindi)}$$

E^1 *what is in a name ?*

नाम में क्या है ?

F^1 *naam meM kya hai ?*

name in what is ?

E^2 *That which we call rose, by any other name will smell as sweet.*

जिसे हम गुलाब कहते हैं, और भी किसी नाम से उसकी कुशबू समान मीठा होगी

F^2 *Jise hum gulab kahte hai, aur bhi kisi naam se uski khushbu samaan mitha hogii*

That which we rose say , any other name by its smell as sweet

That which we call rose, by any other name will smell as sweet.

Vocabulary mapping

Vocabulary

V_E	V_F
what , is , in, a , name , that, which, we , call ,rose, by, any, other, will, smell, as, sweet	*naam, meM, kya, hai, jise, ham, gulab, kahte, aur, bhi, kisi, bhi, uski, khushbu, saman, mitha, hogii*

Key Notations

English vocabulary : V_E
French vocabulary : V_F
No. of observations / sentence pairs : S
Data D which consists of S observations looks like,

$$e^1{}_1, e^1{}_2, \dots, e^1{}_{l^1} \Leftrightarrow f^1{}_1, f^1{}_2, \dots, f^1{}_{m^1}$$

$$e^2{}_1, e^2{}_2, \dots, e^2{}_{l^2} \Leftrightarrow f^2{}_1, f^2{}_2, \dots, f^2{}_{m^2}$$

.....

$$e^s{}_1, e^s{}_2, \dots, e^s{}_{l^s} \Leftrightarrow f^s{}_1, f^s{}_2, \dots, f^s{}_{m^s}$$

.....

$$e^S{}_1, e^S{}_2, \dots, e^S{}_{l^S} \Leftrightarrow f^S{}_1, f^S{}_2, \dots, f^S{}_{m^S}$$

No. words on English side in s^{th} sentence : l^s
No. words on French side in s^{th} sentence : m^s
$index_E(e^s{}_p)$ =Index of English word $e^s{}_p$in English vocabulary/dictionary
$index_F(f^s{}_q)$ =Index of French word $f^s{}_q$in French vocabulary/dictionary

(Thanks to Sachin Pawar for helping with the maths formulae processing)

Hidden variables and parameters

Hidden Variables (Z) :

Total no. of hidden variables = $\sum_{s=1}^{S} l^s\, m^s$ where each hidden variable is as follows:

$z_{pq}^{s} = 1$, if in s^{th} sentence, p^{th} English word is mapped to q^{th} French word.

$z_{pq}^{s} = 0$, otherwise

Parameters (*Θ*) :

Total no. of parameters = $|\, V_E| \times |V_F|$, where each parameter is as follows:

$P_{i,j}$ = Probability that i^{th} word in English vocabulary is mapped to j^{th} word in French vocabulary

Likelihoods

Data Likelihood *L(D; Θ)* :

$$L(D;\Theta) = \prod_{s=1}^{S}\prod_{p=1}^{l^s}\prod_{q=1}^{m^s}\left(P_{index_E(e_p^s),index_F(f_q^s)}\right)^{z_{pq}^s}$$

Data Log-Likelihood LL(D; Θ) :

$$LL(D;\Theta) = \sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s} z_{pq}^s \log\left(P_{index_E(e_p^s),index_F(f_q^s)}\right)$$

Expected value of Data Log-Likelihood E(LL(D; Θ)) :

$$E(LL(D;\Theta)) = \sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s} E(z_{pq}^s)\,\log\left(P_{index_E(e_p^s),index_F(f_q^s)}\right)$$

Constraint and Lagrangian

$$\sum_{j=1}^{|V_F|} P_{i,j} = 1 \ , \forall i$$

$$\sum_{s=1}^{S} \sum_{p=1}^{l^s} \sum_{q=1}^{m^s} E(z_{pq}^s) \, log\left(P_{index_E(e_p^s), index_F(f_q^s)}\right) - \sum_{i=1}^{|V_E|} \lambda_i \left(\sum_{j=1}^{|V_F|} P_{i,j} - 1\right)$$

Differentiating wrt P_{ij}

$$\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}\left(\frac{E(z_{pq}^s)}{P_{i,j}}\right)-\lambda_i=0$$

$$P_{i,j}=\frac{1}{\lambda_i}\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}\,E(z_{pq}^s)$$

$$\sum_{j=1}^{|V_F|}P_{i,j}=1=\sum_{j=1}^{|V_F|}\frac{1}{\lambda_i}\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}\,E(z_{pq}^s)$$

Final E and M steps

M-step

$$P_{i,j} = \frac{\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}E(z_{pq}^s)}{\sum_{j=1}^{|V_F|}\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}E(z_{pq}^s)}, \forall i,j$$

E-step

$$E(z_{pq}^s) = \frac{P_{index_E(e_p^s),index_F(f_q^s)}}{\sum_{q'=1}^{m^s}P_{index_E(e_p^s),index_F(f_{q'}^s)}}, \forall s,p,q$$

Mathematics of EM

From

Pushpak Bhattacharyya, *Machine Translation*, CRC Press, 2015

Maximum Likelihood of Observations

- ***Situation 1: Throw of a Single Coin***
- The parameter is the probability *p* of getting heads in a single toss. Let *N* be the number of tosses. Then the observation *X* and the data or observation likelihood *D* respectively are:

$$X :< x_1, x_2, x_3, \ldots, x_{N-1}, x_N >$$

$$D = \prod_{i=1}^{N} p^{x_i} (1-p)^{1-x_i}, \text{s.t. } \mathrm{x_i} = 1 \text{ or } 0, \text{ and } 0 \le \mathrm{p} \le 1$$

where x_i is an indicator variable assuming values 1 or 0 depending on the *ith* observation being heads or tail. Since there are *N* identically and independently distributed (*i.i.d.*) observations, *D* is the product of probabilities of individual observations each of which is a Bernoulli trial.

Single coin

Since exponents are difficult to manipulate mathematically, we take log of *D*, also called log likelihood of data, and maximize with regard to *p*. This yields

$$p = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{M}{N};\ M = \#Heads, N = \#tosses$$

Throw of 2 coins

- Three parameters: probabilities p_1 and p_2 of heads of the two coins and the probability p of choosing the first coin (automatically, $1\text{-}p$ is the probability of choosing the second coin).
- N tosses and observations of heads and tails. Only, we do not know which observation comes from which coin.
- Indicator variable z_i is introduced to capture coin choice (z_i=1 if coin 1 is chosen, else 0). This variable is hidden, *i.e.*, we do not know its values.
- However, without it the likelihood expression would have been very cumbersome.

Data Likelihood

Data Likelihood,

$D \quad = P_{<p1,p2,p>}(X) = P_\theta(X),\ \theta = <p, p_1, p_2>$

$\quad\ = \Sigma_Z P_\theta(X,Z))$

$$X :< x_1, x_2, x_3, \ldots, x_{N-1}, x_N >$$

$$Z :< z_1, z_2, z_3, \ldots, z_{N-1}, z_N >$$

$$P_\theta(X,Z) = \prod_{i=1}^{N} \left[\left(p p_1^{x_i} (1 - p_1)^{1-x_i} \right)^{z_i} \left((1-p) p_2^{x_i} (1 - p_2)^{1-x_i} \right)^{1-z_i} \right],$$

$$\text{s.t. } z_i, x_i = 1 \text{ or } 0, \text{ and } 0 \le p, p_1, p_2 \le 1$$

Invoke Jensen Inequality

We would like to work with *logP_θ(X).* However, there will be a Σ inside *log*. Fortunately, *log* is a concave function, so that

$$\log\left(\sum_{i=1}^{K} \lambda_i y_i\right) \geq \left(\sum_{i=1}^{K} \lambda_i \log(y_i)\right); \sum_{i=1}^{K} \lambda_i = 1$$

Log likelihood of Data

$LL(D)$= log likelihood of data

$= log(P_\theta(X))= log(\Sigma_Z P_\theta(X,Z))$

$= log[\Sigma_Z \lambda_Z (P_\theta(X,Z)/\lambda_Z)]; \Sigma_Z \lambda_Z=1$

$>= \Sigma_Z[\lambda_Z log[(P_\theta(X,Z)/\lambda_Z)]$

After a number of intricate mathematical steps

$LL(D) >=E_{Z|X,\theta} log(P_\theta(X,Z))$, where E(.) is the expectation function; note that the expectation is conditional on X.

Expectation of log likelihood

$$E_{Z|X}[\log(P_\theta(X,Z)]$$

$$= E_{Z|X}\left[\log\prod_{i=1}^{N}\left[\left(pp_1{}^{x_i}\left(1-p_1\right)^{1-x_i}\right)^{z_i}\left(\left(1-p\right)p_2{}^{x_i}\left(1-p_2\right)^{1-x_i}\right)^{1-z_i}\right],\right]$$

$$= E_{Z|X}\left[\begin{array}{l}\displaystyle\sum_{i=}^{N} z_i\left(\log p + x_i \log p_1 + (1-x_i)\log\left(1-p_1\right)\ \right) \ + \\ (1-z_i)\left(\log(1-p) + x_i \log p_2 + (1-x_i)\log\left(1-p_2\right)\ \right)\end{array}\right]$$

$$= \sum_{i=1}^{N}\left[\begin{array}{l}E(z_i \mid x_i)\left(\log p + x_i \log p_1 + (1-x_i)\log\left(1-p_1\right)\ \right) \ + \\ (1-E(z_i \mid x_i))\left(\log(1-p) + x_i \log p_2 + (1-x_i)\log\left(1-p_2\right)\ \right)\end{array}\right]$$

$$\text{s.t. } \mathrm{z_i}, \mathrm{x_i} = 1 \text{ or } 0, \text{ and } 0 \le \mathrm{p}, \mathrm{p_1}, \mathrm{p_2} \le 1$$

Derivation of E and M steps for 2 coin problem (1/2)- M step

Take partial derivative of $E_{Z|X,\theta}(.)$ (prev. slide) wrt p, p_1, p_2 and equate to 0.

$$p = \frac{\sum_{i=1}^{N} E(z_i \mid x_i)}{N}$$

$$p_1 = \frac{\sum_{i=1}^{N} E(z_i \mid x_i) x_i}{\sum_{i=1}^{N} E(z_i \mid x_i)}$$

$$p_2 = \frac{M - \sum_{i=1}^{N} E(z_i \mid x_i) x_i}{N - \sum_{i=1}^{N} E(z_i \mid x_i)} ; M = \# Heads, N = \# tosses$$

Derivation of E and M steps for 2 coin problem (2/2)- E step

$E(z_i|x_i) = 1.P(z_i=1|x_i) + 0.P(z_i=0|x_i)$

$= P(z_i=1|x_i)$

$$P(z_i = 1 | x_i) = \frac{P(z_i = 1, x_i)}{P(x_i)}$$

$$= \frac{pp_1^{x_i}(1-p_1)^{1-x_i}}{P(x_i, z_i = 1) + P(x_i, z_i = 0)}$$

$$= \frac{pp_1^{x_i}(1-p_1)^{1-x_i}}{pp_1^{x_i}(1-p_1)^{1-x_i} + (1-p)p_2^{x_i}(1-p_2)^{1-x_i}}$$

Generalization into N “throws” using M “things” each having L outcomes

From

Pushpak Bhattacharyya, *Machine Translation*, CRC Press, 2015

Multiple outcomes from multiple entities

- “Throw” of “something” where that something has more than 2 outcomes, e.g., throw of multiple dice
- The observation sequence has a sequence of 1 to 6s
- But we do not know which observation came from which dice
- Gives rise to a multinomial that is extremely useful in NLP ML.

Observation Sequence

- N ‘throws’, 1 of L outcomes from each throw, 1 of the M ‘things’ (called ‘sources’) chosen
- $\Sigma_{k=1,L} x_{ik}=1$, since each x_{ik} is either 1 or 0 and one and only one of them is1.
- D (data):

$$<x_{11}/x_{12}/\ldots x_{1L}>, <x_{21}/x_{22}/\ldots x_{2L}>, \ldots <x_{N1}/x_{N2}/\ldots x_{NL}>$$

Hidden Variable

- Hidden variable for M sources
- $\Sigma_{j=1,M} z_{ij}=1$, since each z_{ij} is either 1 or 0 and one and only one of them is 1.
- Z:

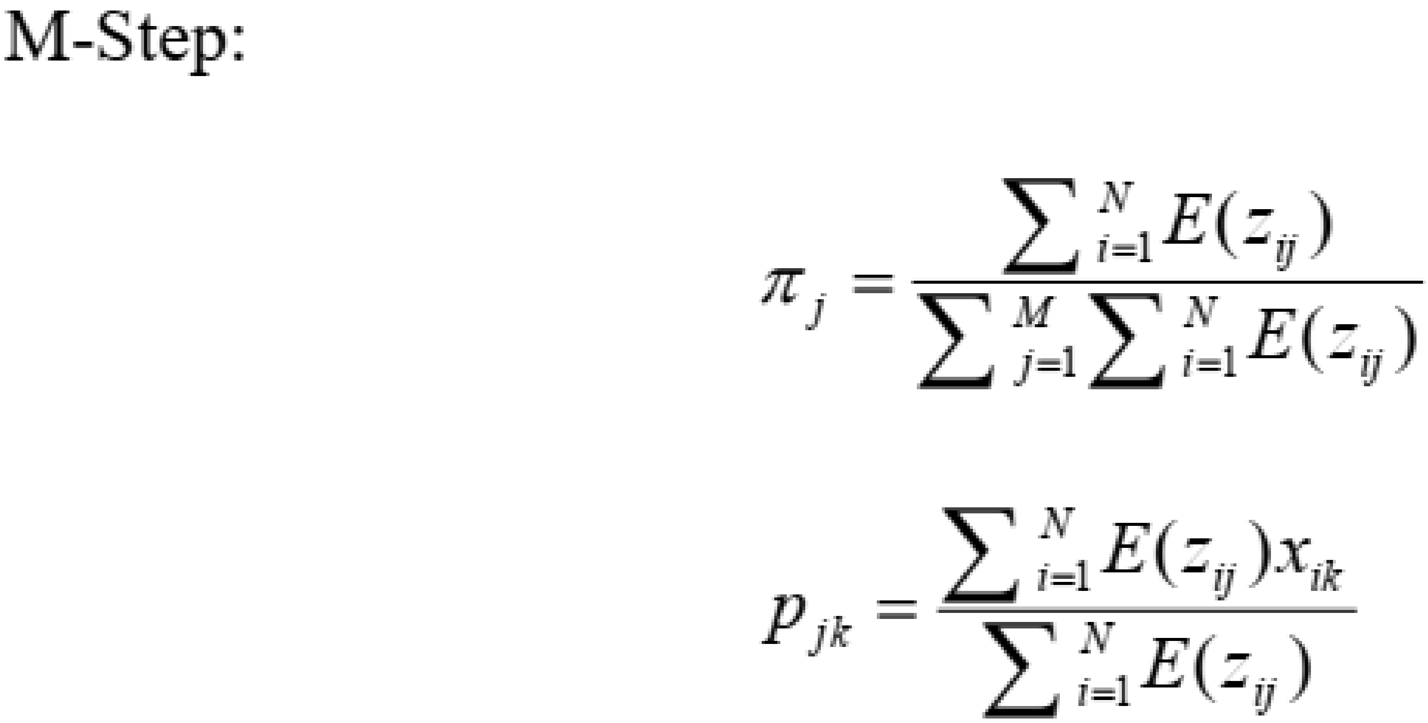
  $<z_{11}/z_{12}/\ldots z_{1M}>, <z_{21}/z_{22}/\ldots z_{2M}>, \ldots$
  $<z_{N1}/z_{N2}/\ldots z_{NM}>$

Parameters

- Parameter set θ:
  - π_j: probability of choosing source j
  - p_{jk}: probability of observing k^{th} outcome from the j^{th} source

This will be elaborated next week; only expressions are given now

M-step

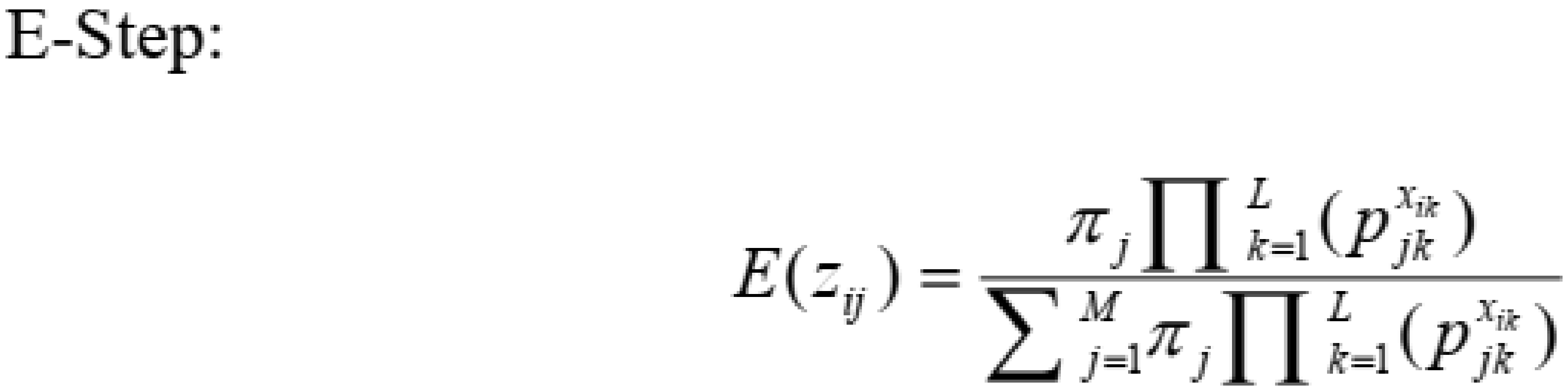


M-Step:

$$\pi_j = \frac{\sum_{i=1}^{N} E(z_{ij})}{\sum_{j=1}^{M} \sum_{i=1}^{N} E(z_{ij})}$$

$$p_{jk} = \frac{\sum_{i=1}^{N} E(z_{ij}) x_{ik}}{\sum_{i=1}^{N} E(z_{ij})}$$

E-step



E-Step:

$$E(z_{ij}) = \frac{\pi_j \prod_{k=1}^{L} (p_{jk}^{x_{ik}})}{\sum_{j=1}^{M} \pi_j \prod_{k=1}^{L} (p_{jk}^{x_{ik}})}$$