CS626: Speech, Natural Language Processing and the Web

Introduction and POS Tagging

Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay

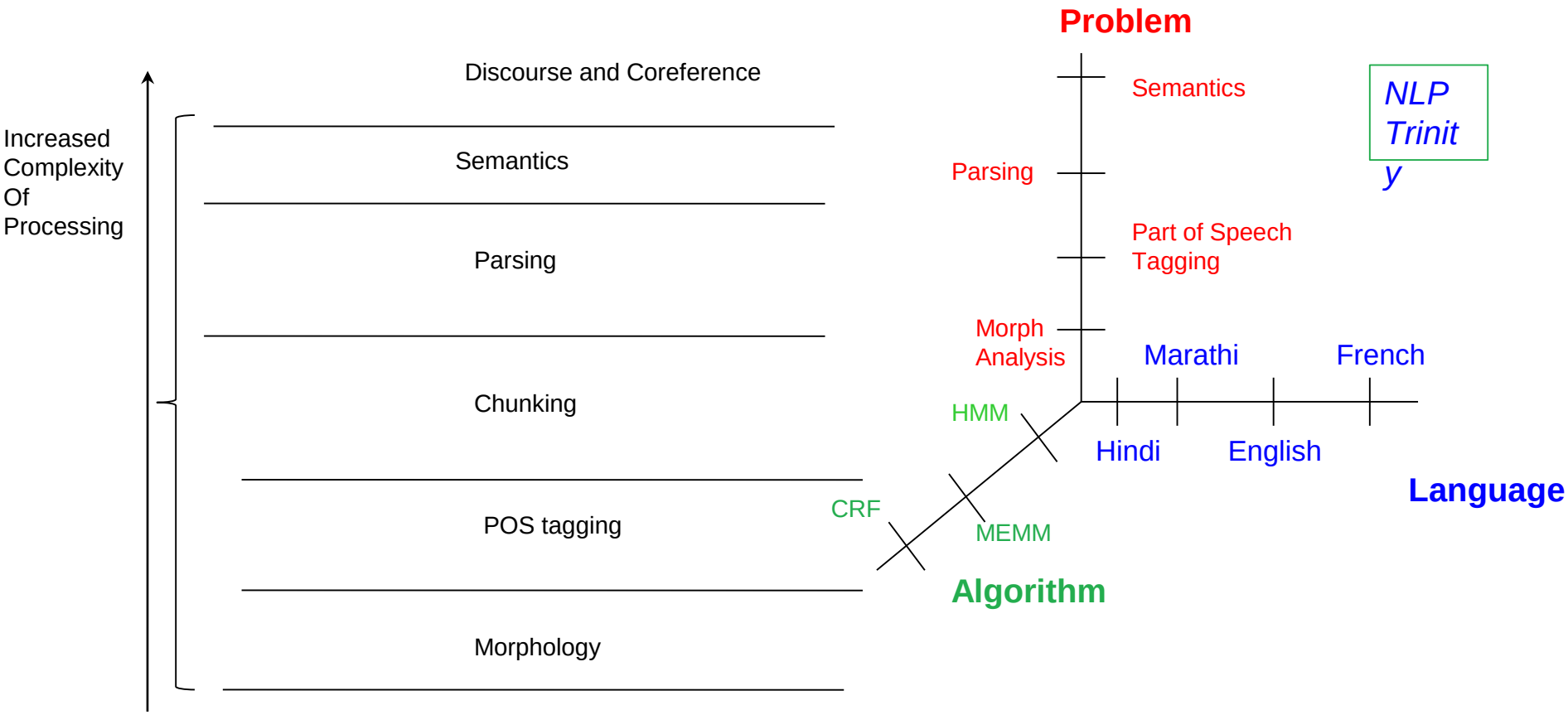
Week 1 of 26th July, 2021

Nature of NLP

Natural Language Processing

Art, science and technique of making computers understand the generate language

NLP is layered Processing, Multidimensional too



Main Challenge: **AMBIGUITY**

Examples (1/2)

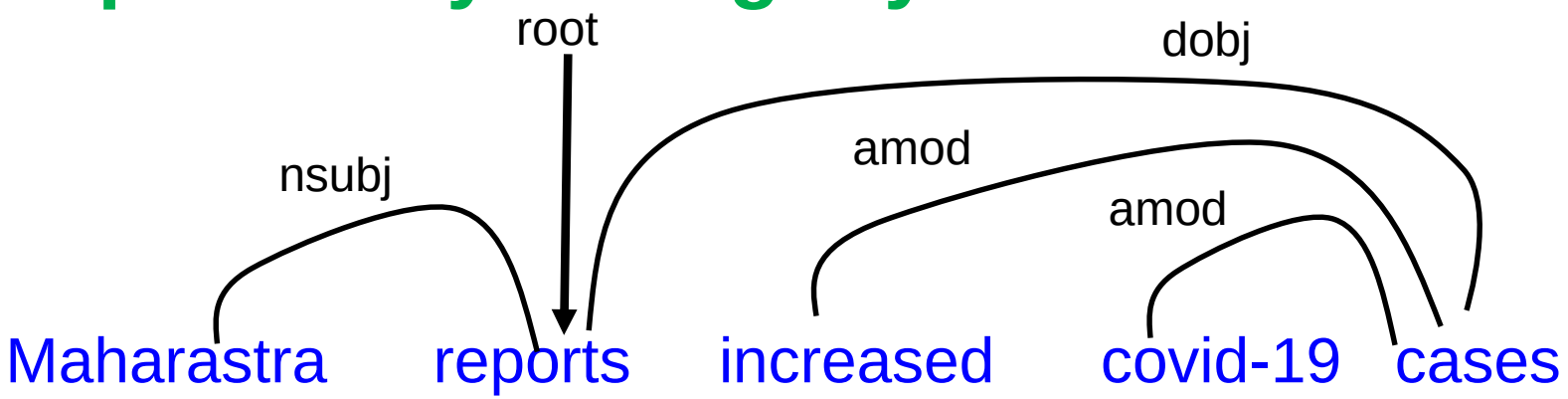
- Covid-19 Origin Probe should Shift to **US**, says Chinese Epidemiologist (Economic Times 18 June 21)
- **Command** Center to Track Best Buses (ToI 30Jan21)
- Why is the teddy bear never hungry; because it is always **stuffed** (ToI Oct 20)

Examples (2/2)

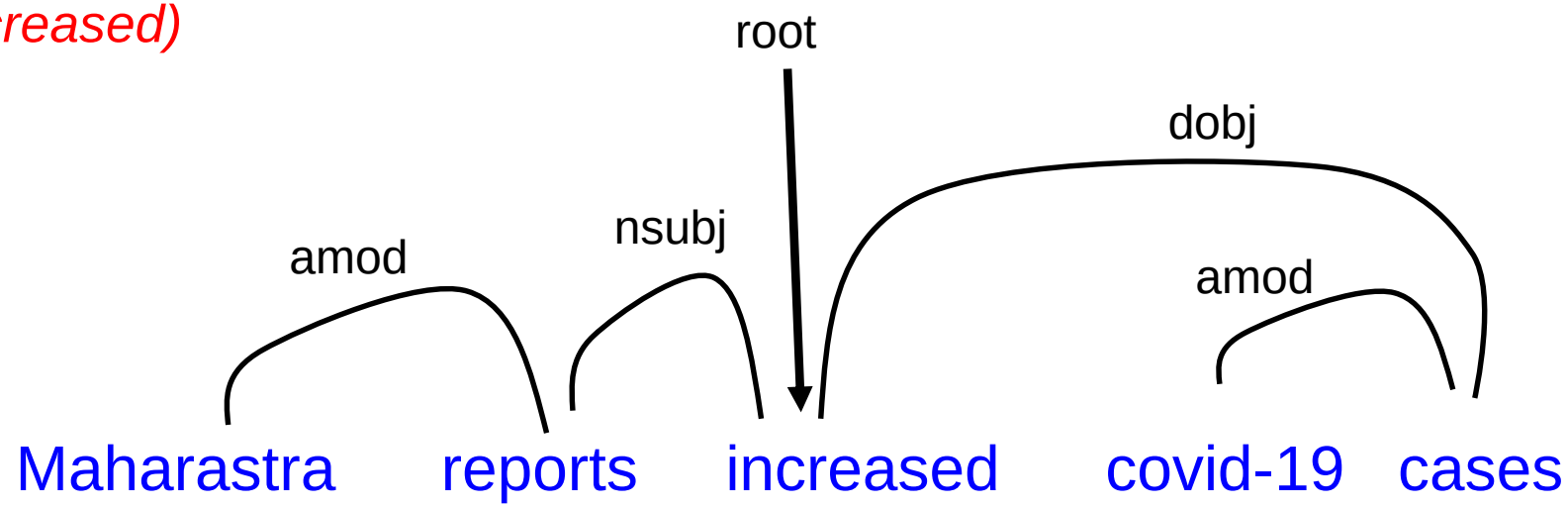
- A truly international collection the book makes **available** stories no longer easily accessible (Book titled "Great Stories by Nobel Prize Winners", Noonday Press, 1959, back cover)

- Elderly with young **face** increased covid 19 risk (ToI Oct 20)

Dependency Ambiguity



(it is reported by Maharastra Govt. that covid-19 cases have increased)



(it is the Maharastra reports that have increased covid-19 cases!!!)

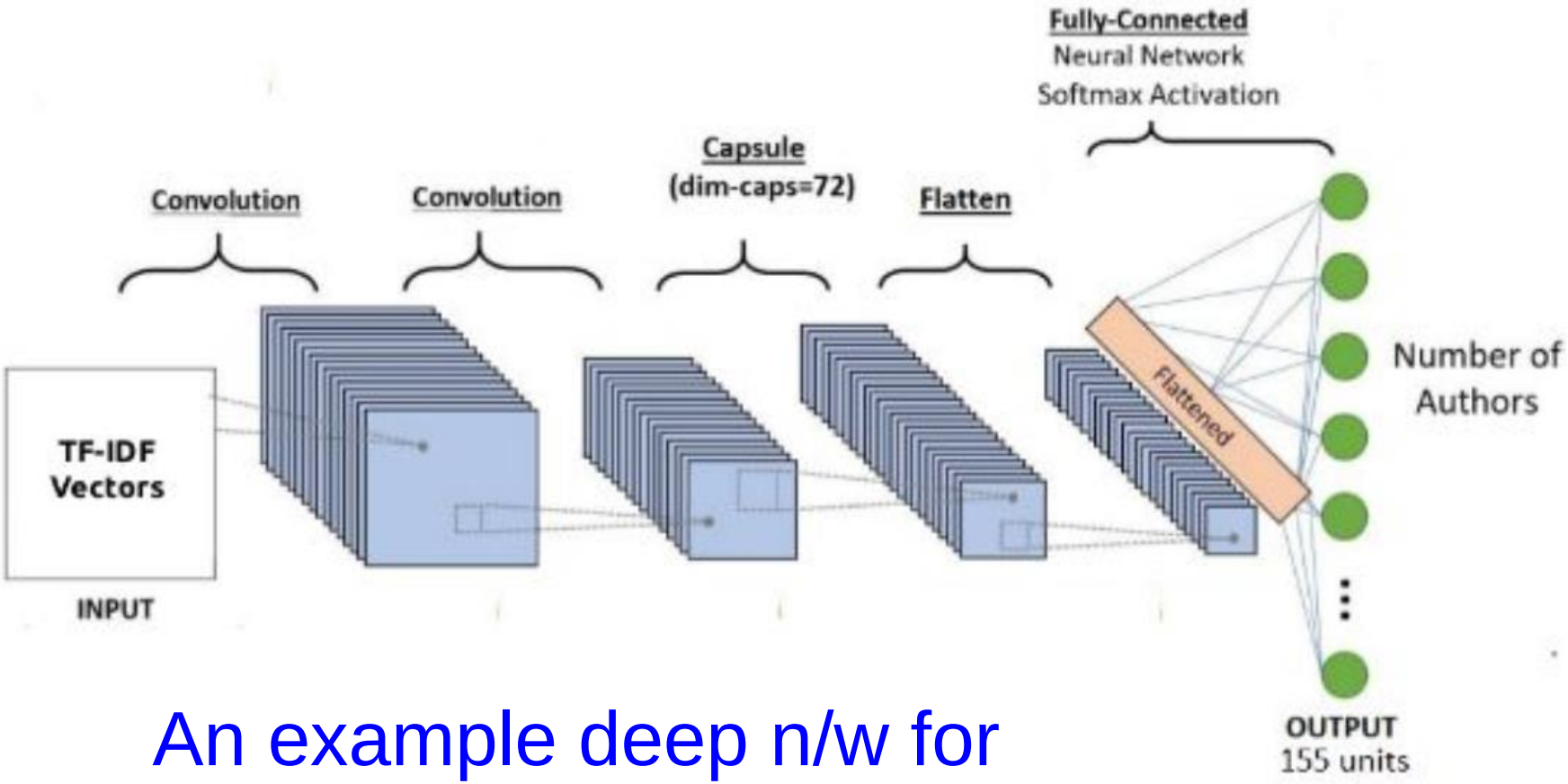
Data + Classifier > Human decision maker !!

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

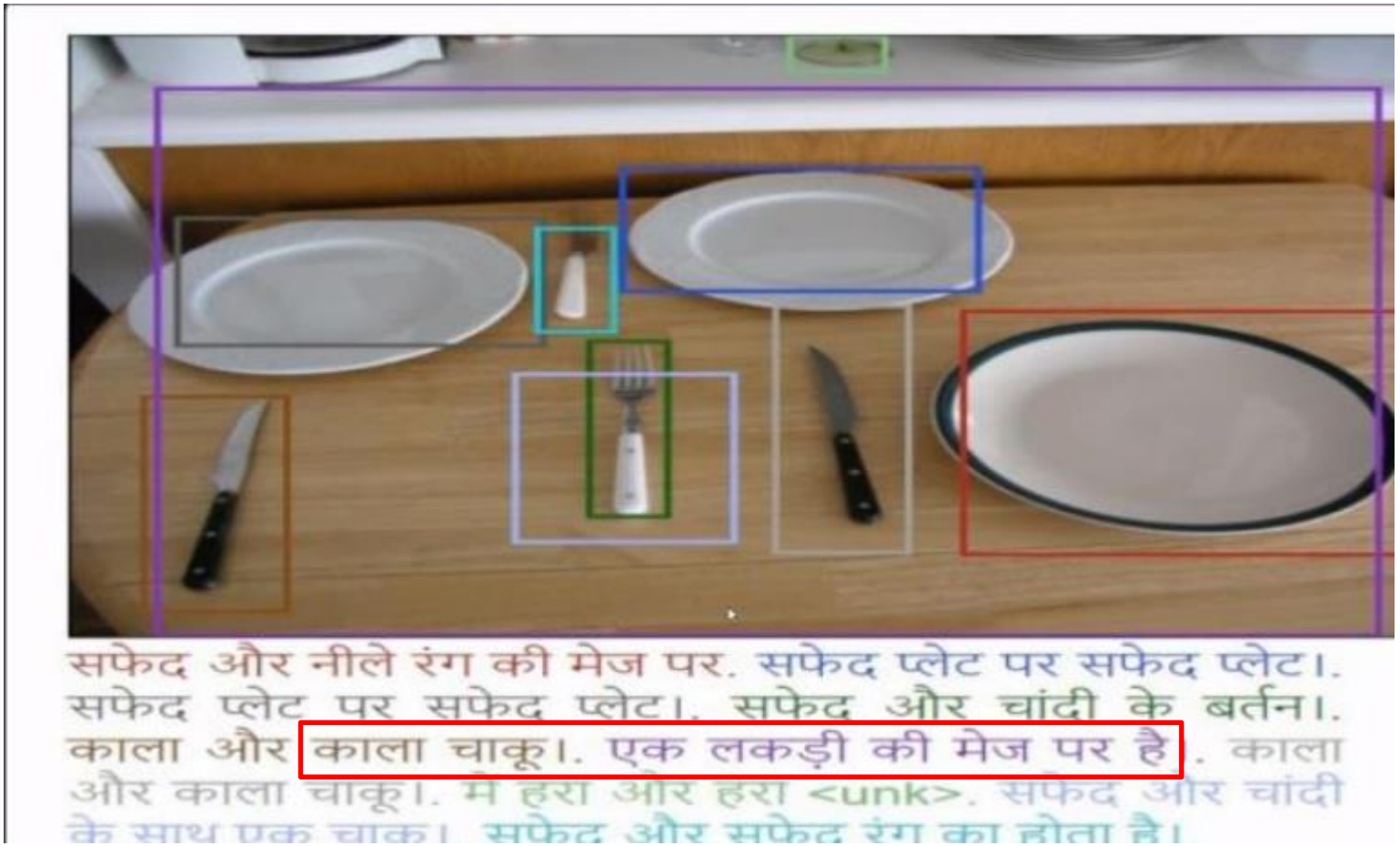
Cutting edge applications

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



Multimodality and Sarcasm

Pax: ***thank you for sending me to Delhi and my luggage to Mumbai ! Brilliant service!!!***

Chatbot: ***Thanks for the appreciation*** 🤦

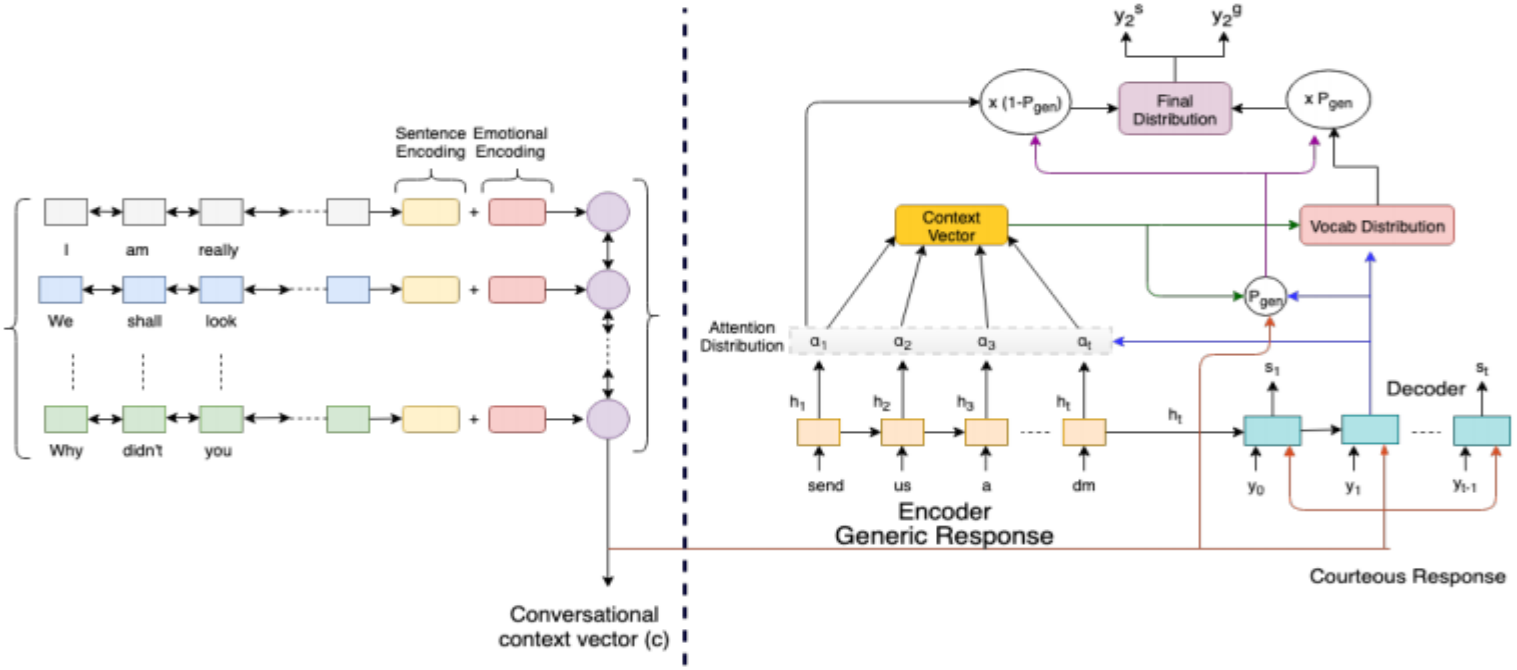
sarcasm

Automatic Politeness Generation

Conversational History	Generic Response	Polite Response
Somebody please help me, I'm having the worst luck with your customer service	How can we help?	Help has arrived! We are sorry to see that you are having trouble, how can we help?
मेरा फोन एक मिनट पहले पूरी तरह से काम कर रहा था और फिर अचानक काम करना बंद कर दिया। (My phone was perfectly working a minute back and then it stopped working suddenly.)	आपको क्या परेशानी हो रही है? (What trouble are you having?)	हम आपके साथ इस पर गौर करेंगे, कृपया हमें बताएं कि आपको क्या परेशानी हो रही है? (We'll look into this with you, please tell us what trouble are you having?)
Is the Radio X radio within the app branded station, or just a station called Radio X?	Can you send us a screenshot of what you're seeing?	No worries at all! Help's here! Please send us a screenshot of what you're seeing? We'd like to take a close look.
पिछले हफ्ते ही घर आया था और मुझे फिर से अपने इंटरनेट से समस्या हो रही है। (Just came home last week and I'm having problems with my internet again.)	आपके इंटरनेट के साथ क्या हो रहा है? (What is happening with your internet?)	अरे नहीं यह अच्छा नहीं है। मैं मदद कर सकता हूँ! कृपया बताएं कि आपके इंटरनेट में क्या समस्या है? (Oh no that's not good. I can help! Please tell whats the problem with your internet?)

Mauajama Firdaus, Asif Ekbal, Pushpak Bhattacharyya; *Incorporating Politeness across Languages in Customer Care Responses: Towards building a Multilingual Empathetic Dialogue Agent*. **LREC 2020**, Marseille, France; 2020.

Polite Response Generation: copy and insert “Polite” Phrase



Inputs to the model: Conversation History (left), Generic Response (centre) Output: Courteous Response (right). The Conversation History is encoded by hierarchical Bi-LSTM to a Conversational Context vector *c*. The encoder encodes the Generic Response into hidden states h_i . Response tokens are decoded one at a time. Attention α_i, and vocabulary distributions (p_{vocab}) are computed, and combined using p_{gen} to produce output distribution. Sampling it yields y^s_i and taking its argmax yields y^g_i

OCR-MT-TTS

- Input image:

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

Course: Basic Info

- Slot 1: Monday 8.30, Tuesday 9.30 and Thursday 10.30
- TA Team: Jyotsna Khatri, Apoorva Nunna, Niteesh Mallela, Kunal Verma
- http://www.cfilt.iitb.ac.in/~cs626-2021
- Channels of communication: MS Teams, Moodle, Course Website

Evaluation Scheme (tentative)

- 50%: Reading, Thinking, Comprehending
  - Quizes (15)
  - Midsem (15)
  - Endsem (20)
- 50%: Doing things, Hands on
  - Assignments (25%)
  - Project (25%)

Course Content: Task *vs.* Technique Matrix

Task (row) vs. Technique (col) Matrix	**Rules Based/Knowledge-Based**	**Classical ML**				**Deep Learning**		
		Perceptron	Logistic Regression	SVM	Graphical Models (HMM, MEMM, CRF)	Dense FF with BP and softmax	RNN-LSTM	CNN
Morphology								
POS								
Chunking								
Parsing								
NER, MWE								
Coref								
WSD								
Machine Translation								
Semantic Role Labeling								
Sentiment								
Question Answering								

Books

- 1. Dan Jurafsky and James Martin, Speech and Language Processing, 3 rd Edition, 2019.
- 2. Ian Goodfellow, Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press, 2016.
- 3. Pushpak Bhattacharyya, Machine Translation, CRC Press, 2017

Books (2/2)

- 4. Christopher Manning and Heinrich Schutze, Foundations of Statistical NaturalLanguage Processing, MIT Press, 1999.

- 5. Pushpak Bhattacharyya, Machine Translation, CRC Press, 2017.

Journals and Conferences

- Journals: Computational Linguistics, Natural Language Engineering, Journal of Machine Learning Research (JMLR), Neural Computation, IEEE Transactions on Neural Networks

- Conferences: ACL, NeuriPS), ICML, EMNLP, NAACL, EACL, AACL

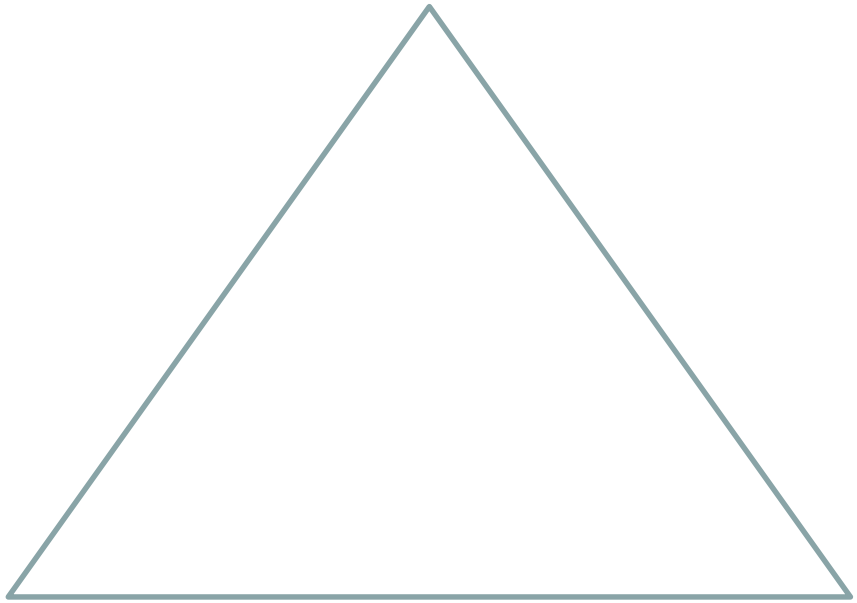
Useful NLP, ML, DL libraries

- NLTK
- Scikit-Learn
- Pytorch
- Tensorflow (Keras)
- Huggingface
- Spacy
- Stanford Core NLP

Part of Speech Tagging

The Trinity

Linguistics



Probability

Coding

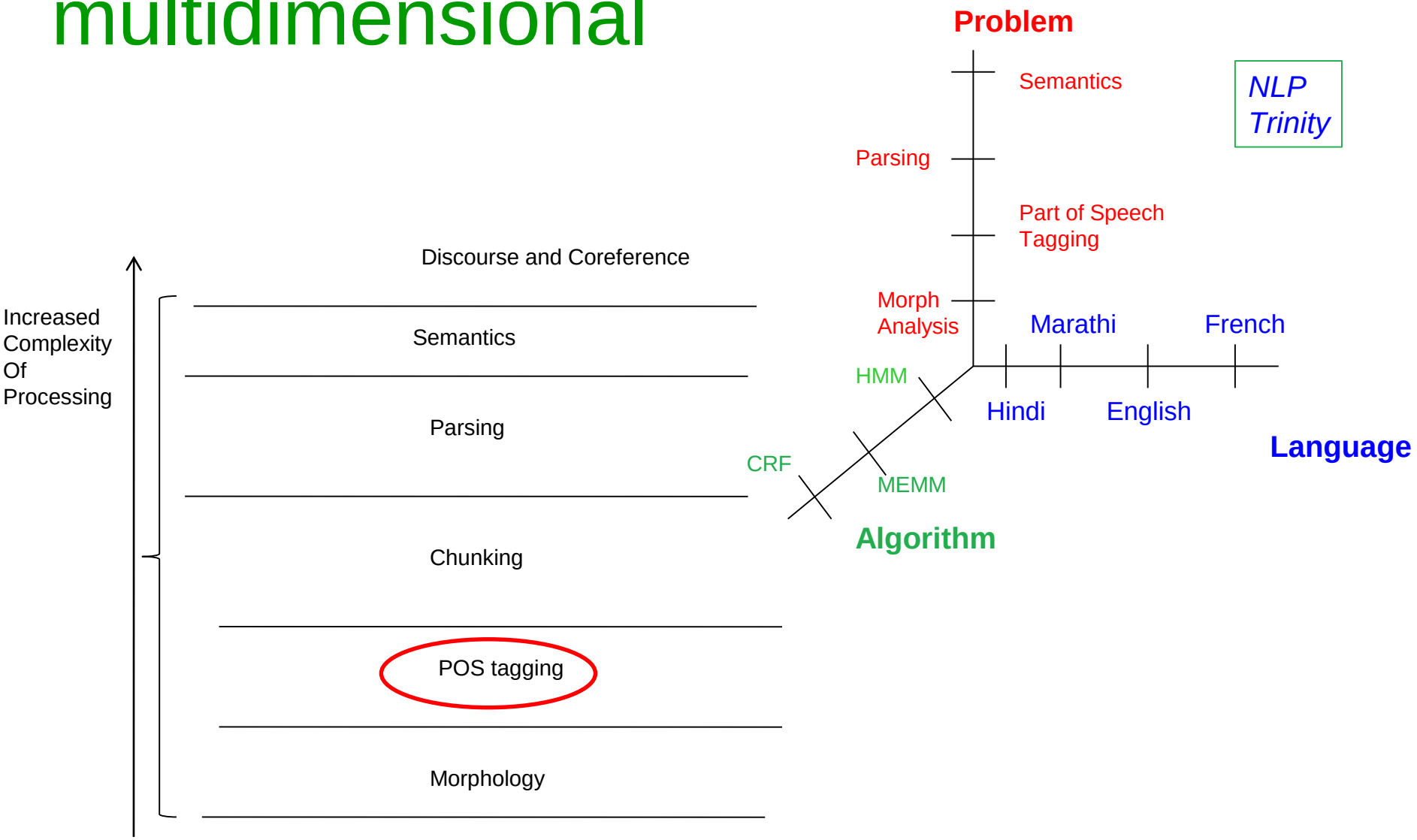
Task *vs.* Technique Matrix

Task (row) vs. Technique (col) Matrix	Rules Based/Knowledge-Based	Classical ML				Deep Learning		
		Perceptron	Logistic Regression	SVM	Graphical Models (HMM, MEMM, CRF)	Dense FF with BP and softmax	RNN-LSTM	CNN
Morphology								
POS								
Chunking								
Parsing								
NER, MWE								
Coref								
WSD								
Machine Translation								
Semantic Role Labeling								
Sentiment								
Question Answering								

Agenda

- Rule Based POS Tagging
- Rule based NLP is also called Model Driven NLP
- Statistical ML based POS Tagging (*Hidden Markov Model, Support Vector Machine*)
- Neural (Deep Learning) based POS Tagging

NLP: multilayered, multidimensional



Useful NLP, ML, DL libraries

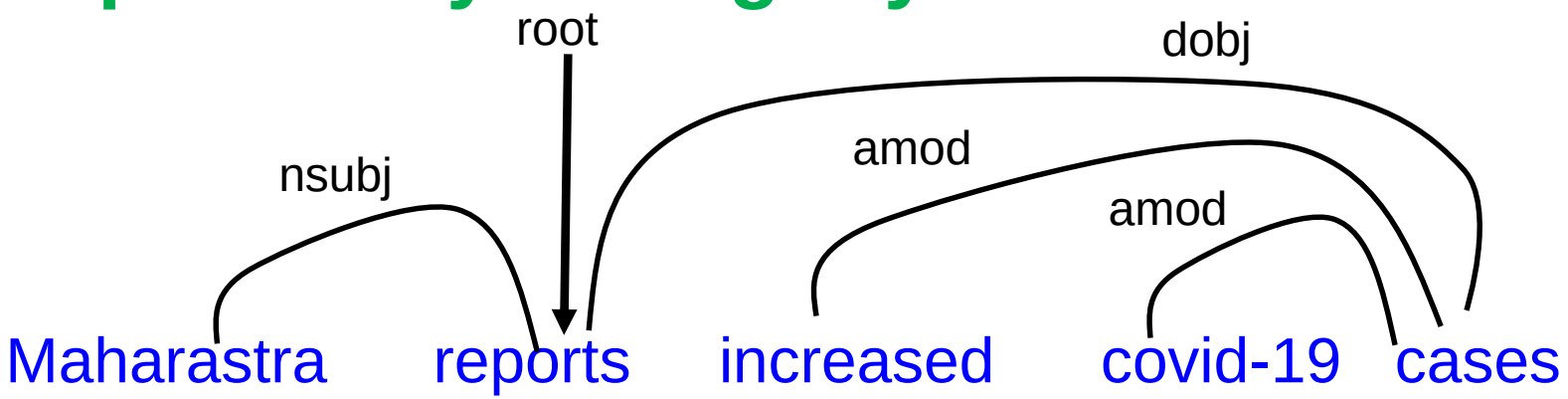
- NLTK
- Scikit-Learn
- Pytorch
- Tensorflow (Keras)
- Huggingface
- Spacy
- Stanford Core NLP

Necessity of POS Tagging (1/2)

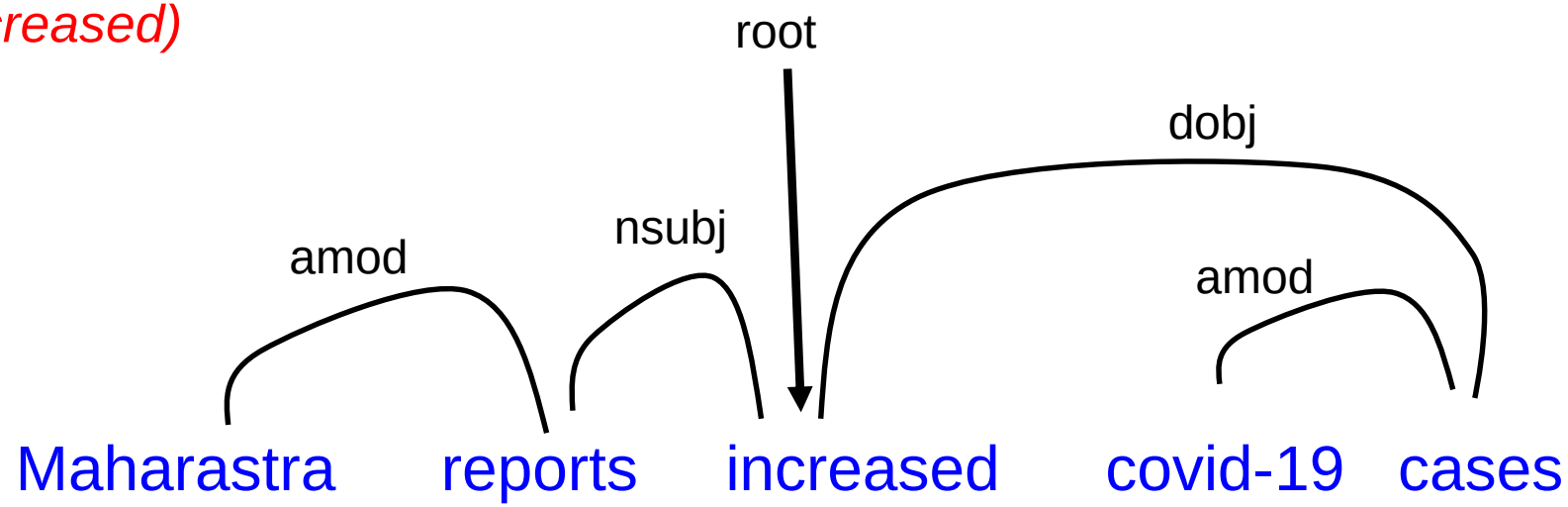
- **Command** Center to Track Best Buses (ToI 30Jan21): POS ambiguity affects

- Elderly with young **face** increased covid 19 risk (ToI Oct 20)

Dependency Ambiguity

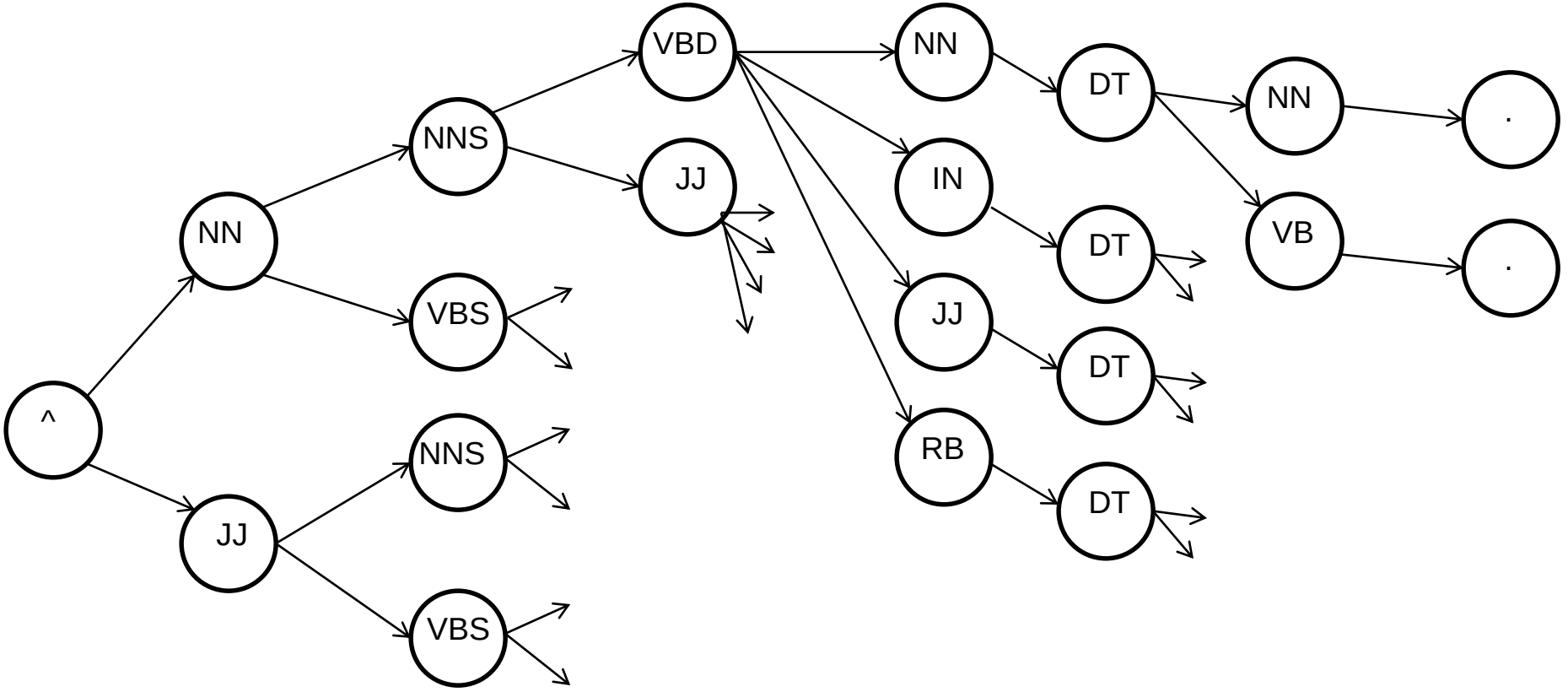


(it is reported by Maharastra Govt. that covid-19 cases have increased)

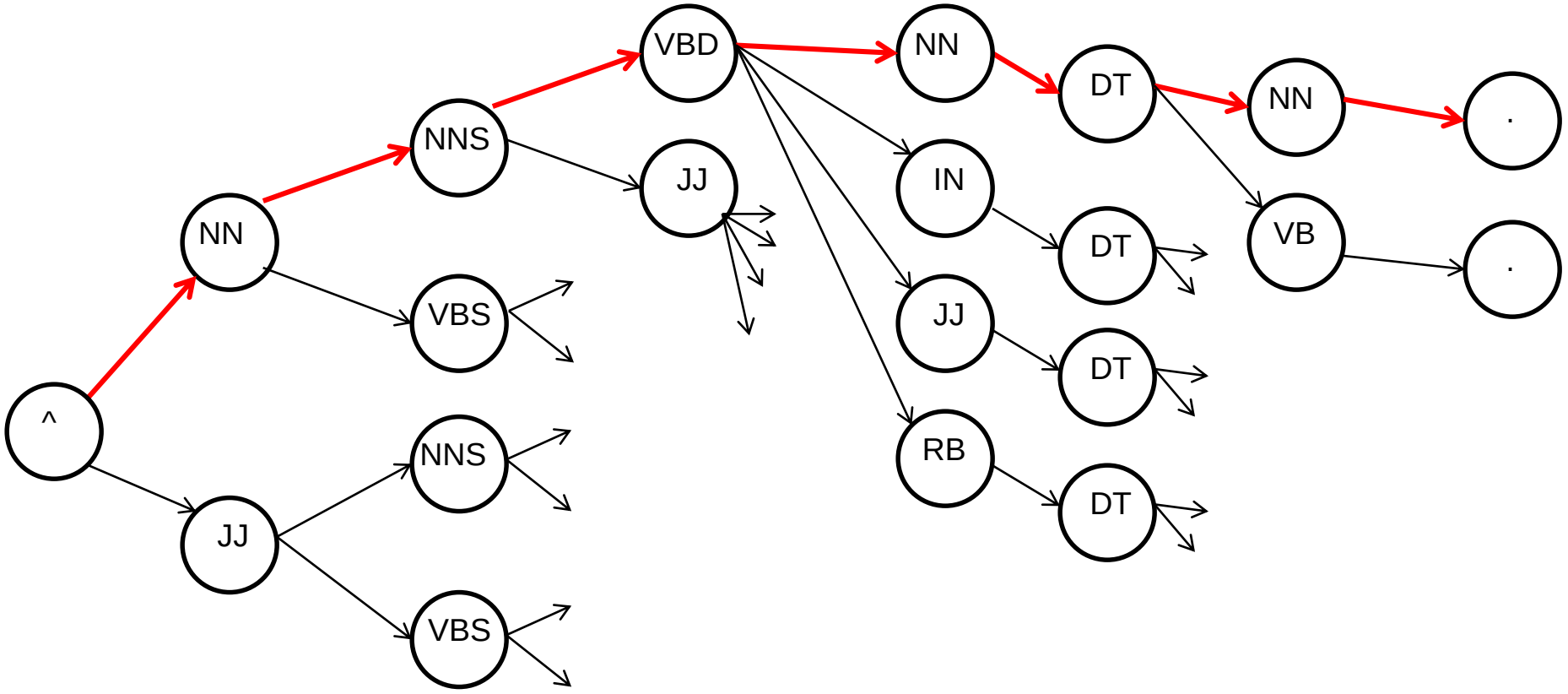


(it is the Maharastra reports that have increased covid-19 cases!!!)

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 .

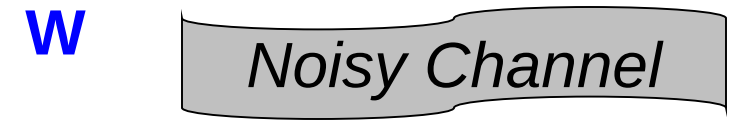


^ Brown foxes jumped over the fence .

Find the PATH with MAX Score.

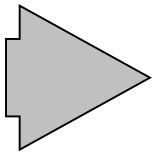
What is the meaning of score?

Noisy Channel Model



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

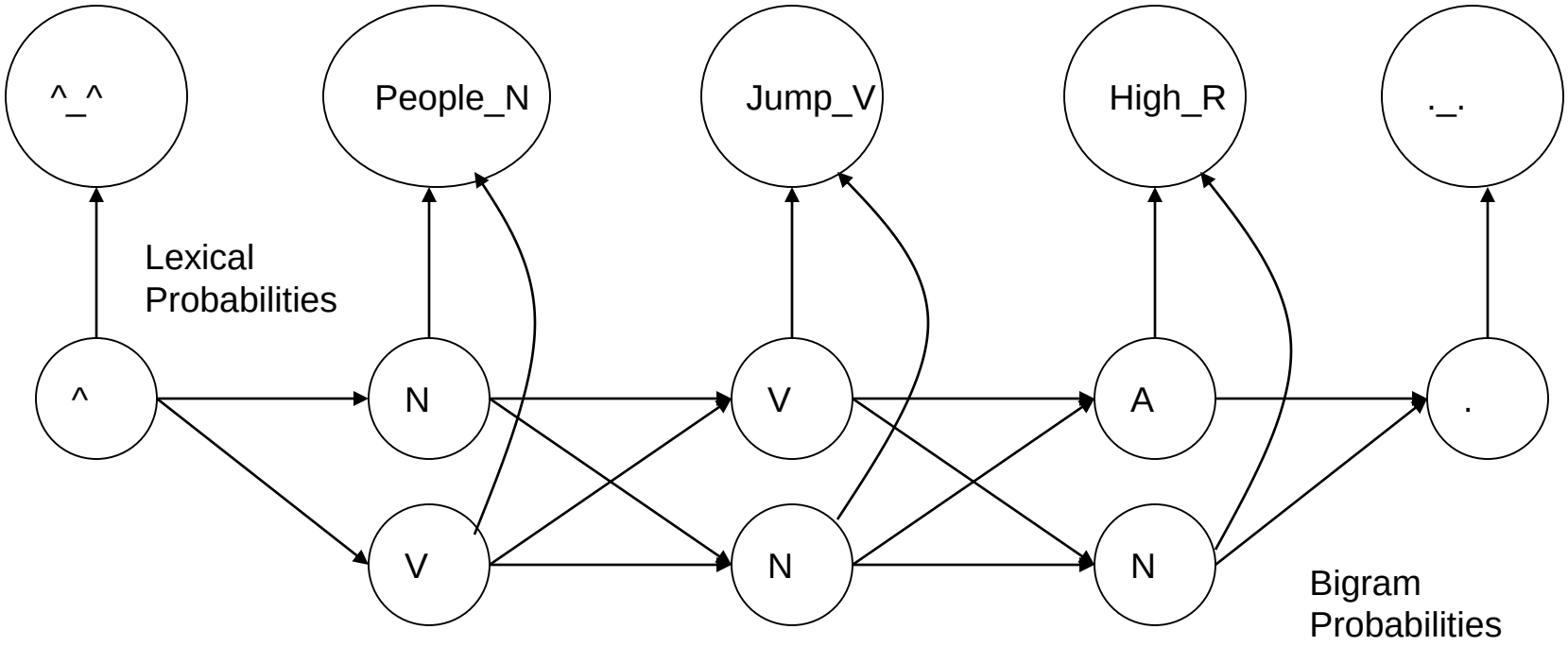
Sequence *W* is transformed into sequence *T*



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

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

HMM: Generative Model



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

Tag Set

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

Penn POS TAG Set

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjun
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative

Penn POS TAG Set (cntd)

22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

A dialogue text POS tagged from Treebank

[SpeakerB1/SYM]
./.
So/UH how/WRB many/JJ ,/, um/UH ,/,
[credit/NN cards/NNS]
do/VBP
[you/PRP]
have/VB ?/.

[SpeakerA2/SYM]
./.
[Um/UH]
,/,
[I/PRP]
think/VBP
[I/PRP]
'm/VBP down/IN to/IN
[one/CD]
/-

https://catalog.ldc.upenn.edu/desc/addenda/LDC99T42.pos.txt

POS ambiguity instances

best ADJ ADV NP V
better ADJ ADV V DET

close RB JJ VB NN (*running close to the competitor, close escape, close the door, towards the close of the play*)

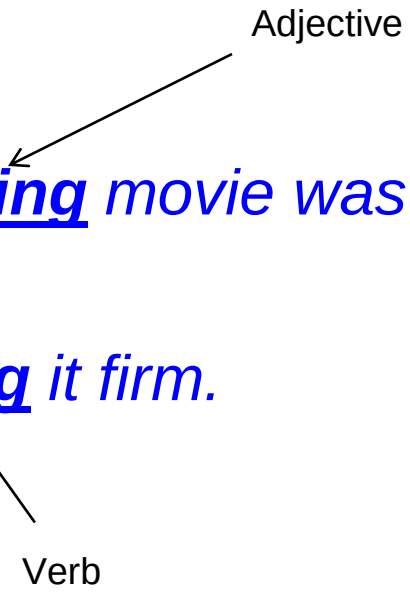
cut V N VN VD
even ADV DET ADJ V
grant NP N V -
hit V VD VN N
lay ADJ V NP VD
left VD ADJ N VN
like CNJ V ADJ P -
near P ADV ADJ DET
open ADJ V N ADV
past N ADJ DET P
present ADJ ADV V N
read V VN VD NP
right ADJ N DET ADV
second NUM ADV DET N
set VN V VD N -
that CNJ V WH DET

POS Ambiguity

- E.g.

  1. *What a **gripping** movie was Abhiman!*

  2. *He is **gripping** it firm.*



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. ***Golf*** *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 → *V*erb (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.

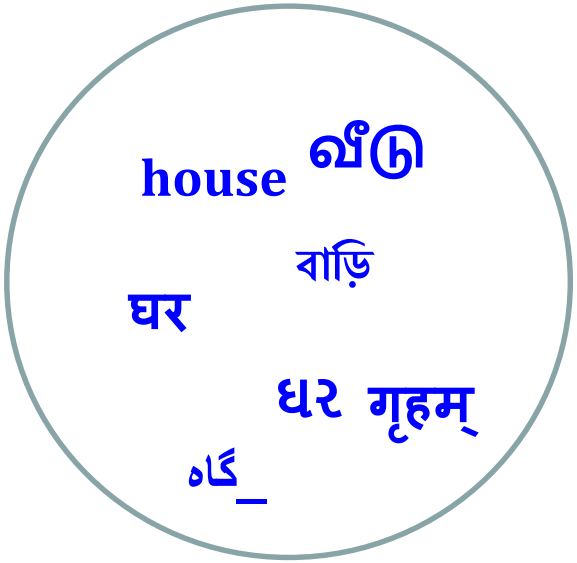
Assignment on POS Tagging

- Implement an HMM based POS tagger
- Training Corpus: Brown
- 80% for training
- 20% for testing
- 5 fold cross validation

More challenging assignment (not to be done now!)

- Use an L1 POS tagger to create an L2 POS tagger
- E.g., use an English POS tagger to create a POS tagger for, say, Telugu/Marathi/Swahili/Manipuri etc.
- Navigate to Universal Word Embedding Space
- “Borrow” the POS tag of host language
- Multilingual Transfer Learning

Multilingual Word Embedding Space



English (en), Tamil (ta), Hindi (hi), Bengali (bn), Gujarati (gu), Sanskrit (sn), Urdu (ur)

English (en), Telugu (te), Punjabi (pu), Malayalam (ml), Konkani (ko)

Facebook: MUSE

Recommended Browsing/Reading (1/2)

- https://www.nltk.org/book/ch05.html (NLTK POS Tagging)
- https://www.nltk.org/_modules/nltk/tag/tnt.html (Trigram Markov Process based tagger)
- Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider and Noah A. Smith, *Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters*, NAACL 2013.

Recommended Browsing/Reading (1/2)

- https://ai.facebook.com/tools/muse/ (multilingual embedding)
- Pushpak Bhattacharyya, *IndoWordNet,* Lexical Resources Engineering Conference 2010 **(LREC 2010**), Malta, May, 2010.
- K.V. Subbarao, *South Asian Languages- A Syntactic Study*, Cambridge University Press, 2012.