CS772: Deep Learning for Natural Language Processing (DL-NLP)

Introduction

Pushpak Bhattacharyya

Computer Science and Engineering Department

IIT Bombay

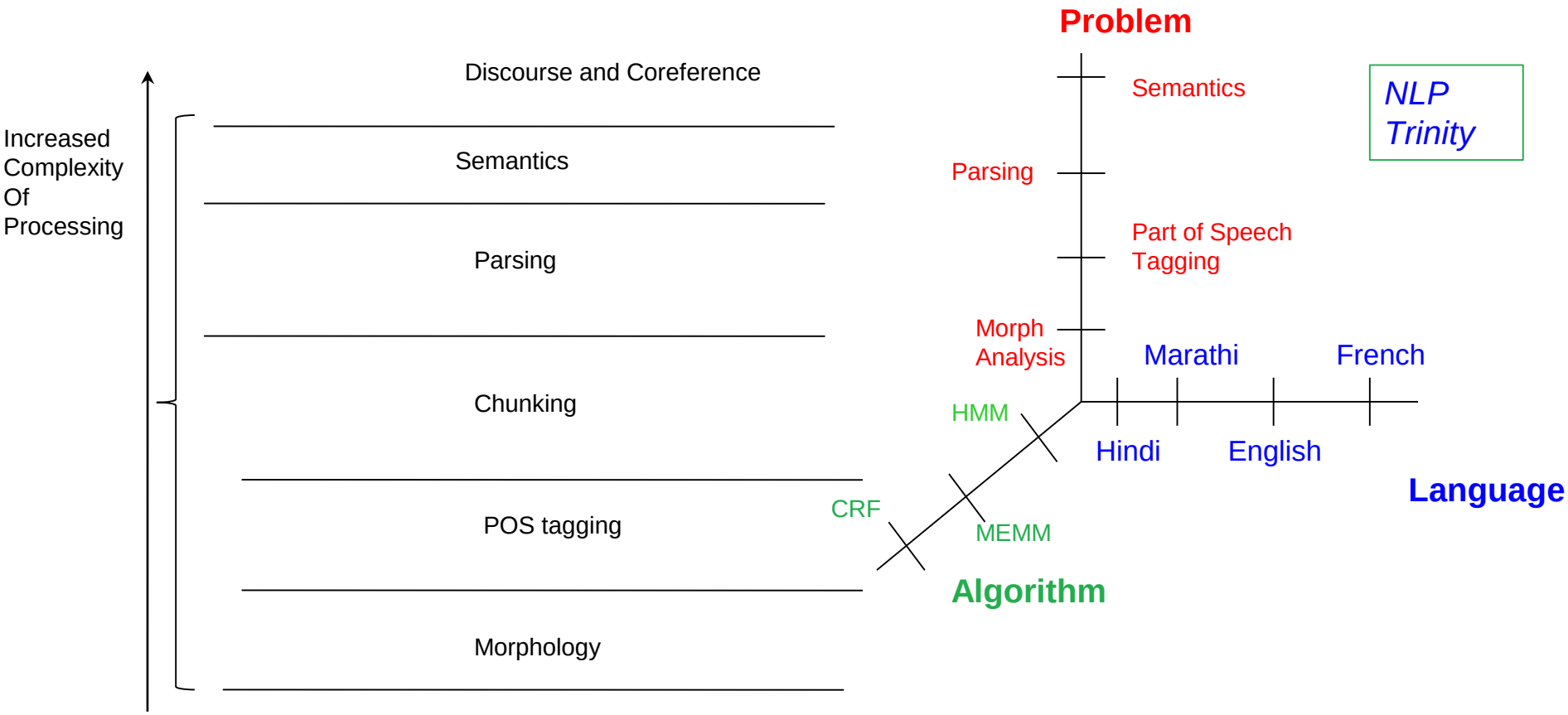
Week 1 of 3rd Jan, 2022

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**

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?

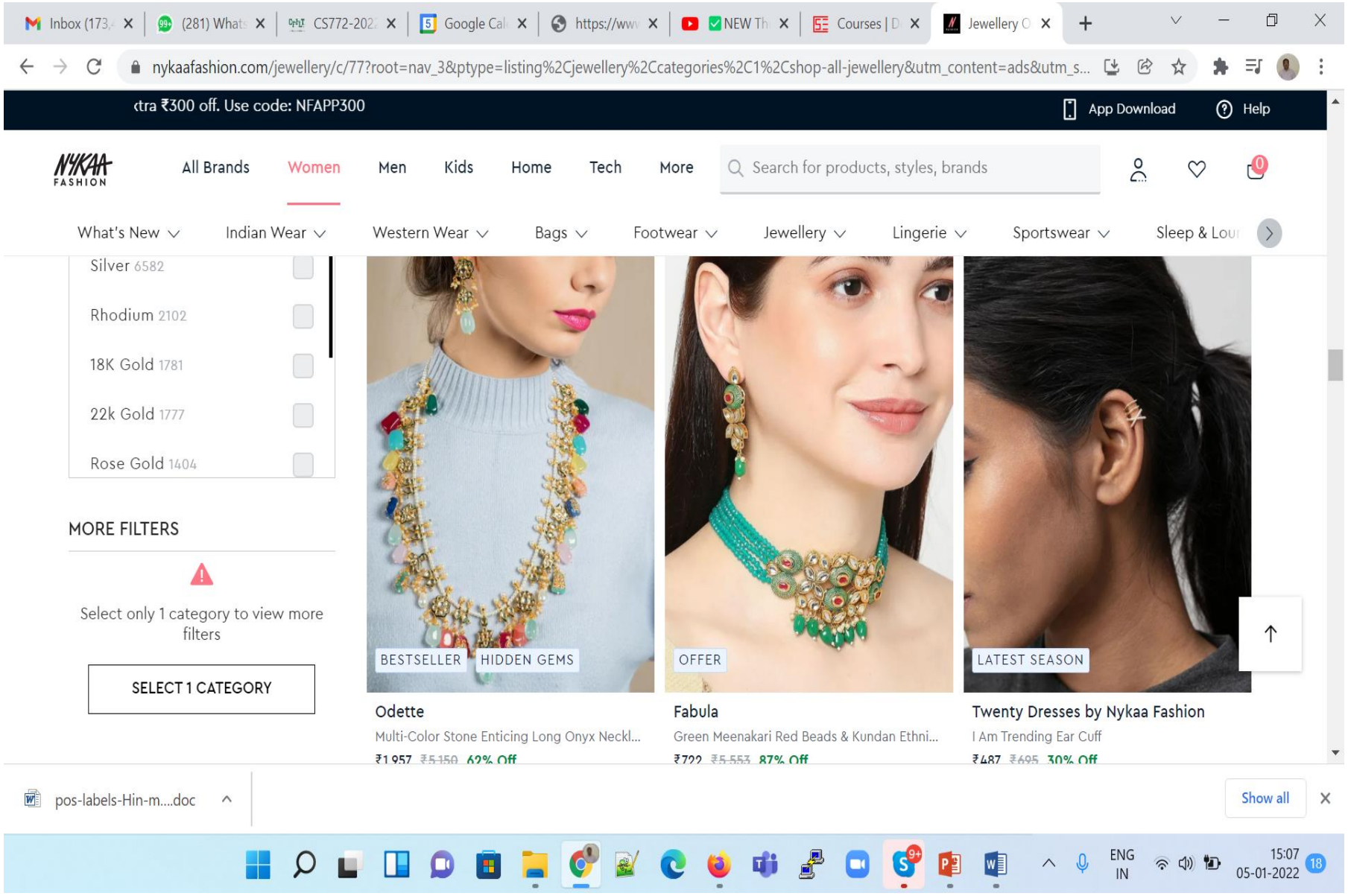
Lady B: <bn> kena ki holo?</bn> (why what happened?)

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

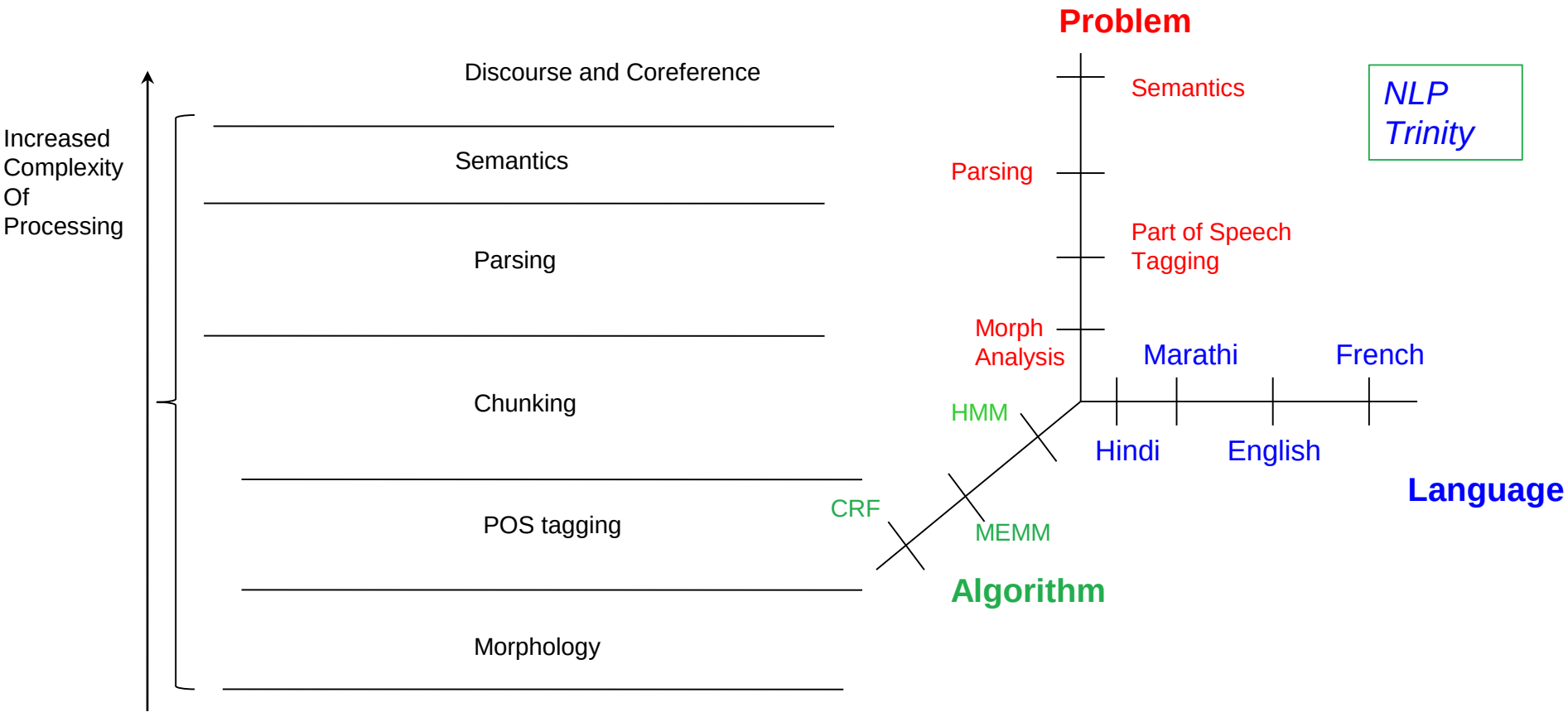
Root cause of the problem: Ambiguity!

- 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



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



Multimodal is important

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

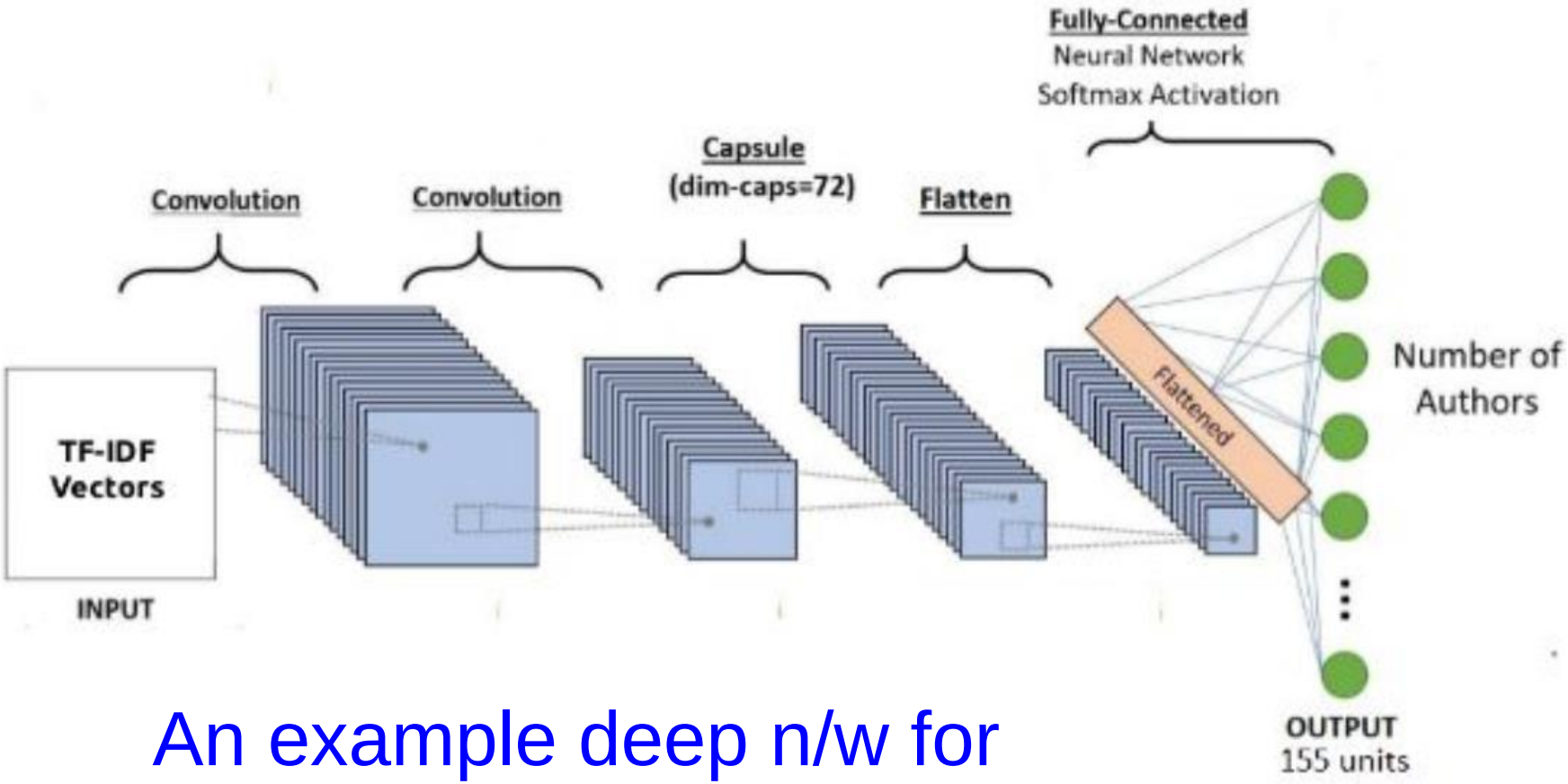
Data + Classifier > Human decision maker !!

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

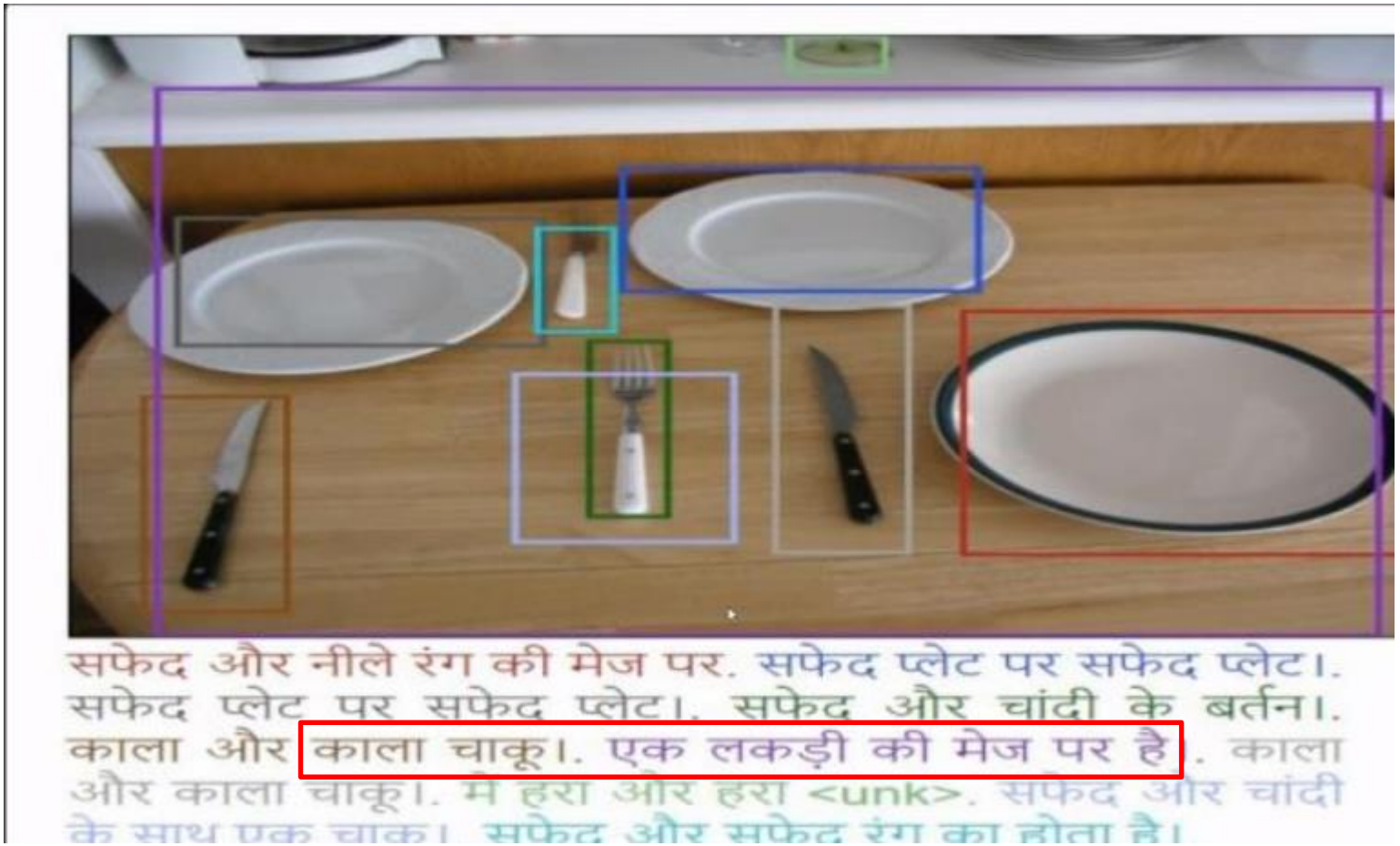
Cutting edge applications

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



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: Nihar Ranjan Sahoo, Apoorva Nunna, Kunal Verma, Vishal Pramanik, Harsh Peswani, Ankush Agrawal
- http://www.cfilt.iitb.ac.in/~cs772-2022
- Channels of communication: MS Teams, Moodle, Course Website

Evaluation Scheme (tentative)

- 50%: Reading, Thinking, Comprehending
  - Quizzes (25) (at least 4)
  - Endsem (25)
- 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.

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, EMNLP, NAACL, EACL, AACL, NeuriPS, ICML

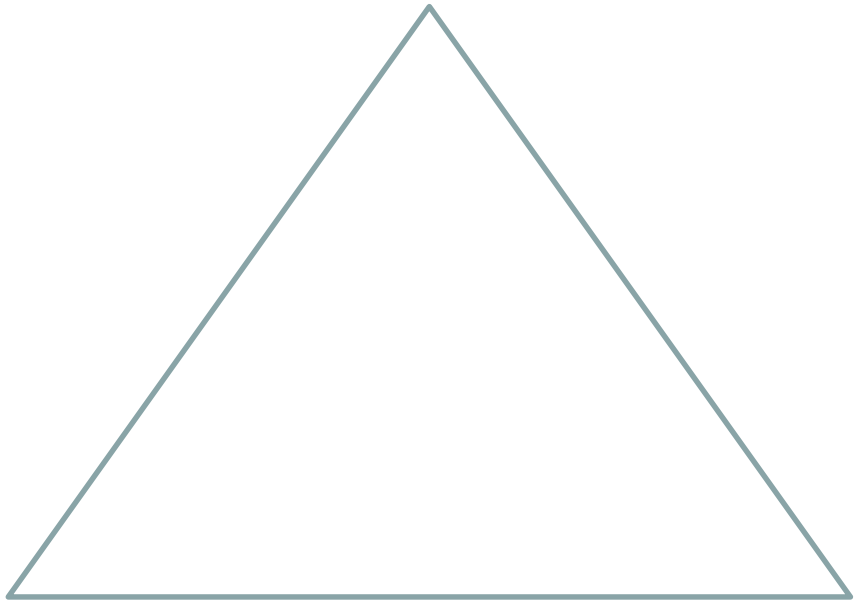
Useful NLP, ML, DL libraries

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

Nature of DL-NLP

The Trinity of NLP

Linguistics



Probability

Coding (DL)

3 Generations of NLP

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

Illustration with POS tagging

Case of “present”

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

*They **present** innovative ideas.*

*He was **present** in the class.*

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.

Table look-up will not do

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

Rule Based POS Tagging

- For Present_NN (look-back)
  - ***If present is preceded by determiner (the/a) or demonstrative (this/that), then the 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"

Rule based POS tagging cumbersome: statistical POS tagging

ML-POS needs training data

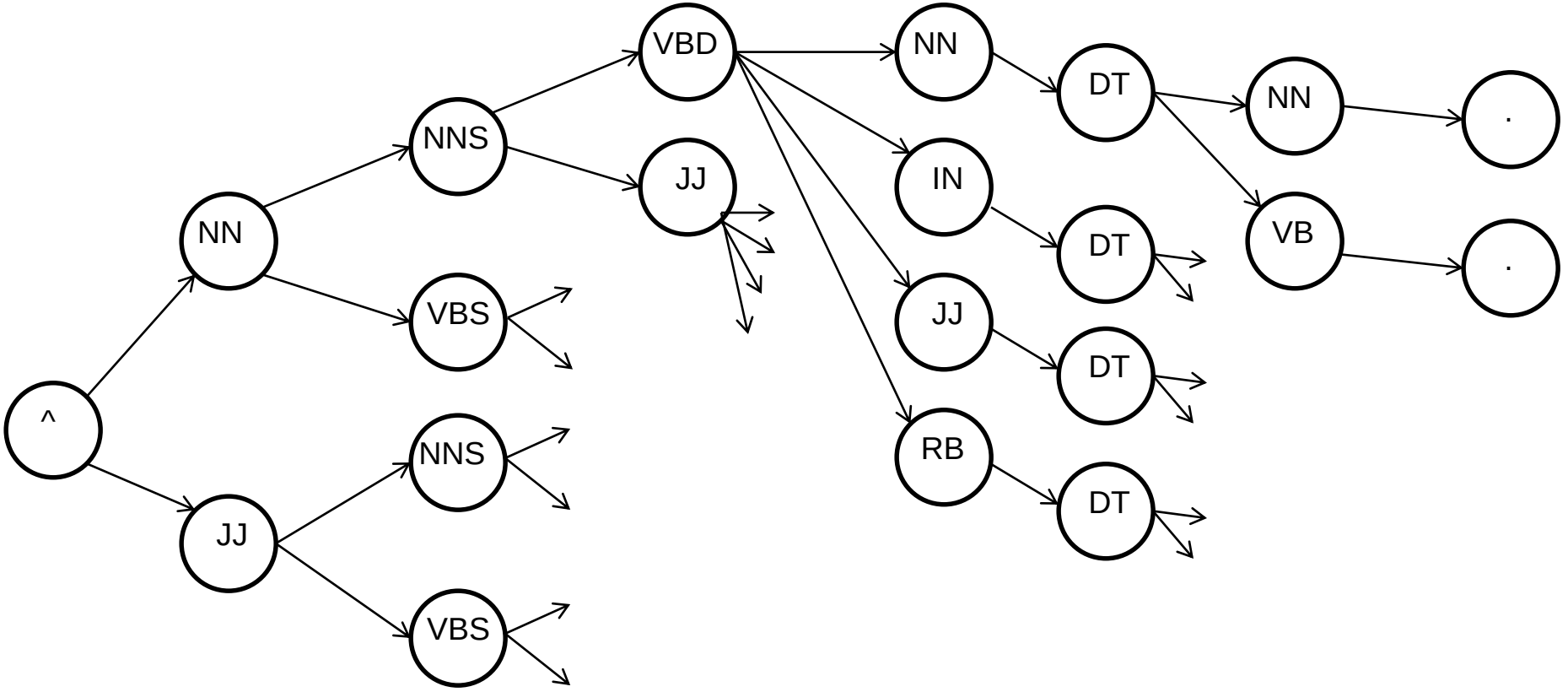
(1) He gifted me the/a/this/that **present_NN**.

(2) They **present_VB** innovative ideas.

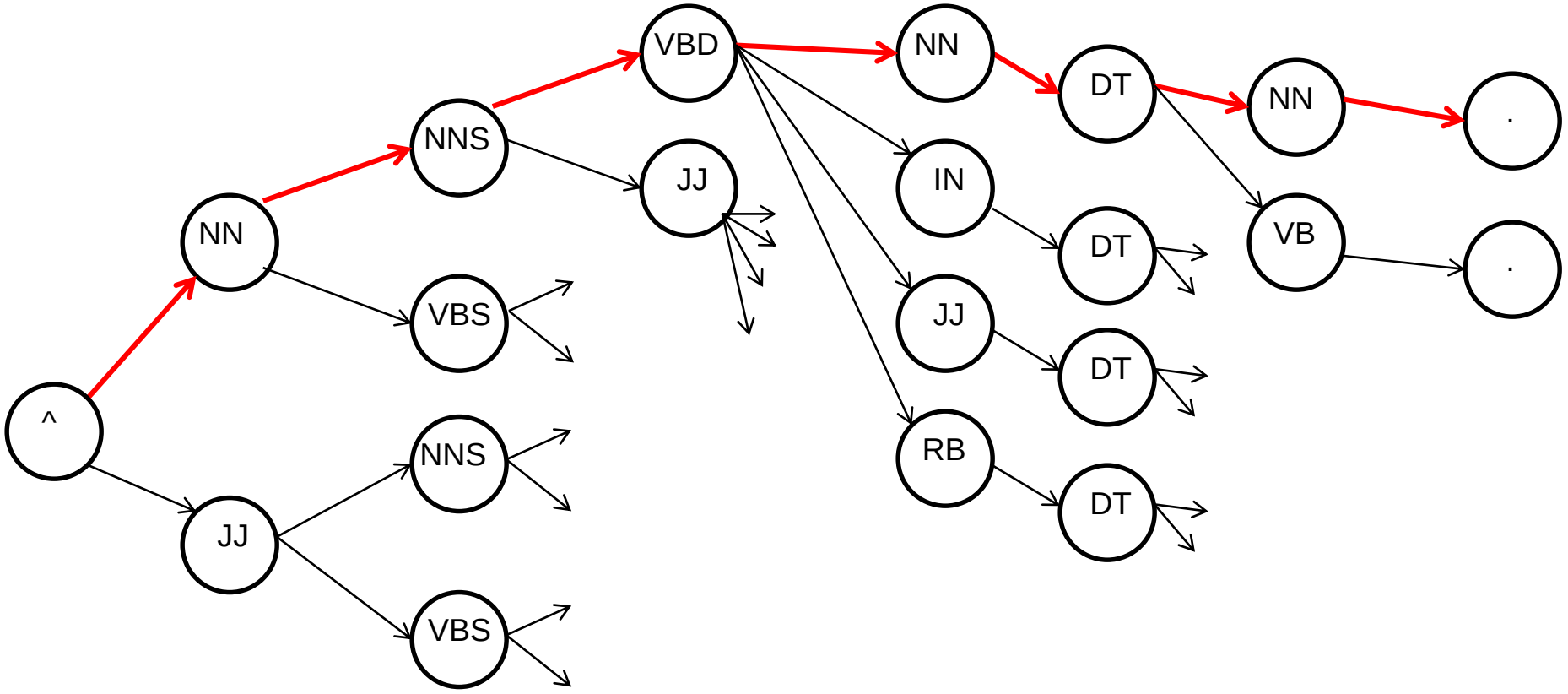
(3) He was **present_JJ** in the class.

POS options form a search graph

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 .

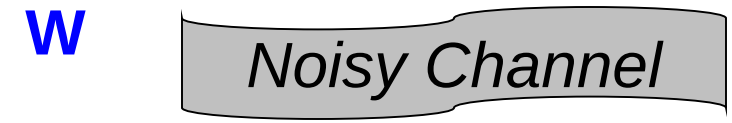


^ Brown foxes jumped over the fence .

Find the PATH with MAX Score.

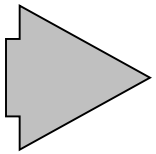
What is the meaning of score?

Noisy Channel Model



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

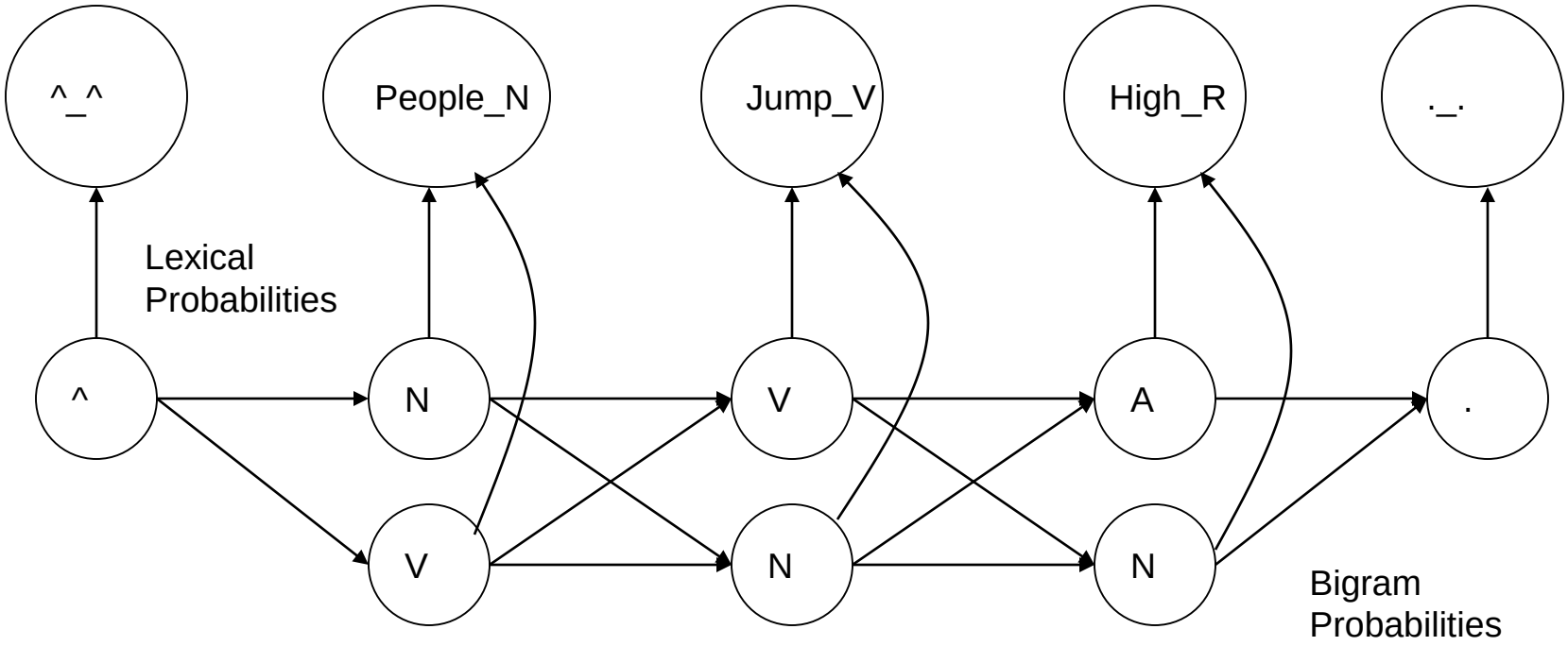
Sequence *W* is transformed into sequence *T*



$$T^* = \arg\max_T (P(T|W))$$

$$W^* = \arg\max_W (P(W|T))$$

HMM: Generative Model



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

CRF Based POS Tagging

Marathi

माणसाने उडण्याचा प्रयत्न केला

NN	VG	NN	VBD
B	B	B	I

Man tried flying

त्याने चालायला सुरुवात केली

PRP	VINF	NN	VBD
B	B	B	I

He started to walk

Harshada Gune, Mugdha Bapat, Mitesh Khapra and Pushpak Bhattacharyya, Verbs are where all the Action Lies: Experiences of Shallow Parsing of a Morphologically Rich Language, Computational Linguistics Conference (COLING 2010), Beijing, China, August 2010.

Decoding for the best Sequence

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} p_{\boldsymbol{\lambda}}(\boldsymbol{y}|\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} \boldsymbol{\lambda} \cdot \boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x})$$

$$p_{\boldsymbol{\lambda}}(\boldsymbol{Y}|\boldsymbol{X}) = \frac{\exp \boldsymbol{\lambda} \cdot \boldsymbol{F}(\boldsymbol{Y}, \boldsymbol{X})}{Z_{\boldsymbol{\lambda}}(\boldsymbol{X})} \qquad (1)$$

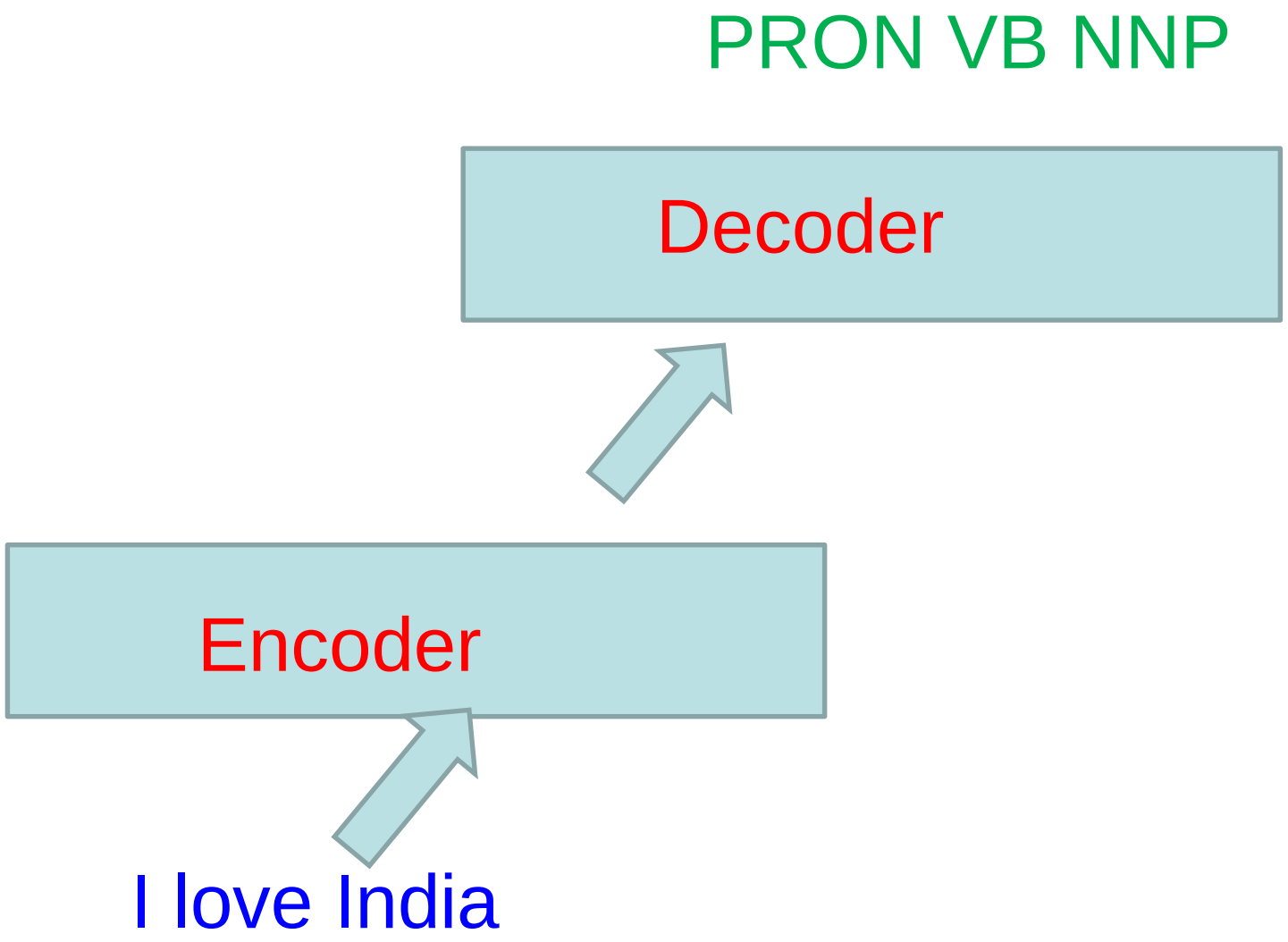
where

$$Z_{\boldsymbol{\lambda}}(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \exp \boldsymbol{\lambda} \cdot \boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x})$$

$$\boldsymbol{F}(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i} \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{x}, i)$$

i ranges over the input positions

DL based POS Tagging



How to input text to neural net? Issue of REPRESENTATION

- Inputs have to be sets of numbers
  - We will soon see why

- These numbers form **REPRESENTATIONS**

- What is a good representation? At what granularity: words, n-grams, phrases, sentences

Issues

- What is a good representation? At what granularity: words, n-grams, phrases, sentences
- Sentence is important- (a) *I bank with SBI; (b) I took a stroll on the river bank; (c) this bank sanctions loans quickly*
- Each ‘bank’ should have a differengt representation
- We have to LEARN these representations

Principle behind representation

- Proverb: “A man is known by the company he keeps”
- Similalry: “A word is known/represented by the company it keeps”
- “Company” → Distributional Similarity

Representation: to learn or not learn?

- 1-hot representation does not capture many nuances, e.g., semantic similarity
  - But is a good starting point
- Collocations also do not fully capture all the facets
  - But is a good starting point

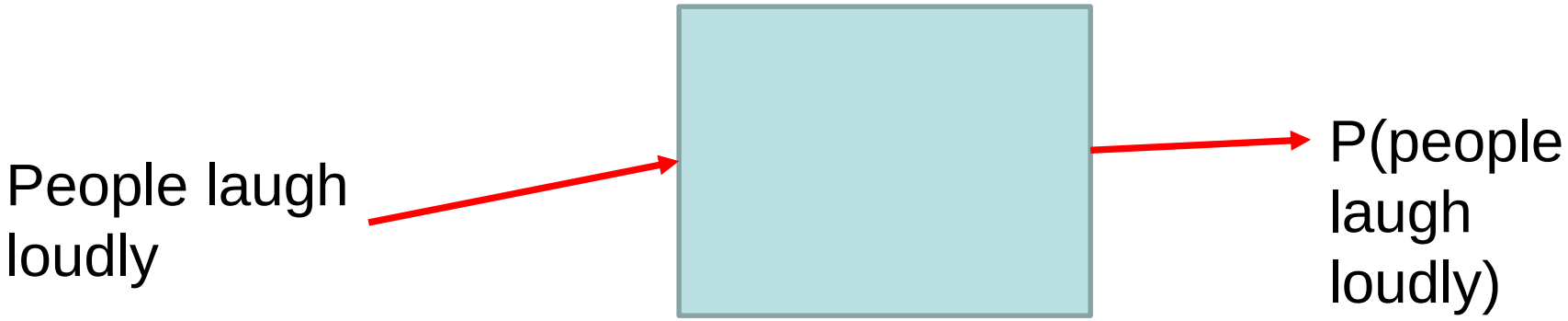
So learn the representation…

- Learning Objective

- ***MAXIMIZE CONTEXT PROBABILITY***

Neural LM

Neural Probability Computer

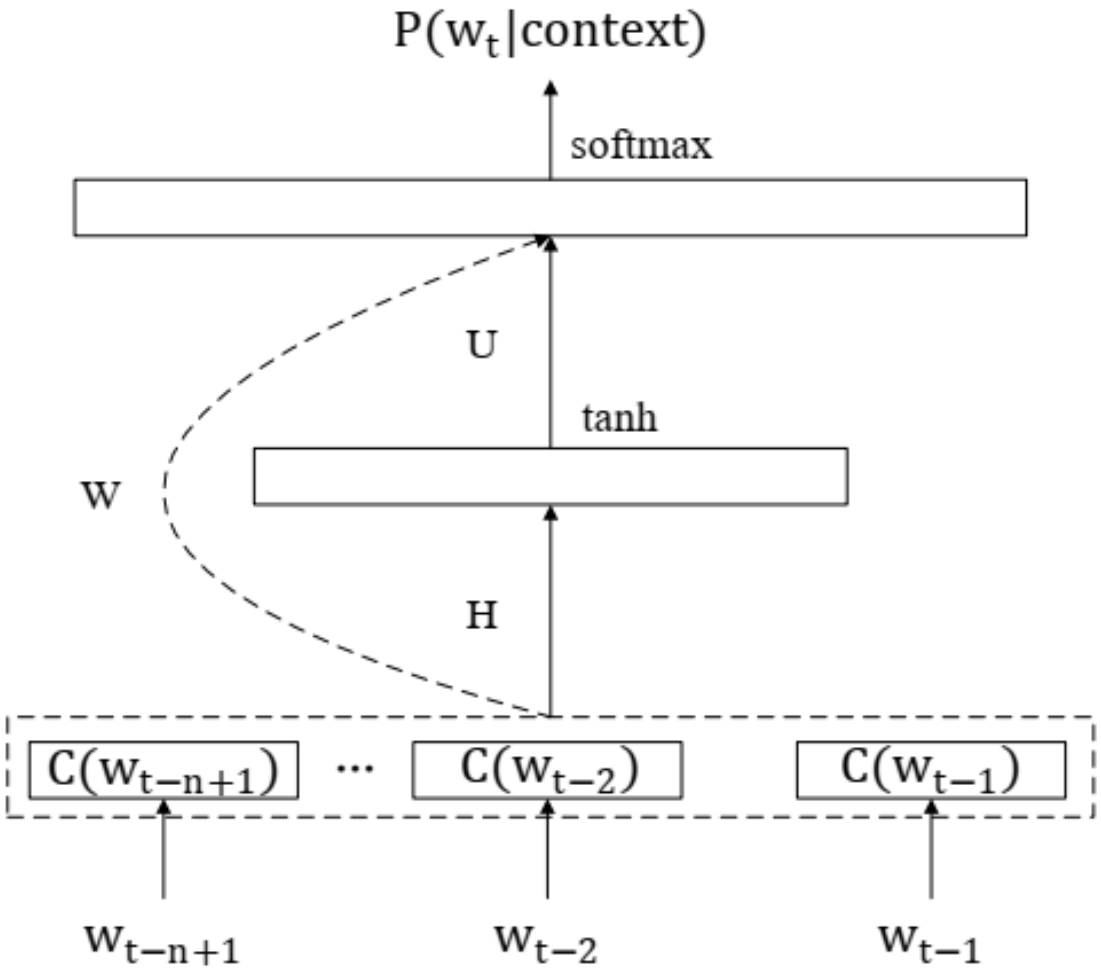


How does this happen

We have to first get the representation in place

- Word representation
- Phrase representation
- Sentence representation
- Long text representation

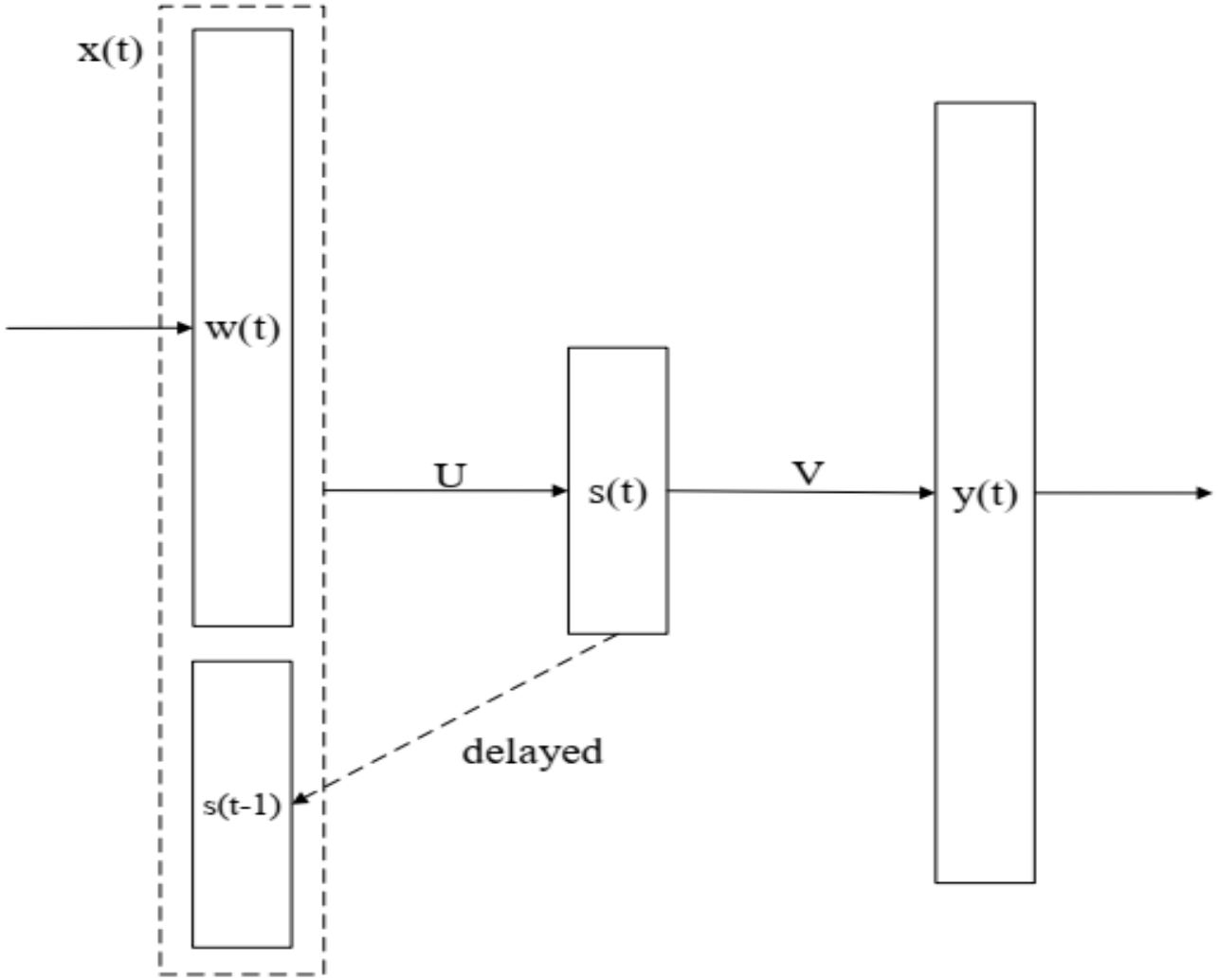
Feedforward Neural Language Model (FFNNLM): *Bengio et al 2003*



FFNNLM

- V is the vocabulary size, m is the dimension of the feature vectors; word w_i is projected as the distributed feature vector $C(w_i) \in R^m$
- The input x of the FFNN is a concatenation of feature vectors of n - 1 words
- Softmax output layer to guarantee all the conditional probabilities of words positive and summing to one
- The learning algorithm is the Stochastic Gradient Descent (SGD) method using the backpropagation (BP) algorithm

Recurrent NN LM (RNNLM)- *Mikolov et al 2010*



RNNLM

- RNN has an internal state that changes with the input on each time step, taking into account all previous contexts

- State s_t can be derived from the input word vector w_t and the state s_{t-1}