CS626: Speech, Natural Language Processing and the Web

Introduction

Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay

28th July, 2022

Sequel Course

- CS772- Deep Learning for Natural Language Processing
- NLP concepts covered in CS626 will see their realization on Deep Neural Nets
- These two courses create a solid platform for launching substantial research and development in NLP

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)

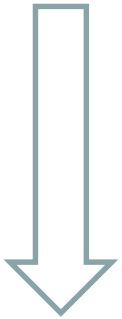
What is “Linguistics”

- 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

What is NLP

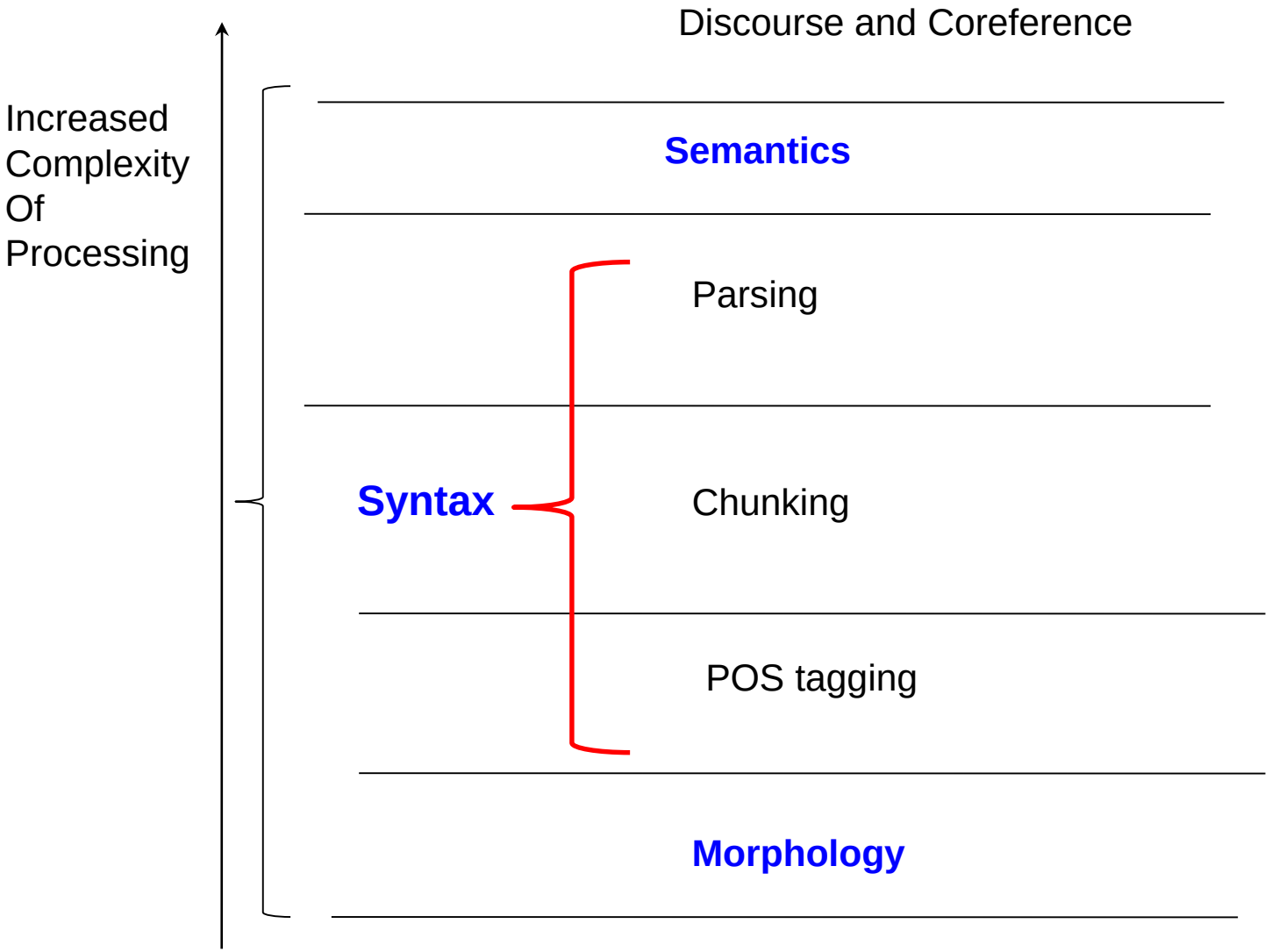
NLP= Language + Computation



(due to ML)

= Linguistics + Probability

NLP Layers


Discourse and Coreference
Increased
Complexity
Of
Processing
Semantics
Parsing
Syntax
Chunking
POS tagging
Morphology

Linguistic Strata *vs.* Languages

	Hindi	**Swahili**	**Tamil**	***etc.***
Sound: Phonetics, Phonology				
Structure: Morphology, Syntax				
Meaning: Semantic, Pragmatics				

Speech-NLP Stack *vs.* Languages

	Hindi	**Swahili**	**Tamil**	***etc.***
Sound: ASR, TTS				
Structure: MA, POS, NE, Chunker, Parser				
Meaning: SRL, Knowledge Nets, SA-EA-OM, QA, Summarizer				

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)

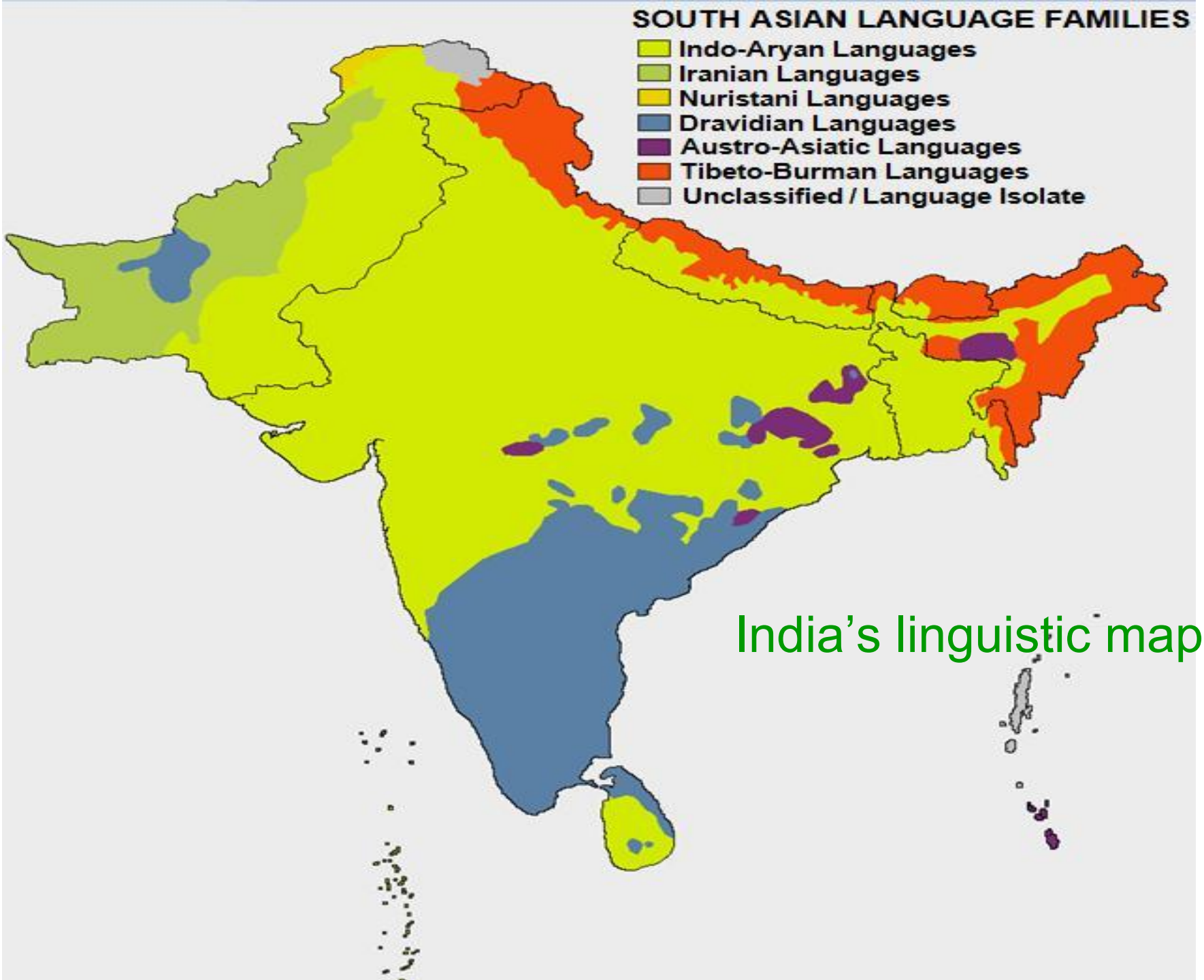
Dominant Word Order Distribution Across Languages (Wikipedia)

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%

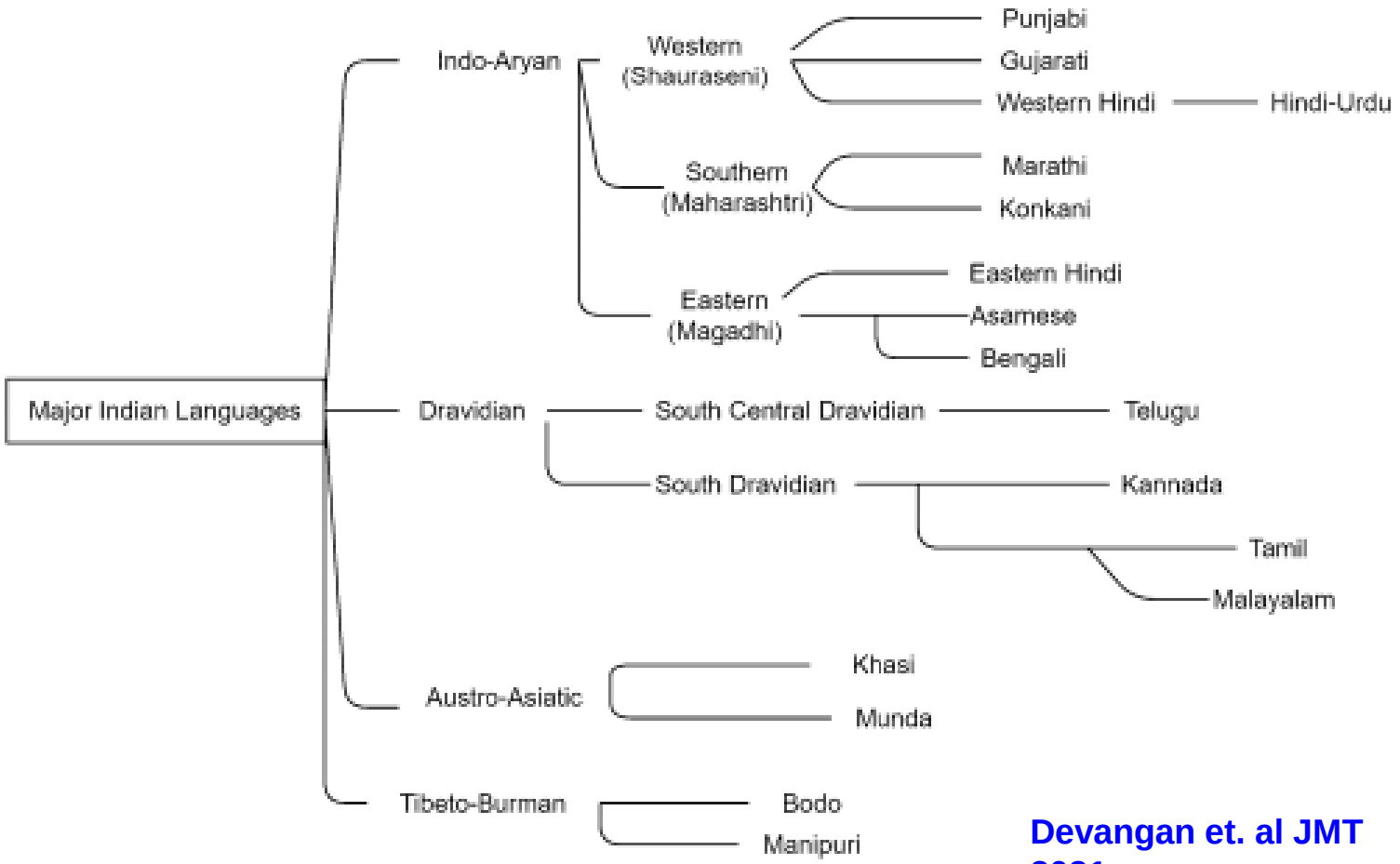
Some interesting cases

- 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



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?

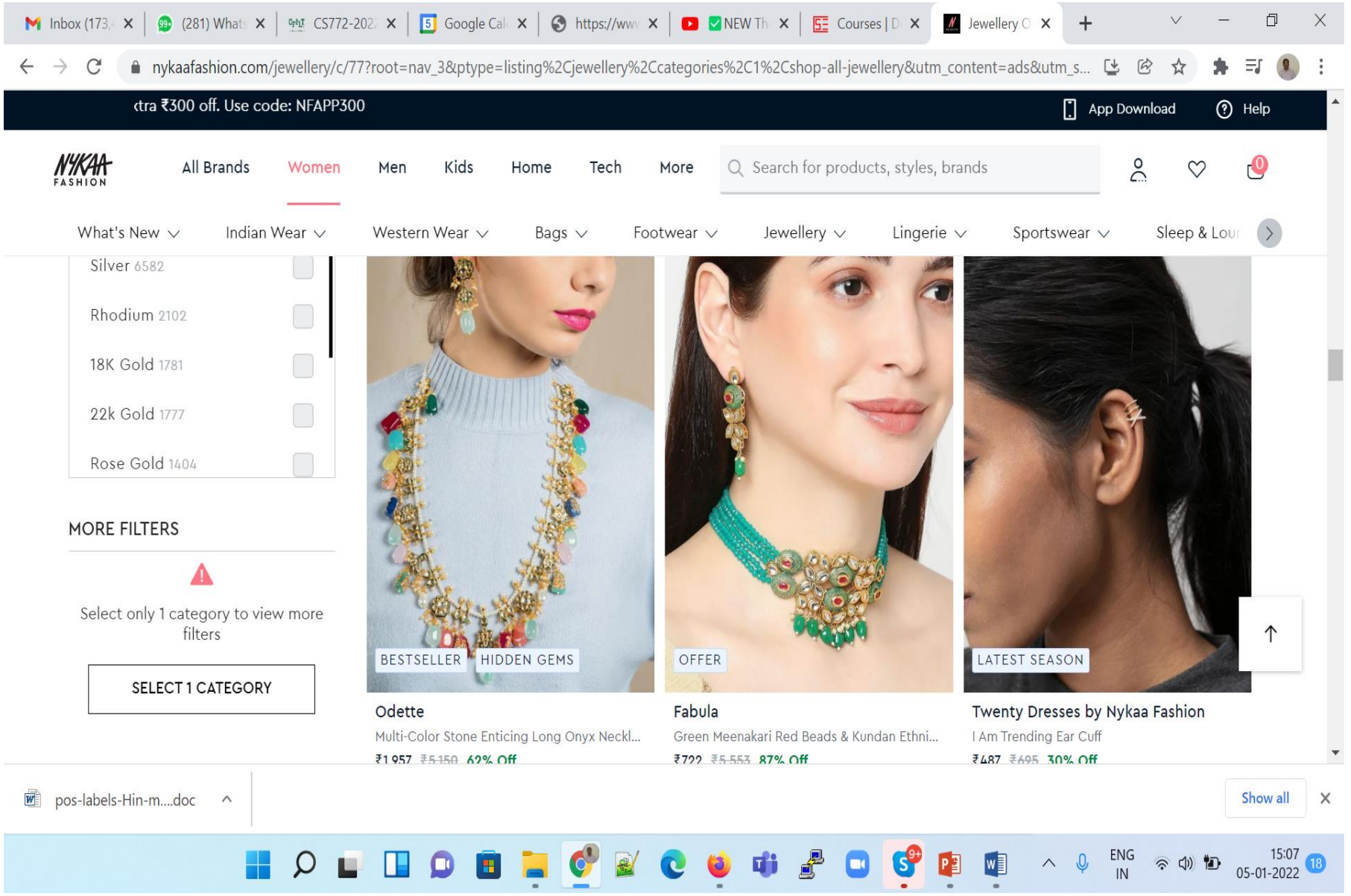
Lady B: <bn> kena ki holo?</bn> (why what happened?)

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

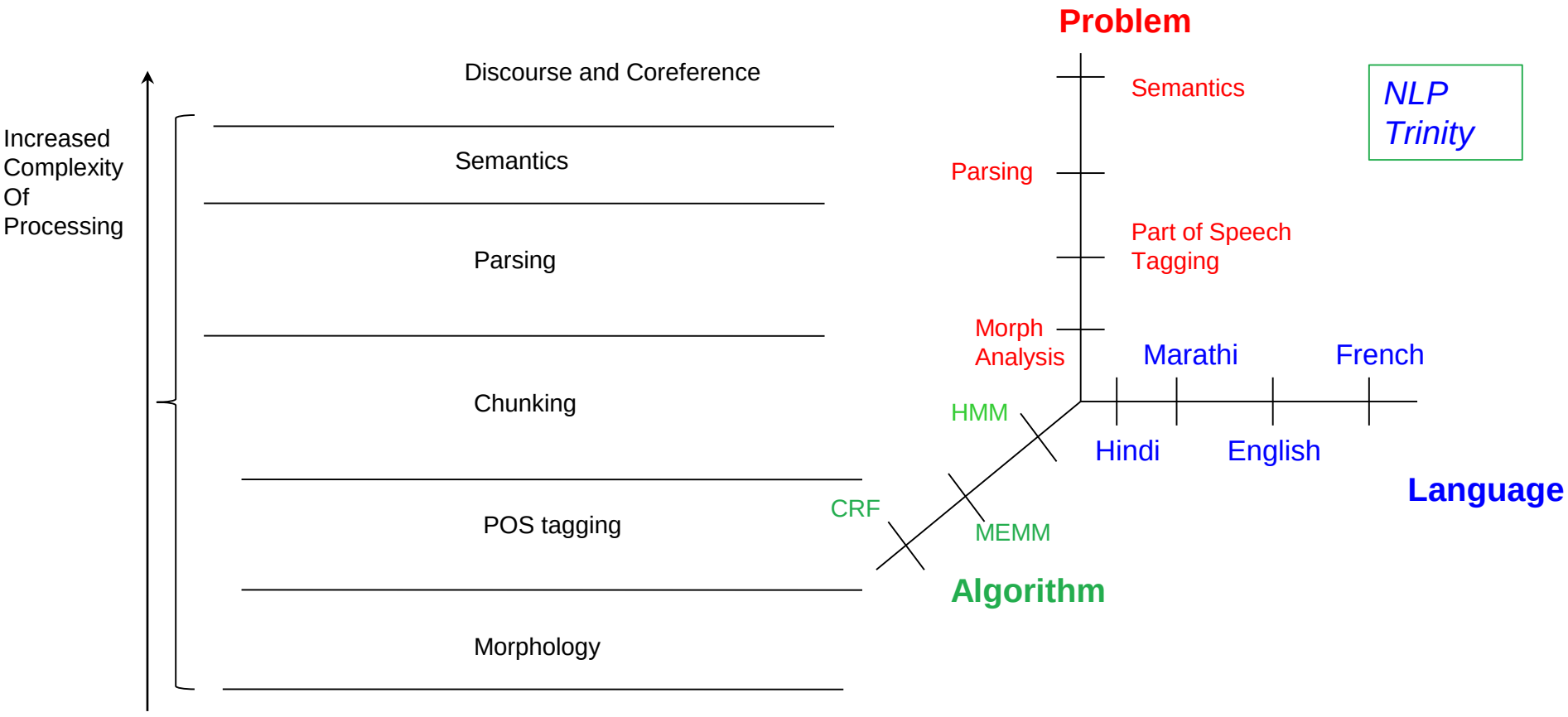
Root cause of the problem: Ambiguity!

- 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



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



Role of Multimodality

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

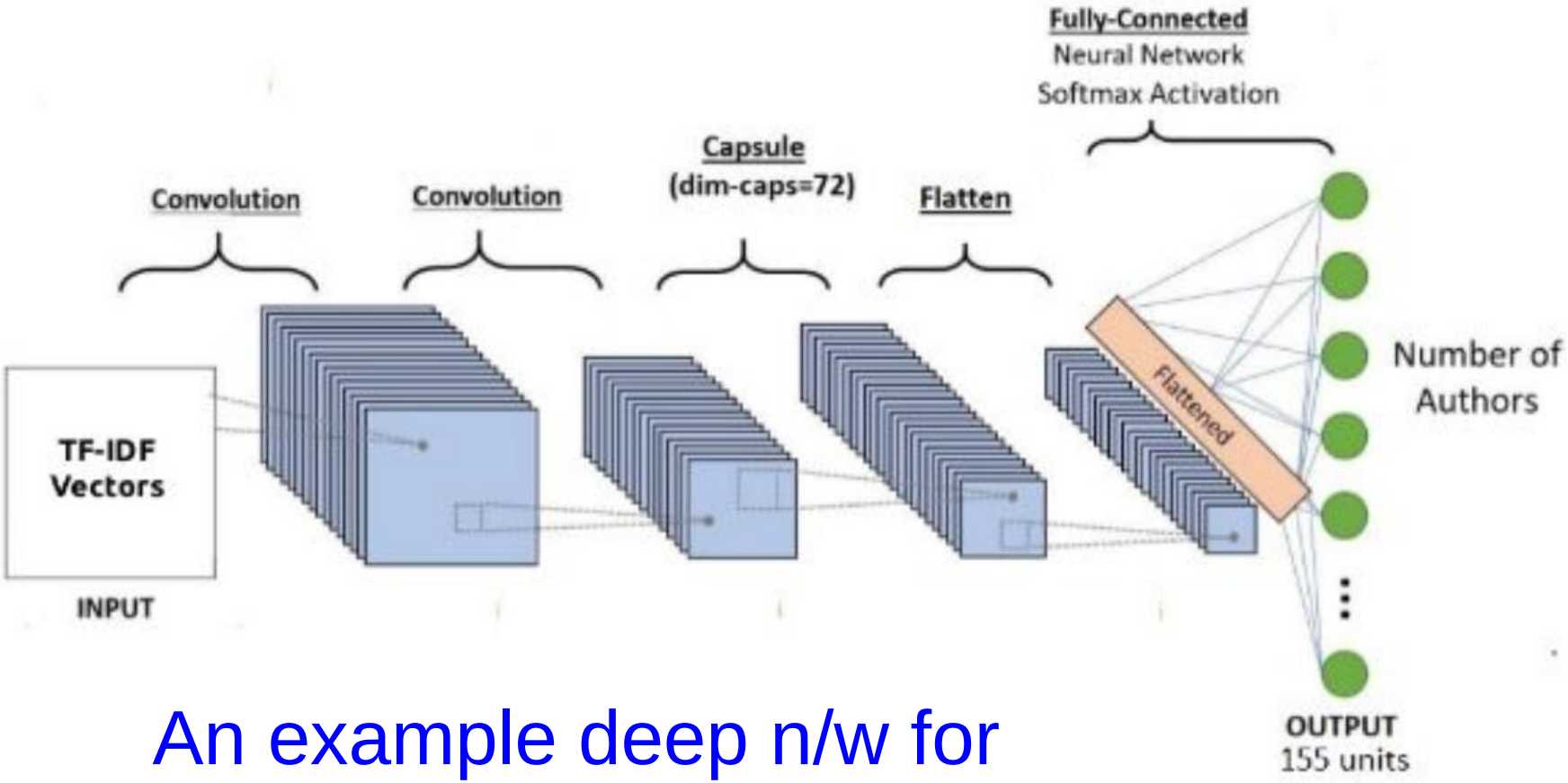
Frequent Observation:
Data + Classifier > Human decision maker !!

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

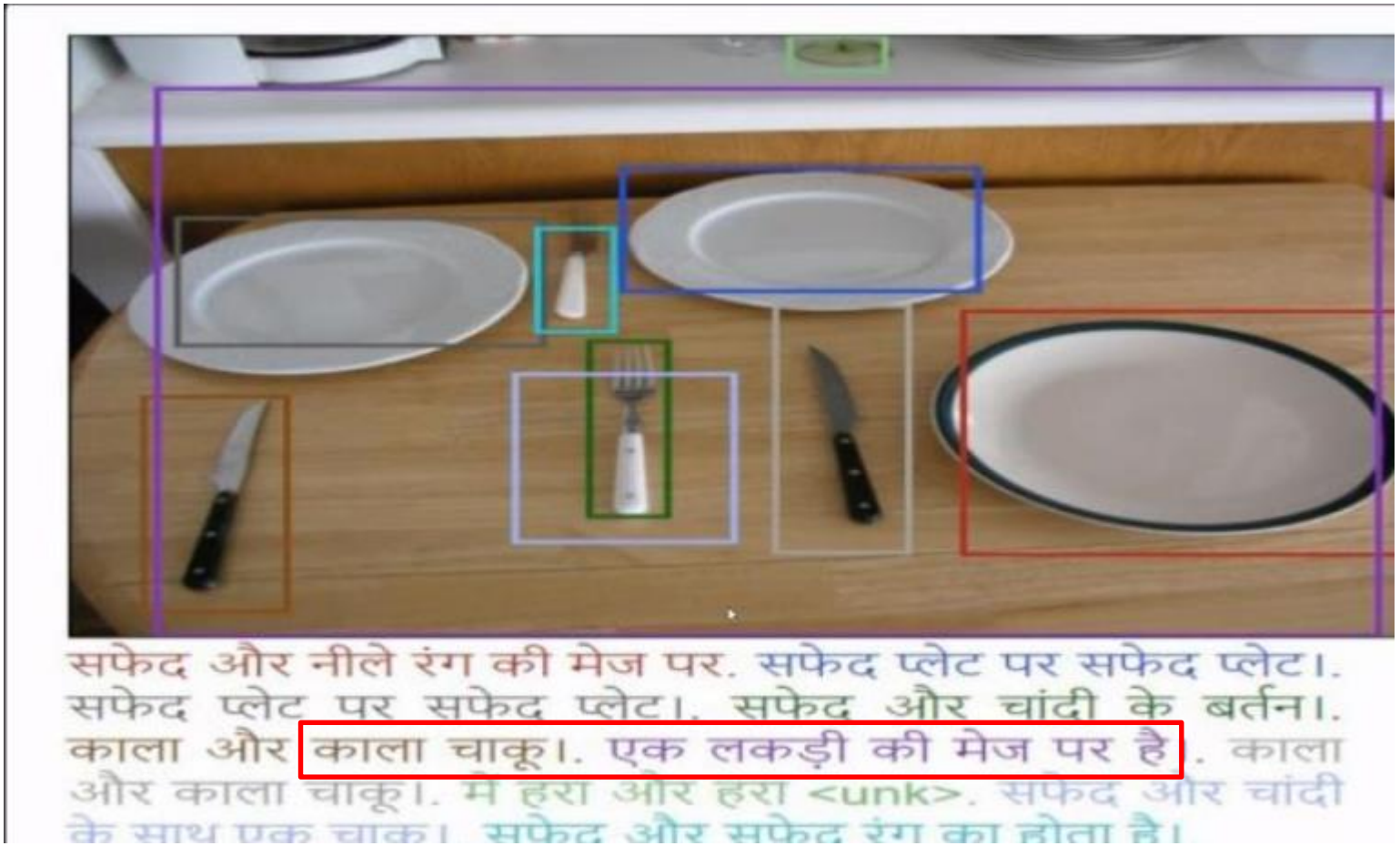
Cutting edge applications

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



OCR-MT-TTS

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

Demoes

Course Logistics

Venue

- First Introductory Lecture: F.C.Kohli Auditorium, Thursday 28th July, 10.30AM

- After that in new CSE building, 101 (first floor)

Course Logistics- MS teams

- Join the MS Teams using the code **jun37pk**
  - Login to MS teams using the LDAP credentails
  - Select **Join or create a team**
  - Choose **Join a team with a code**
  - Join using the code above
  - You will be added to the team CS626-2022
    - General channel: For notifications
    - Live lectures channel: For live lectures

Course website

Website:
https://www.cse.iitb.ac.in/~cs626/2022/#about

Visit the above website for course related information.

For more information about the research in NLP visit CFILT website.

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

Class Schedule

- Monday: 8:30 AM to 9:25 AM
- Tuesday: 9:30 AM to 10:25 AM
- Thursday: 10:30 PM to 11:25 AM

Moodle

Login to Moodle with LDAP credentials.

- Select the course CS626

All course related notifications will be notified via Moodle also.

Course TAs

- Nihar Ranjan Sahoo , PhD CSE
  (nihar@cse.iitb.ac.in)
- Sandeep Singamsetty, M.Tech CSE)
  (213050064@iitb.ac.in)
- NVS Abhishek, M.Tech CSE
  (213050019@iitb.ac.in)

Evaluation Scheme (tentative)

- 50%: Reading, Thinking, Comprehending
  - Quizzes (15)
  - Midsem (15)
  - Endsem (20)
- 50%: Doing things, Hands on
  - Assignments (25%)
  - Reading ONE paper and doing a preliminary implementation of the same (25%)
- Quiz every last Thursday of the month

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

3. Pushpak Bhattacharyya, Machine Translation, CRC Press, 2017.

4. Ian Goodfellow, Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press, 2016.

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