

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

Introduction

Pushpak Bhattacharyya
Computer Science and Engineering
Department
IIT Bombay

Week 1 of 1th Jan, 2024

Nature of Language

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)

Neurophysiology: Broca's Area

Broca's area in frontal lobe, left hemisphere

Damage messes up syntax ✓

a-grammatical speech production; Inability to use syntactic information; telegraphic speech; loss of function words and suffixes

"I eat rice spoon"

→ *I eat rice with spoon*

No difference between

- *Visit to the President vs.*
- *Visit by the President*

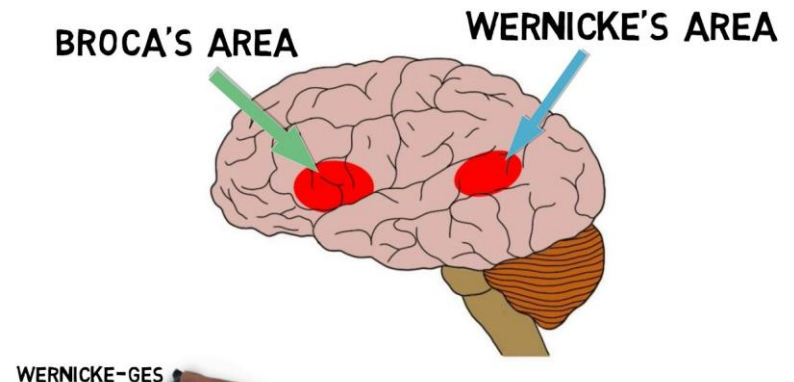


Wernicke's Area

Wernick's area is in superior temporal gyrus, left hemisphere

Damage messes up semantics-pragmatics

Fluent meaningless phrases: *The pink elephant sang and the blue stone danced*

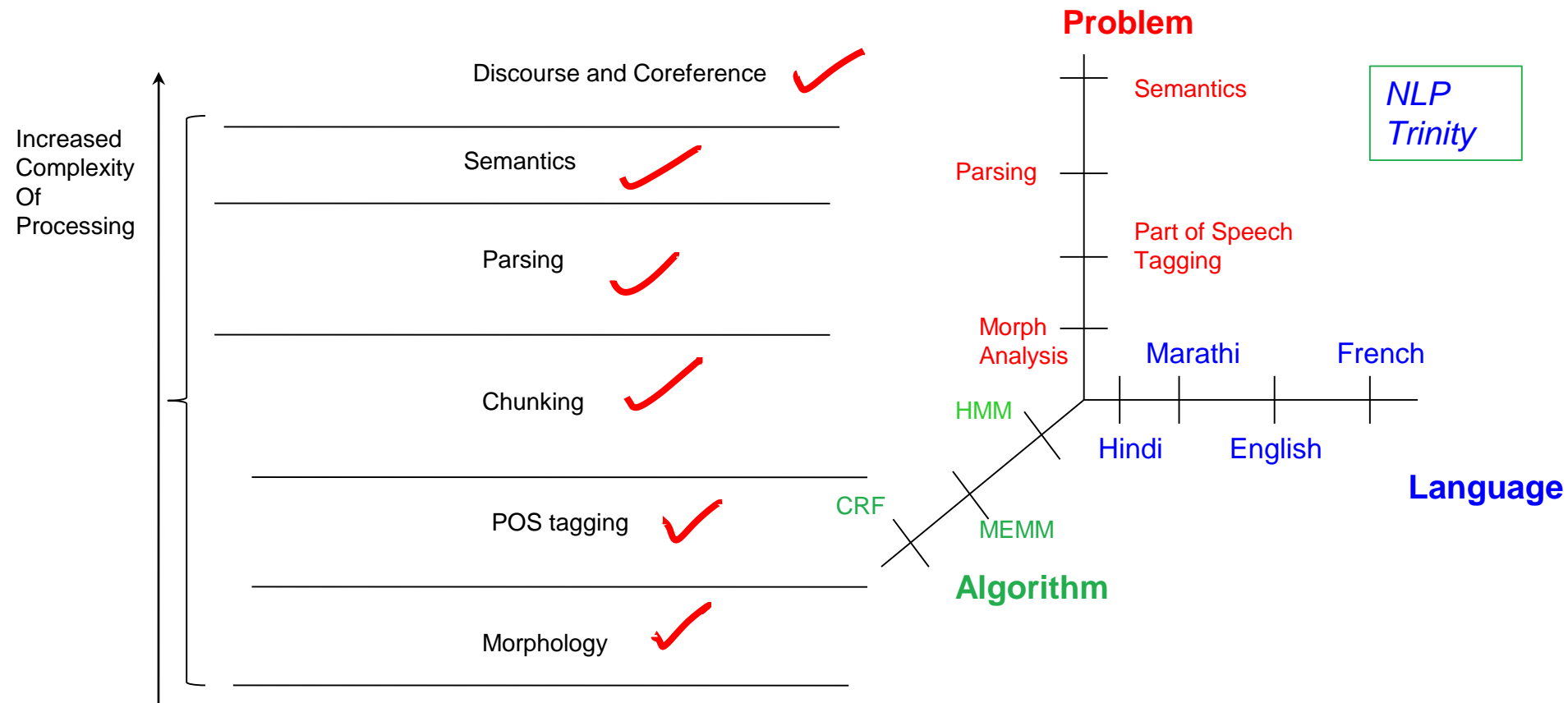


Nature of NLP

Natural Language Processing

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

NLP is layered Processing, Multidimensional too



Course Content: Task vs. Technique Matrix

Task (row) vs. Technique (col) Matrix	Rules Based /Know ledge- Based	Classical ML				Deep Learning, LLM, Stable diffusion		
		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								

What is NLP

NLP= Language + Computation



(due to ML)

= Linguistics + Probability

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

Main Challenge: **AMBIGUITY**

Example

(from a lecture by Prof. Mirella Lapata)

*“Part is history
future is mystery
present is a gift,
so it is called present”*

Example (cntd.)

“mat mat bichhaao!”

1. मत मत बिछाओ (don't don't spread!)
2. चटाई मत बिछाओ (don't spread the mat!)
3. चटाई चटाई बिछाओ (mat, spread the mat!)
4. मत चटाई बिछाओ

ANOTHER Example

“Buffalo buffaloes Buffalo buffaloes buffalo
buffalo Buffalo buffaloes”

Prompt to chatGPT: what do you
understand by the above sentence

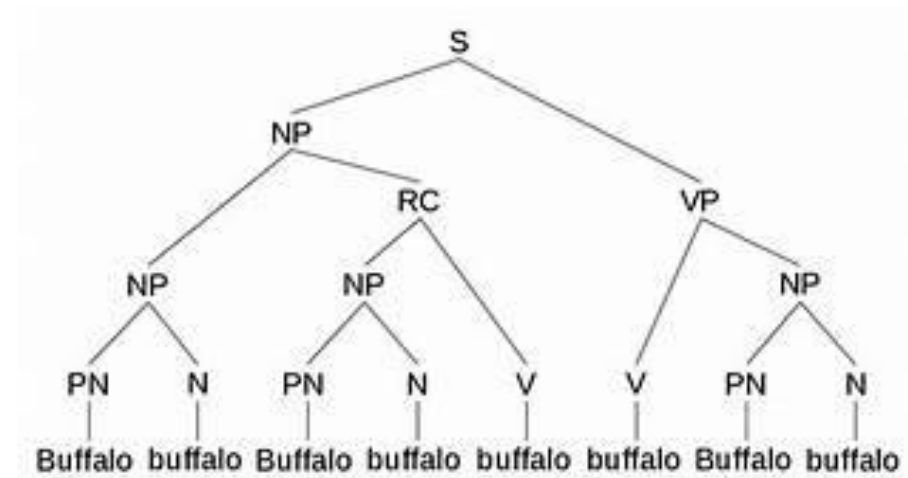
chatGPT response

- Buffalo (noun): Refers to the city of Buffalo, which is in the state of New York in the United States.
- buffaloes (verb): Means to bully, confuse, or baffle.
- Buffalo (noun): Refers to the animal, specifically the American bison.

describes a scenario where bison from Buffalo are being intimidated or confused by other bison from the same place, who, in turn, are intimidating or confusing other bison. It's a playful sentence that plays on the multiple meanings of the word "Buffalo" as a place

CORRECT!!

Bing chat response



The sentence can be parsed as follows: "Buffalo buffalo (Buffalo bison) Buffalo buffalo (Buffalo bison) buffalo (intimidate) buffalo (intimidate) Buffalo buffalo (Buffalo bison)"¹. In other words, the sentence claims that bison from Buffalo, New York, who are intimidated by other bison in their community, in turn intimidate other bison in their community¹.

Back to “*mat mat bichhaao!*”

1. मत मत बिछाओ (don't don't spread!)
2. चटाई मत बिछाओ (don't spread the mat!)
3. चटाई चटाई बिछाओ (mat, spread the mat!)
4. मत चटाई बिछाओ

Which is the correct meaning?

Linguistic approach (model based)- Depends on (a) context, (b) meaning of words, (c) composition

Data approach- depends on the frequency- Maximum Likelihood; **LANGUAGE MODEL**

✓ LLMs

All they do is *predict* next word/sentence and fill in the gaps

Self supervised learning

Set the parameters for prediction

Use the following in increasing order of resource demand for specific tasks

- (a) Prompt engineering (least resource hungry)
- (b) Adapter n/w
- (c) Fine-tuning
- (d) Pre-training (Most resource hungry).

3 stages of LLM based CAI (chatGPT)

- Generative Pretraining (GP)
- Supervised Fine Tuning (SFT)
- Reinforcement Learning based on Human Feedback (RLHF)

Evolution of LLMs

- Perceptron (1960s)
- Feedforward n/w and backpropagation (1984)
- Recurrent Neural Net (RNN; 80s)
- Attention (2015)
- Transformers (2017)
- LLMs (GPTs 2020s)

But language models are very old!! 2000 years- the notion of grammar

Course Info

Topics to be covered

- Single Neuron, perceptron and sigmoid; application to NLP; text classification
- Multilayered FFNN, Backpropagation; Softmax Application to NLP; Multiclass NLP problems
- Recurrent Neural Net (RNN); seq2seq
- Recursive Neural Net; Application to NLP Parsing
- Convolutional Neural Nets; Multimodal NLP
- Transformers; Application to MT, QA, NLG
- LLMs and application

Major Topics covered in CS626, last sem

- NLP and Ambiguity
- POS Tagging
- Parsing
- Maths for NLP
- Linguistics for NLP
- Machine Translation and MT Evaluation
- Medical NLP
- Hypothesis Testing

Course: Basic Info

- Slot 9: Monday 3.30, and Thursday 3.30
- TA Team: Sravanthi Settaluri, Abisek R.K, Tejpal, Tathagata, Sameer, Pavan, Sanyam, Tejomay
- https://www.cfilt.iitb.ac.in/~cs772_2024/
- Channels of communication: MS Teams, Moodle, Course Website
- Teams Name: cs772-2023 Joining code: op38ybr
- Also, visit moodle frequently

Evaluation Scheme (tentative)

- ~50%: Reading, Thinking, Comprehending
 - Quizzes (3 nos., 10%), midsem (15%),
 - Endsem (25%)
- ~50%: Doing things, Hands on
 - Assignments (20%)
 - Course Project (30%)

Question pattern

- Quizzes: only MCQ
- Endsem and midsem
- TWO subjective questions- only one page
- MCQs always on Moodle

Assignments and Project

- Continuous evaluation
- Meeting every two weeks to monitor progress
- Credit for thorough literature survey for the project work

Books

- Ian Goodfellow, Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press, 2016.
- Dan Jurafsky and James Martin, Speech and Language Processing, 3rd Edition, 2019.
- Christopher Manning and Heinrich Schutze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.

Books (2/2)

- Pushpak Bhattacharyya, Machine Translation, CRC Press, 2017.
- Pushpak Bhattacharyya and Aditya Madhav Joshi, *Natural Language Processing*, Wiley, 2023.

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

What can DL-NLP do

Modeling

**$P(e)$: “language”
model**

$$\begin{aligned} e^* &= \arg \max_e P(e \mid f) \\ &= \arg \max_e [P(e)P(f \mid e)] \end{aligned}$$

- Dialogue Act Classification (DAC): $f \rightarrow$ *Dialogue Sequence*, $e \rightarrow$ *Dialogue turn labels*
- Dialogue Intent: $f \rightarrow$ *dialogue sequence*, $e \rightarrow$ *dialogue turns with Intent like ‘question’, ‘elaboration’, ‘affirmation’, ‘command/request’ etc.*

Example of sequence of dialogue and dialogue acts

Fragment of a labeled conversation (from the Switchboard corpus).

Speaker	Dialogue Act	Utterance
A	YES-NO-QUESTION	So do you go to college right now?
A	ABANDONED	Are yo-,
B	YES-ANSWER	<i>Yeah,</i>
B	STATEMENT	<i>it's my last year [laughter].</i>
A	DECLARATIVE-QUESTION	You're a, so you're a senior now.
B	YES-ANSWER	<i>Yeah,</i>
B	STATEMENT	<i>I'm working on my projects trying to graduate [laughter].</i>
A	APPRECIATION	Oh, good for you.
B	BACKCHANNEL	<i>Yeah.</i>
A	APPRECIATION	That's great,
A	YES-NO-QUESTION	um, is, is N C University is that, uh, State,
B	STATEMENT	<i>N C State</i>
A	SIGNAL-NON-UNDERSTANDING	What did you say?
B	STATEMENT	<i>N C State</i>

Digression: What if there are many “abandoned”s

- Raise a flag
- Linguistic Limitation
- Or, Mental Health Problem
- Mental health doctors would like help of chatbots that can give preliminary help to mental health patients by engaging in a dialogue
- Or, they can do a preliminary screening based on disfluencies, abandoned statements and so on

Dialogue Act Classification and Sentiment

T. Saha, A. Patra, S. Saha and P. Bhattacharyya, *Towards Emotion-aided Multi-modal Dialogue Act Classification*, Association of Computational Linguistics Conference (**ACL 2020**), Seattle USA, 5-10 July, 2020.