COMP 554 / CSDS 553 Advanced NLP

Faizad Ullah

About Me

- Faizad Ullah
- Ph.D. Student at LUMS
- **Specialization**
  - Natural Language Processing (NLP)
  - Machine Learning
  - Data Science
- **Contributions**
  - Text Analytics of Low-Resource Languages
  - Medical Image Analysis
  - Graph Analysis

Grading

Quizzes	20%
Assignments	10%
Midterm	20%
Final Term	30%
Project	20%

Programming Tasks

- *3-5 Assignments
  - Programming Assignments
- *One Project
- Programming Environment
  - Python (Pytoch, TensorFlow, Colab)

*Vivas will be conducted for assignments and the project

Policies

- ❑ Sharing
  - Copying is not allowed for assignments. Discussions are encouraged; however, you must submit your own work.
  - Violators would be reported to the Disciplinary Committee or face marks reduction penalties
- ❑ Plagiarism
  - Do NOT pass someone else’s work as your own!
  - Write in your own words and cite the reference if you use someone else’s material.

Policies (2)

- ❑ Submission Policy
  - Submissions are due at the day and time specified
  - Late submissions will result in 10% marks deduction per day from obtained marks.
- ❑ Attendance Policy
  - You are advised to attend all lectures.
  - It's the students' responsibility to recover any information or announcements posted during a lecture from which they were absent.
- ❑ Classroom behavior
  - Maintain classroom sanctity by remaining attentive
  - Asking questions is encouraged.
  - You are not allowed to use a Laptop/mobile phone, etc., during class.

Policies (3)

- Retakes
  - No retakes for quizzes, assignments, exams, or projects
  - In case of any medical emergency or unavoidable circumstances, inform before hand and seek a formal approval. You need to share medical reports for departmental record.
  - **Do not wait for the final exam to seek approval for retakes**

Contact

❑ How to contact me?

- E-mail: faizadullah@fccollege.edu.pk
- Office: 426-G
- Office Hours: Mentioned on office door

Most Important

Don’t be afraid of giving wrong answers!

Let’s start our NLP journey…

Key Areas of NLP

- **Text Processing & Understanding**
  - Tokenization (splitting text into words or sentences), Part-of-Speech Tagging (identifying nouns, verbs, etc.)
  - Named Entity Recognition (extracting names, locations, organizations)
- **Machine Translation**
  - Google Translate, DeepL, and other language translation models
- **Speech Recognition**
  - Voice assistants like Siri, Alexa, and Google Assistant
- **Sentiment Analysis**
  - Detecting emotions in text (positive, negative, neutral)
- **Chatbots & Conversational AI**
  - AI-powered assistants (e.g., ChatGPT, customer support bots)
- **Text Generation**
  - Automated writing tools, AI-generated content
- **Information Retrieval & Search**
  - Search engines like Google understanding user queries
- **Summarization**
  - Extracting key points from long texts (news, reports, articles)

Natural Language Processing

- Study of computational approaches for processing natural languages
  - Processing: acquire, represent, store, understand, characterize etc.
  - Natural Languages: Human Languages

- Other names:
  - Computational Linguistics (CL)
  - Human Language technologies (HLT)

Question Answering

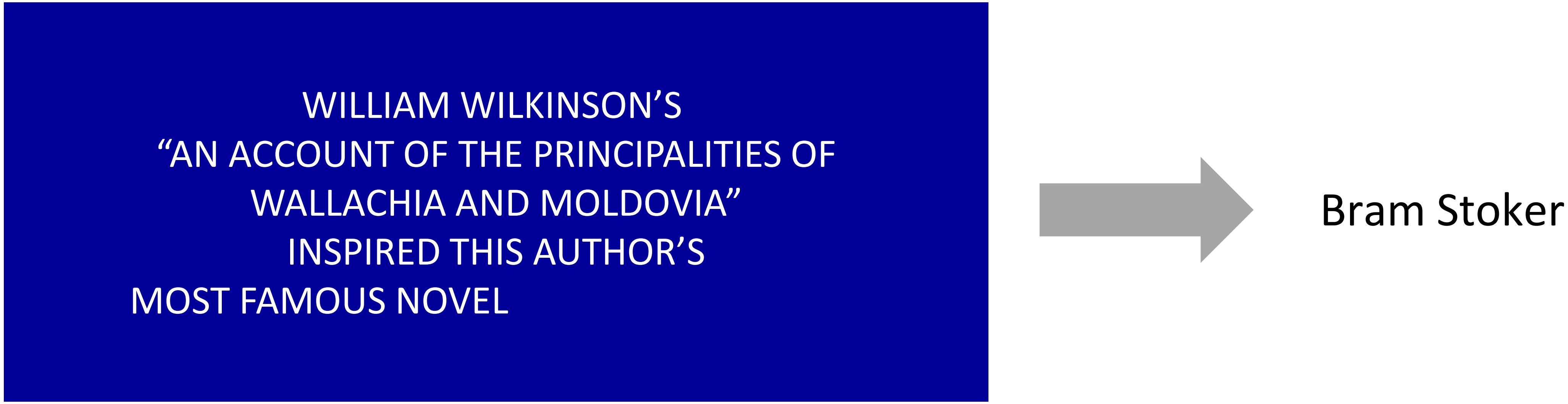
- What is the capital of France?
- Answer

- Is water composed of hydrogen and oxygen?
- Answer

- What is your age?
- Answer

Question Answering: IBM’s Watson

- Won Jeopardy on February 16, 2011!



WILLIAM WILKINSON’S
“AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA”
INSPIRED THIS AUTHOR’S
MOST FAMOUS NOVEL

Bram Stoker

Information Extraction

Event:	**FYP Part-A meeting**
Date:	**Feb-10-2025**
Start:	**10:00am**
End:	**11:30am**
Where:	**S-125**

Subject: **FYP Part-A Meeting**

Date: February 10, 2025

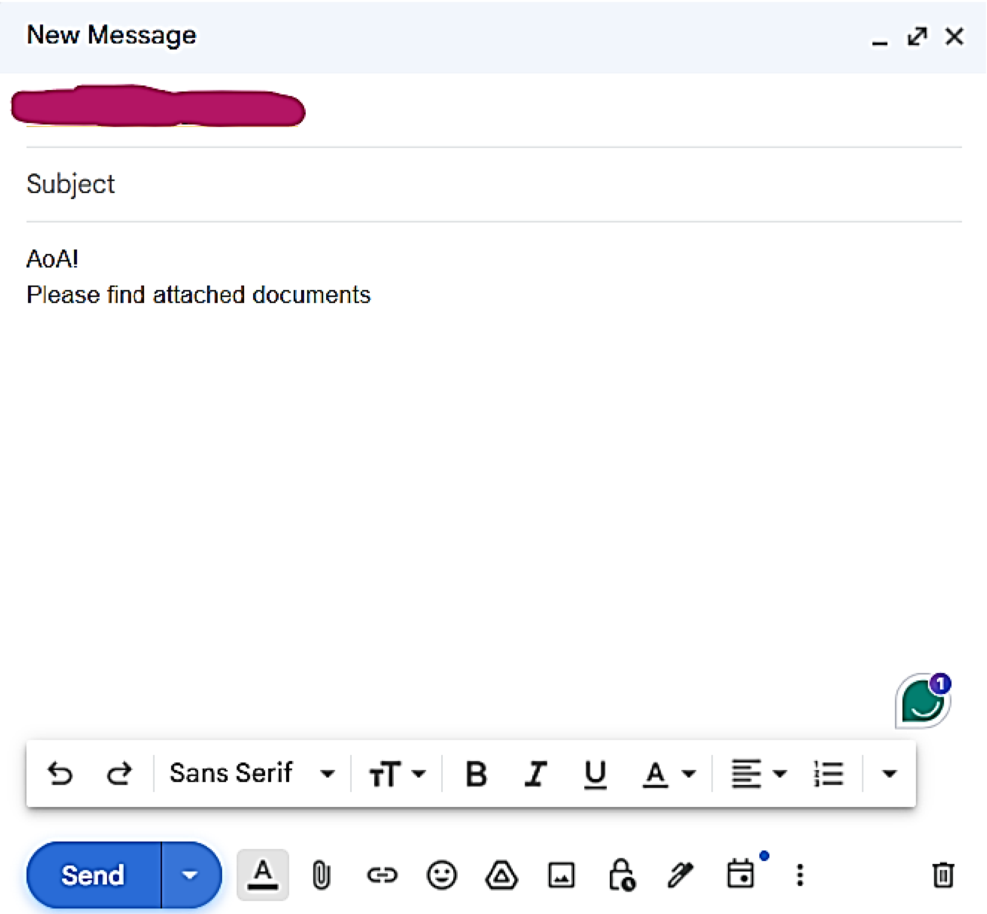
To: Faizad Ullah

Hi Sir, we would like to meet with you to discuss our FYP Part-A presentations. We've scheduled a meeting for tomorrow at S-125 from 10:00 AM to 11:30 AM. Looking forward to your guidance!

Best regards,

Create new Calendar entry

Information Extraction

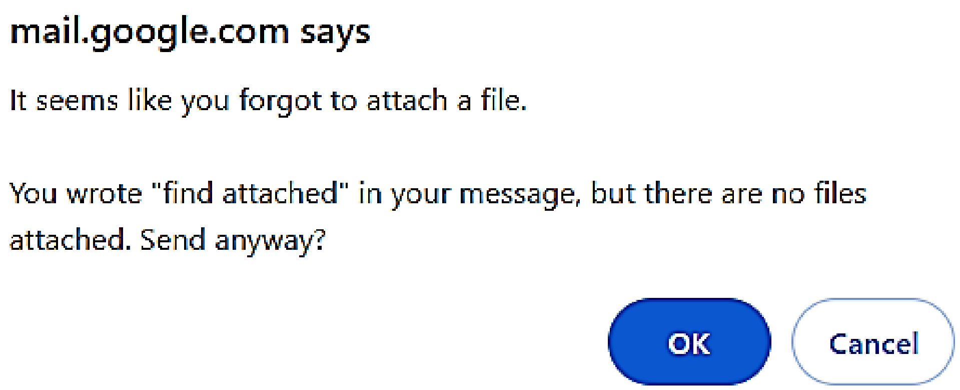


New Message

[redacted]

Subject

AoA!
Please find attached documents

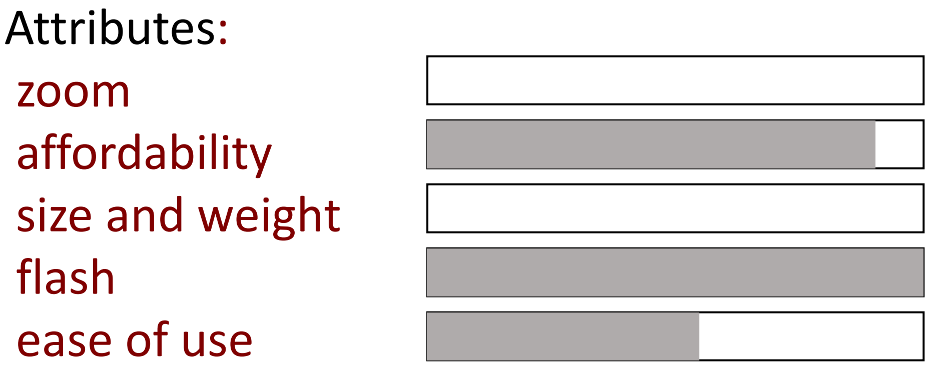


mail.google.com says

It seems like you forgot to attach a file.

You wrote "find attached" in your message, but there are no files attached. Send anyway?

Information Extraction & Sentiment Analysis



Attributes:

zoom

affordability

size and weight

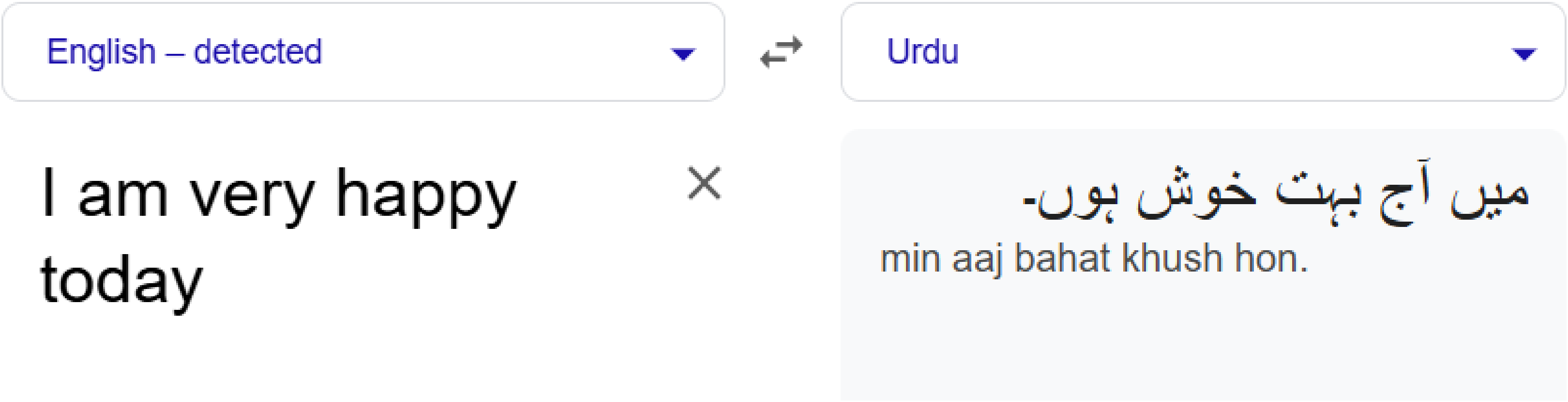
flash

ease of use

Size and weight

- ✓ nice and compact to carry!
- ✓ since the camera is small and light, I won't need to carry around those heavy, bulky professional cameras either!
- ✗ the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

Machine Translation



English – detected

Urdu

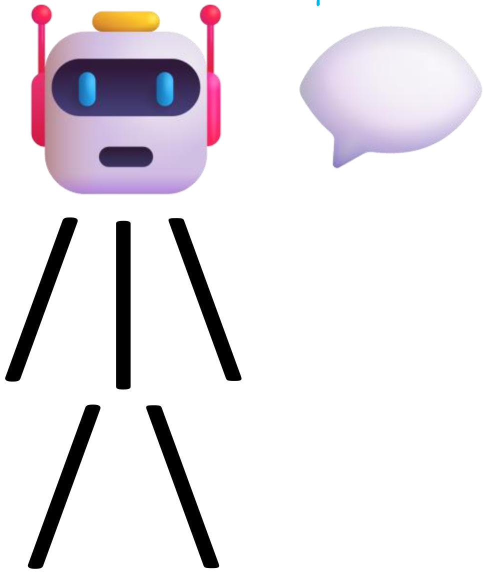
I am very happy today

میں آج بہت خوش ہوں-

min aaj bahat khush hon.

Chatbots

Chatbot is the UI of the future



Language Technology

mostly solved

Spam detection

Let's go to Agra! ✓

You won $100,000 … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON		ORG		LOC
Einstein	met with	UN	officials in	Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

→

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

→

Economy is good

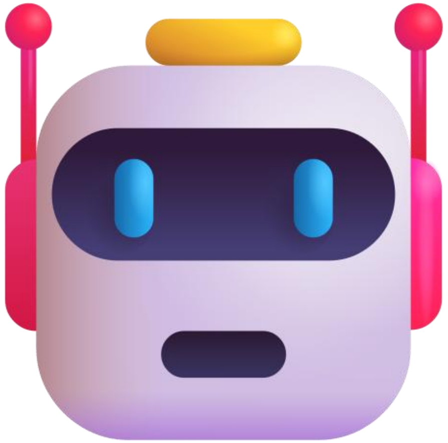
Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

2013 slides from the Stanford University

Can AI Think Like Us?



(1) 🤖 ➡️ 📖 🔍 ❓ vs (2) 🧠 ➡️ 💡 🤔 💬

(1) 🤖 (AI) tries to ➡️ process language (📖 🔍) but still struggles with meaning (❓).

(2) 🧠 (Human) naturally ➡️ understands concepts (💡 🤔) and engages in meaningful conversation (💬).

Break a leg

Good luck!

Hit the nail on the head

Get something exactly right.

Piece of cake

Something very easy.

Spill the beans

Reveal a secret.

Under the weather

Feeling sick.

Bite the bullet

Endure a difficult situation.

The ball is in your court

It’s your turn to decide.

Let the cat out of the bag

Reveal a hidden secret.

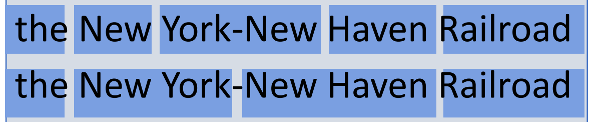
What makes NLU hard?

🤖 ➡️ 📖 🔍 ❓ vs 🧠 ➡️ 💡 🤔 💬

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues



the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing …
Let It Be was recorded …
… a mutation on the *for* gene …

What tools are Important

- Knowledge about language
- Knowledge about the world
- A way to combine knowledge sources

- Probabilistic Models (Language Models) built from language data:
  - P("Forman Christian" -> "College") high
  - P("University College" - > "Forman") low

Linguistics

Linguistics

- Linguistics is the study of languages with respect to its form or structure, meaning, and context.

- Linguistics also deals with the social, cultural, historical, and political factors that influence languages, including their origins and evolution.

- A linguist is a person knowledgeable in linguistics.

Phonetics & Phonology (Sound Patterns)

- Concerned with the sounds of speech, which is important for speech recognition and text-to-speech (TTS) systems.

- Consider the words *“night”* and *“knight”*.

- They are **homophones** (same sound but different meanings).

- A **speech recognition system** must correctly interpret the word based on context.

Morphology

- Studies the structure of words and how they are formed (e.g., prefixes, suffixes, root words).

- This is useful for tokenization and stemming in NLP.

- The words **"running"**, **"runs"**, and **"ran"** share the root word **"run"**.
  - **Stemming** reduces words to their base form:
    - "running" → "run"
    - "happily" → "happi"
  - **Lemmatization** does a more sophisticated reduction based on meaning:
    - "ran" → "run"
    - "better" → "good"

Syntax

- Examines the structure of sentences and grammar rules (e.g., parsing sentences for grammatical correctness).

    1. I am very happy today. ✅ (Correct)
    2. Happy am today I very. ❌ (Incorrect)

- Syntax rules help in POS (Part-of-Speech) tagging

Semantics

- Deals with the meaning of words and sentences, crucial for tasks like machine translation, sentiment analysis, and question-answering.

- The word **"bank"** can mean:
    - **Financial institution** → *"I deposited money in the bank."*
    - **Riverbank** → *"He sat by the bank of the river."*

- An NLP system needs **Word Sense Disambiguation (WSD)** to understand the correct meaning based on context.

Pragmatics

- Focuses on context and how meaning changes depending on the situation, vital for chatbot responses and human-like interactions.

- **"Can you pass the salt?“**
  - Literally, it’s a **yes/no** question.
  - In **pragmatics**, it’s actually a **request**, meaning **“Please pass me the salt.”**

- Chatbots must understand **intent**, not just words.

Discourse Analysis

- Studies how sentences and words connect in longer texts, improving coherence in machine-generated text and summarization tasks.

- **Ali** went to the store. **He** bought some milk.

- “**He**” refers to “**Ali”**, but an NLP model must infer that based on discourse context.

Real-World Example: Google Search

- When you search: **“Why is she eating an apple quickly?”**, NLP techniques help improve search results by applying linguistic concepts:
  - **Morphology** – Google recognizes that *"eating"*, *"eat"*, and *"eats"* are related.
  - **Syntax** –*"she"* is the subject, *"eating"* is the action, and *"an apple"* is the object.
  - **Semantics** – It understands the intent: You are likely looking for reasons why someone eats fast (e.g., hunger, habits).
  - **Pragmatics** – If you meant *"Why do people eat apples quickly?"*, Google may show articles on **health benefits of apples**.
  - **Discourse Analysis** – If you searched *"Why is she eating an apple?"* after searching *"Hunger and eating speed,"* Google considers previous searches to refine results.

Sub-fields of Linguistics

- Historical linguistics
- Cultural linguistics
- Political linguistics
- Social linguistics

- Psycho-linguistics
- Bio-linguistics
- Neuro-linguistics
- Computational linguistics

Grammar

- Rules guiding the composition of clauses, phrases, and words in a language
    - **Clause:** part of a sentence that contain subject and verb.
    - **Phrase:** group of words (that plays a specific role) in a sentence but does not typically represent a complete sentence.
    - **Syntax:** primarily shapes the grammar, but grammar can be influenced by morphology, phonology, and pragmatics as well.

Lexicon

- Collection of words or lexical units in a language
    - Dictionary

Part-of-Speech (POS)

- Category of words that have similar properties and grammatical functions (usage in a sentence)

- Common POS in English: Noun, Verb, Adjective, Adverb, Pronoun, Preposition, Conjunction, and Interjection

Named Entity

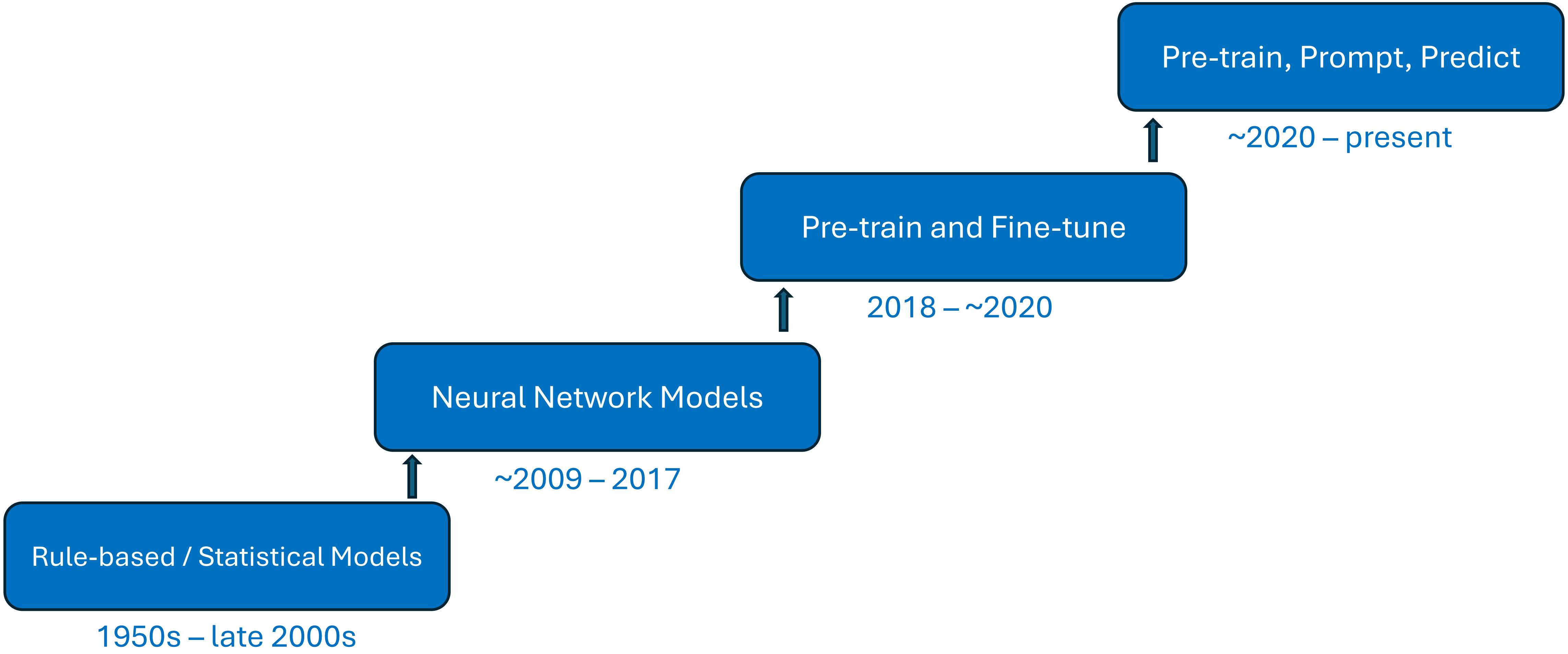
- Entities of specified types (named)
- Person: e.g., Ali
- Location: e.g., Lahore
- Organization: e.g., FCCU
- Date: e.g., 21/02/2025
- Etc

Translation, Transliteration

- Translation: convert from one language to another preserving meaning
- Transliteration: convert from one script to another of a specific language, e.g., Urdu in Perso-Arabic script to Urdu in Roman script

Paradigm Shifts in NLP

Paradigm Shifts in NLP



Pre-train, Prompt, Predict

~2020 – present

Pre-train and Fine-tune

2018 – ~2020

Neural Network Models

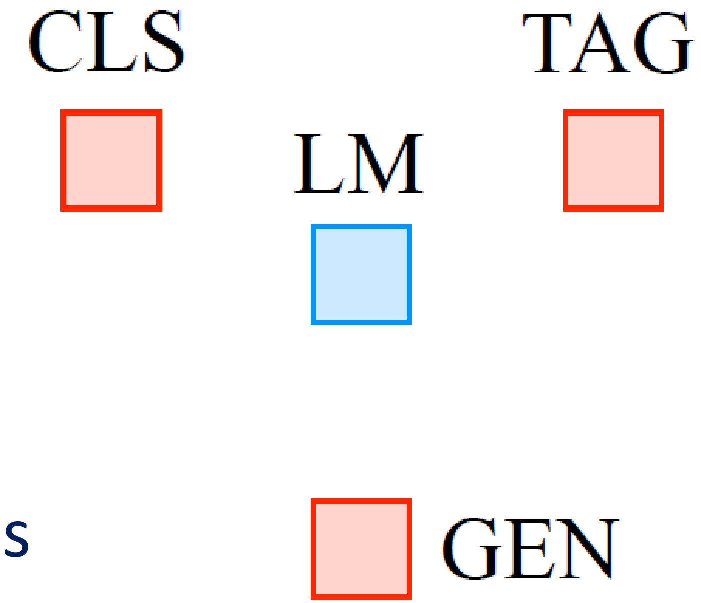
~2009 – 2017

Rule-based / Statistical Models

1950s – late 2000s

Traditional ML Models

- Relied on Feature Engineering (FE)
- Domain knowledge and expertise required
- Task specific datasets
- Insufficient data for quality/generalized models

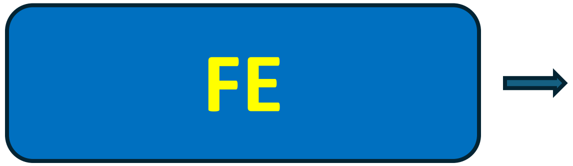


CLS

TAG

LM

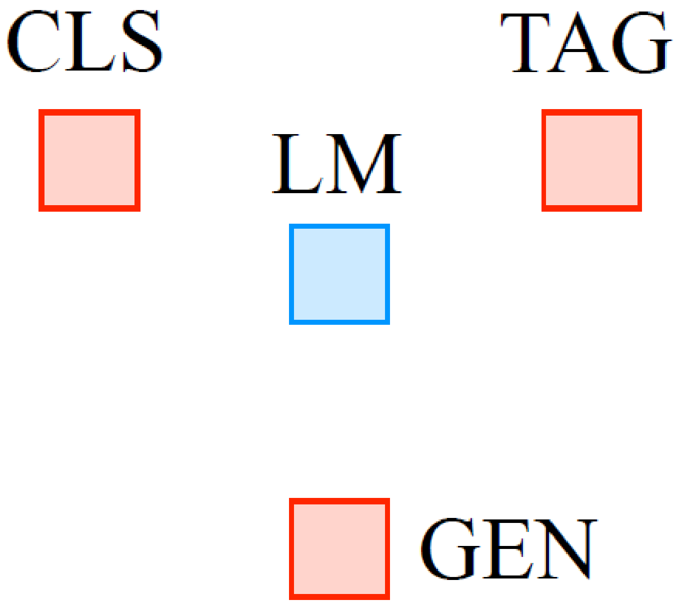
GEN



FE

Neural Network Models

- Features → Architecture Engineering (AE)
- Inductive bias provided → architecture
- Learning features → dataset

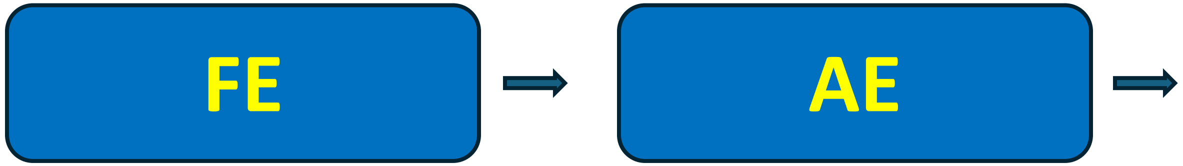


CLS

LM

TAG

GEN

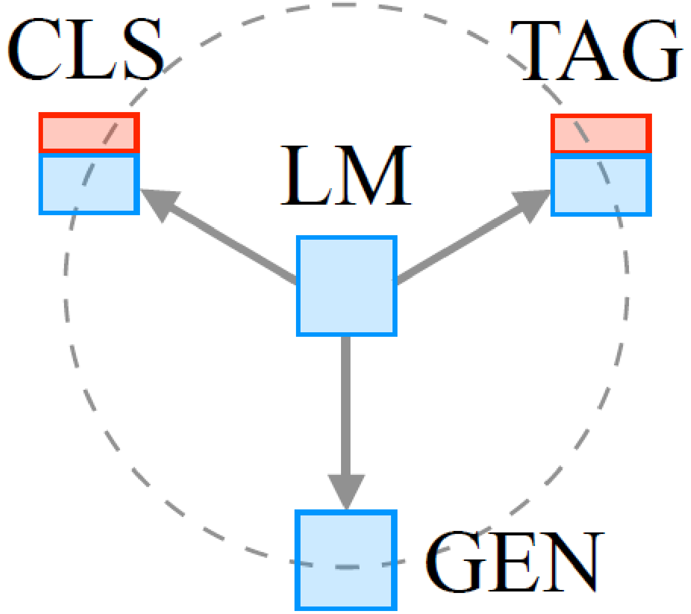


FE

AE

Pre-train and Fine-tune

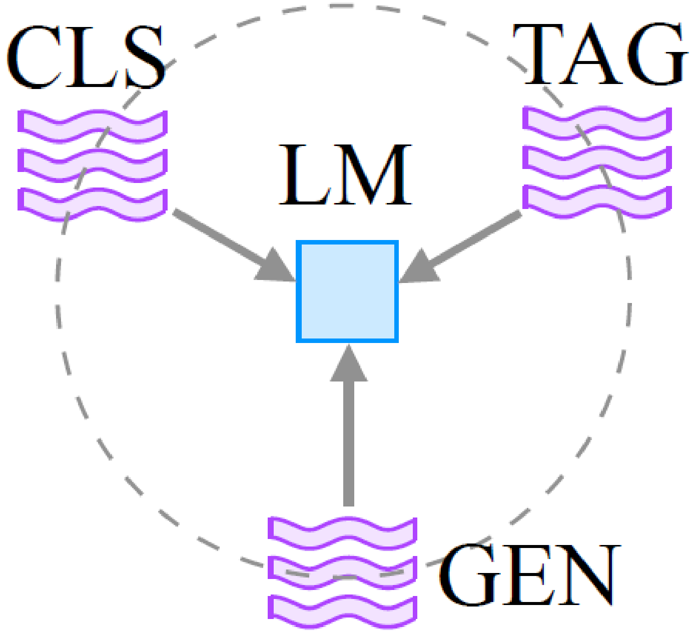
- Pre-trained Language Model (PLMs)
- Fixed architecture, Objective Engineering (OE)
- Easily adapted to downstream NLP tasks
- Standard / Vanilla Fine-tuning



FE → **AE** → **OE** →

Pre-train, Prompt, and Predict

- In-context learning
- Downstream tasks with the help of prompt
- Prompt Engineering (PE)



FE → AE → OE → PE

Key Trends

- Learn a language from large corpora of text
  - No labels are required; just try to predict words in natural language
- Language modeling is driving modern day NLP
  - Traditional probabilistic language models (n-grams) to modern deep learning based models (transformers)
  - Transformer race: GPT, T5, BERT, Turing NLG, …
- Feature representation and end-to-end learning
  - Integrate corpus and knowledge-based information from raw textual input to final desired outcome
- Transfer learning: transfer knowledge in the form of representations from related data
  - Learn rich representations for linguistic units (e.g., word embeddings)
  - Learn entire models (pre-training) on related tasks and adapt them to new task (fine-tuning)

Confluence of Fields

- Statistics and Probability
- Machine Learning / Artificial Intelligence
- Data Structures and Algorithms
- Linguistics
- Psychology

Basic Text Processing

Text

- Text is a sequence of characters arranged in a particular order.

- I am very happy today.

Regular Expressions

- A formal language for specifying text strings
- How can we search for any of these?
  - apple
  - apples
  - Apple
  - Apples

Disjunctions

- Letters inside square brackets []

Pattern	Matches
[aA]pple	apple, Apple
[1234567890]	Any digit

- Ranges [A-Z]

Pattern	Matches	
[A-Z]	An upper case letter	Drenched Blossoms
[a-z]	A lower case letter	my beans were impatient
[0-9]	A single digit	Chapter 1: Down the Rabbit Hole

Negation in Disjunction

- Negations [^Ss]
  - Caret means negation only when first in []

Pattern	Matches	
[^A-Z]	Not an upper case letter	How are you?
[^Ss]	Neither ‘S’ nor ‘s’	I have no exquisite reason
[^e^]	Neither e nor ^	Look here
\^	Looking for a caret ^	Look up a^b now

The Pipe “|” Symbol: More Disjunction

- Woodchucks is another name for groundhog!
- The pipe | for disjunction

Pattern	Matches
groundhog|woodchuck	
yours|mine	yours mine
a|b|c	= [abc]
[gG]roundhog|[Ww]oodchuck	

Regular Expressions: ? * + .

Kleene *, Kleene +

Pattern	Matches	
colou?r	Optional previous char	color colour
oo*h!	0 or more of previous char	oh! ooh! oooh! ooooh!
o+h!	1 or more of previous char	oh! ooh! oooh! ooooh!
baa+		baa baaa baaaa baaaaa
beg.n		begin begun beg3n

Anchors ^ $

^ start of a line, $ end of a line

Pattern	Matches
`^[A-Z]`	`Palo Alto`
`^[^A-Za-z]`	`Hello`
`\.$`	`The end.`
`!$`	`The end!`

Example

- Find me all instances of the word “the” in a text.

  `the` → Misses capitalized examples

  `[tT]he` → Incorrectly returns `other` or `theology`

  `[^a-zA-Z][tT]he[^a-zA-Z]`

Errors

- The process we just went through was based on fixing two kinds of errors
  - Matching strings that we should not have matched (there, then, other)
    - False positives (Type I)
  - Not matching things that we should have matched (The)
    - False negatives (Type II)

Errors

- In NLP we are always dealing with these kinds of errors.
- Reducing the error rate for an application often involves two antagonistic efforts:
    - Increasing accuracy or precision (minimizing false positives)
    - Increasing coverage or recall (minimizing false negatives).

Sources

- https://web.stanford.edu/~jurafsky/slp3/2.pdf
- https://web.stanford.edu/~jurafsky/slp3/3.pdf
- **Machine Learning for Intelligent Systems**, Kilian Weinberger, Cornell, Lectures 3-6, https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecture_note03.html
- **Prof. Mitesh M. Khapra** (https://www.cse.iitm.ac.in/~miteshk/) on NPTEL's (http://nptel.ac.in/) Deep Learning course (https://onlinecourses.nptel.ac.in/noc18_cs41/preview)
- **Perceptrons. An Introduction to Computational Geometry. Marvin Minsky and Seymour Papert. M.I.T. Press, Cambridge, Mass., 1969.** https://science.sciencemag.org/content/165/3895/780