CSDX 236 NATURAL LANGUAGE PROCESSING

Module 1:Origins and challenges of NLP – Language Modeling: Grammar-based LM, Statistical LM - Regular Expressions, Finite-State Automata – English Morphology, Transducers for lexicon and rules, Tokenization, Detecting and Correcting Spelling Errors, Minimum Edit Distance.

By

Mrs. H. FAHEEM NIKHAT , AP/CSE

- Natural Language Processing (NLP) is a subfield of artificial intelligence (AI) that focuses on the interaction between computers and human language. It aims to enable machines to understand, interpret, generate, and respond to human language effectively.

Scope of NLP:

- NLP spans tasks such as text analysis, machine translation, speech recognition, and conversational AI.

Importance of NLP:

- Facilitates communication between humans and machines.
- Powers many real-world applications like chatbots, search engines, and voice assistants.

- Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP) are all related but distinct fields within the broader domain of AI

Machine Learning (ML):

- **What it is:** ML is a subfield of AI that focuses on creating algorithms that allow machines to learn patterns from data and make decisions or predictions without being explicitly programmed.
- **Core Concept:** Models learn from data to generalize and apply knowledge to new, unseen data.

Examples: Linear Regression, Decision Trees, and Support Vector Machines.

- Applications: Predicting house prices, detecting spam, recommendation systems.

Deep Learning (DL):

- **What it is:** DL is a specialized subfield of ML that uses artificial neural networks with multiple layers to process large amounts of data. It is particularly effective for complex, high-dimensional problems.
- **Core Concept:** Models learn representations of data through layers of abstraction, mimicking the structure of the human brain.

- **Examples:**Convolutional Neural Networks (CNNs) for images.
- Recurrent Neural Networks (RNNs) and Transformers for sequences (like text).
- Applications: Image recognition, speech recognition, and NLP tasks.
- **Relationship with ML:** DL is a subset of ML and becomes particularly useful when working with large datasets.

- **Natural Language Processing (NLP):**
- **What it is:** NLP focuses on enabling machines to understand, interpret, and generate human language. It intersects linguistics and computer science.
- **Core Concept:** Process and analyze text or speech data to enable interaction with machines using natural language.
- **Examples:**Tokenization, sentiment analysis, machine translation, and chatbots.
- Applications: Language translation (Google Translate), virtual assistants (Siri, Alexa), text summarization.
- **Relationship with ML and DL:**
  - NLP tasks often rely on ML models (like logistic regression or SVM) or advanced DL models (like Transformers or LSTMs) for better performance.

- ML is the broadest field; DL is a subset of ML; NLP applies ML/DL techniques to language data.

- Learn ML → DL → NLP for a solid, progressive understanding.

Introduction

- Natural language processing (NLP) is a field of
  - Computer science
  - Artificial intelligence
    - Intelligence exhibited by Machines, mimics the cognitive minds
  - Computational linguistics
    - All about language--- kind of expertise in Natural Languages

Can we define?

- The first task is to understand the definition!!
- **NLP is defined as the ability of a machine (i.e. computer program) to understand and interpret the human language as it is spoken!**
- It can be seen as an **"AID provided to computers to understand the human languages!"**
- Now comes the question.
    - Is it not easy to teach computers the human languages??
        - Certainly not!! It is tough, daunting task!

Contd.,

- We are humans and we can speak the languages that we know. Be it English, Tamil or Hindi.
- But, what a computer understands is Machine Language and is certainly not for humans. Humans can't understand BINARY!! 00001010010101010101 = This could be representing Good Morning!! We can't understand.
- This is the real state of computers. But, imagine now!
- We are talking to Google Assistant, Alexa etc. We say "Alexa play the music" and it does. Alexa, order food! It takes it up!! Ok Google, call my friend Sachin! It calls Sachin.
- How is this possible? This interaction is made possible by NLP !! But, NLP is not stand alone, it has Machine learning and Deep Learning Supporting it!

Contd.,

- Why is it important? Is it challenging?
- First, the computers can now talk to humans, in the native language!
- Computers can now hear, interpret, understand and can even meas sentiment through the tone and respond back.
- This will be very consistent and very fast. Also, since done my machine would be without Bias.
- Considering the volume of data we generate everyday, yes, it is challenging.

Contd.,

- The user should talk to the machine.
- The machine gets the audio recorded
- The audio gets converted to Text
- The data is processed now. (here is where the machine understands your language, NLP. ML plays handy here!)
- Then the response – AUDIO or Text comes out!! If it is a chatbot, it would be just text reply. If it is an Audio bot, it would talk to you as well.
- For an instance, ALEXA will reply you with an answer. IRCTC chat bot will give you text reply!

Contd., - The aim ...

- NLP aims to get computers to perform useful tasks involving human language, tasks like enabling human- machine communication, processing of text or speech
- To get the Human and Computer interaction more natural!

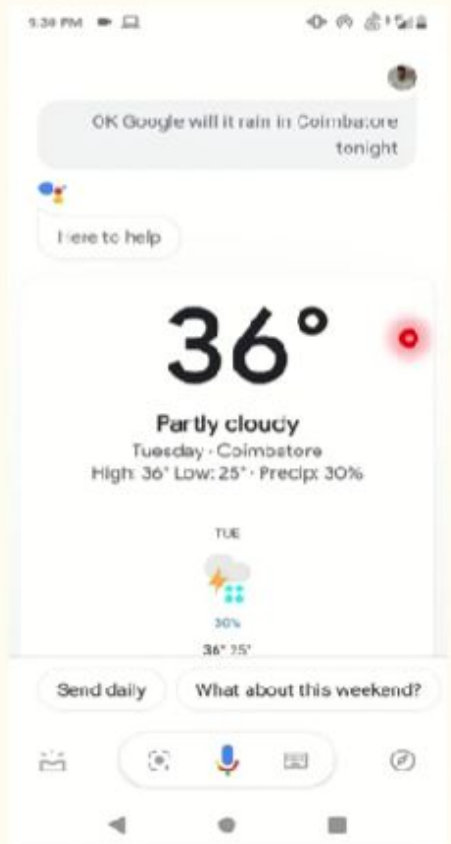
Two Aspects/Goals

- Scientific Goal - Can we make computer understand a human speaking language?
- Practical / Engineering goal – We are living in the data fueled earth. Can we make use of all the data and get real time applications built to constructively help people and to save their time as well?

Real-time Examples.. You and I use them all!

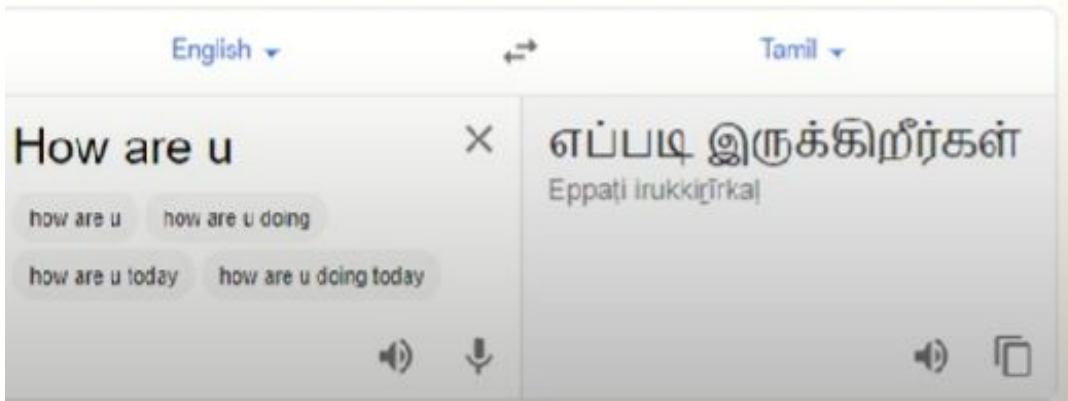
- Google Assistant is a virtual personal assistant developed by Google allows two way conversations. Can we test?
- OK Google! Can you tell me what's the temperature outside?
- OK Google! Can you let me know what's my name?
- OK Google! Will it rain in Coimbatore tonight?

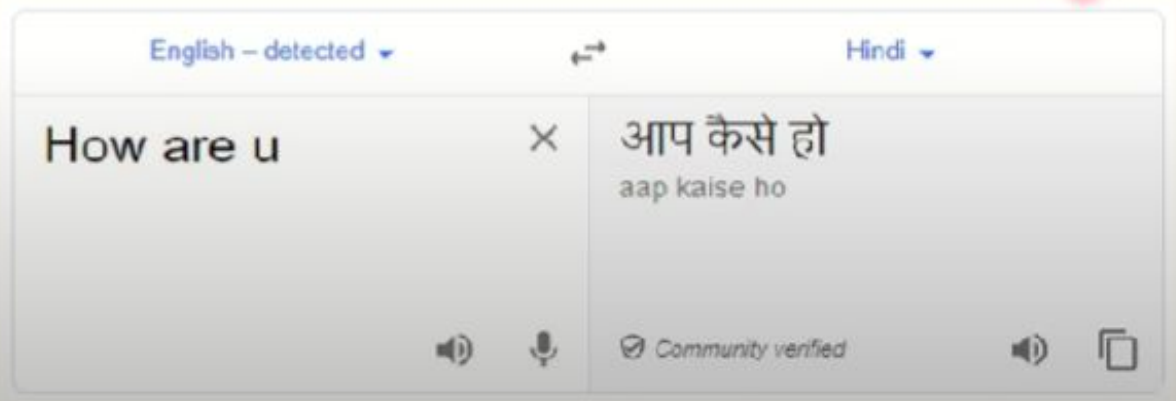
Google Assistant



Language Translation

- This is a 100% NLP – Translation of one language to other.
- Google Translate is the most used example for you and me to understand this better!



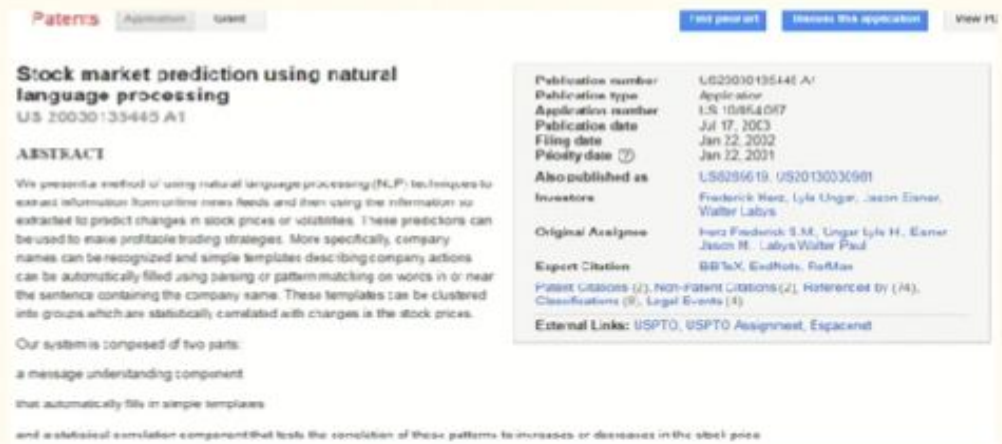


NLP helps in fighting spam. Yes, this is true.

- NLP is useful in detection of mails/msgs as spam or not!
- We shall see a demo for this in near future. (Again, with Python!)

Information Extraction

- The body of the text has to be processed, so that the same can be entered to Relational Database or Analyzed with Data Mining techniques.
- One simple example would be to extract structured relation triple from the plain text as shown below:
- **Dr.Kalam is greatest of all Indian scientists** would get a triple as "Dr.Kalam; is greatest; of Indian Scientists;
- Another instance of information extraction application is the **stock market** prediction from the **tweets** and **reviews.**
- All these are applications of NLP!



Patents

Stock market prediction using natural language processing

US 20030135445 A1

ABSTRACT

We presentia method of using natural language processing (NLP) techniques to extract information from online news feeds and then using the information so extracted to predict changes in stock prices or volatilities. These predictions can be used to make profitable trading strategies. More specifically, company names can be recognized and simple templates describing company actions can be automatically filled using parsing or pattern matching on words in or near the sentence containing the company name. These templates can be clustered into groups which are statistically correlated with changes in the stock prices.

Our system is composed of two parts:

a message understanding component

that automatically fills in simple templates

and a statistical correlation component that tests the correlation of these patterns to increases or decreases in the stock price.

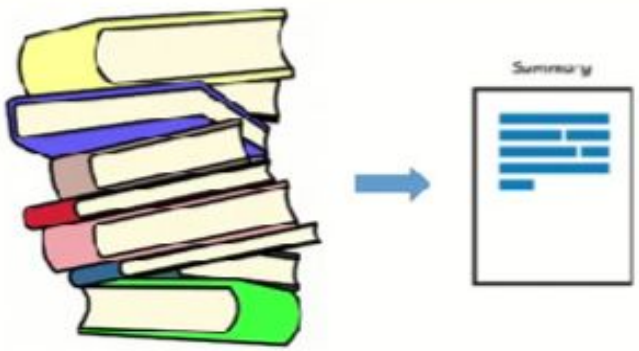
Publication number	US20030135445 A1
Publication type	Application
Application number	US 10/064,057
Publication date	Jul 17, 2003
Filing date	Jan 22, 2002
Priority date	Jan 22, 2001
Also published as	US6826619, US20030033981
Inventors	Frederick Hirz, Lyle Unger, Jason Eisner, Walter Labys
Original Assignee	Hirz Frederick S.M., Unger Lyle H., Eisner Jason M., Labys Walter Paul
Export Citation	BiBTeX, EndNote, RefMan

Patent Citations (2), Non-Patent Citations (2), Referenced by (14), Classifications (8), Legal Events (4)

External Links: USPTO, USPTO Assignment, Espacenet

Text Summarization

- The main idea of summarization is to find a subset of data which contains the "**information**" of the entire set.
- Text summarization refers to the technique of shortening long pieces of text.
- The intention is to create a coherent and fluent summary having only the main points outlined in the document. (Ignoring the unimportant content.)
- Example: I can ask you the review for a book, in a couple of lines you could give a opinion!!



In my language, it is like pitching for a hackathon! The whole product has to be explained in 1 min ☺

Question and Answering

- Any intelligent Chatbots, AI engines which can answer questions from the users is an example! NLP is vital here. IBM Watson is said have performed equal/better than humans in quiz.
- There are challenges, we shall learn them shortly!
- Siri, tell me, whom do I ring frequently?



Origins and challenges of NLP

Evolution of Natural Language Processing

Linguistics, AI, and Computer Science Roots

The foundational disciplines that contributed to NLP

1940s-1950s Machine Translation Efforts

Early rule-based approaches for translating languages

1960s-1970s Chomsky's Grammar Influence

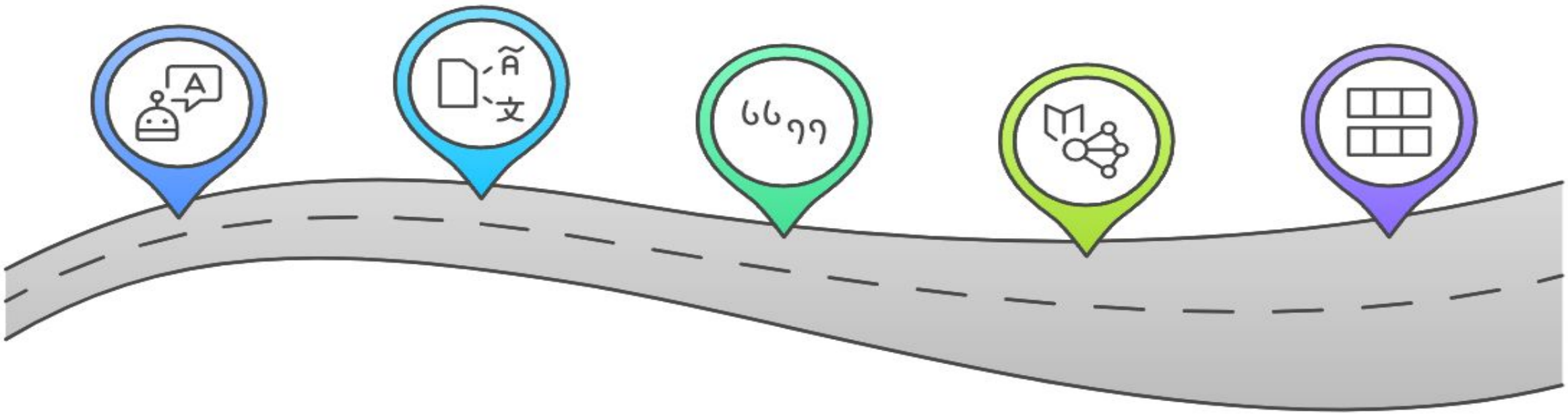
Chomsky's theories shaped computational linguistics

1980s-1990s Machine Learning Rise

Introduction of statistical methods like HMMs

2000s-Present Deep Learning Advances

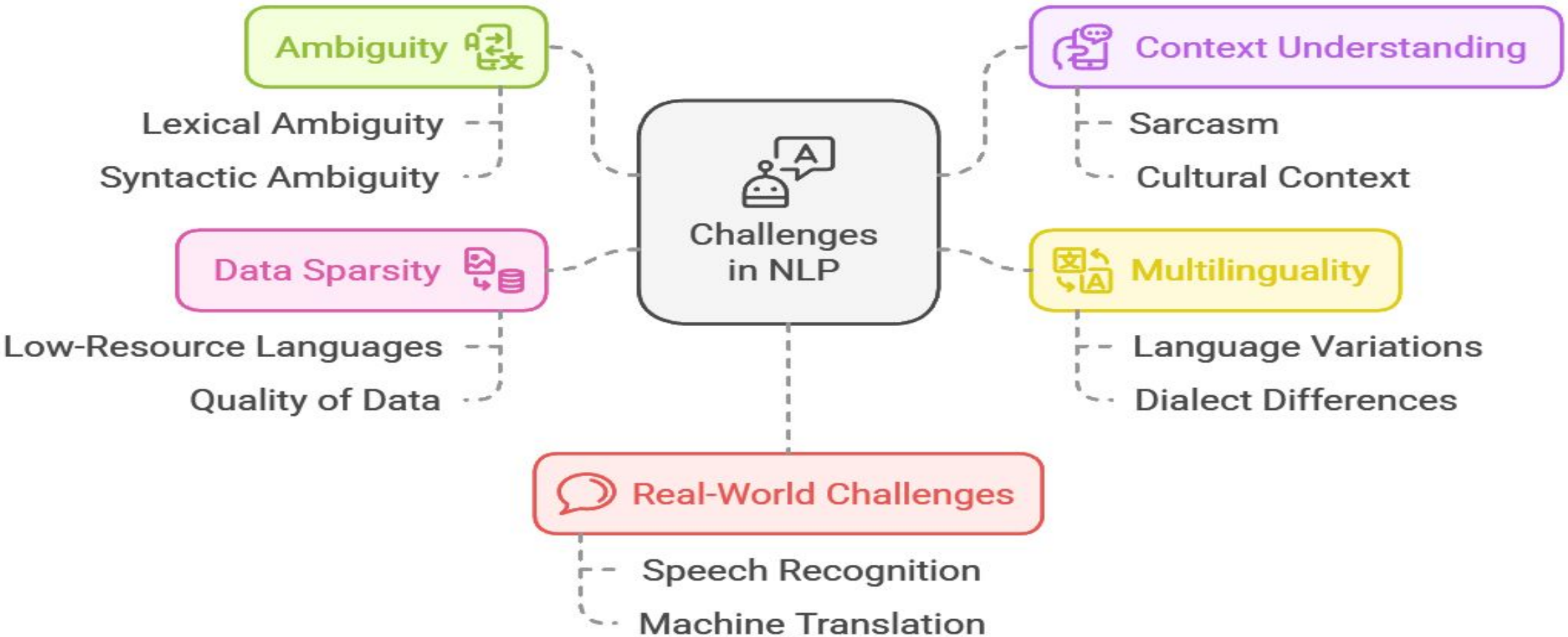
Development of models like GPT and BERT



Origins of NLP

- NLP has its roots in **linguistics**, **artificial intelligence (AI)**, and **computer science**.
- **1940s-1950s:** Early efforts included **rule-based approaches** for machine translation (e.g., translating Russian to English during the Cold War).
- **1960s-1970s:** The emergence of **Chomsky's grammar theories** influenced computational linguistics, focusing on syntactic structures. Early NLP relied heavily on symbolic methods.
- **1980s-1990s:** The rise of **machine learning (ML)** introduced statistical approaches, like Hidden Markov Models (HMMs), for tasks like speech recognition and part-of-speech tagging.
- **2000s-Present:** Advances in **deep learning (DL)** (e.g., Recurrent Neural Networks, Transformers) have revolutionized NLP, enabling models like GPT and BERT to process and generate human language with remarkable accuracy.

Challenges in Natural Language Processing



NLP faces unique challenges due to the complexity and ambiguity of human

1. Ambiguity:

Lexical ambiguity: Words with multiple meanings (e.g., "bank" as a financial institution vs. riverbank).

Syntactic ambiguity: Multiple possible parses for a sentence (e.g., "I saw the man with a telescope").

Possible Parses (Interpretations):

You used the telescope to see the man:**Parse:** *(I [saw [the man] [with a telescope]])*

The man has the telescope:**Parse:** *(I [saw [the man with a telescope]])*

How NLP Models Handle Ambiguity:

- **Parse Trees:** NLP models generate parse trees that show possible sentence structures.
- **Probabilistic Models:** Statistical models (e.g., probabilistic context-free grammars) assign probabilities to different parses and choose the most likely one.
- **Deep Learning:** Modern models like Transformers (e.g., GPT, BERT) use context from surrounding words or sentences to resolve ambiguity. For example:*"I saw the man with a telescope. It was a powerful telescope."* → Suggests you used the telescope.

  "I saw the man with a telescope. He was adjusting its focus." → Suggests the man had the telescope.

2. Context Understanding:

Handling nuances like sarcasm, idioms, or slang.
Understanding the role of cultural or situational context.

Sarcasm:

Text:

"Oh, just perfect! The internet is down again."

Literal Meaning:

The speaker is genuinely saying that it’s perfect.

Sarcastic Meaning:

The speaker actually means that the situation is frustrating or inconvenient, the opposite of "perfect."

Idioms:

Idioms are phrases whose meanings are not derived from their literal words.

Example:

Text: *"It's raining cats and dogs."*

Challenge: NLP needs to recognize this as heavy rain, not a literal downpour of animals.

Slang:

Slang refers to informal, often region-specific language.

Example:

Text: *"That party was lit!"*

Challenge: "Lit" means exciting or fun in slang, which may differ from its dictionary meaning.

Cultural or Situational Context

Cultural and situational context impacts interpretation.
Example:

Text: *"Let's grab a coffee after the meeting."*

Challenge: In some cultures, this might imply a casual chat; in others, it could signal a formal discussion.

How NLP Handles These Challenges(Context Understanding):

1. **Contextual Models:** Transformers like GPT and BERT use large datasets to learn nuanced language patterns.
2. **Sentiment Analysis Tools:** These tools look at surrounding sentences for clues about sarcasm or emotion.
3. **Idiomatic Dictionaries:** Models may rely on specialized datasets for idioms or slang.
4. **Fine-Tuning:** Custom training on culture-specific data improves model understanding.

3. Multilinguality

Challenge:
Adapting NLP models to handle multiple languages, dialects, and script variations is difficult because:

- Languages have different grammar, syntax, and vocabulary.
- Dialects and regional variations add complexity.
- Scripts vary widely (e.g., Latin, Cyrillic, Devanagari).

Solutions:

- **Multilingual Models:** Models like mBERT and XLM-R are pre-trained on data from multiple languages, enabling cross-lingual understanding.
- **Transfer Learning:** Fine-tuning models on specific languages using smaller datasets.
- **Zero-Shot Learning:** Training models in high-resource languages and generalizing to low-resource ones.
- **Script Normalization:** Converting scripts into a unified representation (e.g., Unicode).

4. Data Sparsity

Challenge:
Many languages, especially low-resource ones, lack high-quality labeled datasets for tasks like sentiment analysis or translation.

Solutions:

- **Synthetic Data Generation:** Creating pseudo-labeled data using data augmentation techniques.
- **Crowdsourcing:** Involving native speakers to label datasets.
- **Unsupervised Methods:** Leveraging unlabeled data using unsupervised learning or self-supervised methods.
- **Transfer Learning:** Using knowledge from high-resource languages and adapting it to low-resource ones.

5. Real-World Challenges

Speech Recognition:

- **Problem:** Struggles with accents, regional pronunciations, or noisy environments.
- **Solution:** Training models on diverse datasets with varying accents and background noise. Real-time noise reduction algorithms can also help.

Machine Translation:

- **Problem:** Difficulty with creative expressions, idioms, and cultural nuances.
- **Solution:** Incorporating **contextual embeddings** (e.g., Transformers) to better understand sentence-level meaning. Fine-tuning translation models with culturally specific data.

THANKYOU

Ex: Text preprocessing in NLP

- **Perform the Following Steps:**Convert all text to **lowercase**.
- **Remove punctuation and special characters**.
- **Remove stopwords** (like *the, is, in, and*).
- **Tokenize** the text into words.
- **Lemmatize** or **stem** the words to their root form.

Implement the preprocessing using **NLTK** or **spaCy**.

- import nltk
- import spacy
- import string
- from nltk.corpus import stopwords
- from nltk.tokenize import word_tokenize

- # Download necessary NLTK resources
- nltk.download('punkt')
- nltk.download('stopwords')

- # Load spaCy model for lemmatization
- nlp = spacy.load("en_core_web_sm")
- # Sample text
- text = """Natural Language Processing (NLP) is a fascinating field of Artificial Intelligence (AI)!
- It enables computers to understand human language, whether written or spoken."""

- # Step 1: Convert text to lowercase
- text = text.lower()

- # Step 2: Remove punctuation
- text = text.translate(str.maketrans('', '', string.punctuation))

- # Step 3: Tokenization
- tokens = word_tokenize(text)

- # Step 4: Remove stopwords
- stop_words = set(stopwords.words('english'))
- filtered_tokens = [word for word in tokens if word not in stop_words]

- # Step 5: Lemmatization using spaCy
- doc = nlp(" ".join(filtered_tokens))
- lemmatized_tokens = [token.lemma_ for token in doc]

- # Output Results
- print("Original Text:\n", text)
- print("\nTokenized Text:\n", tokens)
- print("\nFiltered Tokens (Stopwords Removed):\n", filtered_tokens)
- print("\nLemmatized Tokens:\n", lemmatized_tokens)

- Original Text:

natural language processing nlp is a fascinating field of artificial intelligence ai it enables computers to understand human language whether written or spoken

- Tokenized Text:

['natural', 'language', 'processing', 'nlp', 'is', 'a', 'fascinating', 'field', 'of', 'artificial', 'intelligence', 'ai', 'it', 'enables', 'computers', 'to', 'understand', 'human', 'language', 'whether', 'written', 'or', 'spoken']

- Filtered Tokens (Stopwords Removed):

['natural', 'language', 'processing', 'nlp', 'fascinating', 'field', 'artificial', 'intelligence', 'ai', 'enables', 'computers', 'understand', 'human', 'language', 'whether', 'written', 'spoken']

- Lemmatized Tokens:

['natural', 'language', 'processing', 'nlp', 'fascinating', 'field', 'artificial', 'intelligence', 'ai', 'enable', 'computer', 'understand', 'human', 'language', 'whether', 'write', 'speak']