

How Does NLP Work?

Two Key Steps in NLP:

1) Understanding Language (Natural Language Understanding - NLU):

Computers must break down human language into structures they can process.
This involves:

- **Tokenization:** Splitting text into words or smaller units (tokens) for analysis.
- **Part-of-Speech (POS) Tagging:** Labeling words based on their roles in a sentence.
 - (nouns, verbs, adjectives, etc.)
- **Parsing:** Analyzing sentence structure to understand the relationships between words.
- **Named Entity Recognition (NER):** Identifying proper nouns, dates, and other specific information.

FIU

Example: Order me a pizza. 

- Tokenization: ["Order", "me", "a", "pizza"]
- POS Tagging: "Order" (verb), "me" (pronoun), "a" (article), "pizza" (noun)
- Parsing: Subject: "me", Action: "order", Object: "pizza"
- NER: Recognizing "pizza" as a food item.

How Does NLP Work?

2) Generating Responses (Natural Language Generation - NLG):

After understanding the input, it needs to generate a response that makes sense:

- **Text Summarization:** Condensing information into a shorter version, highlighting key points.
- **Text Translation:** Converting text from one language to another while maintaining meaning and context.
- **Dialogue Generation:** Systems like chatbots use NLG to respond to user queries.

Example: Order me a pizza.



The system understands the request and generates a response like “Sure, I’ll order a pizza for you.”

NLP & AI Models

1) Rule-Based Systems (Early NLP):

- Predefined set of rules and linguistic patterns.
- If-then rules are applied to text to detect patterns or keywords.
 - User: "Hi!"
 - System: "Hello!"

2) Machine Learning Models (Statistical NLP)

- ML models are trained on large datasets to recognize patterns in text.
 1. Naive Bayes: Often used for text classification tasks like spam detection.
 2. Support Vector Machines (SVM): Used for tasks like sentiment analysis.
 - Gmail's spam filter

NLP & AI Models

3) Deep Learning Models (Neural Networks):

- Neural Networks, mimic the way the human brain processes information.
- Can automatically learn complex patterns from data
 - including understanding the context of words in a sentence.
 - RNN, LSTM
 - Google Translate

- ✓ handle massive amounts of unstructured text data.
- ✓ can capture context and meaning over longer passages of text.

- ❖ require vast amounts of data and computing power.
- ❖ harder to interpret the "decision-making" process.

NLP & AI Models

4) Transformers and Pre-trained Models (Modern NLP)

Use self-attention mechanisms to process all words in a sentence simultaneously

- highly effective for understanding very long text

1. BERT (Bidirectional Encoder Representations from Transformers):
 - A powerful pre-trained model that can be fine-tuned for various tasks like question answering, text classification, and NER.
2. GPT (Generative Pre-trained Transformer):
 - A model that excels in text generation, capable of generating coherent, contextually relevant text based on a given prompt.

- ✓ superior performance on a wide range of NLP tasks.
- ✓ can be fine-tuned for specific tasks with less data (no extensive retraining)
- ❖ produce biased or incorrect outputs due to biases in the training data.
- ❖ are computationally expensive to train and deploy.

How It Works - ChatGPT

For a text generator to be able to respond so many different prompts, it needs a massive diverse training set

