Natural Language Processing (NLP)

What is NLP?

Natural Language Processing (NLP) is a field at the intersection of linguistics, computer science, and artificial intelligence that enables machines to understand, interpret, and generate human language.

Example:

- Google Translate
- ChatGPT
- Alexa
- spam filters
- sentiment analysis.

NLU (Natural Language Understanding)

Definition:
NLU is the part of NLP that focuses on understanding and interpreting the meaning, structure, and intent behind human language.

Goal: Convert unstructured text into a machine-understandable format.

Key Sub-Tasks of NLU:

Task	**Description**	**Example**
Intent Recognition	Detecting the user's purpose	"Book me a flight" → Intent: BookFlight
Entity Recognition	Identifying important keywords or phrases	"Book a flight to *Delhi*" → Entity: Delhi
Named Entity Recognition (NER)	Extracting names of people, places, dates, etc.	"Meet *John* on *Friday*"
Part-of-Speech Tagging	Assigning word types (noun, verb, adj...)	"She runs" → She/PRONOUN, runs/VERB
Dependency Parsing	Understanding grammatical structure and word relations	"The boy kicked the ball" → Subject → Verb
Coreference Resolution	Resolving pronouns or repeated references	"John said he is coming" → "he" = "John"
Sentiment Analysis	Determining emotional tone	"I love this product" → Positive

Task	Description	Example
Text Classification	Assigning labels to a sentence or document	Spam or Not Spam
Semantic Role Labeling	Identifying roles like who did what to whom	"Mary gave John a book" → Mary=Giver

Tools and Models for NLU:

- **Rule-based**: Regex, spaCy patterns
- **ML-based**: SVM, Random Forest, CRF
- **DL-based**: LSTMs, BiLSTM+CRF
- **Transformer-based**: BERT, RoBERTa, DistilBERT

NLG (Natural Language Generation)

Definition:
NLG is the task of generating human-like text from structured or unstructured data.

Goal: Convert machine-readable data into natural language.

Key Sub-Tasks of NLG:

Task	Description	Example
Content Planning	Deciding what information to include	"User bought 3 items worth $60"
Sentence Planning	Organizing how content will be structured grammatically	"You purchased 3 items for $60."
Surface Realization	Generating the actual sentence from plan	"Thank you for your purchase of $60."
Text Summarization	Generating a short version of a long text	TL;DR of an article
Data-to-Text	Describing structured data in plain language	Weather: "23°C, cloudy" → "It is cloudy today."
Question Generation	Creating questions from context	Input: "Apple is a company." → "What is Apple?"
Conversational Responses	Generating replies in chatbots	User: "How's the weather?" → Bot: "It's sunny!"
Story Generation	Long-form creative writing	Tools like GPT, Claude, etc.

Tools and Models for NLG:

- **Templates**: Predefined sentence structures.
- **Statistical**: N-gram models, Markov chains.
- **Deep Learning**: RNNs, LSTM
- **Transformer-based**: GPT-2/3/4, T5, BART

Relationship Between NLU and NLG

Component	Function	Example
NLU	Understand input	User: "Book a flight to Delhi"
NLG	Generate a meaningful response	Bot: "Sure, when would you like to travel?"

In a chatbot:

- **NLU** identifies intent = `BookFlight`, entity = `Delhi`
- **NLG** constructs the response dynamically.

Example: Voice Assistant Workflow

```
User: "Remind me to call mom at 6 PM"

1. NLU:
  → Intent: CreateReminder
  → Entity: "call mom", "6 PM"

2. Dialogue Manager:
  → Fills reminder slot, asks for date if needed

3. NLG:
  → Response: "Okay, I’ll remind you to call mom at 6 PM."
```

Term	Full Form	Focus
NLU	Natural Language Understanding	**Interpreting** human input
NLG	Natural Language Generation	**Generating** human-like text

Together, **NLU + NLG** power:

- Chatbots
- Voice Assistants
- Virtual Agents
- Customer Support AI
- Smart Home Systems

Key Tasks in NLP

Task	**Description**	**Example**
Tokenization	Splitting text into words/tokens	"I love NLP" → ["I", "love", "NLP"]
Part-of-Speech (POS) Tagging	Identifying grammatical categories	"I love NLP" → [PRP, VBP, NNP]
Named Entity Recognition (NER)	Recognizing real-world entities	"Barack Obama was born in Hawaii" → ["PERSON", "GPE"]
Text Classification	Classifying documents or messages	Spam vs Not Spam
Sentiment Analysis	Detecting emotion in text	"I hate this!" → Negative
Text Summarization	Shortening a long document	News highlights
Machine Translation	Translating between languages	English → French
Question Answering	Answering based on a document	"Who is president of India?"
Chatbots	Interactive conversations	Support assistants

Text Preprocessing Techniques

Tokenization

Breaks text into individual words or sub-words.

- **Word Tokenization:** “I love NLP” → [“I”, “love”, “NLP”]
- **Sentence Tokenization:** Splits paragraph into sentences.

Lowercasing

Converts text to lowercase to avoid case mismatch.

"NLP" and "nlp" → both become "nlp"

Stopword Removal

Removes common words that carry little meaning (e.g., "the", "is", "in").

Stemming

Reduces words to their root form.

"playing" → "play"

Library: PorterStemmer, SnowballStemmer

Lemmatization

More intelligent root-word finding than stemming.

"better" → "good"

Punctuation & Special Character Removal

"I ❤NLP!!!" → "I NLP"

Text Representation (Vectorization)

Computers need numbers — here's how we convert text into numerical form:

Bag of Words (BoW)

Counts word frequency.

Text	**"NLP"**	**"is"**	**"fun"**
NLP is fun	1	1	1
NLP is powerful	1	1	0

TF-IDF (Term Frequency–Inverse Document Frequency)

Highlights **important words** in a document.

Formula:
TF = (No. of times term appears) / (Total terms)
IDF = log(Total docs / Docs containing term)
TF-IDF = TF × IDF

Word Embeddings

Context-aware vector representations.

- **Word2Vec** (Google)
- **GloVe** (Stanford)
- **FastText** (Facebook)
- **BERT Embeddings** (Transformers)

Example:
"king" - "man" + "woman" ≈ "queen"

Language Models

Language Models (LM) predict the next word in a sequence.

N-Gram Models

Probabilistic model based on n previous words.

Formula (Bigram):
P(w2 | w1) = Count(w1 w2) / Count(w1)

Limitation: Sparse and memory-intensive.

Neural Language Models

Use deep learning.

- **RNN (Recurrent Neural Networks)**
- **LSTM / GRU** – Handle long-term dependencies
- **Transformer (BERT, GPT)** – Modern and powerful

Key NLP Algorithms & Techniques

Text Classification

Used in spam filtering, sentiment analysis.

- **Naive Bayes**
  - Assumes features are independent.
  - **Formula (Bayes' Rule):**
    P(class|text) = [P(text|class) × P(class)] / P(text)
- **SVM**
- **Logistic Regression**
- **BERT-based classifiers**

Named Entity Recognition (NER)

Detects entities like PERSON, LOCATION, ORG.

> "Steve Jobs founded Apple" → ["PERSON", "ORG"]

POS Tagging

Assigns grammatical tags.

> "The dog barked" → [DET, NOUN, VERB]

Dependency Parsing

Identifies syntactic relationships between words.

> "He gave her a book"
> → subject: "He", object: "book", indirect object: "her"

Deep Learning in NLP

RNN (Recurrent Neural Networks)

Handles sequences.

Limitations: Vanishing gradients.

LSTM (Long Short-Term Memory)

Improves RNN by remembering long-term dependencies.

GRU (Gated Recurrent Unit)

Simpler, faster than LSTM with similar performance.

Attention Mechanism

Helps the model focus on important words.

In translation: “I am eating” → “Je mange”, focus on “eating” → “mange”.

Transformers

Introduced in "Attention Is All You Need" (Vaswani et al., 2017)

- **Architecture:** No recurrence, only attention
- **Used in:** BERT, GPT, T5, RoBERTa, XLNet

Key Transformer Models:

Model	**Use**
BERT	Bidirectional Encoder Representations (great for understanding)
GPT	Generative Pretrained Transformer (great for text generation)
T5	Text-to-Text Transfer Transformer
DistilBERT	Lightweight BERT
RoBERTa	Robustly Optimized BERT

Semantic Similarity

- Measuring how similar two pieces of text are (used in search engines, chatbots).
- Common metrics: **Cosine Similarity** (on word embeddings).

Formula:

$$\cos(\theta) = (A \cdot B) / (||A|| \ ||B||)$$

Multilingual NLP

- Handling multiple languages in one model (e.g., **mBERT**, **XLM-R**).
- Useful for global applications.

Zero-shot / Few-shot NLP

- NLP without fine-tuning on the target task.
- Example: GPT-4 can perform summarization even without training directly on it.

Prompt Engineering (for LLMs)

- Crafting prompts for better results from generative models like GPT, Claude, etc.
- Very relevant in **Generative AI + NLP**.

Conversational AI Pipelines

- Multi-turn chatbots (e.g., Rasa, Dialogflow)
- Concepts: intent recognition, slot filling, dialogue management

Topic Modeling

- Unsupervised technique to discover hidden topics in a collection of documents.
- Popular algorithms:
  - **LDA (Latent Dirichlet Allocation)**
  - **NMF (Non-negative Matrix Factorization)**

Speech-related NLP

Speech-related NLP focuses on converting spoken language into text and vice versa, enabling natural communication between humans and machines.

ASR (Automatic Speech Recognition)

- **Definition**: Technology that converts human speech into written text.
- **Working**:
  1. **Audio Input** → captured via microphone.
  2. **Preprocessing** → noise reduction, normalization.
  3. **Feature Extraction** → MFCCs, spectrograms.
  4. **Acoustic Model** → maps audio features to phonemes.

5. **Language Model** → predicts the most likely word sequence.

- **Applications**:
  - Voice assistants (Alexa, Siri)
  - Meeting transcription
  - Voice-controlled systems

TTS (Text-to-Speech)

- **Definition**: Technology that converts text into natural-sounding speech.
- **Working**:
  1. **Text Analysis** → splitting into sentences, identifying pronunciation.
  2. **Phonetic Conversion** → converting words to phonemes.
  3. **Prosody Generation** → rhythm, stress, intonation.
  4. **Waveform Generation** → neural vocoders (WaveNet, HiFi-GAN).
- **Applications**:
  - Accessibility tools for visually impaired
  - Audiobook generation
  - Interactive voice bots

Multimodal NLP

Multimodal NLP combines **text** with other modalities such as **images, audio, or video** for richer understanding.

Examples:

- **CLIP** (OpenAI) → Understands images + captions.
- **GPT-4V** → Can process text + images for reasoning.
- **Speech2Text + Image Captioning** → Useful for accessibility.

Applications:

- Content moderation in videos
- Educational tools combining speech + visuals
- Video search with natural language queries

Ethics & Privacy

When processing **audio and text data**, especially in a grammar-scoring or voice-analysis app, ethics and compliance are critical.

GDPR (General Data Protection Regulation) Implications:

- **Data Minimization**: Collect only necessary audio/text data.

- **User Consent**: Must obtain explicit permission before recording.
- **Right to Erasure**: Users can request deletion of their data.
- **Data Security**: Store audio files and transcriptions securely with encryption.
- **Anonymization**: Remove identifiable information from stored text/audio.

Other Considerations:

- **Bias in ASR/TTS**: Ensure models work well across accents and dialects.
- **Misuse Prevention**: Avoid generating harmful or deepfake audio.
- **Transparency**: Inform users how their data will be used.

Evaluation Metrics for NLP

Task	**Metric**	**Formula/Meaning**
Classification	Accuracy, Precision, Recall, F1	Measure correctness of predictions
NER, POS, etc.	F1-score	Balances Precision & Recall
Summarization	ROUGE (Recall-Oriented Understudy for Gisting Evaluation)	Compares overlap of n-grams
Translation	BLEU (Bilingual Evaluation Understudy)	Measures match with reference translations

Popular NLP Libraries & Tools

Purpose	**Libraries**
General NLP	NLTK, spaCy
Deep NLP	Hugging Face Transformers, Flair
Preprocessing	TextBlob, gensim
Word Embeddings	Word2Vec, FastText
OCR	Tesseract, EasyOCR
Datasets	HuggingFace Datasets, Kaggle

NLP Applications

Industry	**Use Cases**
Healthcare	Clinical report summarization, chatbot assistants
Finance	Document classification, sentiment analysis
E-commerce	Product search, reviews analysis

Industry	**Use Cases**
EdTech	Essay scoring, AI tutors
Legal	Contract summarization
Social Media	Trend analysis, content moderation

Challenges in NLP

- **Ambiguity**: "I saw her duck" — what does "duck" mean?
- **Sarcasm**: "Oh great, another Monday!" → tone matters
- **Multilingual NLP**
- **Low-resource languages**
- **Bias & fairness in language models**
- **Context understanding**

Future of NLP

- Multilingual Transformers (mBERT, XLM-R)
- Low-resource NLP & Zero-shot learning
- Real-time translation
- Emotion & sentiment-aware assistants
- Ethical and explainable NLP
- Vision + Language models (e.g., GPT-4V)

Summary Checklist

Stage	**What to Learn**
Basics	Tokenization, stopwords, stemming
Representation	TF-IDF, Word2Vec, BERT
Modeling	Classification, NER, POS
Deep NLP	LSTM, Attention, Transformers
Applications	Chatbots, translation, summarization
Evaluation	F1, BLEU, ROUGE
Tools	spaCy, HuggingFace, NLTK