

Introduction to Natural Language Processing

Introduction to Natural Language Processing (NLP)

Key concepts, goals, and foundational techniques

1. What Is NLP?

Aspect	Description
Definition	A sub-field of artificial intelligence that enables computers to understand, interpret, and generate human language.
Scope	Text & speech processing, sentiment analysis, machine translation, question answering, chatbots, information extraction, summarization, etc.
Why It Matters	Bridges the gap between human communication and digital data, powering applications from search engines to virtual assistants.

2. Core Goals of NLP

Goal	Typical Tasks	Example
Understanding	Parsing, part-of-speech tagging, named-entity recognition (NER)	Identify “Apple” as a company in a news article.
Generation	Text summarization, machine translation, dialogue generation	Produce a concise summary of a research paper.
Interaction	Question answering, sentiment analysis, intent detection	Detect user intent in a chatbot message.
Knowledge Extraction	Relation extraction, event detection, knowledge-base population	Pull out “CEO of” relationships from corporate filings.

3. NLP Pipeline – Typical Stages

1. Data Collection

- Sources: web crawls, APIs, corpora (e.g., Wikipedia, Common Crawl).
- Formats: raw text, PDFs, audio (speech → text).

Introduction to Natural Language Processing

2. Pre-processing

- **Tokenization** – split text into words, sub-words, or characters.
- **Normalization** – lower-casing, removing punctuation, expanding contractions.
- **Stop-word removal** – optional for some tasks (e.g., topic modeling).
- **Stemming / Lemmatization** – reduce words to base forms.

3. Feature Extraction

- **Bag-of-Words (BoW)** – word counts / TF-IDF.
- **Word embeddings** – Word2Vec, GloVe, FastText (static).
- **Contextual embeddings** – BERT, RoBERTa, GPT (dynamic per token).

4. Modeling

- Classical ML: Naïve Bayes, SVM, CRF (sequence labeling).
- Deep Learning: RNN/LSTM, CNN, Transformer architectures.

5. Post-processing & Evaluation

- Decode outputs (e.g., beam search for generation).
- Metrics: accuracy, F1, BLEU, ROUGE, perplexity, exact match.

4. Fundamental Techniques

Technique	Brief Explanation	Typical Use-Cases
Regular Expressions	Pattern matching on character strings.	Simple entity extraction, data cleaning.
Part-of-Speech (POS) Tagging	Assign grammatical categories (noun, verb, etc.).	Syntactic parsing, feature creation.
Named-Entity Recognition (NER)	Detect proper nouns (people, locations, organizations).	Information extraction, knowledge graphs.
Dependency & Constituency Parsing	Build tree structures showing grammatical relations.	Semantic role labeling, relation extraction.
Word Embeddings	Dense vector representations capturing semantic similarity.	Text classification, clustering, similarity search.

Introduction to Natural Language Processing

Technique	Brief Explanation	Typical Use Cases
Transformers	Self-attention mechanism; enables parallel processing of sequences.	State-of-the-art language models (BERT, GPT).
Sequence-to-Sequence (Seq2Seq)	Encoder-decoder models for mapping input sequences to output sequences.	