Artificial Intelligence (Natural Language Processing (NLP))

Prof K R Chowdhary

CSE Dept., MBM University

February 24, 2025
Lecture #8

Introduction to NLP

NLP Definition: NLP is a field of AI focused on the interaction between computers and human language, enabling machines to understand, interpret, and generate human language.

Why is NLP important? • Large amounts of unstructured data (e.g., text data from documents, social media, customer reviews).

• Applications across various industries like healthcare, customer service, engineering, and finance.

• *Challenges:* Ambiguity, variability in language, context understanding, and handling slang.

Text Data in NLP

Types of Text Data: *Structured vs. unstructured text data*: Structured: XML, databases. Unstructured: Natural Language text.

Pre-processing: Raw text is often cleaned and preprocessed before analysis (e.g., removing stop words, punctuation, or special characters).

NLP Workflow

Text acquisition → tokenization → part-of-speech tagging → named entity recognition → syntactic parsing → semantic analysis.

- *Real-world example*: Extracting useful insights from unstructured data, such as classifying civil engineering project reports or analyzing customer service data.

Basic NLP Concepts

a) *Tokenization*

- *Definition*: Tokenization is the process of splitting text into smaller chunks, typically words or sentences.
- *Types of Tokenization:*

Word Tokenization: Splitting text into individual words.

Sentence Tokenization: Splitting text into individual sentences.

- Example: Tokenize a sample sentence: "The robot repaired the engine." → ["The", "robot", "repaired", "the", "engine"].

Tokenization, Parts of Speech (POS)

• *Challenges*: Handling punctuation, contractions (e.g., "don't" → "do" + "n't"), and multilingual text.

b) Part-of-Speech (POS) Tagging

• *Definition*: POS tagging assigns a grammatical category (noun, verb, adjective, etc.) to each word in a sentence.

• Example: "The robot repaired the engine."

'The" → Determiner, "robot" → Noun, "repaired" → Verb, "the" → Determiner, "engine" → Noun.

• *Applications*: POS tagging is used in information extraction, text summarization, and machine translation.

• *Challenges:* Ambiguities in word usage, such as "lead" (noun) vs. "lead" (verb).

c) *Named Entity Recognition* (NER)

• *Definition:* NER is used to identify entities in text, such as names of people, organizations, locations, dates, etc.

NLP Applications, Named entity recognition (NER)

- *Example:* In the sentence "Ram Dev works at XYZ Pvt Ltd since 2020", the entities identified would be:

"Ram Dev" $\rightarrow$ Person

"XYZ Pvt Ld" $\rightarrow$ Organization

"2020" $\rightarrow$ Date

- *Applications:* NER is widely used in document categorization, sentiment analysis, and even legal text extraction.
- *Challenges:* Ambiguities in naming (e.g., distinguishing between persons, places, and other entities like "Apple").
- *d) Combining Techniques*

How tokenization, POS tagging, and NER are typically combined to create more complex text processing pipelines.

- *Example:* Using all techniques to process engineering reports for extracting specific information such as project milestones, technical terms, or engineers' names.

NLP Applications in Engineering

3. Applications of NLP in Engineering

- *Objective:* How NLP is applied in real-world engineering contexts, focusing on applications specific to civil, mechanical, and electrical engineering?
- a) Document Classification in Civil Engineering Reports

Overview: NLP is used to automatically classify engineering documents into categories (e.g., project reports, safety inspections, etc.).

- *Approach:*

Use text classification techniques (e.g., Naive Bayes, SVM, neural networks) to categorize documents based on their content.

Preprocess text (tokenization, stopword removal, and feature extraction).

Pre-process, Classification, Chatbots

- Example: Classifying civil engineering reports into categories like "Structural Reports", "Safety Reports", or "Environmental Reports".
- Challenge: Handling domain-specific terminology and technical language.
- b) Chatbots for Customer Service in Mechanical Engineering
- Overview: NLP-driven chatbots are widely used to automate customer support, reducing response time and increasing efficiency.
- Approach: Build a conversational agent that uses intent recognition and entity extraction to understand user queries.

Integrate with knowledge bases and FAQs to provide answers.

- Example: A chatbot for a mechanical engineering company helping customers troubleshoot common issues with equipment.

Chatbots, automated fault detection

- *Tools:* Intent recognition using NLP models (e.g., BERT or GPT), entity recognition for specific terms like part numbers or technical specifications.
- c) Automated Fault Detection Systems in Electrical Engineering
- Overview: NLP techniques can be applied to analyze logs or reports from electrical engineering systems to automatically detect and diagnose faults.
- Approach: Use text classification and sequence tagging to extract fault information from unstructured text logs.

Analyze patterns, such as repeated error messages or specific terminology associated with faults.

NLP Tools

- Example: An automated system that scans fault detection reports and categorizes issues, such as “power failure”, “overload”, or “sensor malfunction”.
- Challenges: Handling diverse error messages and maintaining the accuracy of automated detection.
- 4. NLP Tools (NLTK and spaCy)

These are NLP libraries and demonstrate basic usage.

a) NLTK (Natural Language Toolkit)

- Overview: NLTK is a widely used library in Python for text processing and NLP tasks.
- Features: Tokenization, POS tagging, NER, stemming, and lemmatization.

NLP Tools, NLTK, spaCy

- *Example:* Using NLTK to tokenize text, POS tagging, and extract named entities.
- Simple Code Example: Tokenize and perform POS tagging on a sample sentence.
- 1.Natural Language Processing (NLP)

b) spaCy

- *Overview*: spaCy is another powerful NLP library designed for large-scale information extraction tasks.
- *Features*: Tokenization, POS tagging, NER, dependency parsing, and more.
- Example: Show how to use spaCy for faster and more efficient NLP tasks compared to NLTK.

Simple Code Example: Load a spaCy model, process text, and perform named entity recognition.

NLP Applications

Natural Language Processing (NLP) has a wide range of applications across various industries and fields. Here are some of the key applications:

1. *Machine Translation*: Automatically translating text or speech from one language to another (e.g., Google Translate, DeepL).

2. *Sentiment Analysis*: Analyzing text to determine the sentiment expressed, such as positive, negative, or neutral. Commonly used in social media monitoring, customer reviews, and brand sentiment analysis.

3. *Chatbots and Virtual Assistants*: NLP powers virtual assistants like Siri, Alexa, and Google Assistant, enabling them to understand and respond to natural language queries.

4. *Speech Recognition*: Converting spoken language into written text (e.g., voice-to-text applications, transcription services).

NLP Applications

5. *Text Summarization*: Automatically generating a shorter version of a given text while retaining key information. This can be extractive (pulling key sentences) or abstractive (generating new content).

6. *Information Retrieval*: Search engines use NLP to process queries and retrieve relevant results. It helps in ranking and understanding the intent behind search queries.

7. *Question Answering Systems*: NLP models can be trained to answer questions posed in natural language, such as in AI-driven help desks, customer support, or research databases.

8. *Named Entity Recognition (NER)*: Identifying and classifying key elements in text such as names of people, locations, organizations, and dates.

NLP Applications

9. *Text Classification*: Categorizing text into predefined categories (e.g., spam detection, topic categorization, and email sorting).

10. *Autocorrection and Text Prediction*: Used in text messaging, email applications, and other writing interfaces to predict and correct typing errors.

11. *Language Generation*: Creating human-like text for creative writing, chatbots, or even news articles (e.g., GPT models for writing).

12. *Information Extraction*: Extracting structured information from unstructured data, such as pulling out data from documents, contracts, and medical records.

NLP Applications

13. *Speech Synthesis (Text-to-Speech)*: Converting written text into spoken language (e.g., reading aloud text on websites, audiobook creation).

14. *Document Classification and Tagging*: Classifying large volumes of documents for better searchability and organization in legal, medical, or corporate environments.

15. *Text-based Recommendation Systems*: Recommending content (e.g., news, articles, movies) based on analyzing the content of text and user preferences.

16. *Healthcare Applications*: Analyzing medical records, extracting important details, and even assisting in diagnosing based on medical language.

NLP continues to evolve with advancements in machine learning and deep learning, leading to even more applications in fields such as education, law, finance, and beyond.

References

[1] Chowdhary, K.R. (2020). Natural Language Processing. In: Fundamentals of Artificial Intelligence. Springer, New Delhi. https://doi.org/10.1007/978-81-322-3972-7_19