Large Language Models

24 CSA 528

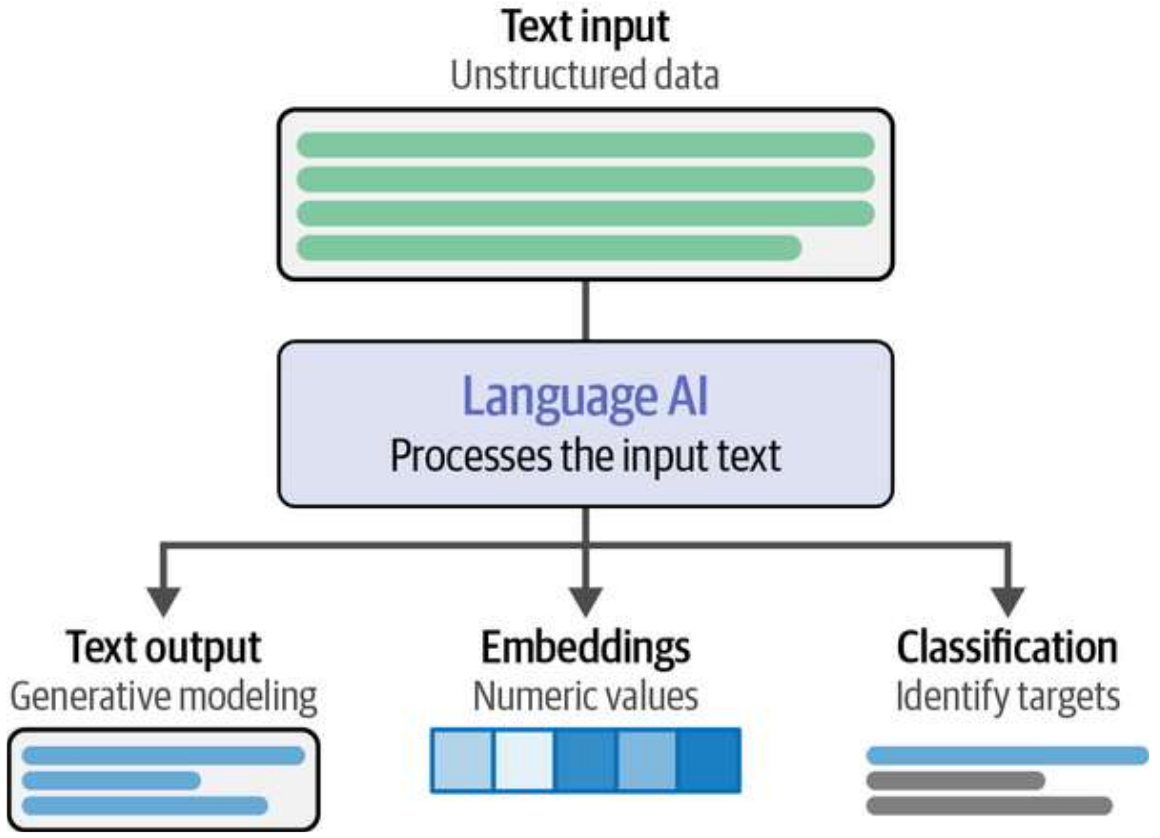
Introduction

What Is Language AI?

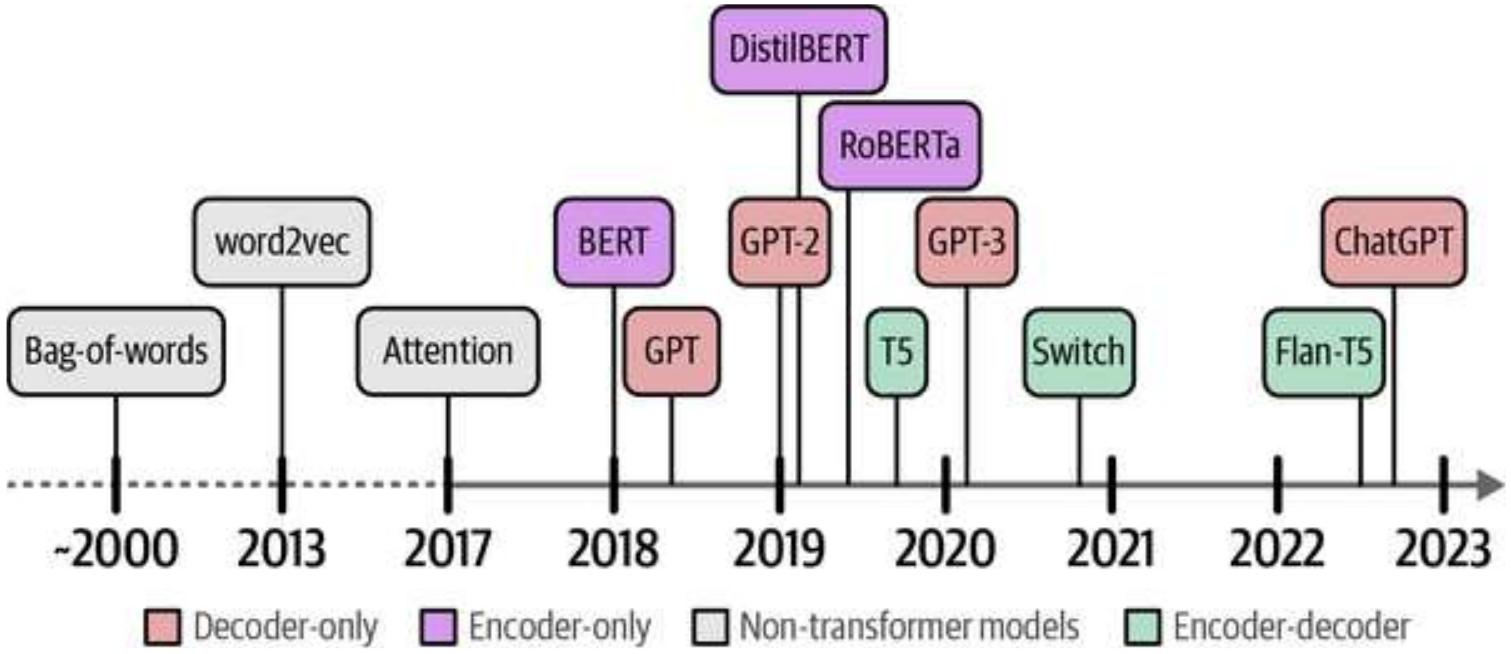
- Language AI refers to a subfield of AI that focuses on developing technologies capable of understanding, processing, and generating human language.
- The term *Language AI* can often be used interchangeably with *natural language processing* (NLP) with the continued success of machine learning methods in tackling language processing problems.
- We use the term *Language AI* to encompass technologies that technically might not be LLMs but still have a significant impact on the field

Why?

- Language is a tricky concept for computers
  - Text is unstructured in nature and loses its meaning when represented by zeros and ones (individual characters).
  - As a result, throughout the history of Language AI, there has been a large focus on representing language in a structured manner so that it can more easily be used by computers.



- The history of Language AI encompasses many developments and models aiming to represent and generate language



Language AI

- Broad Classification of Language AI
  - **Natural Language Processing (NLP)**
  - **Generative Language AI (Natural Language Generation – NLG)**

- **1. Natural Language Processing (NLP)**
  - Focuses on **understanding and analyzing** human language.
- **Key Tasks:**
  - **Text Classification**
    e.g., Sentiment Analysis, Spam Detection
  - **Named Entity Recognition (NER)**
    e.g., Extracting names, locations, organizations
  - **Part-of-Speech Tagging (POS)**
    e.g., Noun, verb, adjective identification
  - **Dependency Parsing / Syntax Analysis**
    e.g., Grammar structures and sentence parsing
  - **Machine Translation** *(rule-based or statistical)*
  - **Text Summarization** *(extractive)*
  - **Speech Recognition** *(ASR)*
    e.g., Converting voice to text
- **Focus:**
  - Language **understanding**
  - Pattern **recognition**
  - Rule-based or statistical **processing**

2. Generative Language AI (Natural Language Generation – NLG)

- Focuses on **producing or generating** natural language text or speech.

- **Key Tasks:**
  - **Text Generation**
    - e.g., Story writing, Chatbots
  - **Abstractive Summarization**
    - e.g., Creating new sentences to summarize
  - **Conversational AI**
    - e.g., ChatGPT, Virtual Assistants
  - **Code Generation**
    - e.g., GitHub Copilot, StarCoder
  - **Machine Translation** *(neural, generative approach)*
  - **Text-to-Speech (TTS)**
    - e.g., Synthesizing speech from text
  - **Multimodal Generation**
    - e.g., Generating image captions, answering image-based questions
- **Focus:**
  - Language **creation**
  - Context-aware **generation**
  - Deep learning and **LLMs**

🧠 Key Difference

Aspect	NLP (Analytical)	Generative Language AI
Goal	Understand and analyze	Generate fluent human-like output
Output	Structured labels, facts	Text, speech, code
Techniques	Tokenization, POS, NER, parsing	Transformers, LLMs, decoding
Example Task	Classify sentiment of a review	Write a review from scratch
Example Model	BERT, spaCy	GPT-4, T5, LLaMA, Claude

Natural Language Processing

NLP enables machines to **understand, interpret, generate, and respond** to human language in a **meaningful and useful way**. Key steps involve:

- **Text Preprocessing/Lexical Analysis**
- **Syntactic Analysis (Parsing)**
- **Semantic Analysis**

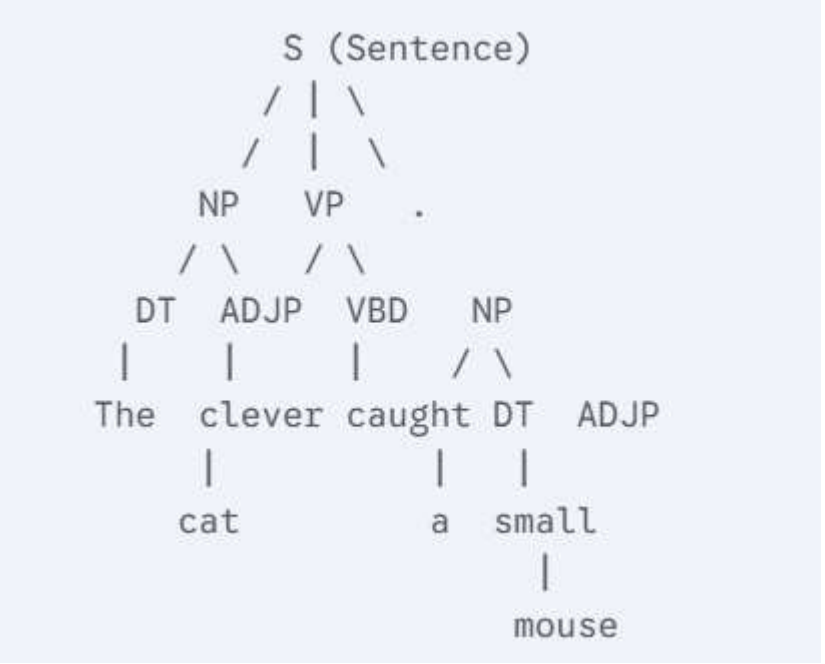
1. **Text Preprocessing/Lexical Analysis:** This is the initial stage where raw text is prepared for further analysis. It involves:

   - **Tokenization:** Breaking down the text into smaller units called "tokens" (words, punctuation, numbers).
     Example Input: San Pedro is a town in Belize.
     Output: Tokens: ['San', 'Pedro', 'is', 'a', 'town', 'in', 'Belize']
   - **Lowercasing:** Converting all text to lowercase to treat "The" and "the" as the same word.
   - **Removing Punctuation:** Eliminating symbols and special characters that may not carry significant meaning.
   - **Removing Stop Words:** Filtering out common words (like "the," "a," "is," "and") that often don't add much semantic value.

- **Morpheme Identification (Morphological analysis):** Once tokens are identified, morphological analysis delves deeper to understand their internal structure. It identifies the root words, prefixes, and suffixes that make up each token.
  - **Stemming and Lemmatization** : These are common tasks within morphological analysis.
    - **Stemming**: reduces words to their root form, often by chopping off suffixes (e.g., "running," "runs," "ran" all become "run").
    - **Lemmatization**: Converting words to their base or dictionary form (lemma), considering their part of speech and context
      - e.g., "better" becomes "good”.
      - This is more linguistically informed than stemming.
      - e.g., "better" becomes "good," not just "bett").
- Handling Numbers and Special Characters: Deciding whether to remove, normalize, or retain numerical data and other special characters based on the task.

- **2. Syntactic Analysis (Parsing):** This step focuses on the grammatical structure of sentences. It involves:
  - **Part-of-Speech (POS) Tagging:** Assigning grammatical categories (e.g., noun, verb, adjective, adverb) to each word in a sentence.
    - "The" - Determiner (DT)
    - "clever" - Adjective (JJ)
    - "cat" - Noun (NN)
    - "caught" - Verb (VBD - past tense)
    - "a" - Determiner (DT)
    - "small" - Adjective (JJ)
    - "mouse" - Noun (NN)
    - "." - Punctuation (P)
  - **Parsing:** Analyzing the grammatical relationships between words to determine the sentence's structure (e.g., identifying subjects, predicates, objects, and their dependencies).
  - **Ambiguity Resolution:** Handling words or phrases that have multiple meanings based on their grammatical role.
    - e.g., identifying if "cuts" is a noun or a verb based on its position in the sentence

Parsing



- ***Example Input2:*** *San Pedro is a town on the southern part of the island of Ambergris Caye. According to 2015 mid-year estimates, the town has a population of about 16,444. It is the second-largest town in the Belize District.*
- ***Output:*** *"It" refers to "San Pedro."*

3. Semantic Analysis: Once the grammatical structure is understood, this step focuses on the literal meaning of words and sentences. Key techniques include:

- **Named Entity Recognition (NER):** Identifying and classifying "named entities" in the text, such as names of people, organizations, locations, dates, and times.
- **Word Sense Disambiguation (WSD):** Determining the correct meaning of a word when it has multiple possible meanings based on its context (e.g., "bank" as a financial institution vs. a river bank).
- **Relationship Extraction:** Identifying relationships between named entities (e.g., "Person X works for Organization Y").

Techniques Used in NLP

- NLP techniques can be broadly categorized into two approaches:
  - **Rule-based Methods**: These involve manually created rules and heuristics to process language data. For example, defining patterns in language to extract meaning.
  - **Machine Learning (ML) and Deep Learning (DL)**: These involve using algorithms to automatically learn from data and improve over time. ML models such as **decision trees**, **support vector machines**, and deep learning models like **recurrent neural networks (RNNs)** and **transformers** are commonly used in modern NLP.