A crash course on Basic NLP

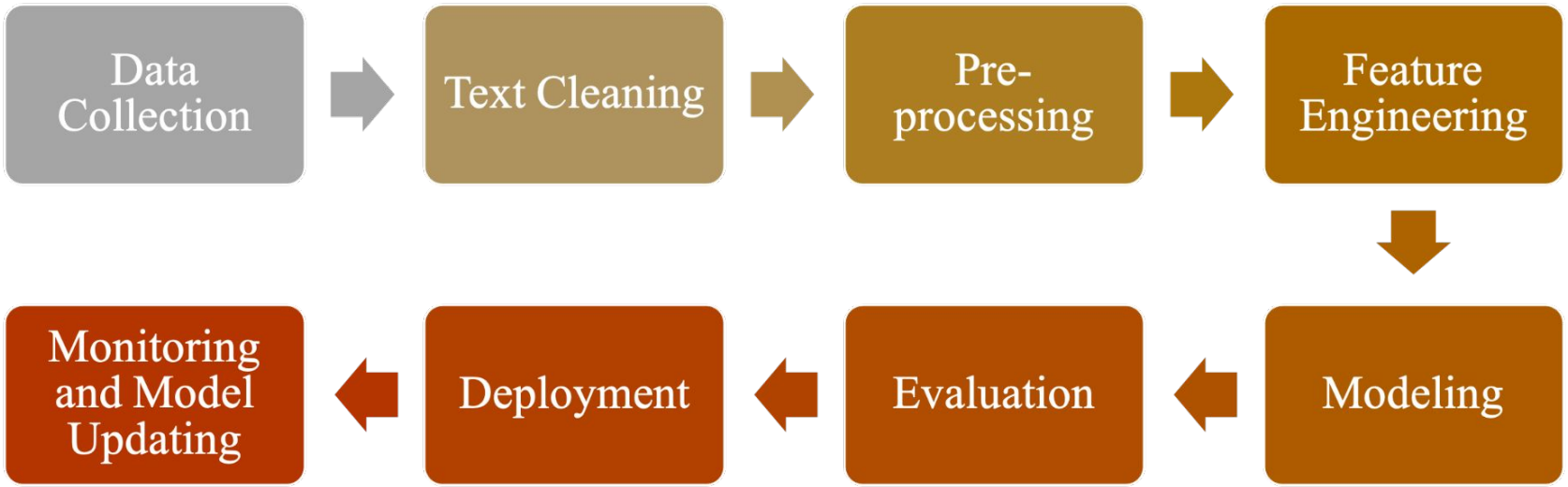
Arun Kumar Rajasekaran

NLP

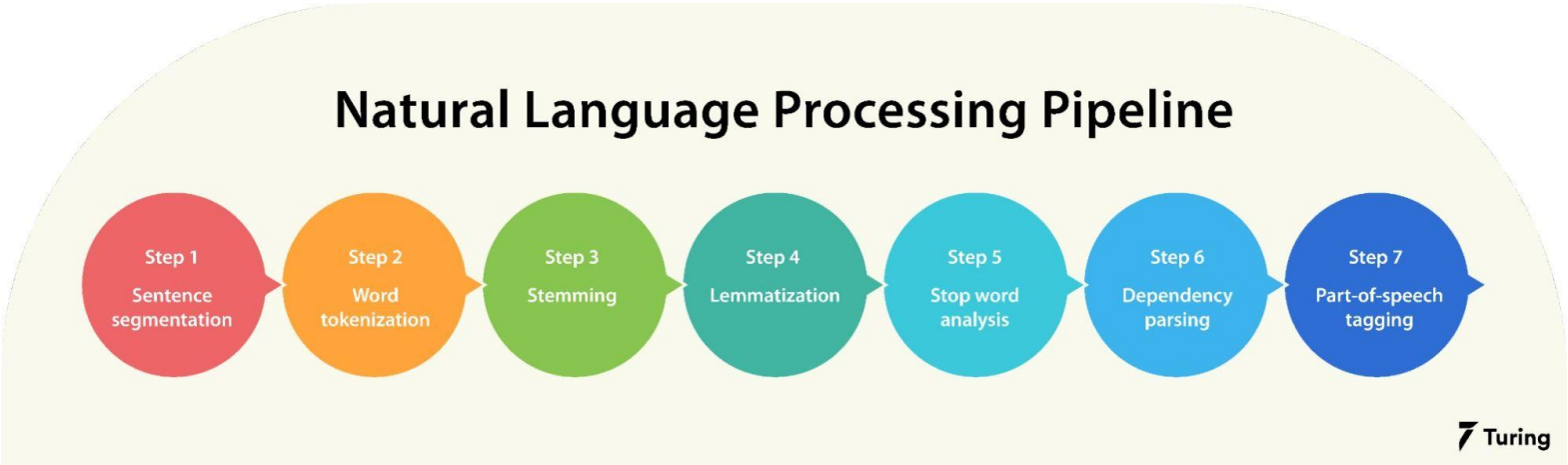
Natural language processing, or NLP, combines computational linguistics—rule-based modeling of human language—with statistical and machine learning models to enable computers and digital devices to recognize, understand and generate text and speech.

translate text from one language to another
respond to typed or spoken commands
recognize or authenticate users based on voice
summarize large volumes of text
assess the intent or sentiment of text or speech
generate text or graphics or other content on demand

NLP Pipeline



More specifically

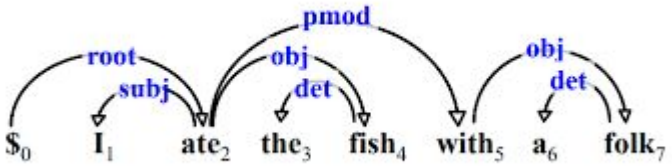


POS tagging?

Part-of-Speech (POS) tagging is a natural language processing technique that involves assigning specific grammatical categories or labels (such as nouns, verbs, adjectives, adverbs, pronouns, etc.) to individual words within a sentence. This process provides insights into the syntactic structure of the text, aiding in understanding word relationships, disambiguating word meanings, and facilitating various linguistic and computational analyses of textual data.

Dependency parsing

The term Dependency Parsing (DP) refers to the process of examining the dependencies between the phrases of a sentence in order to determine its grammatical structure. A sentence is divided into many sections based mostly on this. The process is based on the assumption that there is a direct relationship between each linguistic unit in a sentence. These hyperlinks are called dependencies.



Working by example

"London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the southeast of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium."

Step 1: Sentence segmentation

After using sentence segmentation, we get the following result:

1. “London is the capital and most populous city of England and the United Kingdom.”
2. “Standing on the River Thames in the southeast of the island of Great Britain, London has been a major settlement for two millennia.”
3. “It was founded by the Romans, who named it Londinium.”

Step 2: Word tokenization

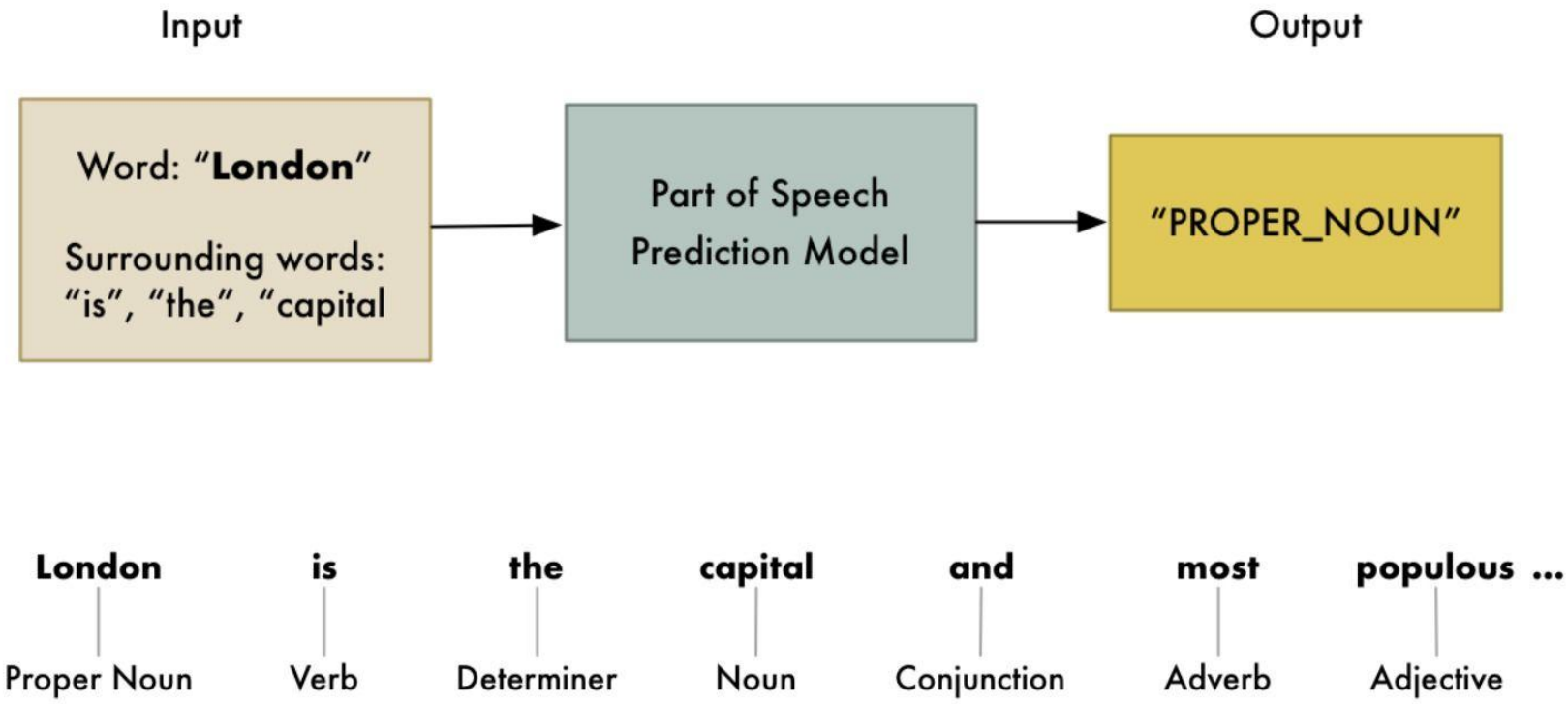
When tokenizing the sentence “London is the capital and most populous city of England and the United Kingdom”, it is broken into separate words, i.e.,

“London”, “is”, “the”, “capital”, “and”, “most”, “populous”, “city”, “of”, “England”, “and”, “the”, “United”, “Kingdom”, “.”

Step 3: Stemming

Stemming helps in preprocessing text. The model analyzes the parts of speech to figure out what exactly the sentence is talking about.

Stemming normalizes words into their base or root form. In other words, it helps to predict the parts of speech for each token. For example, intelligently, intelligence, and intelligent. These words originate from a single root word ‘intelligen’. However, in English there’s no such word as ‘intelligen’.


Input
Output
Word: "London"
Surrounding words: "is", "the", "capital
Part of Speech Prediction Model
"PROPER_NOUN"
London
is
the
capital
and
most
populous
...
Proper Noun
Verb
Determiner
Noun
Conjunction
Adverb
Adjective

Step 4: Lemmatization

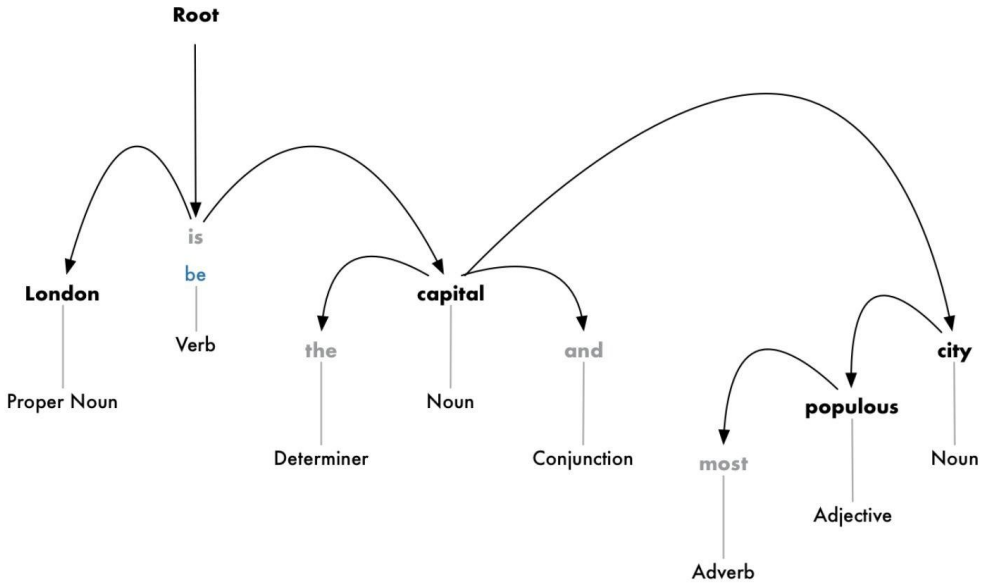
Lemmatization removes inflectional endings and returns the canonical form of a word or lemma. It is similar to stemming except that the lemma is an actual word. For example, ‘playing’ and ‘plays’ are forms of the word ‘play’. Hence, play is the lemma of these words. Unlike a stem (recall ‘intelligen’), ‘play’ is a proper word.

Step 5: Stop word analysis

The next step is to consider the importance of each and every word in a given sentence. In English, some words appear more frequently than others such as "is", "a", "the", "and". As they appear often, the NLP pipeline flags them as stop words. They are filtered out so as to focus on more important words.

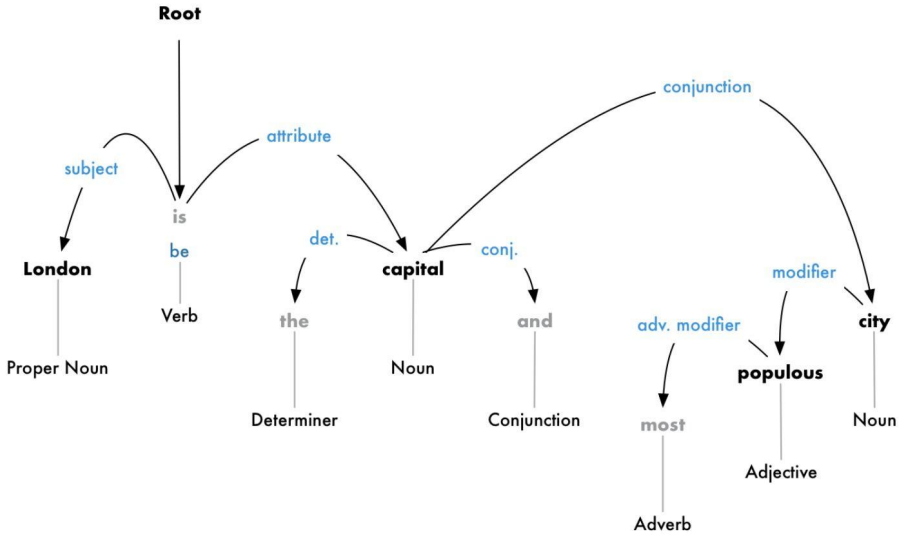
Step 6: Dependency parsing

Next comes dependency parsing which is mainly used to find out how all the words in a sentence are related to each other. To find the dependency, we can build a tree and assign a single word as a parent word. The main verb in the sentence will act as the root node.



Step 7: Part-of-speech (POS) tagging

POS tags contain verbs, adverbs, nouns, and adjectives that help indicate the meaning of words in a grammatically correct way in a sentence.



Large language models

Large language models (LLMs) are a category of foundation models trained on immense amounts of data making them capable of understanding and generating natural language and other types of content to perform a wide range of tasks.

LLMs operate by leveraging deep learning techniques and vast amounts of textual data. These models are typically based on a transformer architecture, like the generative pre-trained transformer, which excels at handling sequential data like text input. LLMs consist of multiple layers of neural networks, each with parameters that can be fine-tuned during training, which are enhanced further by a numerous layer known as the attention mechanism, which dials in on specific parts of data sets.

Improving LLMs

Model performance can also be increased through prompt engineering, prompt-tuning, fine-tuning and other tactics like reinforcement learning with human feedback (RLHF) to remove the biases, hateful speech and factually incorrect answers known as “hallucinations” that are often unwanted byproducts of training on so much unstructured data.

Fine-Tuning: Leveraging Labeled Data

Fine-tuning is the most common training approach for adapting LLMs to new tasks. It works well when you have abundant labeled data available for the task.

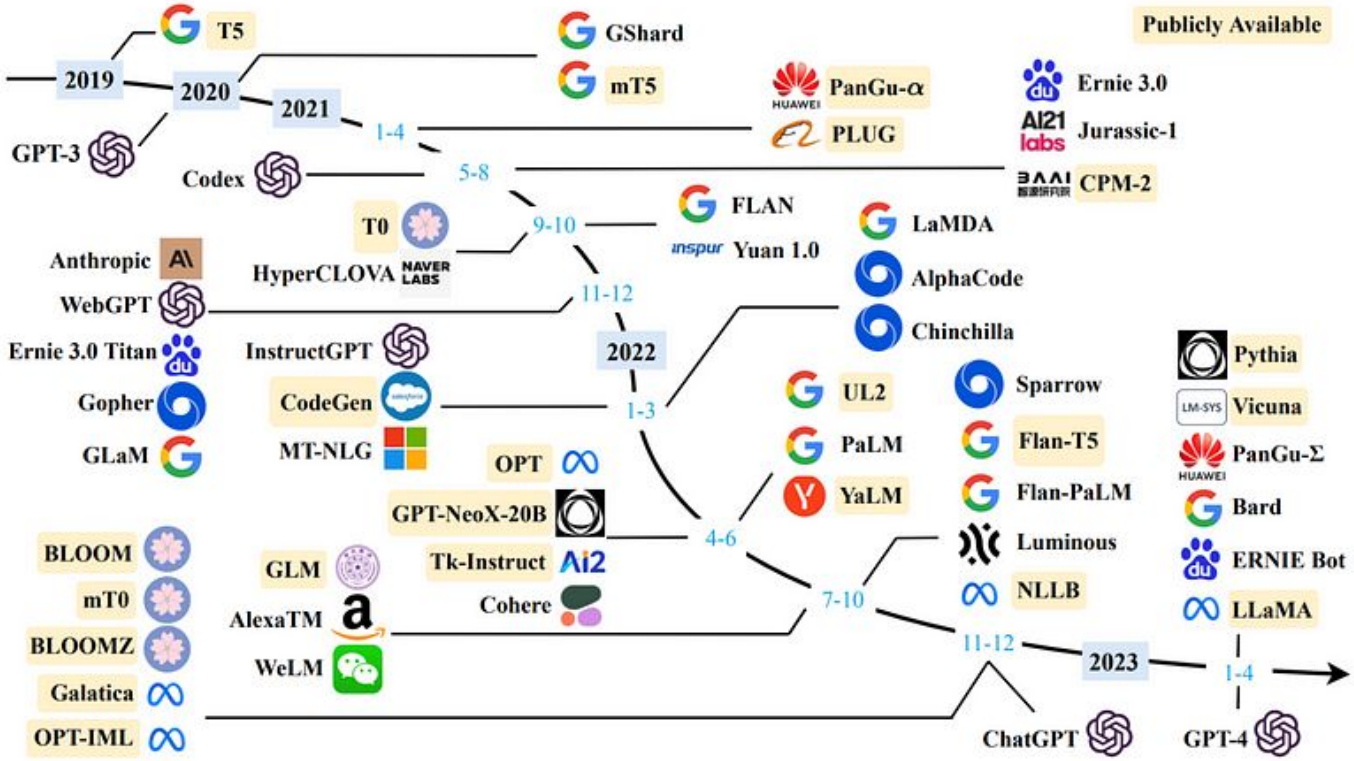
Instruction Tuning: Directing with Natural Language

Instruction tuning provides efficient training without requiring large datasets. Instead of input-output examples, you provide prompt-completion pairs demonstrating the desired behavior.

RLHF: Reinforcement Learning from Human Feedback

RLHF provides subjective training leveraging human judgments instead of labeled data. It is well-suited for open-ended tasks.

Discussions


Publicly Available
2019
2020
2021
2022
2023
T5
GPT-3
GShard
mT5
PanGu-α
PLUG
Ernie 3.0
Jurassic-1
CPM-2
1-4
Codex
5-8
T0
HyperCLOVA
9-10
FLAN
Yuan 1.0
LaMDA
AlphaCode
Chinchilla
Anthropic
WebGPT
11-12
Ernie 3.0 Titan
Gopher
GLaM
InstructGPT
CodeGen
MT-NLG
1-3
UL2
PaLM
YaLM
Sparrow
Flan-T5
Flan-PaLM
Luminous
NLLB
OPT
GPT-NeoX-20B
Tk-Instruct
Cohere
4-6
7-10
11-12
Pythia
Vicuna
PanGu-Σ
Bard
ERNIE Bot
LLaMA
1-4
BLOOM
mT0
BLOOMZ
Galatica
OPT-IML
GLM
AlexaTM
WeLM
ChatGPT
GPT-4