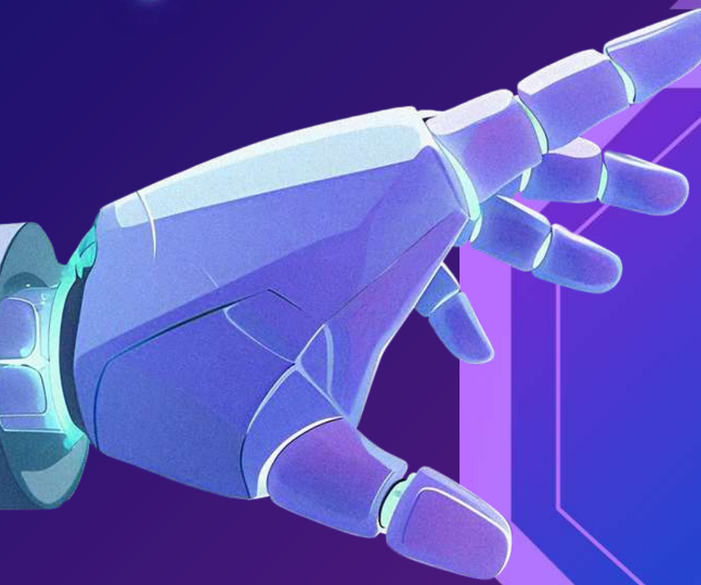
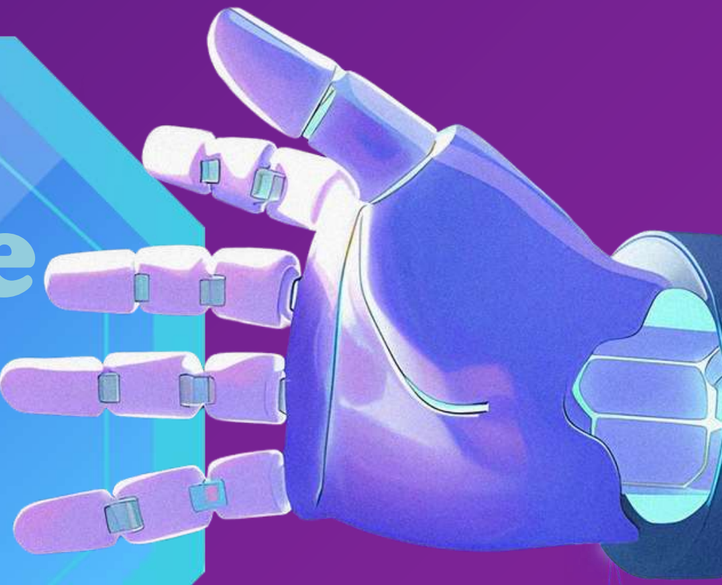
NLP

Natural Language Processing





By: Abdelrhman Khalil

What is NLP?

Natural Language Processing (NLP)is a branch of artificial intelligence (AI) that enables machines to understand human language. The main intention of NLP is to build systems that are able to make sense of text and then automatically execute tasks like spell - check, text translation, topic classification, etc. Companies today use NLP in artificial intelligence to gain insights from data and automate routine tasks.

This article will look at how natural language processing functions in AI.

Components of natural language processing in AI

NLP has two components as underlined below.

1. Natural language generation (NLG)

NLG is a method of creating meaningful phrases and sentences (natural language) from data. It comprises three stages: text planning, sentence planning, and text realization.

Text planning: Retrieving applicable content.

Sentence planning: Forming meaningful phrases and setting the sentence tone.

Text realization: Mapping sentence plans to sentence structures.

Chatbots, machine translation tools, analytics platforms, voice assistants, sentiment analysis platforms, and AI - powered transcription tools are some applications of NLG.

Components of natural language processing in AI

2. Natural language understanding (NLU)

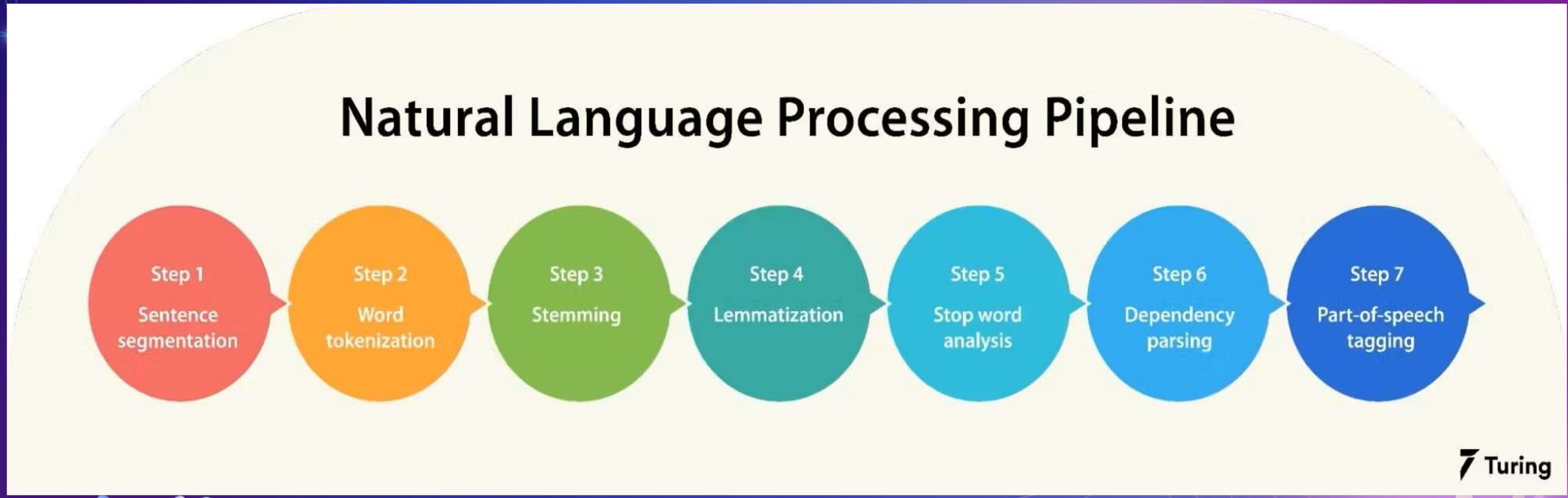
NLU enables machines to understand and interpret human language by extracting metadata from content. It performs the following tasks:

- Helps analyze different aspects of language.
- Helps map the input in natural language into valid representations.

✓ NLU is more difficult than NLG tasks owing to referential, lexical, and syntactic ambiguity.

- **Lexical ambiguity:** This means that one word holds several meanings. For example, "The man is looking for the **match**." The sentence is ambiguous as ‘match’ could mean different things such as a partner or a competition.
- **Syntactic ambiguity:** This refers to a sequence of words with more than one meaning. For example, "The fish **is ready** to eat.” The ambiguity here is whether the fish is ready to eat its food or whether the fish is ready for someone else to eat. This ambiguity can be resolved with the help of the part - of - speech tagging technique.
- **Referential ambiguity:** This involves a word or a phrase that could refer to two or more properties. For example, Tom met Jerry and John. **They** went to the movies. Here, the pronoun ‘they’ causes ambiguity as it isn’t clear who it refers to.

Pipeline of natural language processing in artificial intelligence



Step 1: Sentence segmentation

Sentence segmentation is the first step in the NLP pipeline. It divides the entire paragraph into different sentences for better understanding. For 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."

After using sentence segmentation, we get the following result:

“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.”

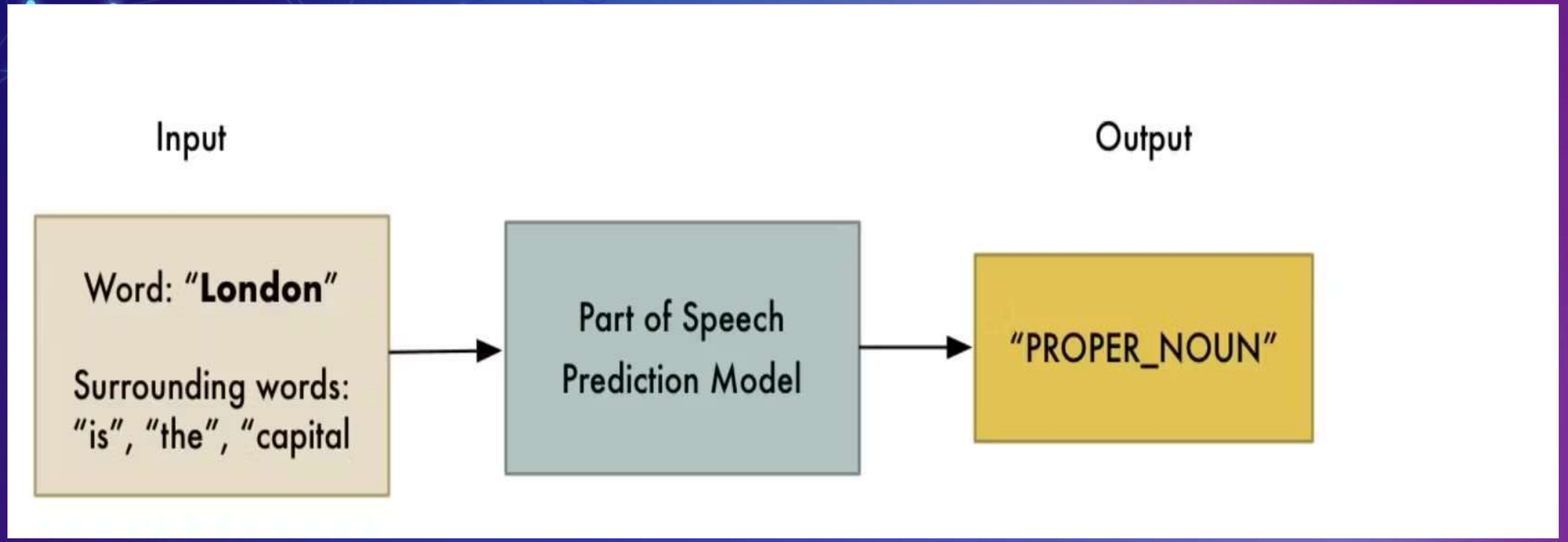
Step 2: Word tokenization

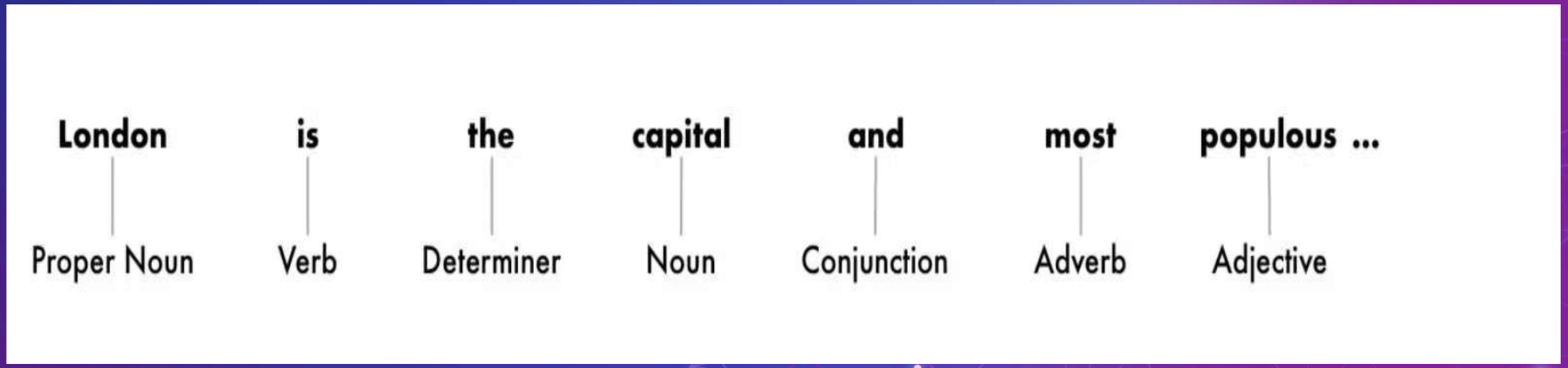
Word tokenization breaks the sentence into separate words or tokens. This helps understand the context of the text. When tokenizing the sentence "London is the capital and most populous city of England and the United Kingdom", it is broken into separate words, i.e., "London", "is", "the", "capital", "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

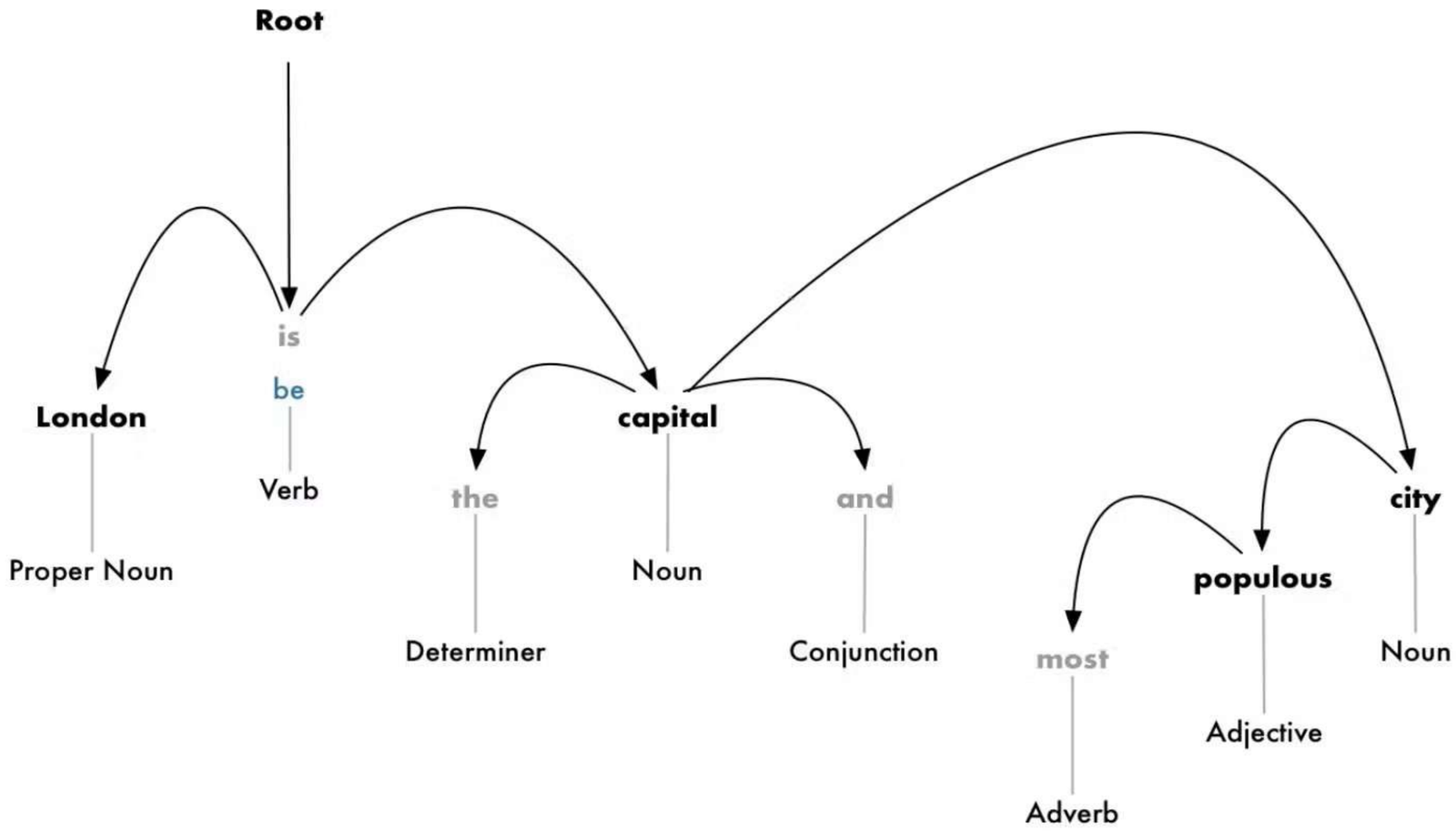
- Lemmatization removes inflectional endings and returns the canonical form of a word or lemma. It is similar to stemming except that the lemma is an actual word. For example, ‘playing’ and ‘plays’ are forms of the word ‘play’. Hence, play is the lemma of these words. Unlike a stem (recall ‘intelligen’),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

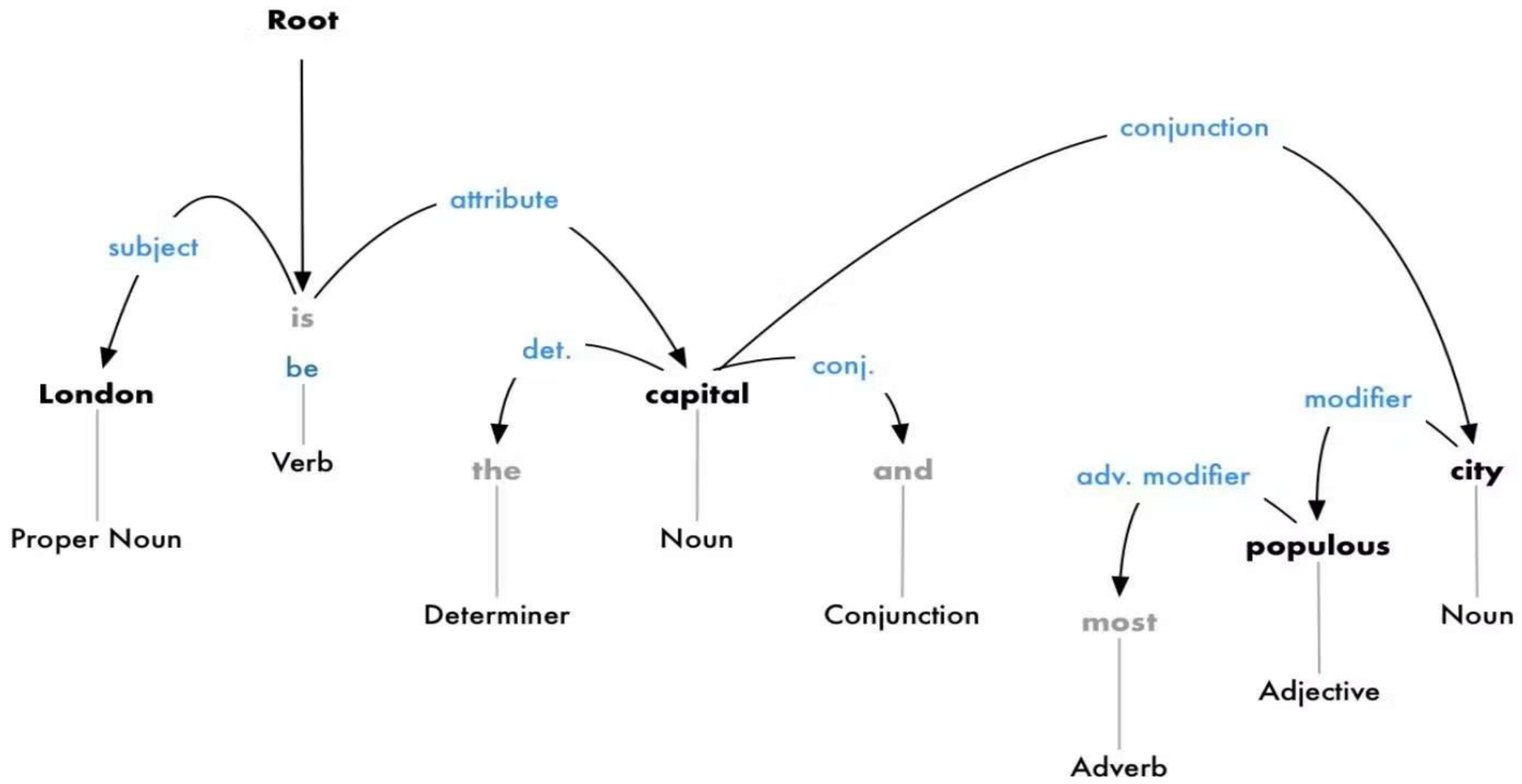
Step 6: Dependency parsing

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.


Root
London
Proper Noun
is
be
Verb
the
Determiner
capital
Noun
and
Conjunction
most
Adverb
populous
Adjective
city
Noun

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.


Root
subject
attribute
conjunction
is
be
Verb
London
Proper Noun
det.
conj.
capital
Noun
the
Determiner
and
Conjunction
modifier
adv. modifier
city
Noun
populous
Adjective
most
Adverb

Thank You

Abdelrhman Khalil

www.linkedin.com/in/abdulrahman-khalil-ba64272a3