Information Retrieval

-MUKESH KUMAR

Barack Obama is an American politician who served as the 44th President of the United States from 2009 to 2017. He is the first African American to have served as president, as well as the first born outside the contiguous United States.

Extracting information

1. Who are the people that are being talked about in this paragraph?
2. What geography or locations are mentioned?
3. What time period is mentioned in this article?

Barack Obama is an American politician who served as the 44th President of the United States from 2009 to 2017. He is the first African American to have served as president, as well as the first born outside the contiguous United States.

What will it take to build an ML model to extract information?

Barack Obama is an American politician who served as the 44th President of the United States from 2009 to 2017. He is the first African American to have served as president, as well as the first born outside the contiguous United States.

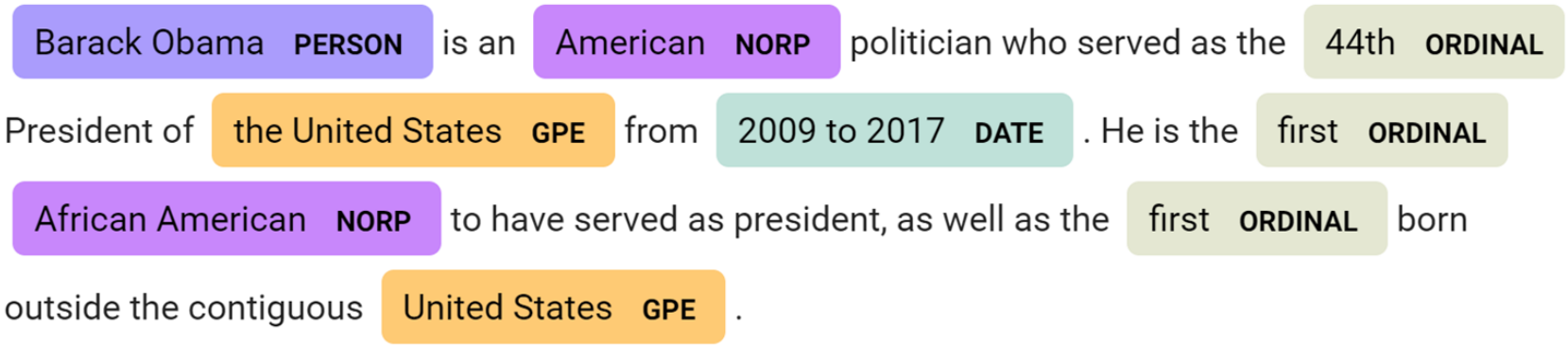
Named Entity Recognition Model

Classifies text into predefined categories or real world entities.

Named Entity Recognition Model

Some of the common labels/entities

TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.
CARDINAL	Numerals that do not fall under another type.



Named Entity Recognition (NER) model's output

Good News!!!

There are lots of pre-trained Named Entity Recognition models are available which we can use out of the box.

NAMED ENTITY RECOGNITION (NER) MODEL

- Named Entity Recognition (NER) is a sub-task of information extraction that seeks to locate and classify named entities mentioned in unstructured text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

Applications

- **Information Retrieval**: Enhancing search engines to provide more accurate results.
- **Question Answering**: Improving the understanding of context in QA systems.
- **Content Categorization**: Organizing and tagging content for better management and retrieval.

Challenges

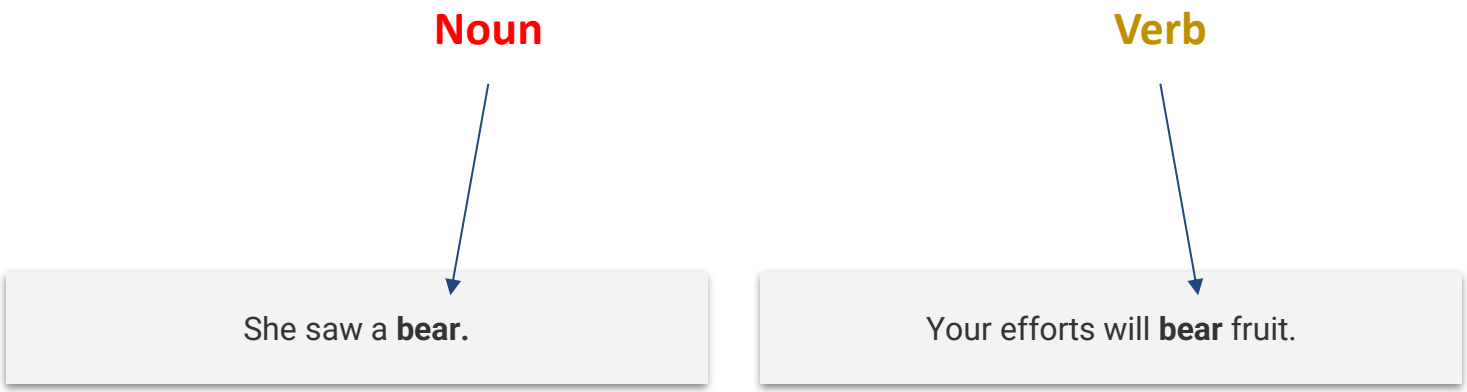
- **Ambiguity**: Words can have different meanings based on context (e.g., "Apple" can refer to a fruit or a company).
- **Variability**: Named entities can appear in various forms (e.g., "New York", "NYC").
- **Out-of-Vocabulary Entities**: Handling entities not seen during training.

PART OF SPEECH (POS) TAGGER

She saw a **bear.**

Your efforts will **bear** fruit.

How the word ‘bear’ different in two sentences?



Words play different roles in a sentence.

How do we find extract role a word plays in a sentence?

Part of Speech (PoS) tagger

- A Part of Speech (PoS) tagger is a tool in Natural Language Processing (NLP) that reads text in a language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc. PoS tagging is a fundamental preprocessing step in many NLP applications as it provides syntactic information about the text, which can be useful for further analysis.

POS Key Concepts

- **Parts of Speech**: Categories that define the syntactic roles of words in sentences. Common parts of speech include:
  - **Noun (NN)**: Person, place, thing (e.g., "dog", "city").
  - **Verb (VB)**: Action or state (e.g., "run", "is").
  - **Adjective (JJ)**: Describes a noun (e.g., "big", "happy").
  - **Adverb (RB)**: Modifies a verb, adjective, or other adverb (e.g., "quickly", "very").
  - **Pronoun (PRP)**: Replaces a noun (e.g., "he", "they").
  - **Preposition (IN)**: Links nouns, pronouns, and phrases to other words (e.g., "in", "on").
  - **Conjunction (CC)**: Connects words or groups of words (e.g., "and", "but").
  - **Determiner (DT)**: Introduces nouns (e.g., "the", "a").
  - **Interjection (UH)**: Expresses emotion (e.g., "oh", "wow").

- **Tagging**: The process of labeling each word in a sentence with its corresponding part of speech.

Hyderabad	is	the	capital	of	Telangana.
PROPN	VERB	DET	NOUN	ADP	PROPN

Applications

- **Syntactic Parsing**: Understanding sentence structure.
- **Information Retrieval**: Improving search accuracy.
- **Text-to-Speech Systems**: Ensuring proper pronunciation and intonation.
- **Named Entity Recognition**: Improving entity extraction by understanding context.

DEPENDENCY PARSING

Hyderabad is the capital of the Indian state of Telangana occupying 650 square kilometres (250 sq mi) along the banks of the Musi River. Hyderabad City has a population of about 6.9 million, making it the fourth-most populous city in India.

Established in 1591 by Muhammad Quli Qutb Shah, Hyderabad remained under the rule of the Qutb Shahi dynasty for nearly a century before the Mughals captured the region.

Extracting facts about Hyderabad

- Capital of which state?
- How populated is Hyderabad?
- Who established Hyderabad?

Information Retrieval

- Dependency parsing is a technique in natural language processing (NLP) that analyzes the grammatical structure of a sentence by establishing relationships between "head" words and words which modify those heads.

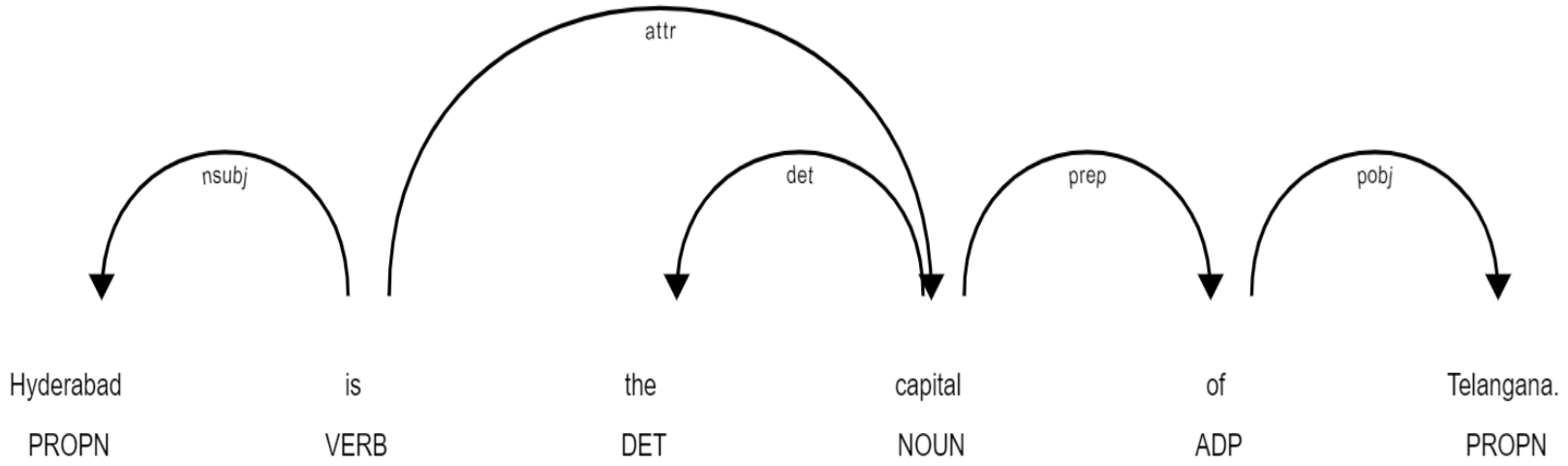
- The output of dependency parsing is a dependency tree, where nodes are words, and directed edges represent syntactic dependencies between them.

Understanding Language Structure & Syntax

1. Part of Speech tagging (PoS)
2. Dependency Parsing

How are the words in a sentence related to each other?

Dependency Parsing



1. Shows how words in a sentence relate to each other.
2. Allows further understanding of language structure and syntax.

Types of Dependencies

- Common dependency labels include:
- **nsubj**: Nominal subject
- **dobj**: Direct object
- **iobj**: Indirect object
- **amod**: Adjectival modifier
- **det**: Determiner

PoS Tagging, Dependency Parsing & Named Entity Recognition

Using spaCy