# Normalization

# Normalization

- Morpheme : base form of a word
- Structure of token : <prefix> <morpheme> <suffix>

  Example : Antinationalist : Anti + national + ist

# Normalization

- Morpheme : base form of a word
- Structure of token : <prefix> <morpheme> <suffix>

  Example : Antinationalist : Anti + national + ist
- Normalization : Process of converting a token into its base form

  (morpheme)

# Normalization

- Morpheme : base form of a word
- Structure of token : <prefix> <morpheme> <suffix>

  Example : Antinationalist : Anti + national + ist
- Normalization : Process of converting a token into its base form

  (morpheme)
- Helpful in reducing data dimensionality, text cleaning
- Types : Stemming and Lemmatization

# Normalization: Stemming

- Elementary rule based process of removal of inflectional forms from a token
- Outputs the stem of a word

"laughing", "laughed", "laughs", "laugh" >>> "laugh"

| Form | Suffix | Stem |
|---|---|---|
| studies | -es | studi |
| studying | -ing | study |
| niñas | -as | niñ |
| niñez | -ez | niñ |

# Normalization: Stemming

- Elementary rule based process of removal of inflectional forms from a token
- Outputs the stem of a word

  “laughing”, “laughed”, “laughs”, “laugh” >>> “laugh”
- May generate non-meaningful terms

  his teams are not winning

  >> hi team are not winn

| Form | Suffix | Stem |
|---|---|---|
| studies | -es | studi |
| studying | -ing | study |
| niñas | -as | niñ |
| niñez | -ez | niñ |

# Normalization: Lemmatization

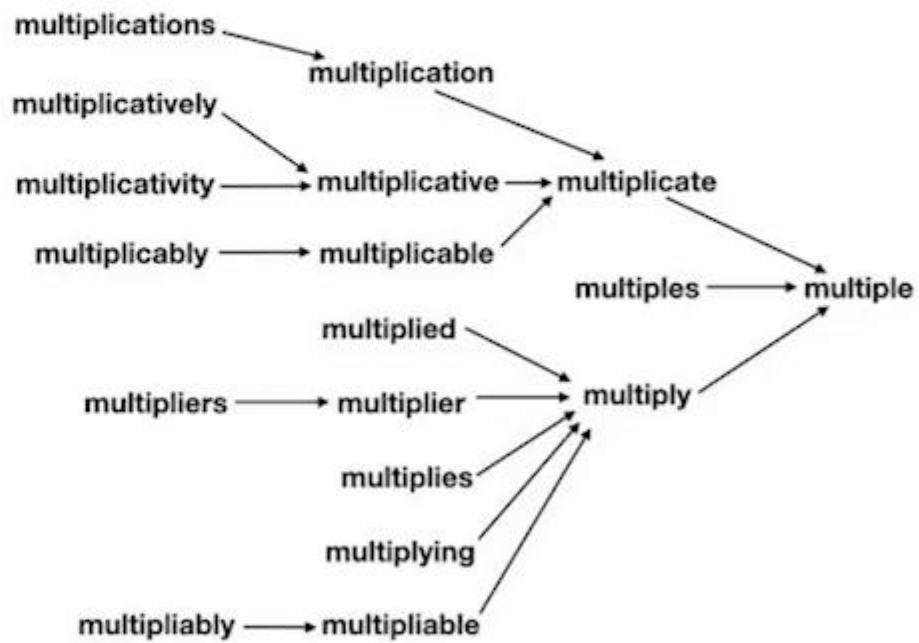
- Systematic process for reducing a token to its lemma

# Normalization: Lemmatization

- Systematic process for reducing a token to its lemma
- Makes use of vocabulary, and morphological analysis
- Example :

  am, are, is >> be

  running , ran , run , rans >> run



# Information Extraction

# About Module

1. What is Information Extraction?
2. Part-of-Speech Tagging (POS)

# About Module

1. What is Information Extraction?
2. Part-of-Speech Tagging (POS)
3. Dependency Parsing
4. Named Entity Recognition (NER)
5. Relation Extraction

# What is Information Extraction?

Information Extraction is the process of extracting meaningful information from text data.

# What is Information Extraction?

Eg, The symptoms of COVID-19 are fever, cough, shortness of breath, sore throat and fatigue.

# What is Information Extraction?

Eg, The symptoms of COVID-19 are fever, cough, shortness of breath, sore throat and fatigue.

Information:

- Disease: COVID-19
- Symptoms: [ “fever”, “cough”, “shortness of breath”, “sore throat”, “fatigue” ]

# Applications of Information Extraction

- Business Intelligence

# Applications of Information Extraction

- Business Intelligence
- Resume Harvesting

**John Doe**

San Francisco CA • (123) 456-7891

john_doe@email.com

SUMMARY

Analytical professional with 8+ years of experience in generating and analyzing data reports for management. A strategic mindset that focuses on problem-solving tasks and maintaining priorities on strict deadlines.

EDUCATION

Temp Tech
Aug '08 - May '12
Computer Science/Programming

EXPERIENCE

XYZ Analytics, Data Analyst
2012 - 2015

- Manage department metrics report and database
- Test and develop a database management plan

ABC.ai, Lead Data Analyst
2015- Current

- Research existing database methods and create a study for suggested changes
- Lead database project with 5 team members

SKILLS

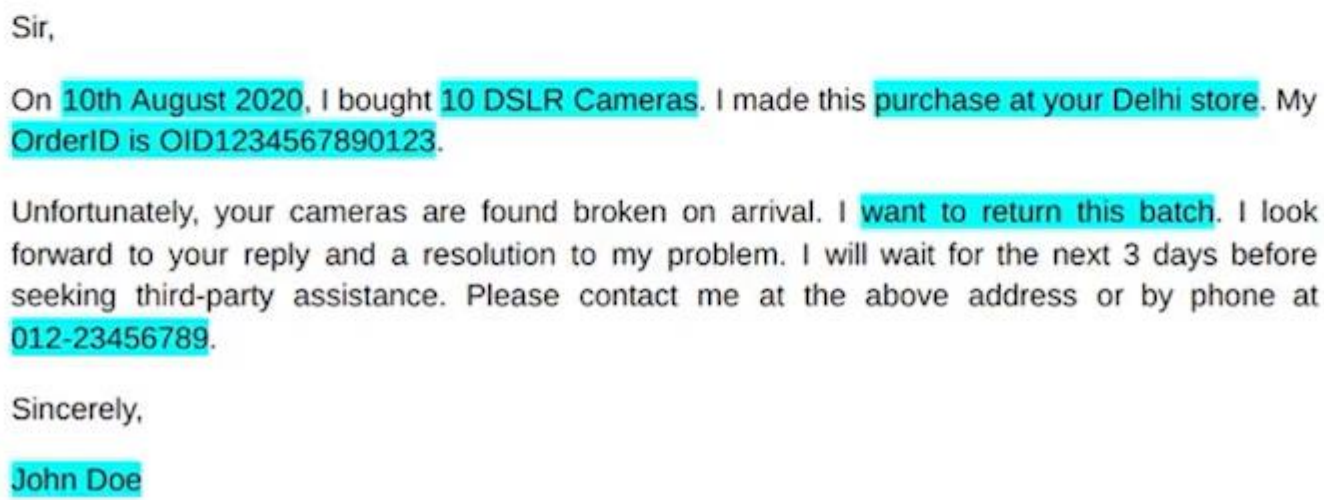
- Project management
- Data analysis

# Applications of Information Extraction

- Business Intelligence
- Resume Harvesting
- Criminal Justice Information Systems

# Applications of Information Extraction

- Business Intelligence
- Resume Harvesting
- Criminal Justice Information Systems
- Scanning Email and Chat Conversations



Sir,

On 10th August 2020, I bought 10 DSLR Cameras. I made this purchase at your Delhi store. My OrderID is OID1234567890123.

Unfortunately, your cameras are found broken on arrival. I want to return this batch. I look forward to your reply and a resolution to my problem. I will wait for the next 3 days before seeking third-party assistance. Please contact me at the above address or by phone at 012-23456789.

Sincerely,

John Doe

# Applications of Information Extraction

- Business Intelligence
- Resume Harvesting
- Criminal Justice Information Systems
- Scanning Email and Chat Conversations

Sir,

On 10th August 2020, I bought 10 DSLR Cameras. I made this purchase at your Delhi store. My OrderID is OID1234567890123.

Unfortunately, your cameras are found broken on arrival. I want to return this batch. I look forward to your reply and a resolution to my problem. I will wait for the next 3 days before seeking third-party assistance. Please contact me at the above address or by phone at 012-23456789.

Sincerely,

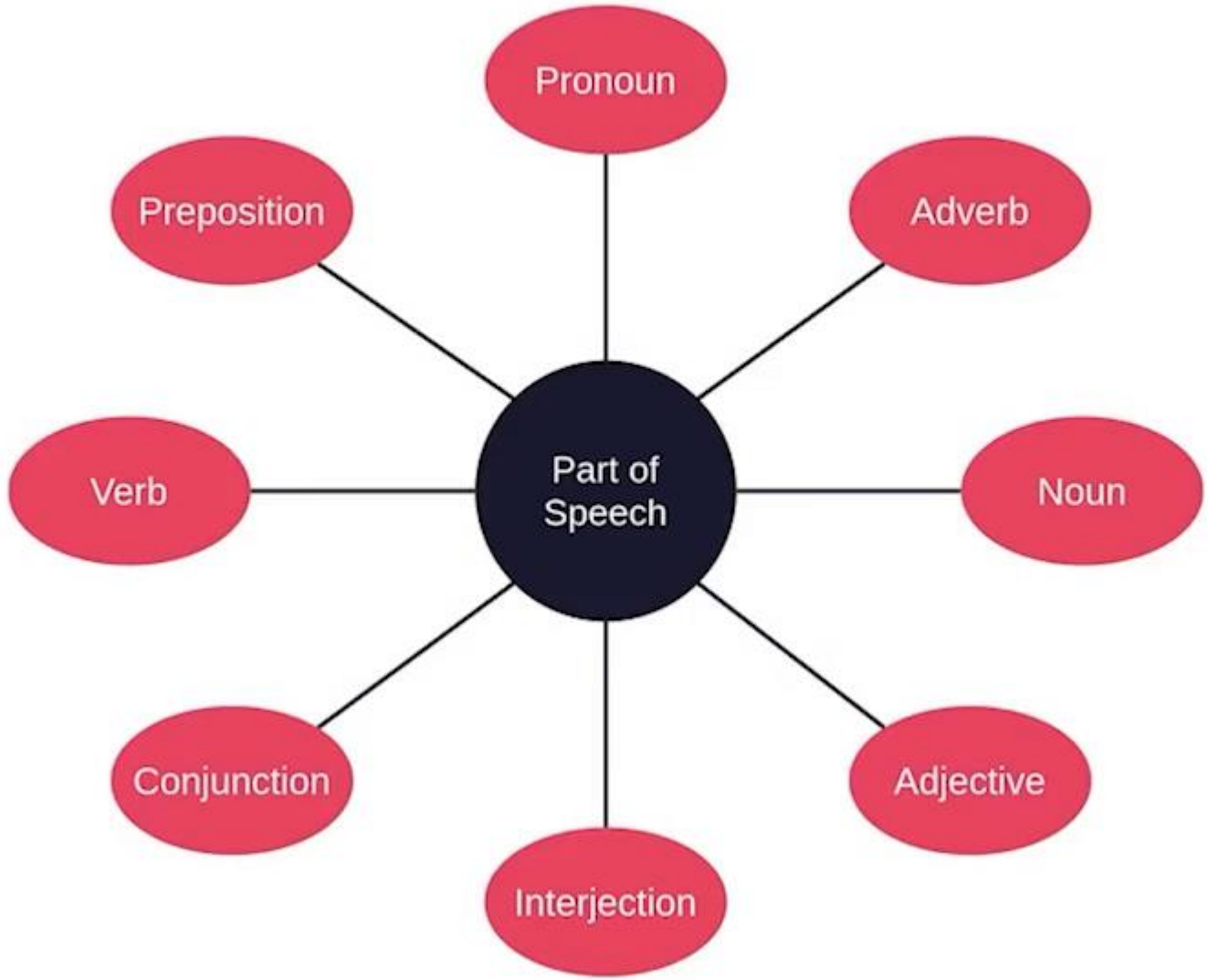
John Doe

# Applications of Information Extraction

- Business Intelligence
- Resume Harvesting
- Criminal Justice Information Systems
- Scanning Email and Chat Conversations
- Extracting Research Findings

# Part-of-Speech (POS) Tagging

# Part of Speech



# Part-of-Speech (POS) Tagging

- Process of assigning different labels known as POS tags

# Part-of-Speech (POS) Tagging

- Process of assigning different labels known as POS tags
- Tells about Part-of-Speech of the word

# Part-of-Speech (POS) Tagging

- Process of assigning different labels known as POS tags
- Tells about Part-of-Speech of the word
- Example tags,

  PROPN: Proper Noun => John, Netflix, Amazon

  DET: Determiner => a, an, the

# Part-of-Speech (POS) Tagging

- Process of assigning different labels known as POS tags
- Tells about Part-of-Speech of the word
- Example tags,

  PROPN: Proper Noun => John, Netflix, Amazon

  DET: Determiner => a, an, the

  NUM: Numerals => 1, two, III
- Example, Usain Bolt is the fastest man on earth.

# Types of POS Tags

1. Universal POS Tags:
   a. Used in Universal Dependencies version 2
   b. Core Part-of Speech tags for various languages

   E.g., ADJ: Adjective
   NOUN: Common Noun

# Types of POS Tags

1. Universal POS Tags:
   a. Used in Universal Dependencies version 2
   b. Core Part-of Speech tags for various languages

   E.g., ADJ: Adjective
   NOUN: Common Noun

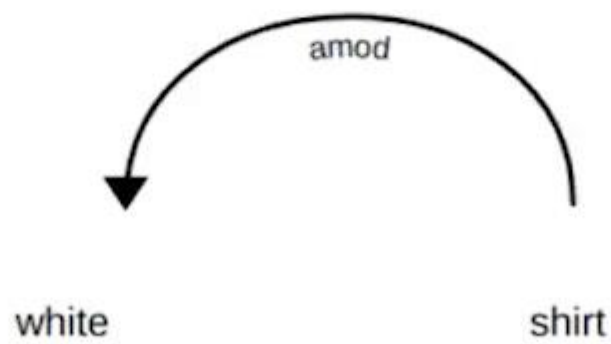
2. Detailed POS tags:
   a. Language-specific
   b. Division of Universal POS Tags into multiple tags

   E.g., PROPN: Proper Noun, e.g., Americans, Airlines, Google, Microsoft
   NNP: Singular Proper Noun, e.g., Microsoft, Google
   NNPS: Plural Proper Noun, e.g., Americans, Airlines

# Dependency Parsing

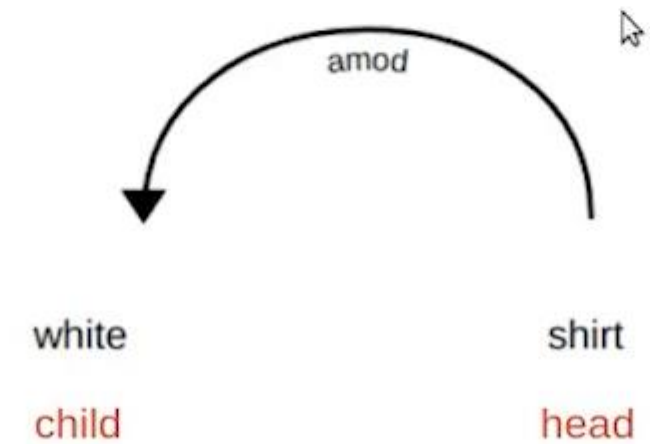
# Dependency Parsing

- Analyze the grammatical structure of a sentence
- Based on dependencies between the words in a sentence
- Dependency tags represent relationship between words
- Example,



# Dependency Parsing

- Analyze the grammatical structure of a sentence
- Based on dependencies between the words in a sentence
- Dependency tags represent relationship between words
- Example,

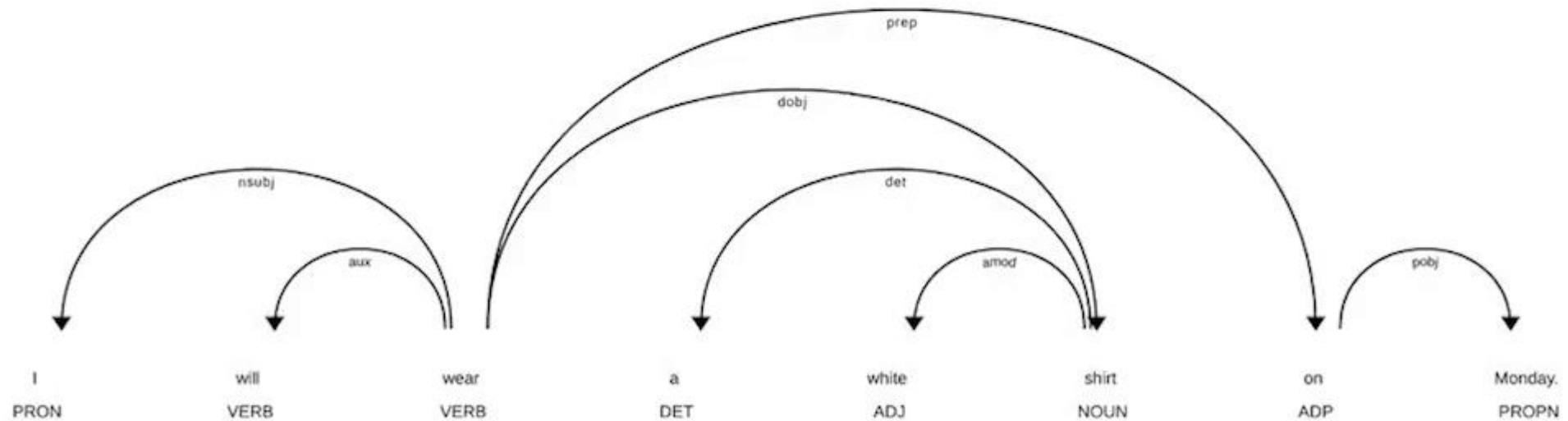


# Dependency Tree

- Parse tree generated in dependency parsing

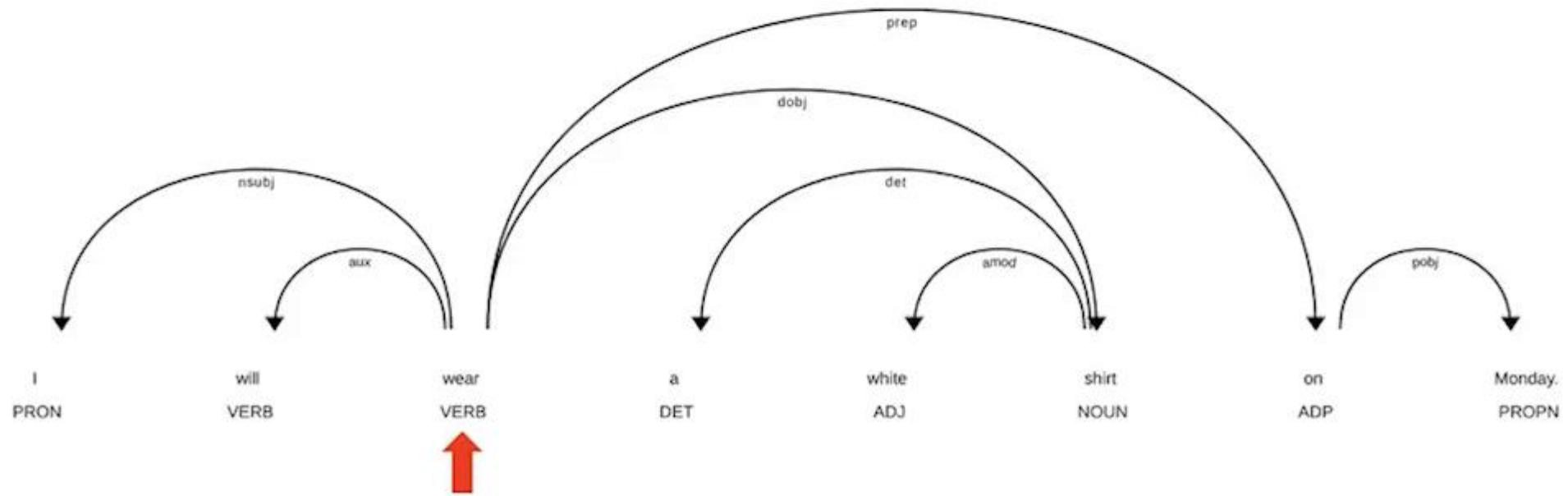
# Dependency Tree

- Parse tree generated in dependency parsing
- E.g., “I will wear a white shirt on Monday.”



# Dependency Tree

- Parse tree generated in dependency parsing
- E.g., “I will wear a white shirt on Monday.”



# Entities and Named Entities

# Entities and Named Entities

- Entities: common things, belongs to noun family

# Entities and Named Entities

- Entities: common things, belongs to noun family

The president will meet the chairman of the company in the capital city

# Entities and Named Entities

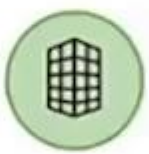
- Entities: common things, belongs to noun family

The president will meet the chairman of the company in the capital city

- Named Entities: entities having proper name


People


Places


Companies


Products

Donald Trump will meet the chairman of Google in New York City

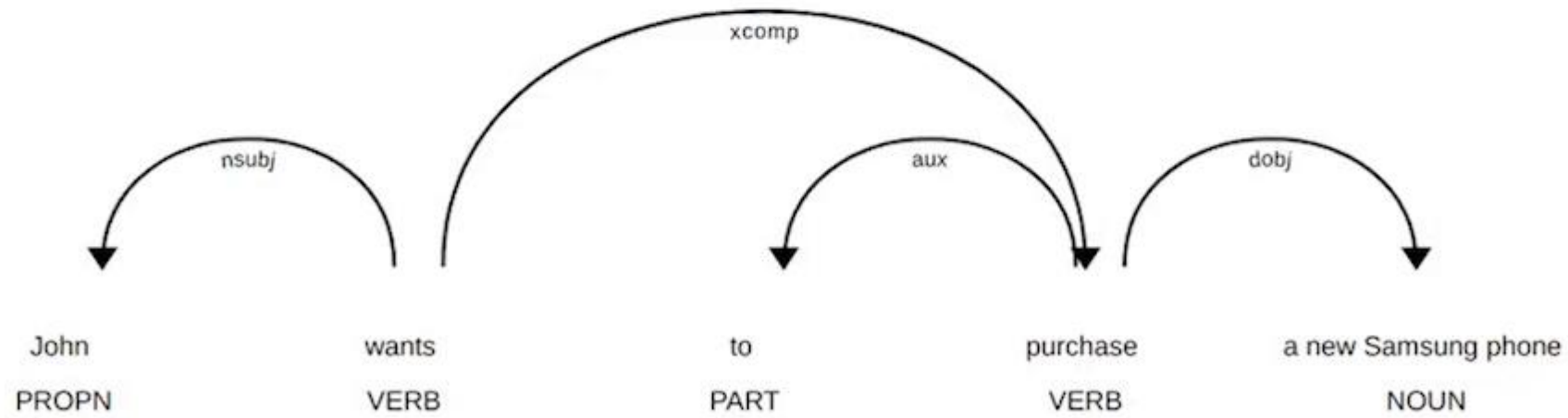
# Named Entity Recognition (NER)

# Named Entity Extraction

- Identification of noun phrases
- Noun phrases: connected by direct subject or object relationships

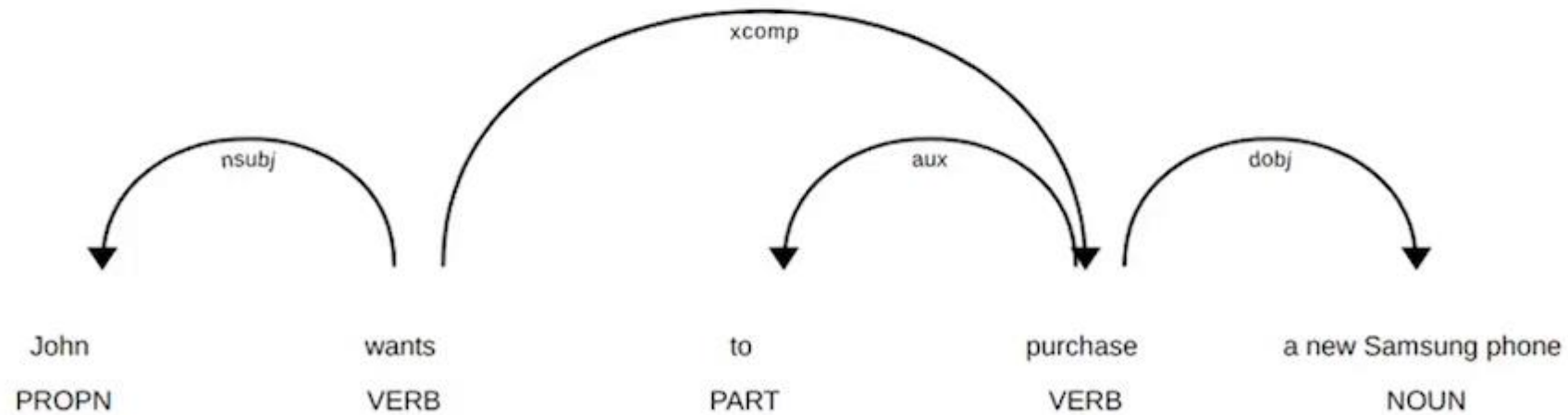
# Named Entity Extraction

- Identification of noun phrases
- Noun phrases: connected by direct subject or object relationships
- Sentence: John wants to purchase a new Samsung phone

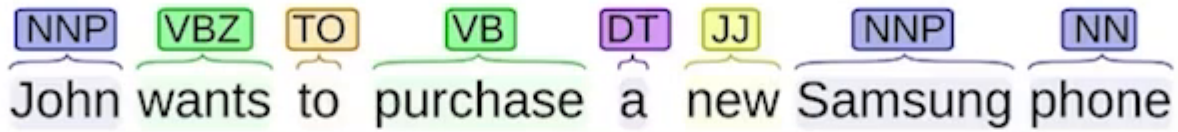


# Named Entity Extraction

- Identification of noun phrases
- Noun phrases: connected by direct subject or object relationships
- Sentence: John wants to purchase a new Samsung phone

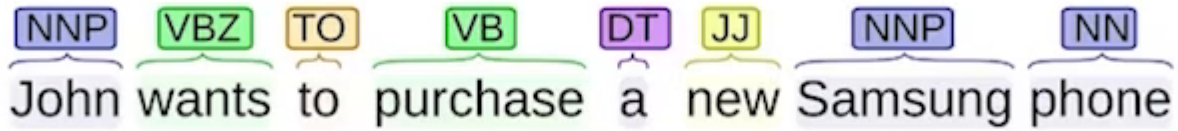


# Named Entity Extraction

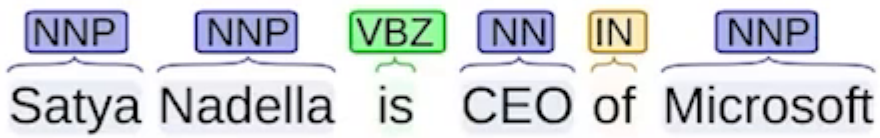



Named Entities: John, Samsung

# Named Entity Extraction




Named Entities: John, Samsung



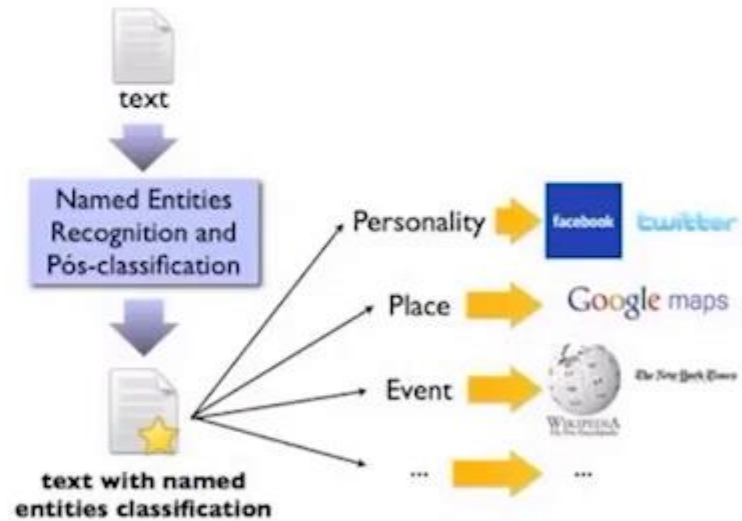

Named Entities: Satya Nadella, Microsoft

# Named Entity Classification / Linking

- Assigning the class / category of the named entity

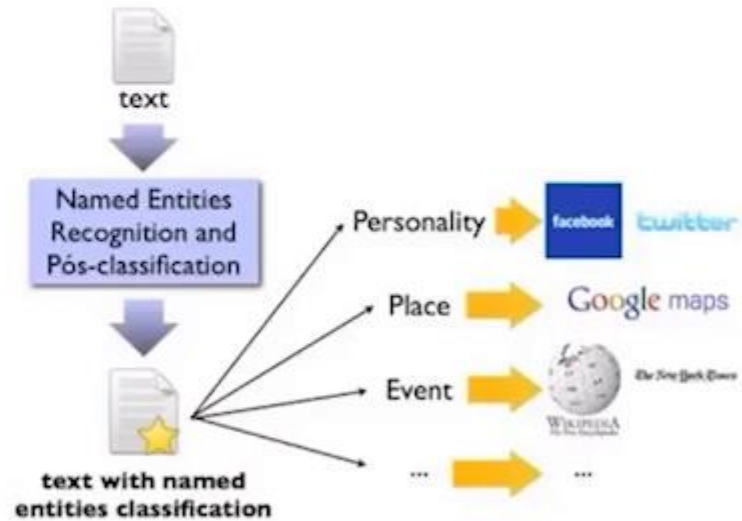
# Named Entity Classification / Linking

- Assigning the class / category of the named entity
- Classes: Person, Organization, Location, Drug etc.
- Entity Linking with a Knowledge Base



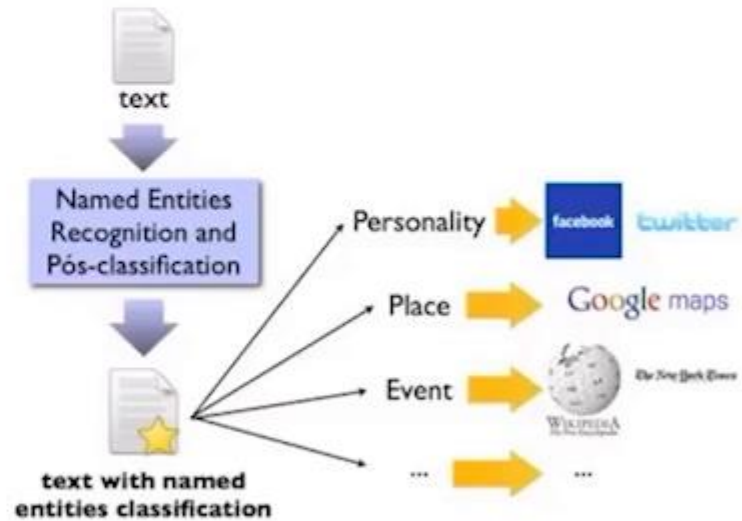
# Named Entity Classification / Linking

- Assigning the class / category of the named entity
- Classes: Person, Organization, Location, Drug etc.
- Entity Linking with a Knowledge Base
- Linking Approaches:
  - Domain Dictionary Lookup (Ontologies)



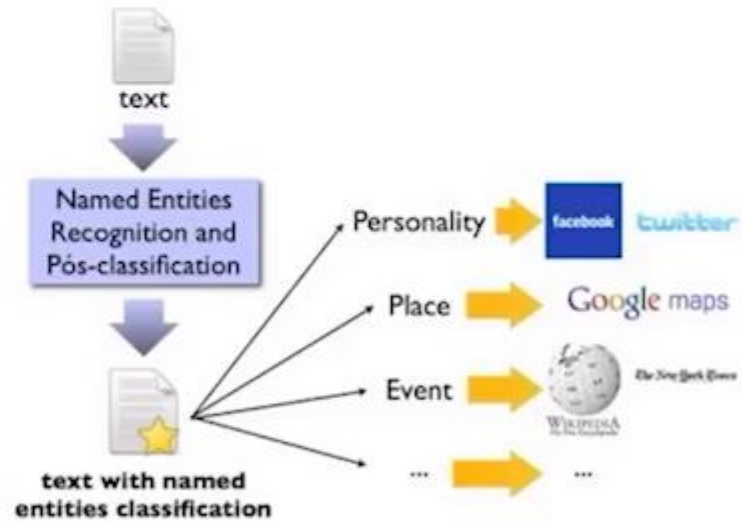
# Named Entity Classification / Linking

- Assigning the class / category of the named entity
- Classes: Person, Organization, Location, Drug etc.
- Entity Linking with a Knowledge Base
- Linking Approaches:
  - Domain Dictionary Lookup (Ontologies)
  - Knowledge Base: Wikipedia, Google



# Named Entity Classification / Linking

- Assigning the class / category of the named entity
- Classes: Person, Organization, Location, Drug etc.
- Entity Linking with a Knowledge Base
- Linking Approaches:
  - Domain Dictionary Lookup (Ontologies)
  - Knowledge Base: Wikipedia, Google
  - APIs - Google Maps, News API



# Challenges

- Updated Ontologies
- Name Variation
  - Singh Yuvraj, Yuvraj Kr Singh

# Challenges

- Updated Ontologies
- Name Variation
  - Singh Yuvraj, Yuvraj Kr Singh
- Multiple Categories
  - London: Name of a person, Name of a city
  - Donald Trump Park: Name of a park(location), Name of a person

# Challenges

- Updated Ontologies
- Name Variation
  - Singh Yuvraj, Yuvraj Kr Singh
- Multiple Categories
  - London: Name of a person, Name of a city
  - Donald Trump Park: Name of a park(location), Name of a person
- Non Capitalization

# Relation Extraction

- Process of extracting relational triples from natural language text.

# Relation Extraction

- Process of extracting relational triples from natural language text.
- Triple represents the relation between a couple of entities.
- E.g., (USA, President, Donald Trump)

# Relation Extraction

- Process of extracting relational triples from natural language text.
- Triple represents the relation between a couple of entities.
- E.g., (USA, President, Donald Trump)
  (Russia, President, Vladimir Putin)
- John Doe was born in England. John is working as a Data Scientist. John also founded ABC Inc. and currently owns a Dodge Hellcat, which is a beautiful car.

# Relation Extraction

- Process of extracting relational triples from natural language text.
- Triple represents the relation between a couple of entities.
- E.g., (USA, President, Donald Trump)
  (Russia, President, Vladimir Putin)
- John Doe was born in England. John is working as a Data Scientist. John also founded ABC Inc. and currently owns a Dodge Hellcat, which is a beautiful car.

  (John Doe, born in, England)
  (John, working, Data Scientist)

# Relation Extraction

- Process of extracting relational triples from natural language text.
- Triple represents the relation between a couple of entities.
- E.g., (USA, President, Donald Trump)
  (Russia, President, Vladimir Putin)
- John Doe was born in England. John is working as a Data Scientist. John also founded ABC Inc. and currently owns a Dodge Hellcat, which is a beautiful car.

  (John Doe, born in, England)
  (John, working, Data Scientist)
  (John, founded, ABC Inc.)

# Types of Relation Extraction

1. Rule-based Relation Extraction:
    a. Uses hand-crafted patterns
    b. Can be made according to particular domains
    c. Very high precision
    d. A lot of Manual Labor
    e. Low Recall

# Types of Relation Extraction

1. Rule-based Relation Extraction:
    a. Uses hand-crafted patterns
    b. Can be made according to particular domains
    c. Very high precision
    d. A lot of Manual Labor
    e. Low Recall

2. Supervised Relation Extraction:
    a. Consider two entities E1 and E2
    b. Detect if a relation exists between E1 and E2
    c. The task becomes of Relation Detection
    d. Needs a lot of labelled data
    e. Expensive to label data

# Types of Relation Extraction

3. Semi-Supervised Relation Extraction:
    a. Used when don't have enough labelled data
    b. Starts with a set of seed tuples
    c. Extract relations from text
    d. Iterative Process
    e. Higher Recall than Rule-based Relation Extraction

# Types of Relation Extraction

3. Semi-Supervised Relation Extraction:
    a. Used when don't have enough labelled data
    b. Starts with a set of seed tuples
    c. Extract relations from text
    d. Iterative Process
    e. Higher Recall than Rule-based Relation Extraction
4. Distantly Supervised Relation Extraction:
    a. Combines the idea of seed with relation detection
    b. Uses a set of tuples from existing knowledge bases
    c. Checks if entities are present in a sentence
    d. Extracts features and trains a classifier
    e. Allows to work with the large labelled dataset
    f. Produces noisy annotation and restricted to the knowledge

# Types of Relation Extraction

5. Unsupervised Relation Extraction:
    a. Also known as Open Relation Extraction
    b. No training data & pre-defined rules required
    c. Extract new relations from the web