



Co-reference resolution

What is Reference Resolution?

Reference Resolution

- Reference Resolution:
 - Which words/phrases refer to some other word/phrase?
 - How are they related?
- Anaphora vs. Cataphora
 - Anaphora: an *anaphor* is a word/phrase that refers back to another phrase: the *antecedent* of the anaphor
 - *Mary thought that she lost her keys.*

 - Cataphora (less common): a *cataphor* is a word/phrase that refers forward to another phrase: its *precedent*.
 - *She was at NYU, when Mary realized that she lost her keys.*

 - *Anaphora* is often used as a synonym for *Reference Resolution* and the term *antecedent* is often used instead of *precedent*.

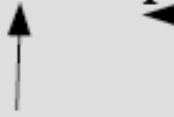
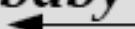

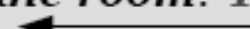

Types of Anaphora

Types of Anaphora I

- Coreference: Antecedent = Anaphor
 - Though **Big Blue** won the contract, this official is suspicious of **IBM**.
 - **Mary** could not believe what **she** heard.
- Similar to Coreference
 - Type Coreference (vs. Token)
 - AKA, identify of sense (vs. identify of reference)
 - John ate **a sandwich** and Mary ate **one** also.
 - Bound variable
 - Every **lioness** guards **its** cubs
 - $(\forall \text{lioness } L)(L \text{ guards } L\text{'s cubs})$
- Predication and Apposition: some (not all) specs label as coreference
 - *Mary is a basketball player*
 - *Mary, a basketball player from NYU*

Types of Anaphora (Contd..)

Types of Anaphora II

- **Bridging Anaphora:** links between “related” objects
 - *The amusement park is very dangerous. The gate has sharp edges. The rides have not been inspected for years.*

- Some IE relation instances can be viewed as bridging
 - When *the baby* cried, *the parents* rushed into the room.

 - ACE Relation: **Per-Social.family**(*the baby, the parents*)
- **“Other” Anaphora:** words including *other* and *another* invoke an “other instance of type” relation
 - *This book* is valuable, but *the other book* is not.

- **Non-NP Anaphora**, e.g., events/propositions
 - *Mary left the room. This upset her parents.*

 - *John read the dictionary. Then Mary did it too.*


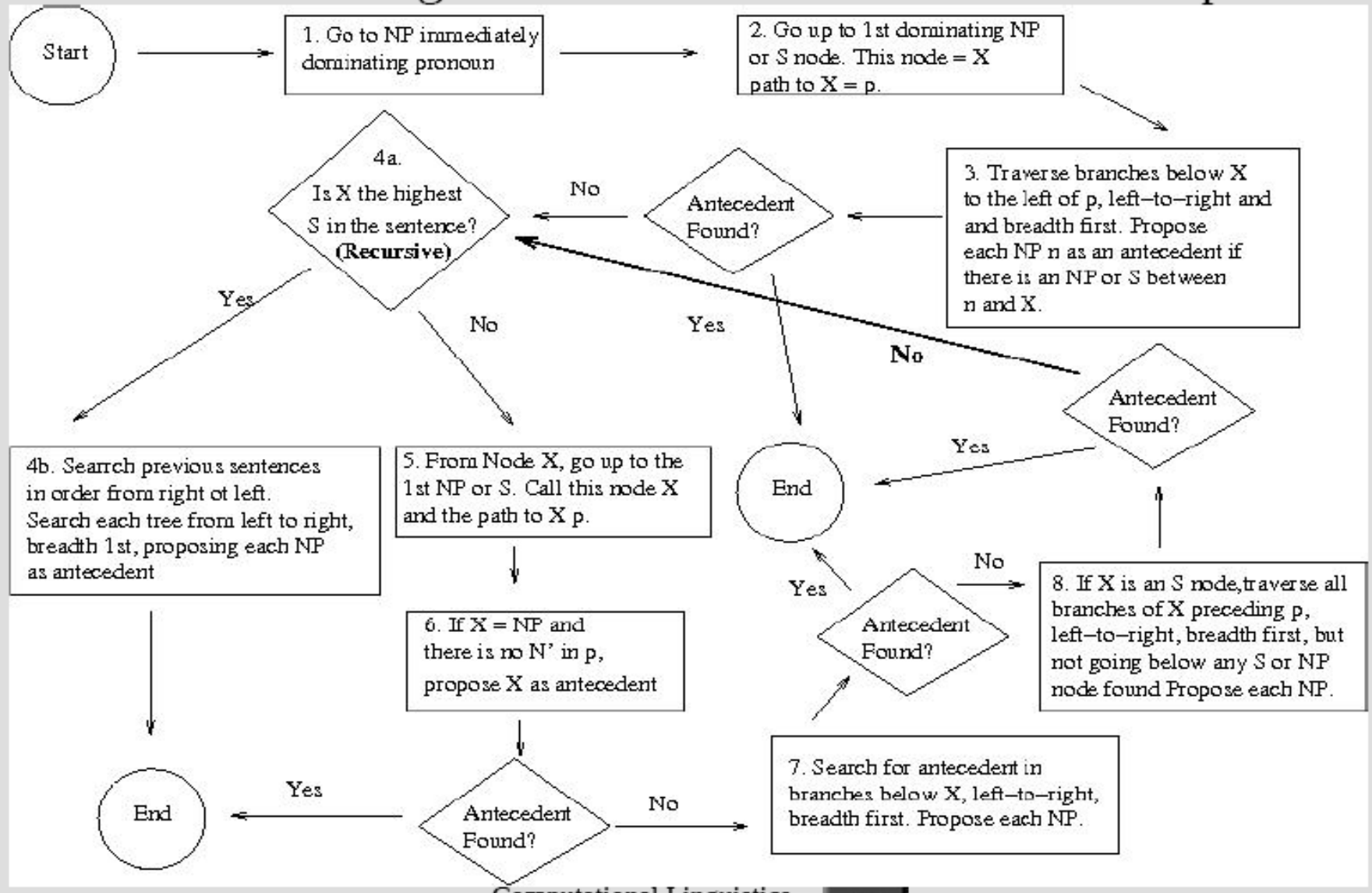
Coreferencing Algorithms

Pronoun Resolution Methodology

- Hobbs search:
 - a simple system that provides a high baseline
 - Lappin and Leas (1994) report 82% F-score for Hobbs Search
- Sets a High Baseline for Pronoun Coreference
- Higher Scoring Systems Tend to be Much More Complex

Example of Coreference algo

Hobbs Search Algorithm to Find Antecedent of Anaphors



Named Entity Recognition

Named Entity(NE) Recognition

- What is Named Entity (NE) ?

Named Entity(NE) Recognition

- What is Named Entity (NE) ?
 - Named entities are proper nouns.
 - Named entity tasks often include :
 - expressions for date and time,
 - names of persons, organization, location, sports adventure activities, etc
 - terms for biological species and substances

Categorization of Named Entity

Categories and subcategories of Named Entities:

- 1) Entity (ENAMEX): person, organization, location
- 2) Time expression (TIMEX): date, time
- 3) Numeric expression (NUMEX): money, percent.

Named Entity Recognition and Classification (NER)

- Recognition of information units like names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions required for various Information Extraction and NLP tasks.
- Identifying references to these entities in text is called as Named Entity Recognition and Classification (NER)

- Most NER systems has been structured as taking an unannotated block of text,

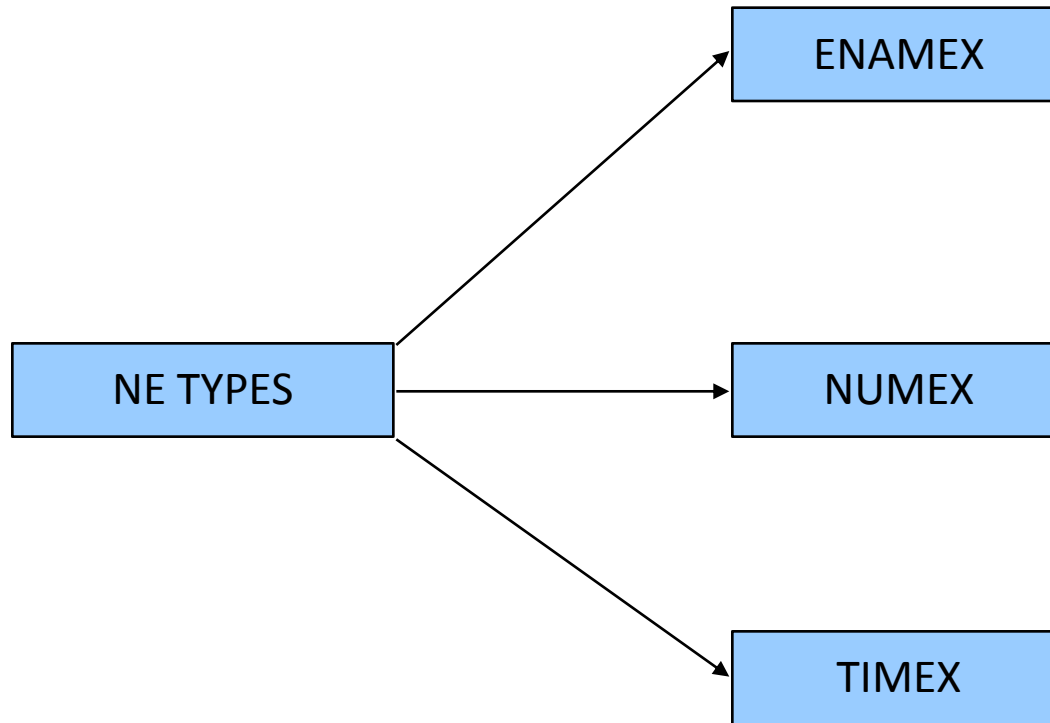
Example : *“The delegation, which included the commander of the U .N. troops in Bosnia , Lt. Gen. Sir Michael Rose reached Sarajevo on 13th October .”*

Annotated block of text : *“The delegation, which included the commander of the <ORG> U .N. </ORG> troops in <LOC> Bosnia </LOC>, <PERS>Lt. Gen. Sir Michael Rose </PERS> reached <LOC> Sarajevo </LOC> on <TIME>13th October </TIME>”.*

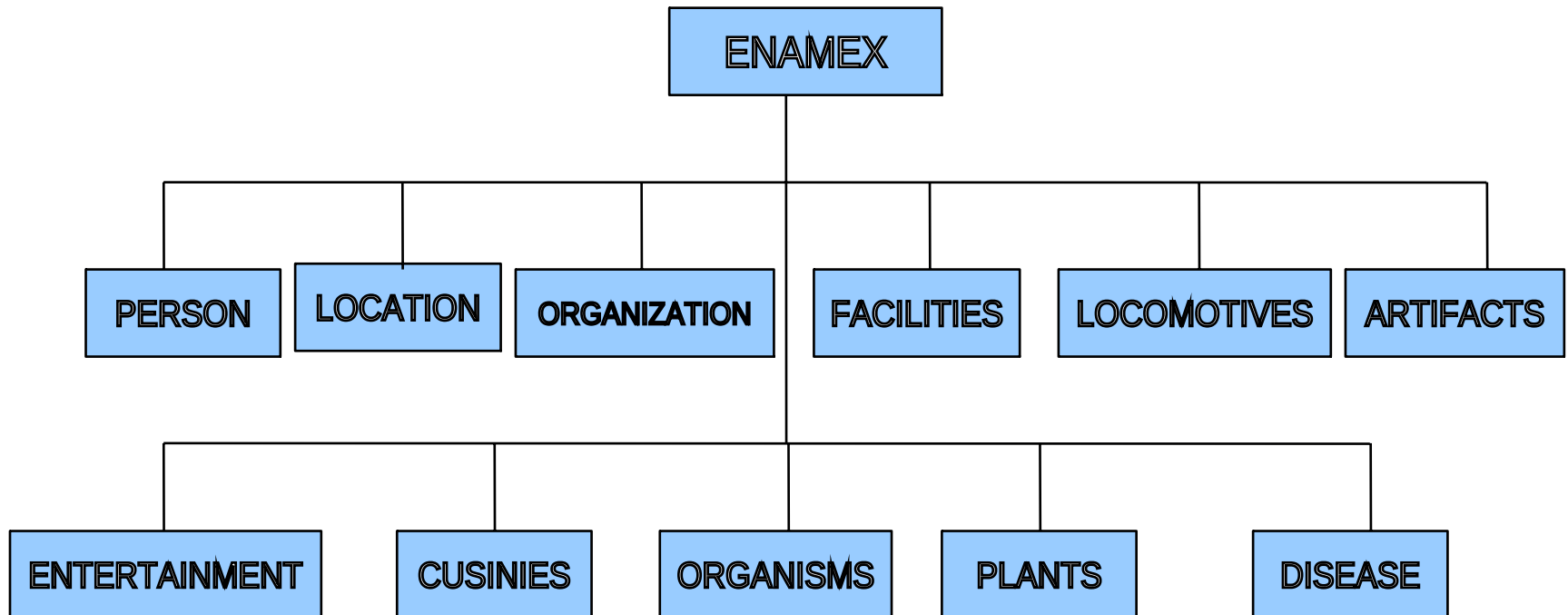
Note: Both the boundaries of an expression and its label must be marked

NE Types

The Named entity hierarchy is divided into three major classes Entity Name, Time and Numerical expressions.



Entity Types



Entity Name Types

- ❑ **Persons** are entities limited to humans. A person may be a single individual or a group. Individual refer to names of each individual person. Group refers to set of individual
- ❑ **Location** entities are limited to geographical entities such as geographical areas like names of countries, cities, continents and landmasses, bodies of water, and geological formations.
- ❑ **Organization** entities are limited to corporations, agencies, and other groups of people defined by an established organizational structure

Examples for Entity Name Types

- En: [Sita]_{PERSON} is working at [HCL]_{ORGANIZATION}, which is in [Chennai]_{LOCATION}
- Hi: [Seetha]_{PERSON} [HCL]_{ORGANIZATION} main kaam kar rahi hai, jo [chennai]_{LOCATION} main hain.

En: Seetha HCL work is which is in
Chennai

Entity Name Types

Facilities are limited to buildings and other permanent man-made structures and real estate improvements like hospitals, airport, colleges, libraries etc.

En: [Appolo Hospital]_{FACILITY} is in Chennai_{LOCATION}

Hi: [Appolo aspathaal]_{FACILITY} [chennai]_{LOCATION} mein hain.

Entity Name Types

A **locomotive** entity is a physical device primarily designed to move an object from one location to another, by carrying, pulling, or pushing the transported object.

En: [Ananthapuri Express]_{LOCOMOTIVE} departs from [Chennai]_{LOCATION} at [7.30pm]_{Time}.

Hi: [Ananthapuri express]_{LOCOMOTIVE} [Chennai]_{LOCATION} se [rAth
7.30]_{TIME} ko ravana hogi

Entity Name Types

Artifact entities are objects or things, produced or shaped by human craft, such as tools, weapons/ammunition, art paintings, clothes, ornaments, medicines, etc

En: [Vinayaka Statue]_{ARTIFACT} is looking beautiful

Hi: [Vinayaka moorthi]_{ARTIFACT} achi lagh rahi hain.

Entity Name Types

Entertainment entities denote activities, which are diverting and hold human attention or interest, giving pleasure, happiness, amusement especially performance of some kind such as dance, music, sports, events.

En: [Flower Exhibition]_{ENTERTAINMENT} is held at
[Hyderabad]_{LOCATION}

Hi: [phool pradarshnii]_{ENTERTAINMENT} [hyderabad]_{LOCATION} mein
Ayojith kiyaa jatha hai

Entity Name Types

Materials refer to the names of food items, cuisines, chemicals and cosmetics

En: [Honey]_{MATERIALS} is good for face

Hi: [Shahad]_{MATERIALS} chehare ke liye achcha hai.

Entity Name Types

ORGANISMS: These are the names of different animal species including birds, reptiles, viruses, bacteria and names of herbs, medicinal plants, shrubs, trees, fruits, flowers etc.

En: [Peacock]_{ORGANISM} is the national bird of [India]_{LOCATION}

Hi: [Mor]_{ORGANISM} [bhaarath]_{LOCATION} kaa raashtriya pakshi hai.

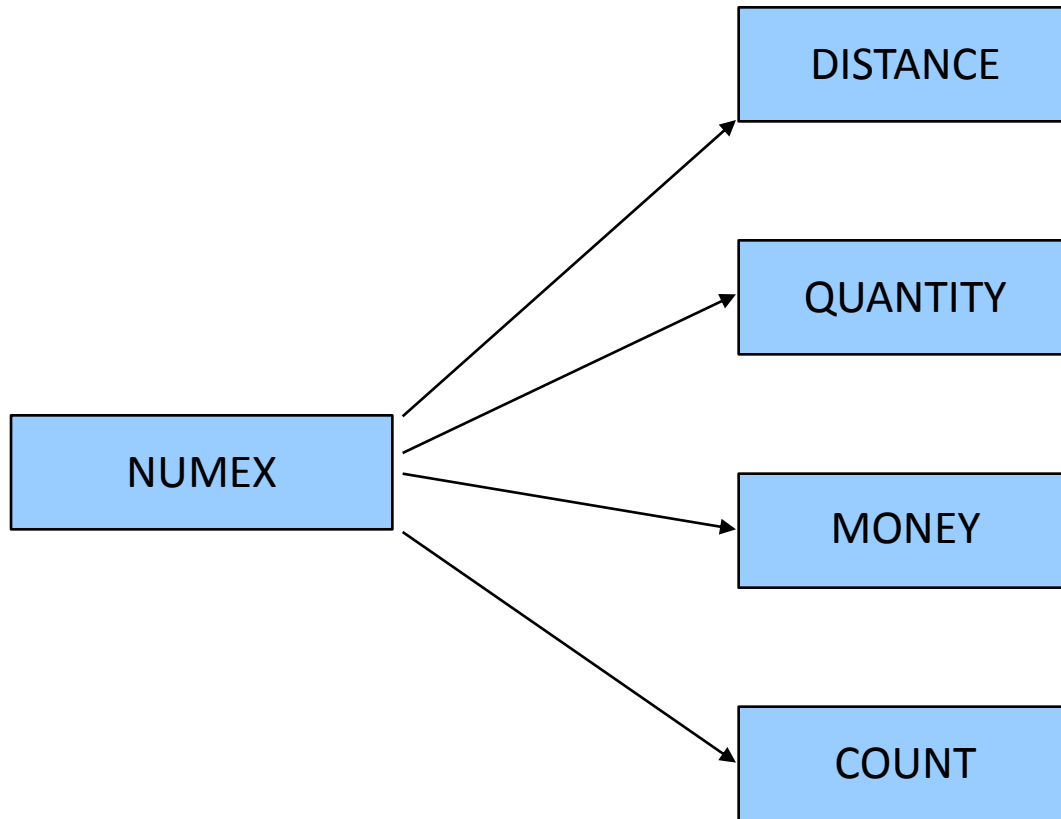
Entity Name Types

Disease: Names of disease, symptoms, diagnosis and treatment are comes under this type.

En: Smoking Causes [Cancer]_{DISEASE}

Hi: dhumrapan [kaansar]_{DISEASE} ka kaaran banaatha hai.

Numerical Expressions



Numerical Expressions

- Distance refers to the distance measures such as kilometers, Centimeters, meters, acres, feet etc.

Example: 10 cm., twenty feet, 15 hectares

- Money specifies the different currency value such as rupee, euro, Dinar, dollar etc.

Example: Rs. 1000, 250 Euro, \$160

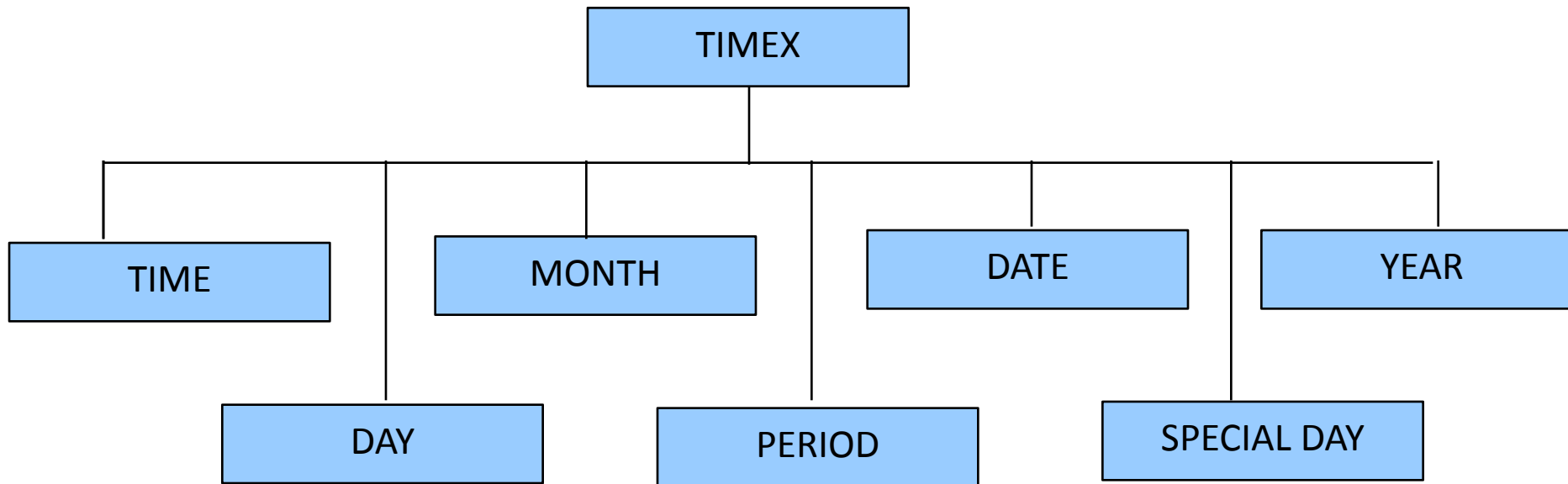
- Count denotes the number (or counts) of Items/ articles/things etc.

Example: 5 subjects, 12 students, 20 books

- Quantity measurements like liters, tons, grams, volts etc. are comes under this category.

Example: 20 litres, 22 kg, 50g, 100 volts

Time Expressions



Temporal Expressions

- Temporal expressions are the entities refers to time, date, year, month and day
- Time: These refer to expressions of time, includes different forms of expressing time. This also includes Hours, minutes and seconds.
- Example:
 - *5 'o clock in the morning*
 - 9.30 a.m.
 - Evening 6.30 p.m.
- Date: This refers to expressions of Date such as 13/12/2001 etc in different forms. This also includes month, date and year
- Example
 - August 15 1947 , 1956 , September 11

Temporal Expressions

Day: These are expressions, which convey days in a year. Also it can include days occurring weekly /fortnightly/ monthly /quarterly/ biennial etc.

Example

- Sunday
- Tomorrow
- Today
- Yesterday

Special Day: refers to special days in a year

Example

- Gandhi Jayanthi
- Rama Navami

Temporal Expressions

Period: refers to expressions, which express duration of time or time periods or time intervals.

Example

- 17 th century
- 10 minutes
- 10 a.m. to 12 p.m.
- One year

Challenges of NER in Indian Languages

Following are the major challenges encountering in Indian Languages.

- Ambiguity
 - Between Proper and common nouns
 - Between named entities
- Lack of Capitalization

Some problems in indentifying NE

- Variation of NEs.
 - Manmohan Singh, Manmohan, Dr. Manmohan Singh
- Ambiguity of NE types:
 - 1945 (date vs. time)
 - Washington (location vs. person)
 - May (person vs. month)
 - Tata (person vs. organization)

Challenges of NER in Indian Languages

- Ambiguity
- Comparatively Indian languages suffer more due to the ambiguity that exists between common & proper nouns and between named entities itself. In some cases same word can refer to different named entity types. Those instances can be recognized by contextual information.
- Examples:
 - Hi: Akash - Person name and Sky
 - Hi: Sooraj - Person name and Sun
 - Hi: Chaandh – Moon and Silver
 - Hi: Aam – Mango and Common
 - Ml: Roopa – Person name and Rupee
 - Ml: Madhu – Person name and Honey
 - Ml: Mala – Person name and Garland

Challenges of NER in Indian Languages

Spell Variation: Due to the different writing styles same entity is represented in various word forms. In Tamil, sanskrit letters such as “ja”, “sha”, “sri” “Ha” are replaced by “sa”, “ciri”, “ka”

Example:

Roja can be written as Rosa

Srimathi	-	cirimathi
----------	---	-----------

Raja	-	rasa
------	---	------

ShajahAn	-	sajakAn
----------	---	---------

Challenges of NER in Indian Languages

Lack of Capitalization

- In English and some other European languages capitalization is considered as the important feature to identify proper noun.
- It plays a major role in NE identification.
- Unlike English capitalization concept is not found in Indian languages.

Nested Entities

Nested Entities: Refers to the named entities which occurs within another named entities. Also called as embedded entities.

Ta: [[Mathurai]_{LOCATION} [MeenAtchi Amman]_{PERSON} Koyil]_{RELPLACE}
En: Mathurai Meenakshi Amman Temple

MI: [[Nittoor]_{PERSON} Srinivasa rao]_{PERSON}
En : Nittoor Srinivasa rao

Hi: [[Rajeev]_{PERSON} MArg]_{ROAD}
En : Rajeev Road

Applications of NER

1. Search Engines- NER helps in structuring textual information
2. Cross-Lingual Information Access Retrieval (CLIR)- given a query word, it is very important to find if it is a named entity or not.
3. News aggregation platforms- powered by named entity recognition- plotting the popularity of entities over time and generating geospatial heat maps
4. Machine translation
5. Automatic indexing of Books: Most of the words indexed in the back index of a book are Named Entities.
6. Useful in Biomedical domain to identify Proteins, medicines, diseases, *etc.*

[Message Understanding Conference]

Methodologies

Methods:

1) Rule Based

2) Machine Learning

Hidden Markov Model (HMM)

Naïve Bayes Classifier

Maximum Entropy Markov Model (MEMM)

Conditional random Fields (CRF)

4) Hybrid Approach (Statistical+ Linguistics)

Dictionary (Gazetteers) Look-up Approach

- Uses Dictionaries for identifying NERs (Gazetteers)
- Gazetteer contains NEs from all domains
- Advantage
 - Very simple approach
 - Gives very high precision

Disadvantages of Dictionary Approach

- Preparation of exhaustive dictionary is a tedious and expensive process.
- The dictionary should cover the different spellings of the same place.