LECTURE 1

☐ Sentence Segmentation:

Detecting a sentence correctly.

- CHALLENGES
  - '.' cannot always detect (USA)
  - '!', → "Alas! he died"
  - '?' → "What!?"
- SOLUTION
  - Rules, &
    - Large set of rules are hard to maintain.
  - Machine Learning
    - 99% perfect.

☐ Tokenization: Detect a word?

But on what?

- Challenges
  - punctuation, C++, C#,
  - emoticons. ☺ :-
  - New York vs York!
  - Phrasal verb → Work out
  - Different Language.

Solution:

- Each language needs it's own soln. So no general soln.

☐ Stemming

অবার নেড এ আঙুল লাগায় দেওয়া.

Equivalence ⟶ Equival.

Fast, but innacurate.

Organization ⟶ Organ

⚠ UN নেড হয়ে গেছে!

☐ Lemmatization.

- Hand built lexicon for all word forms. (Walked → Walk).
- Accurate but slower.
- Chicken egg scenario with POS tagging.

// Lemmatization is like figuring out what kind (the root word) an egg (the word in the sentence) will become. But to do that accurately, you need to know the context in which is used, which is provided by POS tagging. //

Embedding

- Word as a vector.

→ Dog = [0.5; 0.4; 0.1]

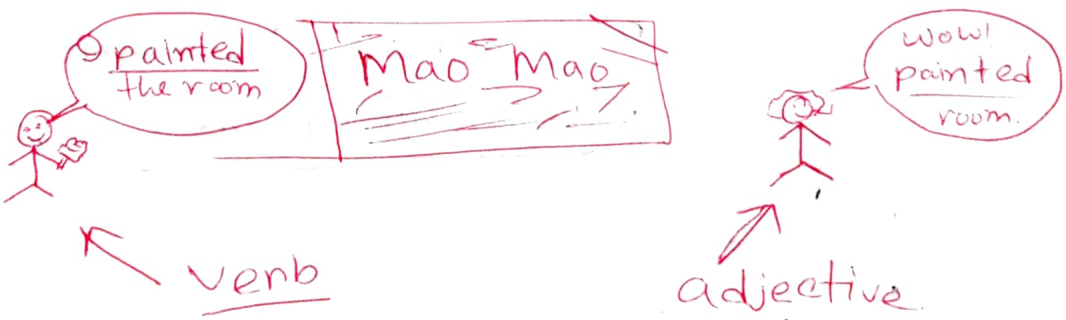
Dogs = [0.5; 0.4; 0.2]

Parts of speech.

- Closed class → Suppose pronouns.
- Open class
  → Names! → "Eshita the gach goru"

Challenge

* One word can have diff. POS tag based on it's use.



Named Entity Recognition

Indentify phrases that are named people, locations, organizations.

Person: Turing is the. . . .

Org: ACM is a. . .

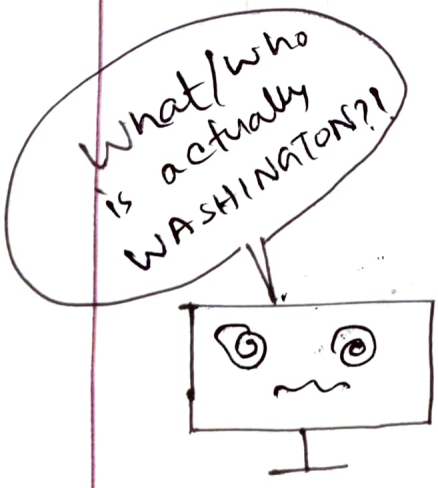
location: The Mt. Everest is

Geo pol. entity: U.S.A is the . . .

Challenge

* Ambiguity

- Washington was born into slavery
- Washington went up 2 games
- I arrived Washington today
- Washingto passed a pri. . . law.

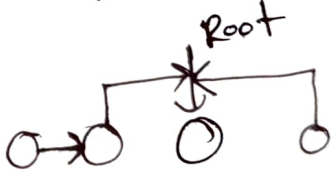


Soln

BIO → Begin name
Inside "
Outside "

BILOU → Begin "
Inside "
Last "
Outside "
Unit - length.

Parsing & Syntactic Representation.

* Contituency tree 
* Dependency tree 

Challenge:

Attachment Ambiguity.

One morning, I saw Messi in my Pajamas.

Who was in my PAJAMAS?

— Me?

— Messi?

Coordination Ambiguity.

Old men and Women.

- old(men and Women)?
- old (men) and Women ?

Soln

* Probabilistic grammar parsing
* Transitin based parsing.