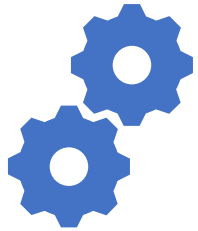
Natural Language Processing

Lec 2: Introduction

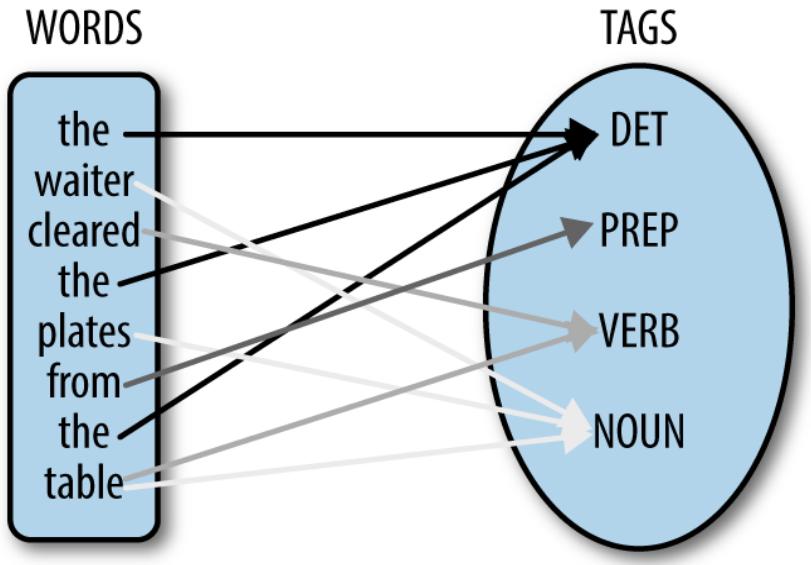
Some slides in this lecture are inspired by Dr. Kai-wei Chang UCL, Standford course CS124.

NLP main components and tasks



Parts of Speech (POS) Tagging

- POS tagging is a process of assigning a POS or lexical class marker to each word in a sentence (and all sentences in a corpus).



NLP Tasks

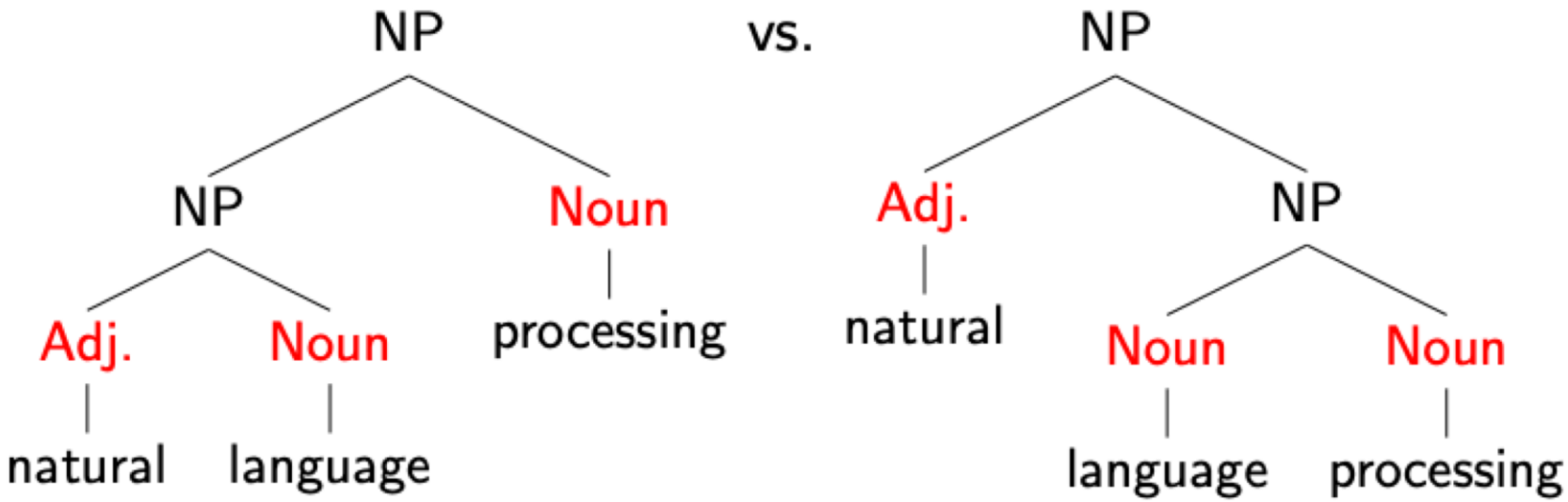
- NLP applications require several NLP analyses:
  - Word tokenization
  - Sentence boundary detection
  - Part-of-speech (POS) tagging
    - to identify the part-of-speech (e.g. noun, verb) of each word
  - Named Entity (NE) recognition
    - to identify proper nouns (e.g. names of person, location, organization; domain terminologies)
  - Parsing
    - to identify the syntactic structure of a sentence
  - Semantic analysis
    - to derive the meaning of a sentence

Named Entity Recognition (NER)

- NER is to process a text and identify named entities in a sentence
  - e.g. “U.N. official Ekeus heads for Baghdad.”

[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad] .

Parsing


NP
NP
Noun
Adj.
Noun
processing
natural
language
vs.
NP
Adj.
NP
natural
Noun
Noun
language
processing

Semantics

- Every fifteen minutes a woman in this country gives birth.
- Our job is to find this woman, and stop her!

Making progress on this problem…

- The task is difficult! What tools do we need?
  - Knowledge about language
  - Knowledge about the world
  - A way to combine knowledge sources
- How we generally do this:
  - probabilistic models built from language data
    - P(“maison” → “house”) high
    - P(“L’avocat général” → “the general avocado”) low
  - Luckily, rough text features can often do half the job.

Where Are We Now ?

deepseek

Basic Text Preprocessing

Regular expressions

- A formal language for specifying text strings
- How can we search for any of these?
  - woodchuck
  - woodchucks
  - Woodchuck
  - Woodchucks

Basic ways

Regex	Example Patterns Matched
/woodchucks/	"interesting links to woodchucks and lemurs"
/a/	"Mary Ann stopped by Mona's"
/!/	"You've left the burglar behind again!" said Nori

Regex	Match	Example Patterns Matched
/[A-Z]/	an upper case letter	"we should call it 'Drenched Blossoms' "
/[a-z]/	a lower case letter	"my beans were impatient to be hoed!"
/[0-9]/	a single digit	"Chapter 1: Down the Rabbit Hole"

Regular Expressions: Disjunctions

- Letters inside square brackets []

Pattern	Matches
`[wW]oodchuck`	Woodchuck, woodchuck
`[1234567890]`	Any digit

- Ranges `[A-Z]`

-

Pattern	Matches	
`[A-Z]`	An upper case letter	Drenched Blossoms
`[a-z]`	A lower case letter	my beans were impatient
`[0-9]`	A single digit	Chapter 1: Down the Rabbit Hole

Regular Expressions: Negation in Disjunction

- Negations [^Ss]
  - Carat means negation only when first in []

Pattern	Matches	
[^A-Z]	Not an upper case letter	Oyfn pripetchik
[^Ss]	Neither 'S' nor 's'	I have no exquisite reason"
[^e^]	Neither e nor ^	Look here
a^b	The pattern a carat b	Look up a^b now

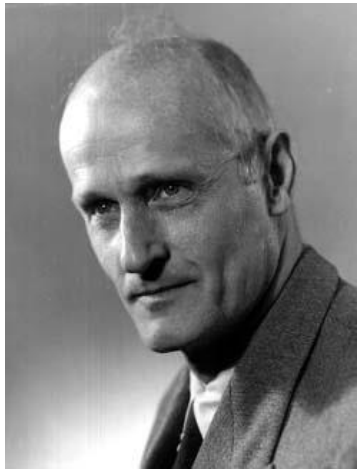
Regular Expressions: More Disjunction

- Woodchucks is another name for groundhog!
- The pipe | for disjunction

Pattern	Matches
`groundhog|woodchuck`	
`yours|mine`	yours mine
`a|b|c`	= [abc]
`[gG]roundhog|[Ww]oodchuck`	

Regular Expressions: ? * + .

Pattern	Matches	
colou?r	Optional previous char	color colour
oo*h!	0 or more of previous char	oh! ooh! oooh! ooooh!
o+h!	1 or more of previous char	oh! ooh! oooh! ooooh!
baa+		baa baaa baaaa baaaaa
beg.n	Exactly 1 char	begin begun begun beg3n



Stephen C Kleene

Kleene *, Kleene +

Regular Expressions: Anchors ^ $

Pattern	Matches
^[A-Z]	Palo Alto
^[^A-Za-z]	1 "Hello"
\.$	The end.
.$	The end? The end!

Example

- Find me all instances of the word “the” in a text.
  - `the` Misses capitalized examples
  - `[tT]he` Incorrectly returns `other` or `theology`
  - **`[^a-zA-Z][tT]he[^a-zA-Z]`**

Errors

- The process we just went through was based on fixing two kinds of errors
  - Matching strings that we should not have matched (there, then, other)
    - False positives (Type I)
  - Not matching things that we should have matched (The)
    - False negatives (Type II)

Errors cont.

- In NLP we are always dealing with these kinds of errors.
- Reducing the error rate for an application often involves two efforts:
  - Increasing accuracy or precision (minimizing false positives)
  - Increasing coverage or recall (minimizing false negatives).

Summary

- Regular expressions play a surprisingly large role
  - Sophisticated sequences of regular expressions are often the first model for any text processing tasks
- For many hard tasks, we use machine learning classifiers
  - But regular expressions are used as features in the classifiers
  - Can be very useful in capturing generalizations

Enough?!

- Enough! OK…any questions?!