Introduction to **natural language processing**

Hutchatai Chanlekha

Department of Computer Engineering, Faculty of Engineering
Kasetsart University

Why we need NLP?

- We want computers to do things for us
  - Read our email
  - Selecting interesting articles for us
  - Spell check, grammar check, summarizing documents, …
  - Do library search
  - Etc.
- All of these things involve dealing with natural language (NL)
  - A lot of human communication is my means of NL
  - A lot of information that is possessed by human beings is in NL

NLP and Computer

- Computers are good at dealing with machine language that are made for them
    - Precisely specify
    - Unambiguous
- But natural languages aren’t like that!
- Many researchers try to find different ways, which are easy for computers, to avoid dealing with NL
    - XML
    - Semantic web
    - Designing applications with things like menus or drop-down boxes
    - Etc.
- But we want to tackle some of the hard problems …

What is NLP?

- NLP is a subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

  https://en.wikipedia.org/wiki/Natural_language_processing

- NLP is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language.

  The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable.

  https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32

- NLP is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language. NLP draws from many disciplines, including computer science and computational linguistics, in its pursuit to fill the gap between human communication and computer understanding.

  https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html

Can you give examples of daily applications that involve NLP?

- MS office, Spelling checker, Grammar checker
- Spam detection, Automatic Calendar Entries,
- Google search, Google translation
- Siri, some call center services
- Chat bot
- Etc.

Goal of NLP

- Can be far reaching …
  - Understanding text
  - Summarizing text
  - Reasoning about the consequences of text
  - Real-time spoken dialogues

- Can be down-to-earth
  - Context sensitive spelling correction
  - Keyword search
  - Extracting company names and locations from news articles
  - Text categorization

- Currently, the later systems still predominate
  - as NLP becomes increasingly practical, focused on performing measurably useful task now.
- Although language complex and ambiguity, NLP can also be surprisingly easy sometimes:
  - Rough text features often do half the job.
  - Just looking through a large amount of text and counting things, then predict base on counting

Natural Language Processing is difficult because:

- Natural language is:
  - highly ambiguous, complex, and subtle
  - interpretation involves *combining evidence*
    - Same sentence but was spoken in different situation/intention can have different meaning
      - “ดังมาก” “กินได้”, etc.
  - involves reasoning about the world
      - (Food recommendation chat bot) User: “ฉันแพ้ถั่วกับแป้งสาลี”
  - embedded a social system of people interacting
  - persuading, insulting and amusing people
  - changing over time

NL is ambiguous

Natural language Understanding depends on making complex and subtle use of context, both the context of the rest of the sentence, and context of prior discourse and thing around us in the world.

- Ambiguity
  - One word can have more than one meaning. Many words can have the same meaning (“ขัน” “หมัด”)
  - Complex/complicated sentence structure
    - “เขาใส่เสื้อยืดลายเสือถือปืน”
    - “ขอให้นิสิตมาส่งงานที่สั่งในวันที่ 20 สิงหาคม ที่ห้องประชุม ที่ชั้น 3 ของภาควิชา”
  - Flexibility of the language (omission, movement, etc.)
    - “ข้าวนี้กินได้นะ” “กินได้นะ ข้าวนี้” “กินข้าวนี้ได้นะ”
  - Idiom, sarcasm
    - อธิบดีสั่งล้อมคอกทั่วประเทศ, ลิ้นเป็นปลาไหล, พูดไปก็เหมือนสีซอให้ควายฟัง
  - Etc.

What computer sees …

ขณะนี้สถานการณ์น้ำท่วมในพื้นที่ภาคใต้
ส่วนใหญ่เริ่มคลี่คลายลงแล้ว

35863603363236093637365736263606
36343609358536343619360336603609
36573635360736563623361736513609
36143639365736093607363736563616
36343588365136053657362636563623
36093651362735973656364836193636
36563617358836213637365635883621
36343618362135913649362136573623

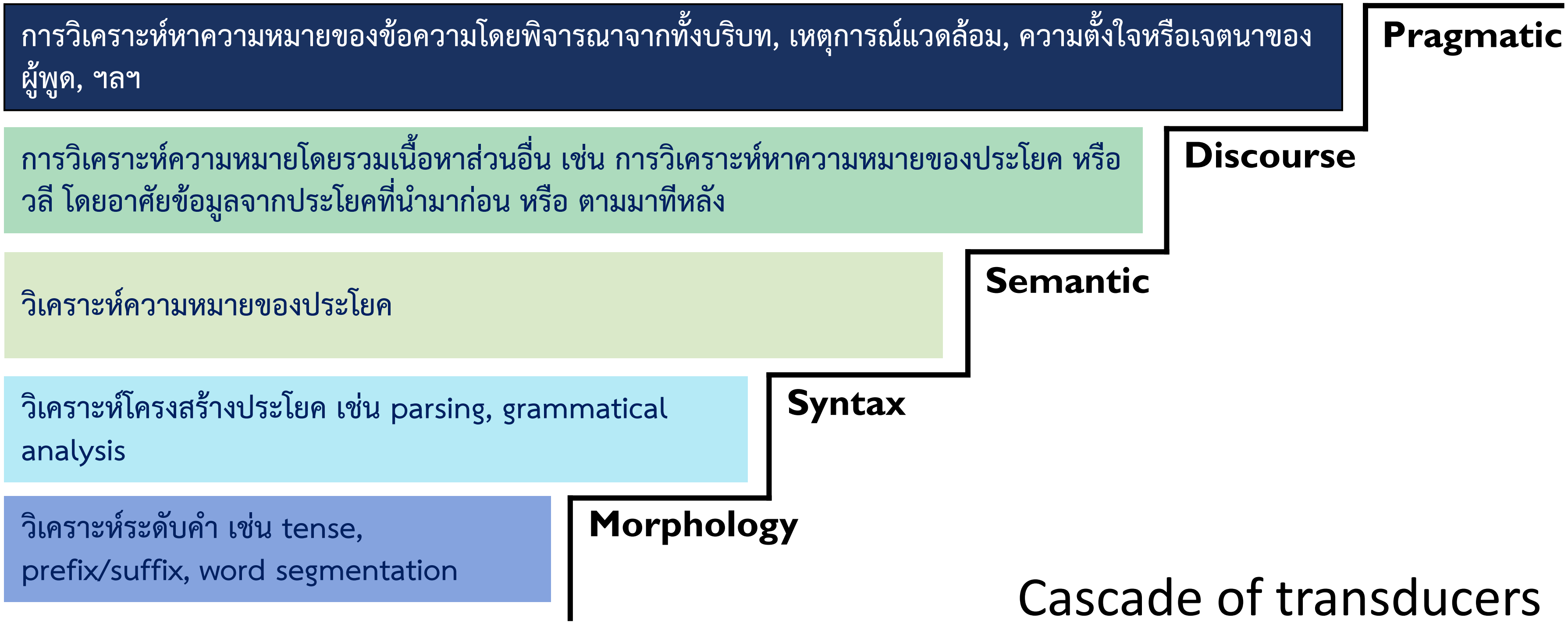
Many people stayed in their
homes throughout the battle

77 97 110 121 32 112 101 111 112 108
101 32 115 116 97 121 101 100 32 105
110 32 116 104 101 105 114 32 104 111
109 101 115 32 116 104 114 111 117 103
104 111 117 116 32 116 104 101 32 98
97 116 116 108 101

Layers of Computational Linguistic

- Phonetics & Phonology
- Morphology
- Syntax
- Semantics
- Discourse
- Pragmatics

Stages in Text Processing



Pragmatic

การวิเคราะห์หาความหมายของข้อความโดยพิจารณาจากทั้งบริบท, เหตุการณ์แวดล้อม, ความตั้งใจหรือเจตนาของผู้พูด, ฯลฯ

Discourse

การวิเคราะห์ความหมายโดยรวมเนื้อหาส่วนอื่น เช่น การวิเคราะห์หาความหมายของประโยค หรือวลี โดยอาศัยข้อมูลจากประโยคที่นำมาก่อน หรือ ตามมาทีหลัง

Semantic

วิเคราะห์ความหมายของประโยค

Syntax

วิเคราะห์โครงสร้างประโยค เช่น parsing, grammatical analysis

Morphology

วิเคราะห์ระดับคำ เช่น tense, prefix/suffix, word segmentation

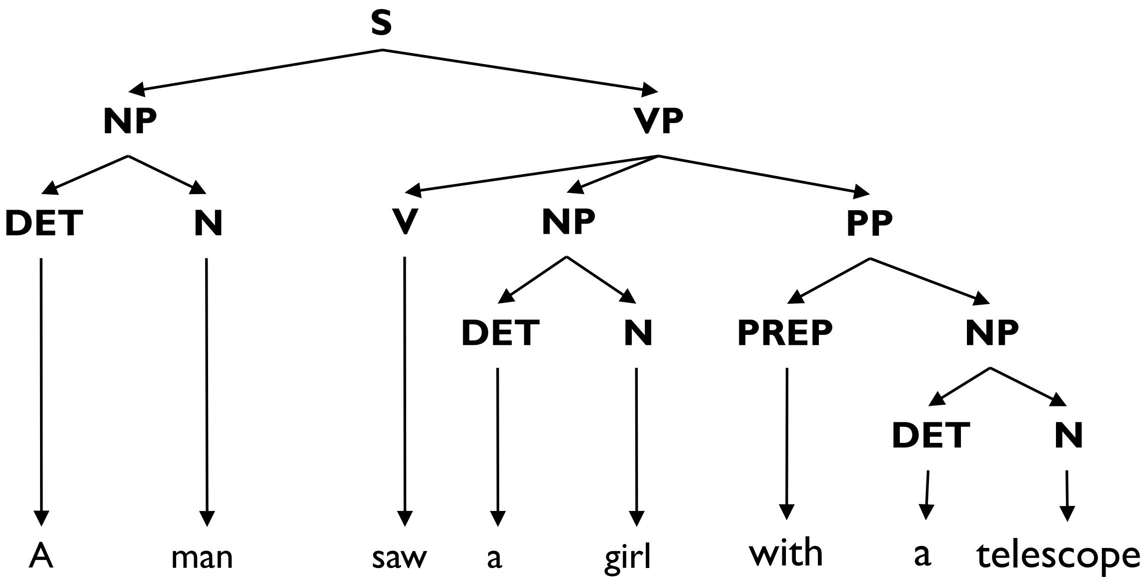
Cascade of transducers

Morphology

- The study of word
  - The study of the sub-word units of meaning (word derivation)
    - English: “disconnect” = dis (not) + connect (to attach)
    - Even more necessary in some other languages,
      - e.g. Turkish: uygarlastiramadiklarimizdanmissinizcasina
        - = uygar las tir ama dik lar imiz dan mis siniz casina
        - = (behaving) as if you are among those whom we could not civilize
  - The study of word inflection
    - English: plural -> singular, past tense -> present tense, etc.
  - Stemming, lemmatization
- What about Thai?
    - No derivation, no inflection
    - “ลูกเสือ” → boyscout or cub?
    - “หนอนแมลงวันเจาะยอดข้าว” → noun phrase or name of worm?

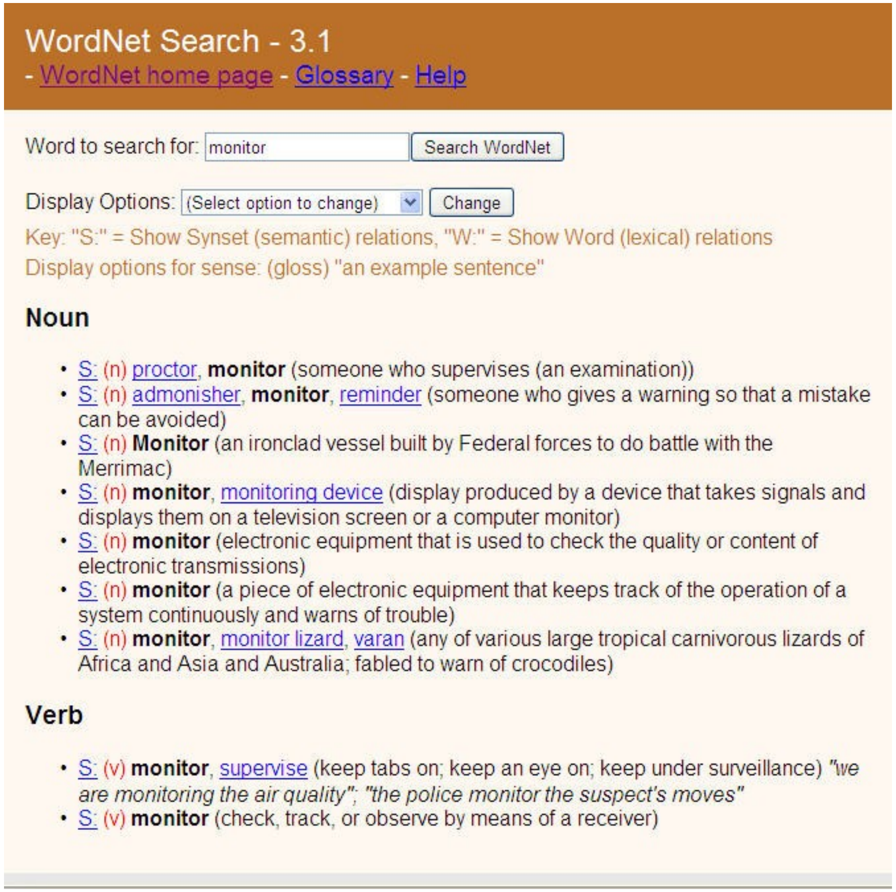
Syntax

- The study of the structural relationships between words.
  - Such as how to generate grammatical-correct sentence
  - Syntactic relationships between words
    - Relationship between verb, its subject and object
    - Relation between modifier and modifiee
    - Etc.



Semantic

- The study of the literal meaning



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: monitor [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) proctor, **monitor** (someone who supervises (an examination))
- S: (n) admonisher, **monitor**, reminder (someone who gives a warning so that a mistake can be avoided)
- S: (n) **Monitor** (an ironclad vessel built by Federal forces to do battle with the Merrimac)
- S: (n) **monitor**, monitoring device (display produced by a device that takes signals and displays them on a television screen or a computer monitor)
- S: (n) **monitor** (electronic equipment that is used to check the quality or content of electronic transmissions)
- S: (n) **monitor** (a piece of electronic equipment that keeps track of the operation of a system continuously and warns of trouble)
- S: (n) **monitor**, monitor lizard, varan (any of various large tropical carnivorous lizards of Africa and Asia and Australia; fabled to warn of crocodiles)

Verb

- S: (v) **monitor**, supervise (keep tabs on; keep an eye on; keep under surveillance) *"we are monitoring the air quality"; "the police monitor the suspect's moves"*
- S: (v) **monitor** (check, track, or observe by means of a receiver)

http://wordnetweb.princeton.edu/perl/webwn

GIVE

Core:

Donor : The person that begins in possession of the Theme and causes it to be in the possession of the Recipient.

Recipient : The entity that ends up in possession of the Theme.

Theme : (**Semantic Type** Physical_object) The object that changes ownership.

I gave John a book. → Action: Give
Actor: I
Recipient: John
Theme: a book

$$\mathrm{ISA}(e, \mathrm{Giving}) \wedge \mathrm{Giver}(e, \mathrm{Speaker}) \wedge \mathrm{GiveTo}(e, \mathrm{John}) \wedge \mathrm{GivenThing}(e, b_1) \wedge \mathrm{ISA}(b_1, \mathrm{Book})$$

Discourse

- The study of linguistic units larger than a single utterance.

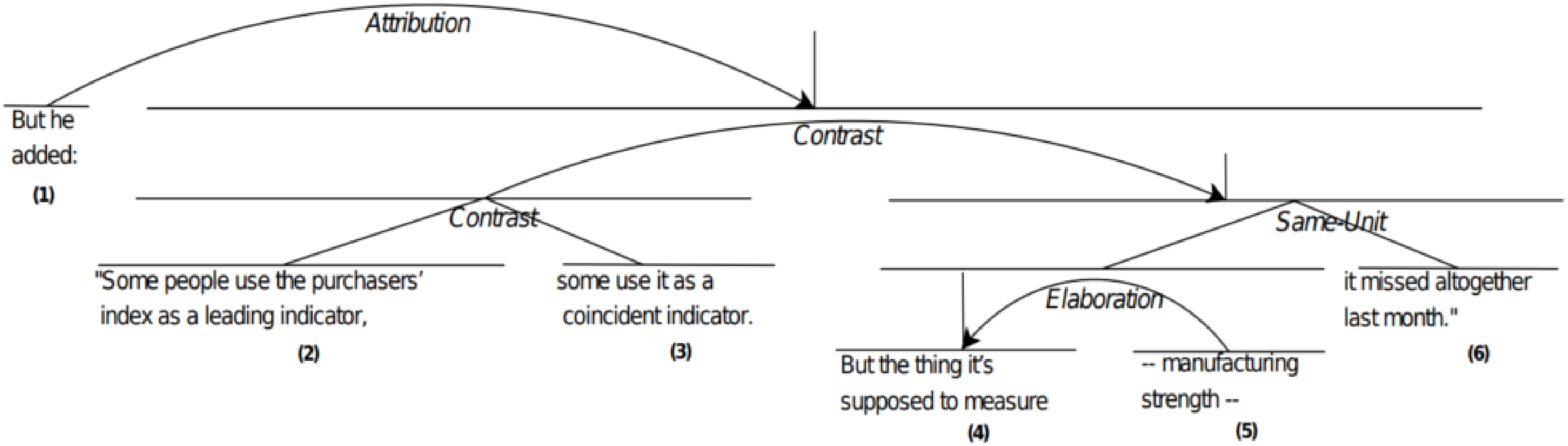
ฟ้า: ไปไหน

แดง: ไปตลาด เอาอะไรไหม

ส้ม: เอาก๋วยเตี๋ยว

ฟ้า: เอาด้วย

“John saw a beautiful mini cooper at the dealership. He showed it to Bob. He bought it.”



Attribution

But he added: (1)

Contrast

Contrast

"Some people use the purchasers' index as a leading indicator, (2)

some use it as a coincident indicator. (3)

Same-Unit

Elaboration

But the thing it's supposed to measure (4)

-- manufacturing strength -- (5)

it missed altogether last month." (6)

Pragmatic

- The study of how language is used to accomplish goals.
  - What should you conclude from the fact I said something? How should you react?
    - **กินข้าว**
      - บอกเล่า?
      - สั่ง?
      - ชักชวน?

Common Pre-processing in NLP Application

ต้นซื้อกล่องไม้ให้มะลิ

- **Word segmentation/tokenization**
  - Finding word boundary

ต้น ซื้อ กล่อง ไม้ ให้ มะลิ

- **Part-of-speech tagging**
  - Part of speech is a category to which a word is assigned in accordance with its syntactic functions
  - Marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context

ต้น/NNP ซื้อ/VBD กล่อง/NN ไม้/JJ ให้/IN มะลิ/NNP

Ton/NNP bought/VBD a/DT wooden/JJ box/NN for/IN Mali/NNP ./.

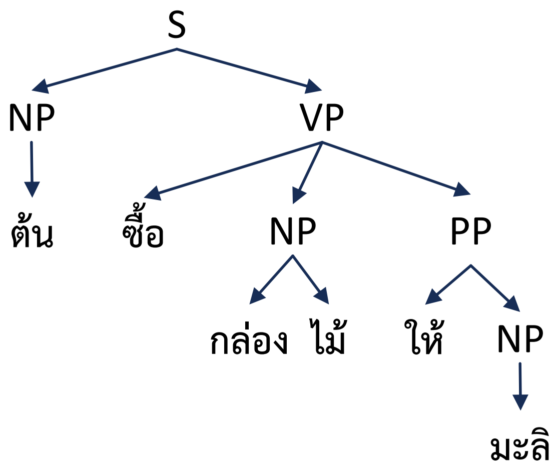
Common Pre-processing in NLP Application

- **Chunking and Parsing**
  - Determining the syntactic structure of a text by analyzing its constituent words based on an *underlying grammar* (of the language)

ต้น ซื้อ กล่อง ไม้ ให้ มะลิ

[[ต้น/NNP]$_{NP}$ [ซื้อ/VBD [กล่อง/NN ไม้/JJ]$_{NP}$ [ให้/IN [มะลิ/NNP]$_{NP}$]$_{PP}$]$_{VP}$]$_{SENTENCE}$

```
(ROOT
 (S
   (NP (NNP ต้น))
   (VP (VBD ซื้อ)
     (NP (NN กล่อง) (JJ ไม้))
     (PP (IN ให้)
       (NP (NNP มะลิ))))
   (. .)))
```



nsubj(ซื้อ-2, ต้น-1)
root(ROOT-0, ซื้อ-2)
amod(กล่อง-3, ไม้-4)
dobj(ซื้อ-2, กล่อง-3)
case(พิไล-6, ให้-5)
nmod:for(ซื้อ-2, มะลิ-6)

Common Pre-processing in NLP Application

- **Named Entity Recognition**
  - Identifying and marking certain classes of names in a document

ต้น ซื้อ กล่อง ไม้ ให้ มะลิ

<PERSON>ต้น</PERSON> ซื้อ กล่อง ไม้ ให้ <PERSON>มะลิ</PERSON>

หลังได้รับรายงาน ร.อ. ภูรีวรรธน์ โชคเกิด นายแพทย์สาธารณสุขจังหวัดเชียงใหม่ ได้เดินทางเข้าตรวจสอบที่เกิดเหตุ ก่อนระบุว่า รถพยาบาลที่เกิดอุบัติเหตุเป็นรถของโรงพยาบาลอมก๋อย โดยเมื่อคืนที่ผ่านมา เวลาประมาณ 00.30 น. ได้นำผู้ป่วยส่งต่อจากโรงพยาบาลอมก๋อยไปที่โรงพยาบาลนครพิงค์ อ.แม่ริม จ.เชียงใหม่ ถึงโรงพยาบาลนครพิงค์เวลาประมาณ 03.00 น.

Common Pre-processing in NLP Application

- Stemming & Lemmatization
  - Finding root word or lexeme
  - What are the differences between stemming and lemmatization?

Sentence I	**It**	**was**	**very**	**fantastic**	**experience**
Lemma	it	be	very	fantastic	experience
Stemming	it	wa	veri	fantast	experi

Sentence II	**John**	**sent**	**an**	**email**	**to**	**Mary**
Lemma	John	send	a	email	to	Mary
Stemming	John	sent	an	email	to	Mari

SentenceIII	**He**	**didn't**			**get**	**a**	**reply**
Lemma	he	do		not	get	a	reply
Stemming	he	didn	t		get	a	repli

A bit higher-level pre-processing

- Word sense disambiguation
  - Identifying which sense (meaning) of a word is used in a sentence
- Coreference resolution
  - Coreference resolution is the task of finding all expressions that refer to the same entity in a text.
  - Important step for a lot of higher level NLP tasks that involve natural language understanding
- Etc.

Not all NLP applications need every preprocessing step

Tools

- Online tools
  - Stanford corenlp: https://corenlp.run/
  - Stanford online parser: http://nlp.stanford.edu:8080/parser/
- Tools and library
  - NLTK
  - Stanford corenlp