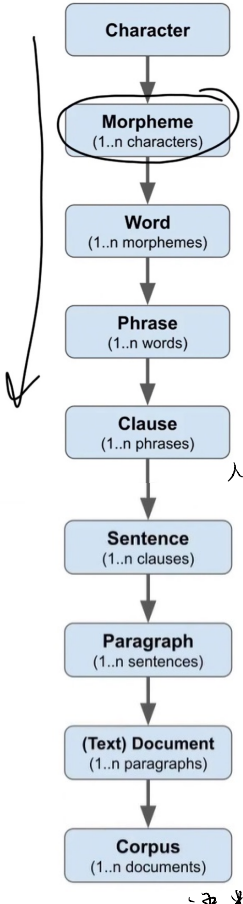
- Chapter 1 Why NLP so hard ?



- Basic symbol of written language (letter, numeral, punctuation mark, etc.)
- Smallest meaning-bearing unit in a language
- Single independent unit of language that can be represented
- Group of words expressing a particular idea or meaning
- Phrase with a subject and verb
- Expresses an independent statement, question, request, exclamation, etc.
- Self-contained unit of discourse in writing dealing with a particular point or idea.
- Written representation of thought
- Collection of writings (i.e., written texts)

r, e, a, c, t, i, o, n

re-act-ion

reaction

his quick reaction

his quick reaction saved him

His quick reaction saved him from the oncoming traffic.

Bob lost control of his car. His quick reaction saved him from the oncoming traffic. Luckily nobody was hurt and the damage to the cae was minimal.

chapter → book

book → library

Morphemes

"Shape"

- Morpheme
  - Smallest meaning-bearing unit in a language ➜ word = 1..n morphemes
- Example: Prefixes & Suffixes
  - Change the semantic meaning or the part of speech of the affected word

un-happy *de-frost-er* *hope-less*

  - Assign a particular grammatical property to that word (e.g., tense, number, possession, comparison)

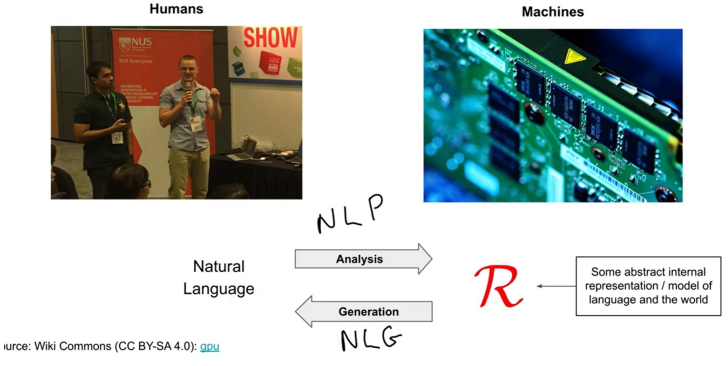
walk-ed *elephant-s* *Bob-'s* *fast-er*

Since 1950: Communication with Machines

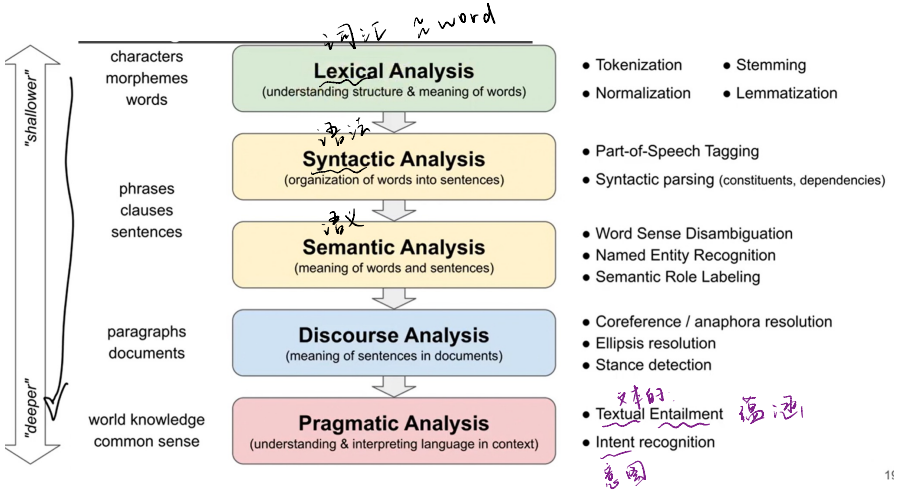
Basic symbolic languages
(e.g., punch cards)

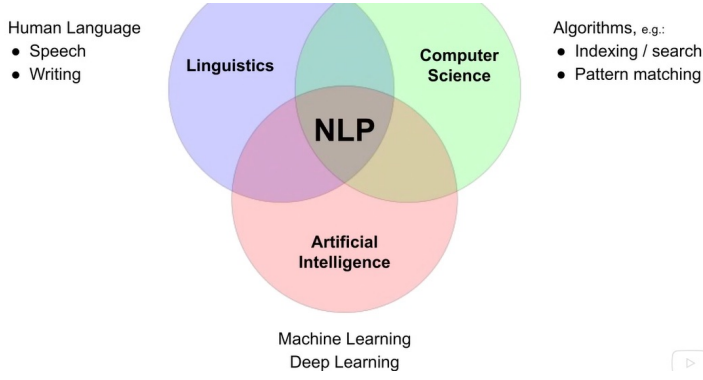
Formal languages
(e.g., programming languages)

Natural language
(e.g., conversational agents / chatbots)



ıurce: Wiki Commons (CB BY-SA 4.0): gpu



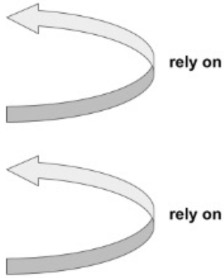


- Lexical Analysis: Find the stem of a word
- Syntactic Analysis: Find the part-of-speech of a certain word in a sentence
- Semantic Analysis: Does this word means A or B in this sentence?
- Discourse Analysis: What is the correct subject that is omitted by this sentence 遗漏
- Pragmatic Analysis: Does this comment posted by the user indicates he/she like or dislike the produce?

Important NLP Tasks

"shallower" ↕ "deeper"

- **Lower-level tasks,** e.g.:
  - Text preprocessing (e.g., tokenization, normalization)
  - Part-of-Speech Tagging
- **Mid-level tasks,** e.g.:
  - Word Sense Disambiguation
  - Named Entity Recognition
- **Higher-level tasks,** e.g.:
  - Coreference / anaphora resolution
  - Intent recognition



1.

Lexical Analysis — Tokenization ✓

- Tokenization
  - Splitting a sentence or text into meaningful / useful units
  - Different levels of granularity applied in practice

character-based ✓	S h e ' s		d r i v i n g		f a s t e r		t h a n	a l l o w e d		.
subword-based ✓ (≈ morpheme.)	She	's	driv	ing	fast	er	than	allow	ed	.
word-based ✓	She's		driving		faster		than	allowed		.

2.

Syntactic Analysis — Part-of-Speech Tagging ✓

词性

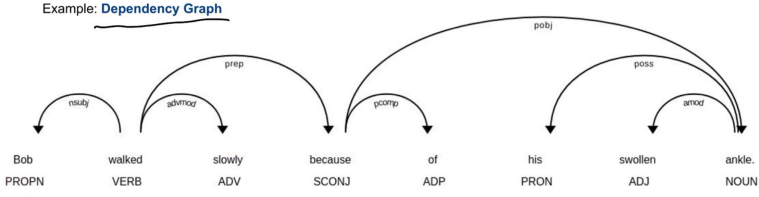
- Part-of-Speech (POS) tagging
  - Labeling each word in a text corresponding to a part of speech
  - Basic POS tags: noun, verb, article, adjective, preposition, pronoun, adverb, conjunction, interjection

Bob	walked	slowly	because	of	his	swollen	ankle	.
NNP	VBD	RB	IN	IN	PRP$	JJ	NN	.
proper noun singular	verb past tense	adverb	preposition or subordinating conjunction	preposition or subordinating conjunction	possessive pronoun	adjective	noun, singular or mass	punctuation

Syntactic Analysis — Syntactic Parsing ✓

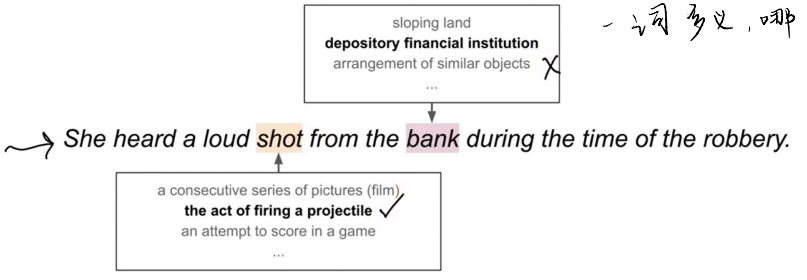
- Dependency parsing
  - Analyze the grammatical structure in a sentence
  - Find related words & the type of the relationship between them

Example: **Dependency Graph**



Semantic Analysis — Word Sense Disambiguation ✓

- Word Sense Disambiguation (WSD)
  - Identification of the right **sense** of a word among all possible senses
  - Semantic ambiguity: many words have multiples meanings (i.e., senses)



Semantic Analysis — Named Entity Recognition

- Named Entity Recognition (NER)
  - Identification of **named entities**: terms that represent real-world objects
  - Examples: persons, locations, organizations, time, money, etc.

Semantic Analysis — Semantic Role Labeling ✓

- Semantic Role Labeling (SRL)
  - Identification of the semantic roles of these words or phrases in sentences
  - Express semantic roles as predicate-argument structures

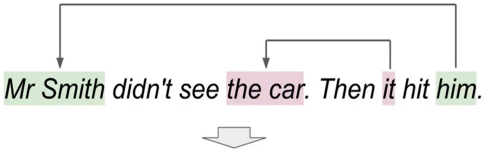
The teacher sent the class the assignment last week.

4.

Discourse Analysis — Coreference Resolution

指代关系.

- Coreference Resolution
  - Identification of expressions that refer to the same entity in a text
  - Entities can be referred to by named entities, noun phrases, pronouns, etc.



Discourse Analysis — Ellipsis Resolution

省略. 消解策略.

- Ellipsis Resolution
  - Inference of ellipses using the surrounding context
  - **Ellipsis**: omission of a word or phrases in sentence

省略.

He studied at NUS, his brother at NTU.

↳ *He studied at NUS, his brother studied at NTU.*

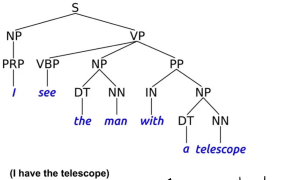
She's very funny. Her sister is not

↳ *She's very funny. Her sister is not very funny.*

Ambiguity: same text → different meaning.

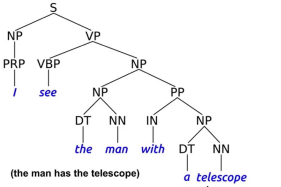
bank { 河岸 / 银行

- Syntactic structure ["I see the man with a telescope"] → affects semantics!



(I have the telescope)

我拿 telescope 看到一个人



(the man has the telescope)

我看到一个人拿 telescope

- Anaphoric ambiguity
  - Ambiguous resolution of anaphora / coreferences (without additional context)

Alice and Sarah went for dinner. She invited her. (??? ???)

Who is "she" and "her" referring to?
Useful context: It was Sarah's birthday.

The box didn't fit in the car because it was too big. (???)

vs.

The box didn't fit in the car because it was too small. (???)

What is "it" referring to?
Resolution requires the understanding of
- Objects can contain other objects
- Physical size of objects
- Physical limitations due to size

- Pragmatic Ambiguity
  - Unclear semantics if context is unknown

"Do you know what time it is?"
- "Yes!" (interpreted as Yes/No question)
- "It's 4.30 pm." (interpreted as question asking for the time) ← in daily life
- (interpreted as a rhetorical question, e.g., the lecturer scolding a student for being late for the lecture)

Expressivity: same meaning different form

- Alice gave book the to Bob
- vs. Alice gave Bob the book

Expressivity

- Idioms
  - *It's raining cats and dogs today.*
  - *He was over the moon to see her.*
- Neologisms (new word)
  - May be added to the dictionary over time
  - *selfie, retweet, photobomb, staycation, binge-watching, crowdfunding, adulting, chillax, noob, kudos,* etc.
- Literary devices, e.g.:
  - Humor
  - Sarcasm
  - Irony
  - Satire
  - Exaggeration

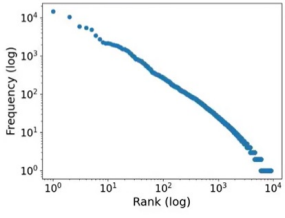
"Oh yeah...studying NLP 24/7 is reeeally my favorite way to spend a weekend!"

Variation

- No one-size-fits-all NLP solutions
  - Difference in underlying task
    (tokenizing, stemming, syntax parsing, part-of-speech tagging, entity recognizing, etc.)
  - ~6.500 languages and ~150 language families
    (different phonetics/phonology, morphology, syntax, grammar)
  - Different domains: news articles, social media, scientific papers, ancient literature, etc.
    (particularly: different vocabularies, formal vs. informal language (e.g., slang), narrative vs. dialogue)
  - Cultural differences and biases
    (example: *"I'm over 40 and live alone."* — perceived sentiment affected by cultural background)

Sparsity

- Sparsity in text corpora
  - Word frequencies inversely proportional to their rank ➜ Zipf's Law
  - Example: *"On the Origin of Species"*
    (Charles Darwin, 1859; 212k+ words)



Rank	Word	Freq.
1	*the*	14,767
2	*of*	10567
3	*and*	5920
4	*in*	5477
5	*to*	4837
6	*a*	3460
7	*that*	2764
8	*as*	2242
9	*have*	2121
10	*be*	2116
...	...	...
101	*mr*	263
102	*parts*	260
103	*often*	260
104	*period*	259
105	*common*	256
...	...	...
1001	*increasing*	25
1002	*expected*	25
1003	*egg*	25
1004	*fly*	25
1005	*aquatic*	25
...	...	

no useful (by meaning)

Unmodeled Representation

negative sentiment.

- The meaning / interpretation of a sentence often depends on:
  - The current context or situation
  - Shared understanding about the world

➜ How to capture this in $\mathcal{R}$?

["I killed all the children."]

Serial killer or **Linux administrator**?

? young persons

= OS processes

["I slipped and fell hard on the floor."]

Arguably a negative sentiment, but **WHY**?

→ pain, injury

→ hospital

→ suffering ⇒ BAD

NLP — Ethical Questions & Challenges

- "Could" vs. "Should" — e.g.:
  - ***Should*** *we build a classifier to identify social media user if they suffer from depression based on their posts?* } privacy
  - ***Should*** *we organize users news feed based on their interests and likings to maximize user engagement?* } echo chambers
  - ***Should*** *we build chatbots that can perfectly mimic humans?*

"If you torture the data long enough, it will confess to anything"
(Ronald Coase; 1981 — paraphrased)

"With great power comes great responsibility!"
(Spider-Man's Uncle Ben)

- Fundamental challenges in NLP
  - Most NLP techniques rely on statistical models
    (never 100% correct and often difficult to quantify errors)
  - Most NLP techniques learn from data
    (Is the data representative? Is the data biased?)

*"Your scientists were so preoccupied with whether they **could**, they didn't stop to think if they **should**."*

网上的 bias 对练 NLP 会导致不好的结果

What is NLP? — The Big Picture

- NLP as machine learning → learn from data
  - Symbolic, probabilistic, and connectionist ML have found their way into NLP
  - Good ML needs bias and assumptions ➜ NLP: linguistic theory & representations
- NLP as linguistics ← word order, style, genre, ...
  - NLP must contend with NL data as found in the world
  - NLP ≈ computational linguistics
  - Linguists now use tools that originated from NLP!

NLP is Changing

- Increases in computing power
  - Deep Learning = matrix operations ➜ Game changer: GPUs
- The rise of the Web, then the social web
  - More "food" for data hungry algorithms
  - User generated content = informal, natural, lively text

网上半鬼蛇神的内容对 NLP 训练影响很大

- Advances in machine learning
  - Continuously growing model zoo (LSTM/GRU, CNN, VAE, Transformers, etc.)
- Advances in understanding of language in social context

Which stage of NLP would sentiment analysis go? That is, deciding whether a sentence or clause has a positive or negative?

- [ ] syntactic analysis
- [x] discourse analysis
- [x] pragmatic analysis

What are some of the key factors that led to the large changes in NLP since 2012?

- [x] the rise of graphics processing units that made massive parallel computation possible
- [ ] climate changes that caused people to favor working from home over commuting
- [ ] the large increase in use of private messaging applications
- [x] the open availablity of large text corpora on the Web
- [x] the algorithmic evolution of efficient methods to learn complex representations from deep learning

✔