NUS | Computing

National University
of Singapore

CS4248 Natural Language Processing

Lecture 1 — What is NLP and Why is it so Hard?

Outline

🏃‍♀️ 🏃 🏃‍♂️ What *is* NLP?

(3 minutes)

What is

Natural Language Processing

anyways? (ahem, that subject that you are taking now)

Go ahead and define it yourselves. Post a reply here:
Canvas > Discussions > [In-Lecture] L1 Fri 9:00–12:00 (Jan 19)

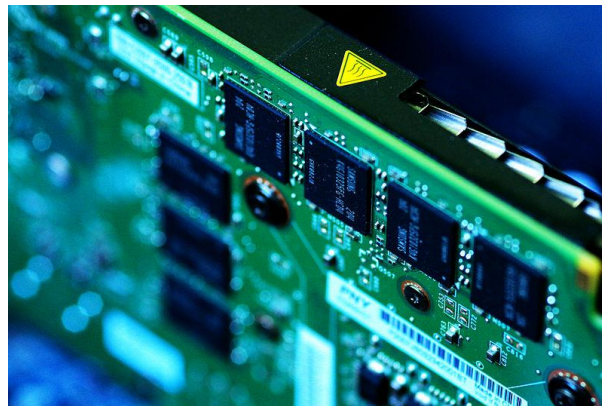And like those replies you think are good too!

Communication with Machines

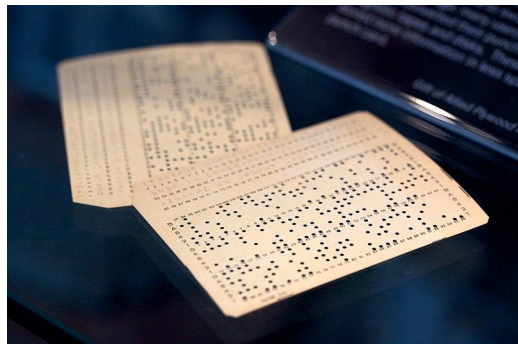**Humans**

Machines

Natural
Language

Analysis →

← Generation

$\mathcal{R}$

← Some abstract internal representation / model of language and the world

Communication with Machines

~50s-70s

~80s

Today

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

Formal languages
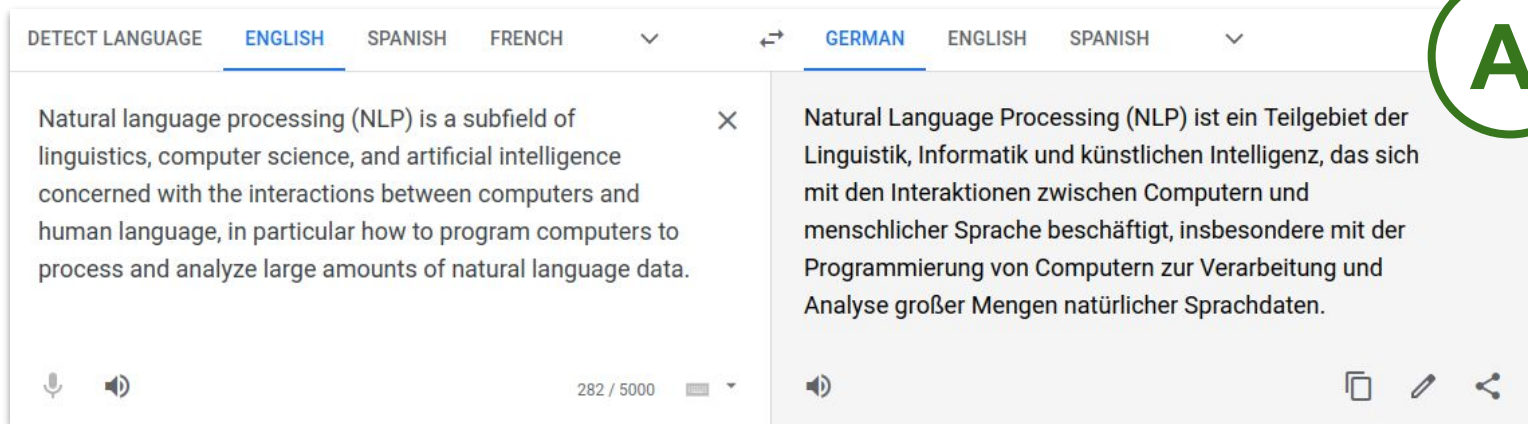(e.g., programming languages)

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

Outline

- **What is NLP?**
  - Basic definition
  - **Prominent applications**
  - Core building blocks
  - Fundamental tasks

- Why is NLP so hard?
  - Characteristics of language
  - When NLP goes wrong

- The Big Picture
  - NLP as a research field
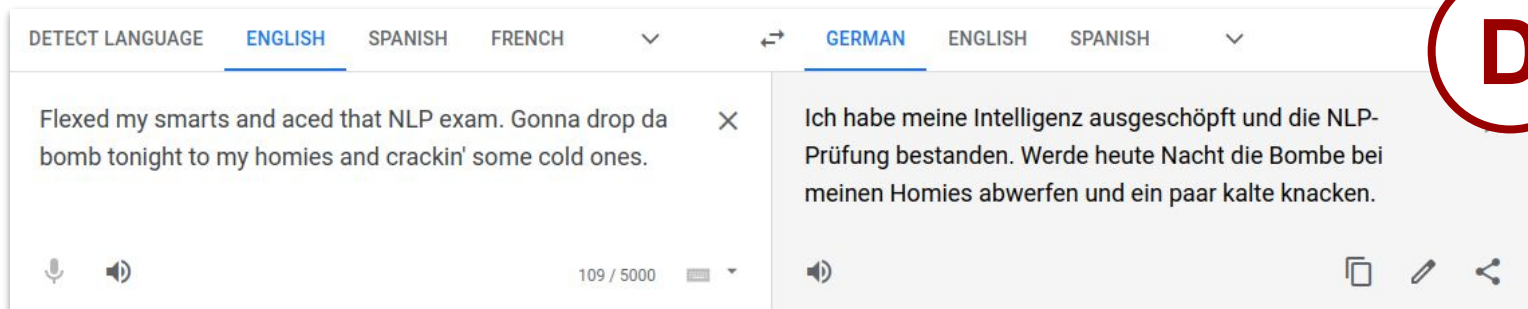 - Topics covered by CS4248

Machine Translation

A-

DETECT LANGUAGE | ENGLISH | SPANISH | FRENCH | ⌄ | ⇄ | GERMAN | ENGLISH | SPANISH | ⌄

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

Natural Language Processing (NLP) ist ein Teilgebiet der Linguistik, Informatik und künstlichen Intelligenz, das sich mit den Interaktionen zwischen Computern und menschlicher Sprache beschäftigt, insbesondere mit der Programmierung von Computern zur Verarbeitung und Analyse großer Mengen natürlicher Sprachdaten.

282 / 5000

D

DETECT LANGUAGE | ENGLISH | SPANISH | FRENCH | ⌄ | ⇄ | GERMAN | ENGLISH | SPANISH | ⌄

Flexed my smarts and aced that NLP exam. Gonna drop da bomb tonight to my homies and crackin' some cold ones.
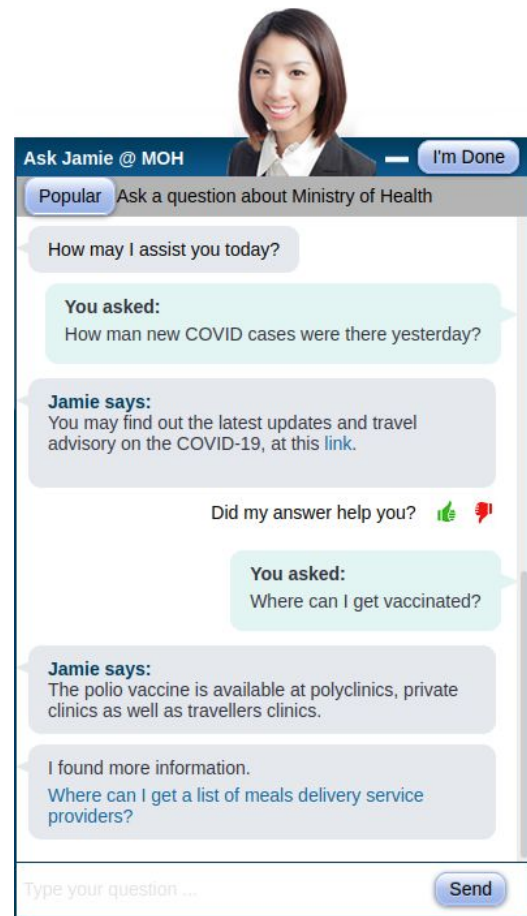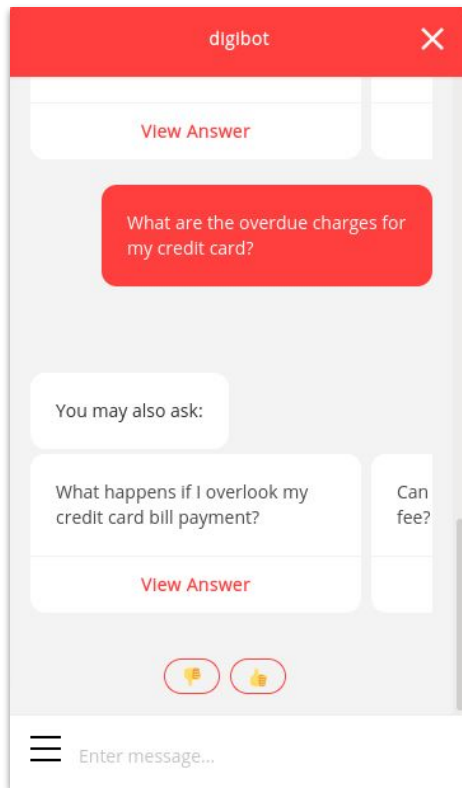
Ich habe meine Intelligenz ausgeschöpft und die NLP-Prüfung bestanden. Werde heute Nacht die Bombe bei meinen Homies abwerfen und ein paar kalte knacken.
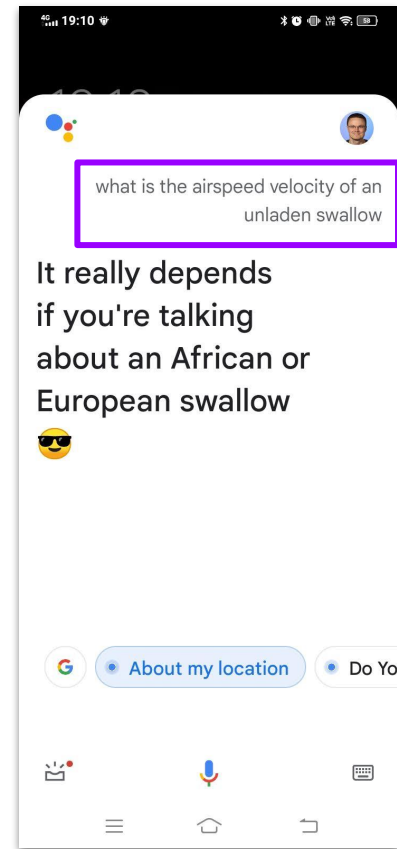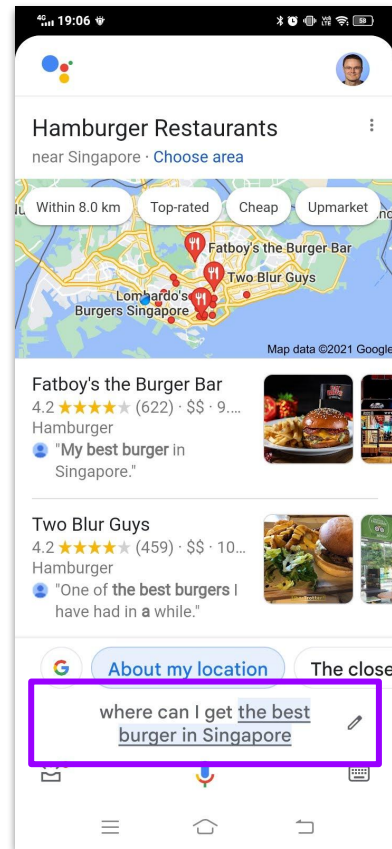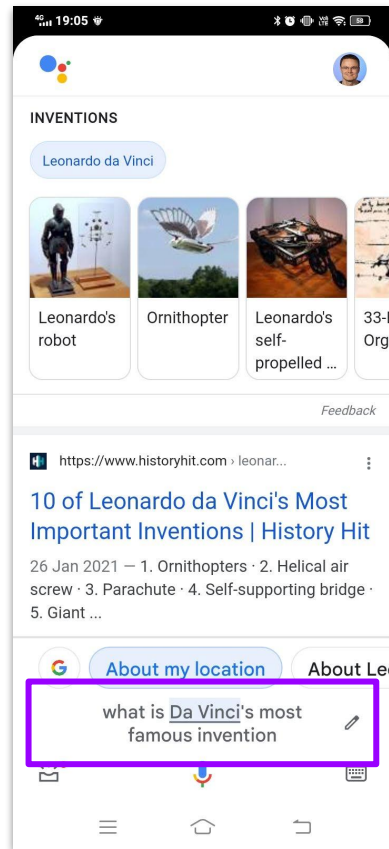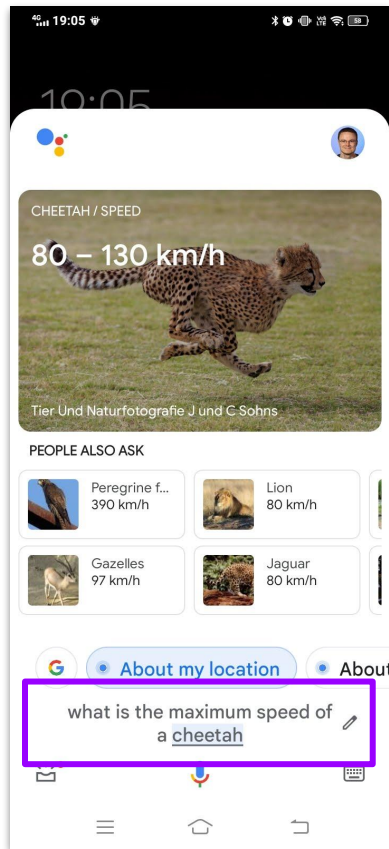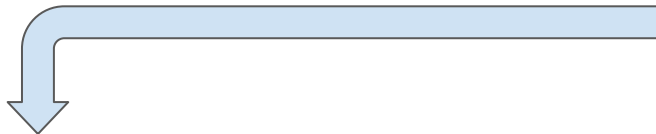
109 / 5000

Conversational Agents

- Conversational agents — core components
  - Speech recognition
  - Language analysis
  - Dialogue processing
  - Information Retrieval
  - Text-to-Speech

Conversational Agents — Question Answering

Text Summarization

Google's cloud unit looked into using artificial intelligence to help a financial firm decide whom to lend money to. It turned down the client's idea after weeks of internal discussions, deeming the project too ethically dicey. Google has also blocked new AI features analysing emotions, fearing cultural insensitivity. Microsoft restricted software mimicking voices and IBM rejected a client request for an advanced facial-recognition system.

THE STRAITS TIMES

Money and mind control: Big Tech slams ethics brakes on AI

PUBLISHED SEP 14, 2021, 5:00 PM SGT

SAN FRANCISCO (REUTERS) - In September last year, Google's cloud unit looked into using artificial intelligence (AI) to help a financial firm decide whom to lend money to.

It turned down the client's idea after weeks of internal discussions, deeming the project too ethically dicey because the AI technology could perpetuate biases like those around race and gender.

Since early last year, Google has also blocked new AI features analysing emotions, fearing cultural insensitivity, while Microsoft restricted software mimicking voices and IBM rejected a client request for an advanced facial-recognition system.

All these technologies were curbed by panels of executives or other leaders, according to interviews with AI ethics chiefs at the three US technology giants.

Reported here for the first time, their vetoes and the deliberations that led to them reflect a nascent industry-wide drive to balance the pursuit of lucrative AI systems with a greater consideration of social responsibility.

"There are opportunities and harms, and our job is to maximise opportunities and minimise harms," said Ms

Text Generation

- Example: Image Captioning

➜ *"A man riding a red bicycle."*

Other Applications

- Spelling correction

- Document clustering

- Document classification, e.g.:
    - Spam detection
    - Sentiment analysis
    - Authorship attribution

Outline

- **What is NLP?**
  - Basic definition
  - Prominent applications
  - **Core building blocks**
  - Fundamental tasks

- Why is NLP so hard?
  - Characteristics of language
  - When NLP goes wrong

- The Big Picture
  - NLP as a research field
  - Topics covered by CS4248

NLP in One Slide

"shallower"

"deeper"

characters
morphemes
words

phrases
clauses
sentences

paragraphs
documents

world knowledge
common sense

Lexical Analysis
(understanding structure & meaning of words)

- Tokenization
- Normalization
- Stemming
- Lemmatization

Syntactic Analysis
(organization of words into sentences)

- Part-of-Speech Tagging
- Syntactic parsing (constituents, dependencies)

Semantic Analysis
(meaning of words and sentences)

- Word Sense Disambiguation
- Named Entity Recognition
- Semantic Role Labeling

Discourse Analysis
(meaning of sentences in documents)

- Coreference / anaphora resolution
- Ellipsis resolution
- Stance detection

Pragmatic Analysis
(understanding & interpreting language in context)

- Textual Entailment
- Intent recognition

14

Core Building Blocks of (Written) Language

Character	• Basic symbol of written language (letter, numeral, punctuation marks, etc.)

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

Morpheme (1..n characters)	• Smallest meaning-bearing unit in a language

re-act-ion

Word (1..n morphemes)	• Single unit of language that can be represented

reaction

Phrase (1..n words)	• Group of words expressing a particular idea or meaning

his quick reaction

Clause (1..n morphemes)	• Phrase with a subject and verb

his quick reaction saved him

Core Building Blocks of (Written) Language

Sentence
(1..n clauses)

- Expresses an independent statement, question, request, exclamation, etc.

Paragraph
(1..n sentences)

- Self-contained unit of discourse in writing dealing with a particular point or idea.

(Text) Document
(1..n paragraphs)

- Written representation of thought

Corpus
(1..n documents)

- Collection of writings (i.e., written texts)
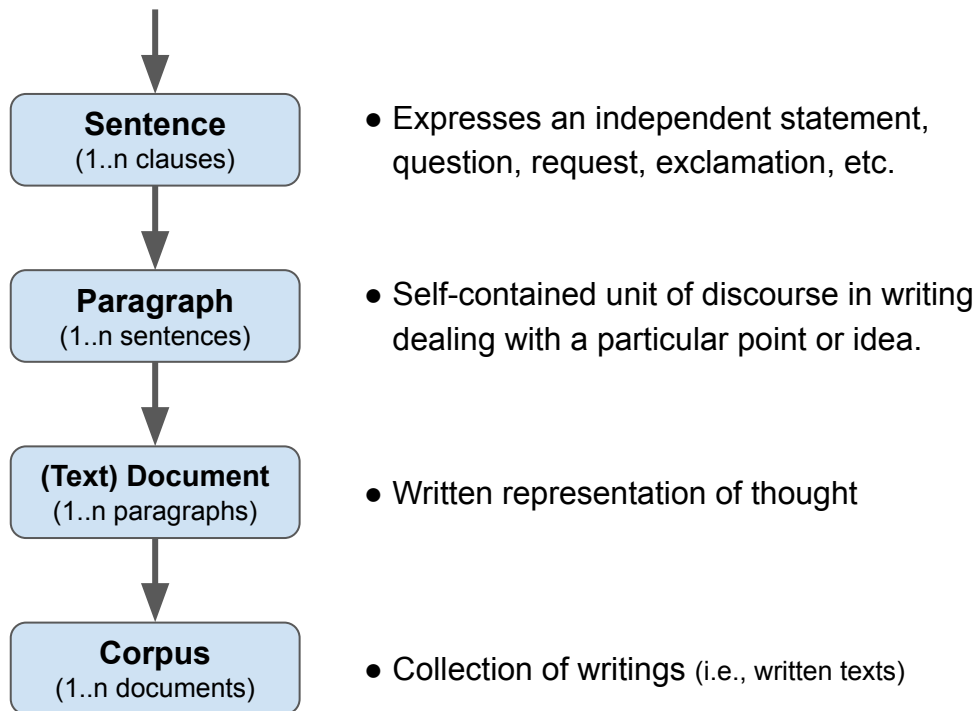
His quick reaction saved him from the oncoming traffic.

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.

Morphology

- Morphology (definition):
  - Study of the forms & formation of words in a language
  - Words are built of **morphemes**

- Morpheme
  - Smallest meaning-bearing unit in a language
  - Word
    = 1..n morphemes
    = 1..n stems + 0..n affixes

  (affix: prefix or suffix)
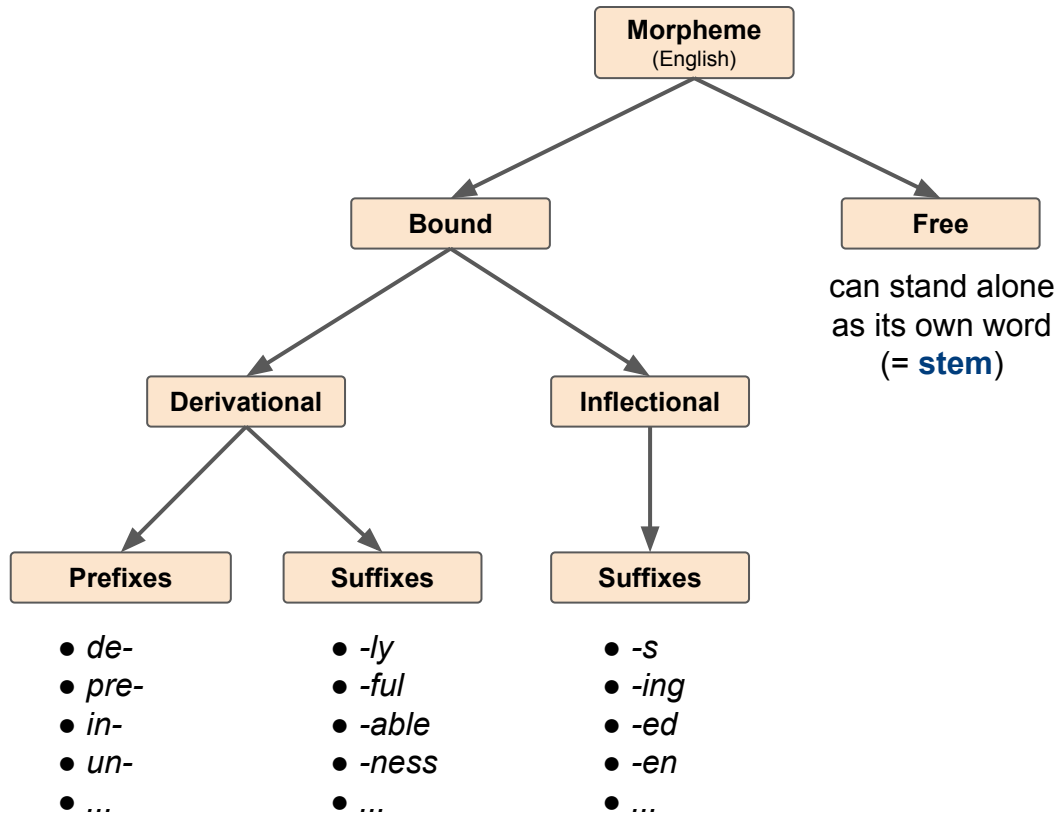
```
                    ┌─────────────────┐
                    │    Morpheme     │
                    │    (English)    │
                    └─────────────────┘
                     ╱               ╲
            ┌──────────┐         ┌──────────┐
            │  Bound   │         │   Free   │
            └──────────┘         └──────────┘
```

Free: can stand alone as its own word (= **stem**)

```
        ┌───────────────┐   ┌───────────────┐
        │ Derivational  │   │ Inflectional  │
        └───────────────┘   └───────────────┘
         ╱           ╲              │
  ┌──────────┐  ┌──────────┐  ┌──────────┐
  │ Prefixes │  │ Suffixes │  │ Suffixes │
  └──────────┘  └──────────┘  └──────────┘
```

Prefixes
- *de-*
- *pre-*
- *in-*
- *un-*
- *...*

Suffixes
- *-ly*
- *-ful*
- *-able*
- *-ness*
- *...*

Suffixes
- *-s*
- *-ing*
- *-ed*
- *-en*
- *...*

🏃‍♀️🏃🏃‍♂️ Does English have Complex Morphology?

(2 minutes)

Canvas > Discussion > [In-Lecture] L1 Fri 9:00–12:00 (Jan 19)
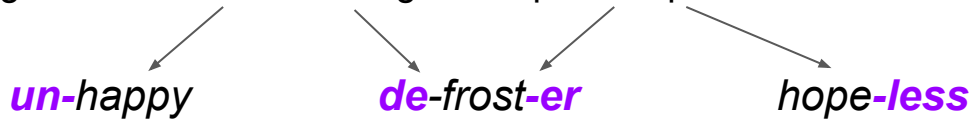
First: 👆**Yes**✅ or **No**❌?

Then: Name a language with **more** morphology?

And one with **less**?

Bound Morphemes

- Derivational morphemes (prefix or suffix)
  - Change the semantic meaning or the part of speech of the affected word

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

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

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

Examples

	Prefix	Prefix	Stem	Suffix	Suffix	Suffix
dogs			***dog***	*-s*		
walked			***walk***	*-ed*		
imperfection		*im-*	***perfect***	*-ion*		
hopelessness			***hope***	*-less*	*-ness*	
undesirability		*un-*	***desire***	*-able*	*-ity*	
unpremeditated	*un-*	*pre-*	***mediate***	*-ed*		
antidisestablishmentarianism	*anti-*	*dis-*	***establish***	*-ment*	*-arian*	*-ism*

Examples with multiple stems: *daydream-ing*, *paycheck-s*, *skydive-er*

Morphology — Challenges

- Combining morphemes — effects on syntax
  - Words often not simply concatenations of morphemes

  read-*able*-*ity* ➜ *readability*

- Imprecise meanings   *flammable* vs. *inflammable* *vs.* *non-flammable*

- Complex morphology
  - Many languages have a more complex morphology (compared to English)

  Example (Turkish):   *Avrupalılaştıramadıklarımızdan mısınız?*

  "Are you one of those whom we could not Europeanize?"

Outline

- **What is NLP?**
  - Basic definition
  - Prominent applications
  - Core building blocks
  - **Fundamental tasks**

- Why is NLP so hard?
  - Characteristics of language
  - When NLP goes wrong

- The Big Picture
  - NLP as a research field
  - Topics covered by CS4248

AI Tools: The Next Spreadsheet

Calculator analogy, upgraded

How do you know whether a spreadsheet's output is correct?

The basis for modern finance.

More innovation outside academia.

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 | She | 's | driv | ing | fast | er | than | allow | ed | . |

| word-based | She's | driving | faster | than | allowed | . |

Syntactic Analysis — Part-of-Speech Tagging

- 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

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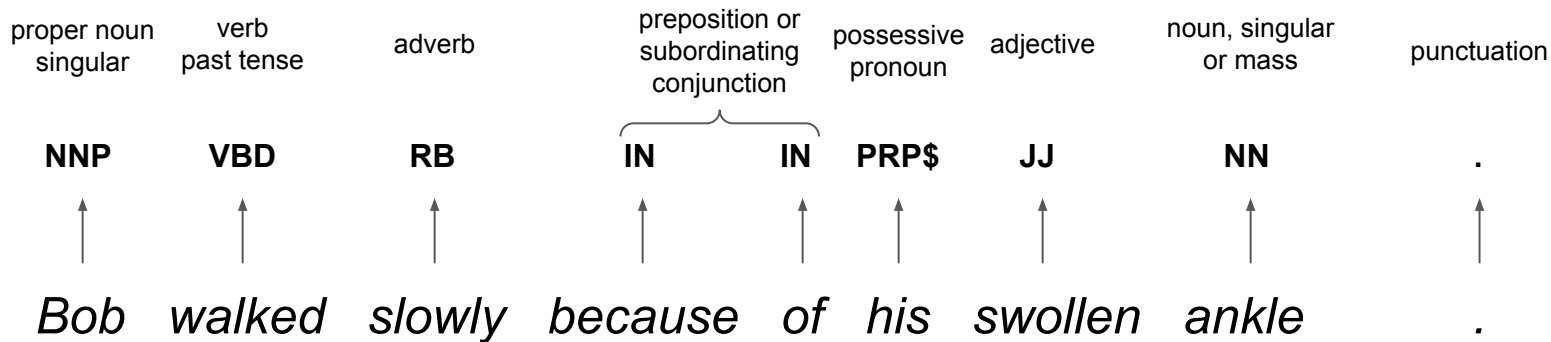
Syntactic Analysis — Syntactic Parsing

- ## Dependency parsing
    - ■ Analyze the grammatical structure in a sentence

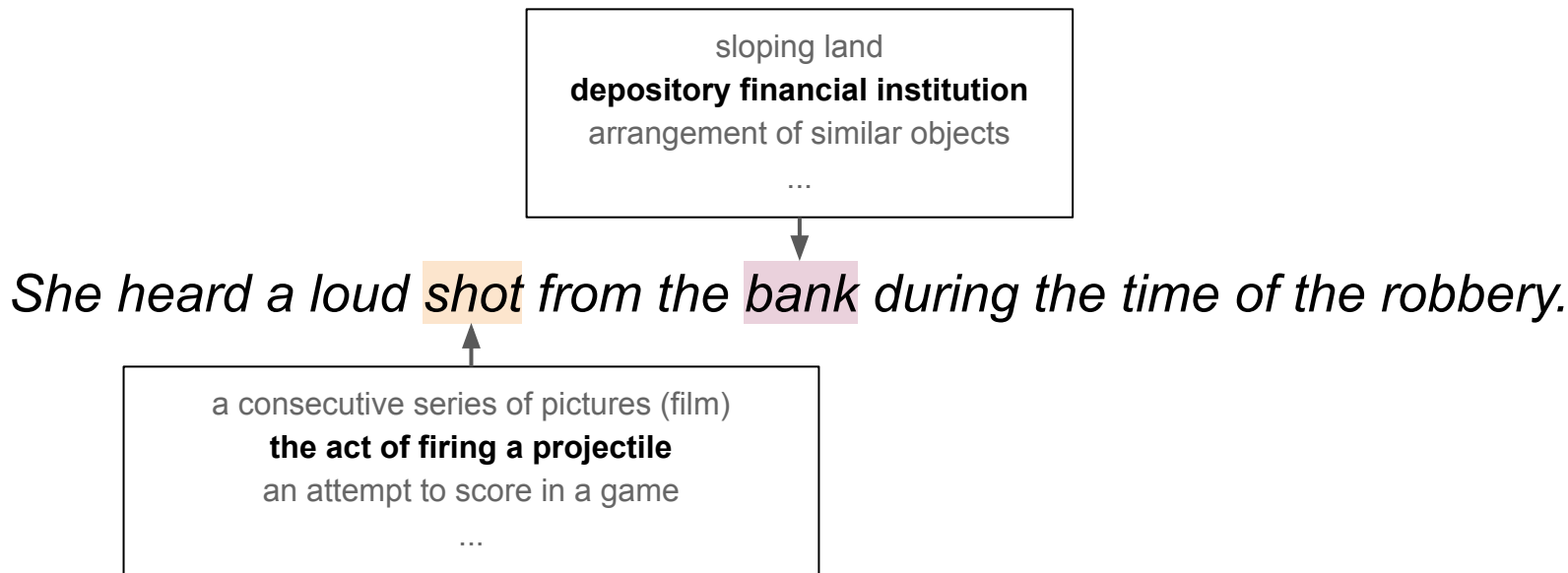   - ■ Find related words & the type of the relationship between them

Example: **Dependency Graph**
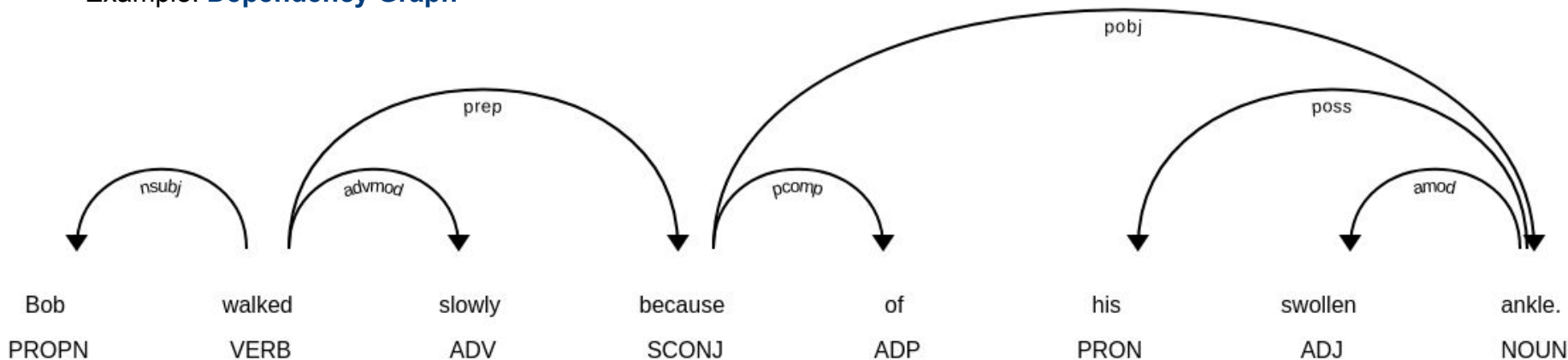
Semantic Analysis — Word Sense Disambiguation

- 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)

> sloping land
> **depository financial institution**
> arrangement of similar objects
>
> ...

She heard a loud shot from the bank during the time of the robbery.

> a consecutive series of pictures (film)
> **the act of firing a projectile**
> an attempt to score in a game
>
> ...

🏃‍♀️ 🏃 🏃‍♂️ **Weird Cents This Am Big Ovation**

(3 minutes)

Canvas > Discussion > [In-Lecture] L1 Fri 9:00–12:00 (Jan 19)

Give examples of two words that are ambiguous

The more 🆒 , the more 🔥 .

Bonus: Why are they ambiguous?

Why did you pick these words?

28

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.

PERSON	ORGANIZATION	LOCATION	MONEY

Chris booked a Singapore Airlines flight to Germany for S$1,200.

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

Who   did **What** to **Whom**   **What** exactly   at **When**

The teacher sent the class the assignment last week.

Discourse Analysis — Coreference Resolution

- 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.

Mr Smith didn't see the car. Then it hit him.

⬇

Mr Smith didn't see the car. Then the car hit Mr Smith.

Discourse Analysis — Ellipsis Resolution

- ## Ellipsis Resolution
  - Inference of ellipses using the surrounding context

  - **Ellipsis**: omission of a word or phrases in sentence

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.*

Pragmatic Analysis — Textual Entailment

- ## Textual Entailment
  - Determining the inference relation between two short, ordered texts

  - Given a text *t* and hypothesis *h*, "t entails h" (t ⇒ h)
    - ➜ someone reading *t* would infer that *h* is most likely true

t: *A mixed choir is performing at the National Day parade.*

h: *The anthem is sung by a group of men and women.*

t ⇒ h

Required world knowledge:
- Mixed choir: male and female members
- Singing a song is a performance
- "anthem" typically refers to "national anthem"

Pragmatic Analysis — Intent Recognition

- Intent Recognition
    - Classification of an utterance based on what the speaker/writer is trying to achieve
    - Core component of sophisticated chat bots

"I'm hungry!"

Additional context:
- The writer is vegetarian
- The writer is near VivoCity
- It's 1pm: lunch time
- ...

➜ **Intent:**
Writer is looking for a place to eat

➜ **Action:**
Search for vegetarian restaurants in and around VivoCity that are open.

Outline

- **What is NLP?**
  - Basic definition
  - Prominent applications
  - Core building blocks
  - Fundamental tasks

- **Why is NLP so hard?**
  - **Characteristics of language**
  - When NLP goes wrong

- **The Big Picture**
  - NLP as a research field
 - Topics covered by CS4248

What Makes NLP so Hard?

- Main challenges
  - Ambiguity

  - Expressivity

  - Variation

  - Scale

  - Sparsity

Ambiguity

- Ambiguity at different levels, e.g.:
    - Word senses: *bank* (financial institute or edge of river?), *cancer* (disease or zodiac sign?)

    - Part of Speech: *run* (verb or noun?), *fast* (verb or noun or adjective or adverb?)

    - Syntactic structure: *"I see the man with a telescope"* ➜ affects semantic!

(I have the telescope)   **(the man has the telescope)**
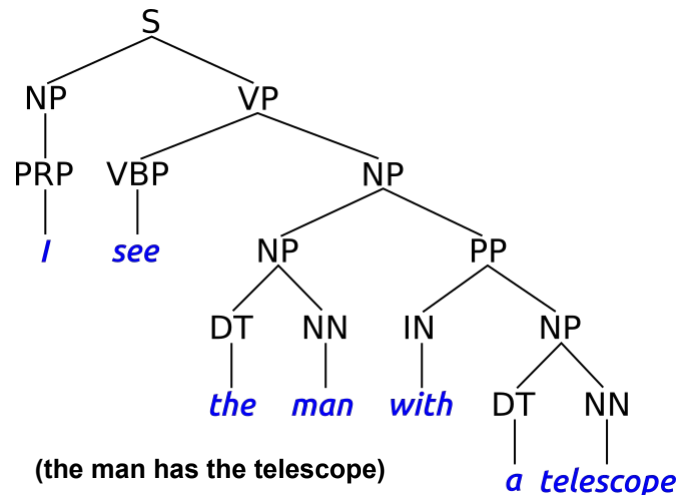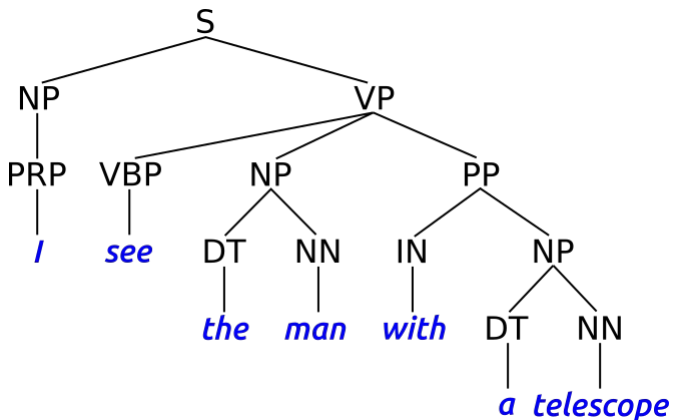
Ambiguity

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

Alice and Sarah went for dinner. She invited her.
??? (above "She") **???** (above "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.
??? (above "it")

vs.

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

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

Ambiguity

- Winograd Schema (Challenge)
  - A pair of sentences differing in only one or two words and containing an ambiguity that is resolved in opposite ways

  - Resolution requires the use of world knowledge & reasoning

- Example (see also previous slide)

???

I poured water from the bottle into the cup until it was full.

vs.

???

I poured water from the bottle into the cup until it was empty.

🏃‍♀️ 🏃 🏃‍♂️ DYI Winograd Schema

(5 minutes)

Find and post your own examples for Winograd Schema(s)

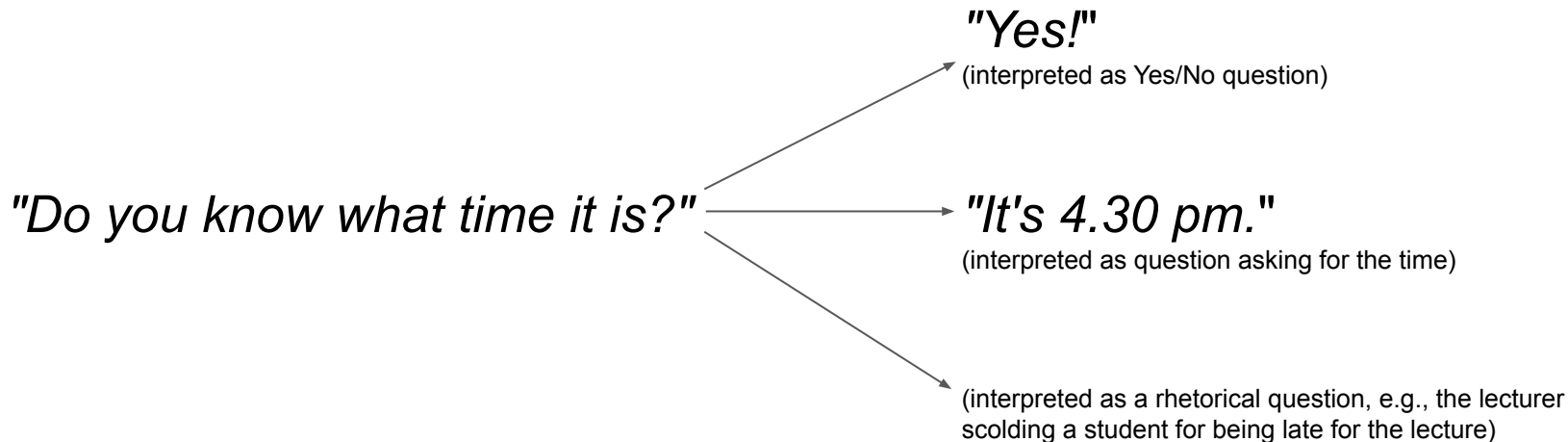Explain what is ambiguous about it.

Go wild! Find good interesting ones.

Give kudos and upvote your peers!

Ambiguity

- ## 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)

(interpreted as a rhetorical question, e.g., the lecturer scolding a student for being late for the lecture)

Expressivity

- In general, the same meaning can be expressed with very different forms

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

This burger is very delicious.   **vs.**   *This burger is a banger!*

Please stop talking and pay close attention to what I want to tell you!   **vs.**   *Shut up and listen to me!*

Expressivity

- Idioms

  It's raining cats and dogs today.

  He was over the moon to see her.

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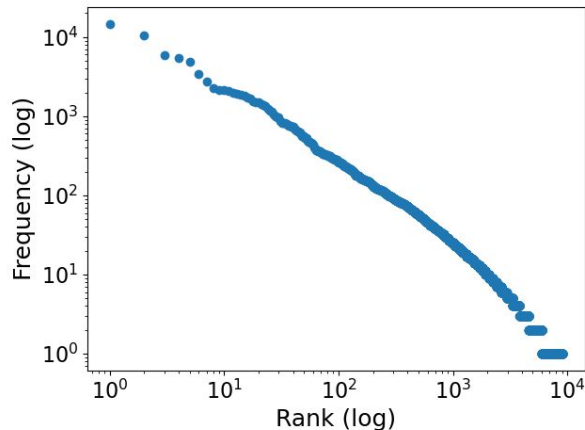
| Rank	Word	Freq.
1	*the*	14,767
2	*of*	10567
3	*and*	5920
4	*in*	5477
5	*to*	4837
6	*a*	3460
7	*that*	2764
8	*as*	2242
9	*have*	2121
10	*be*	2116
...	...	...
101	*mr*	263
102	*parts*	260
103	*often*	260
104	*period*	259
105	*common*	256
...	...	...
1001	*increasing*	25
1002	*expected*	25
1003	*egg*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

Sparsity

The Hound of the Baskervilles (63k+ words)
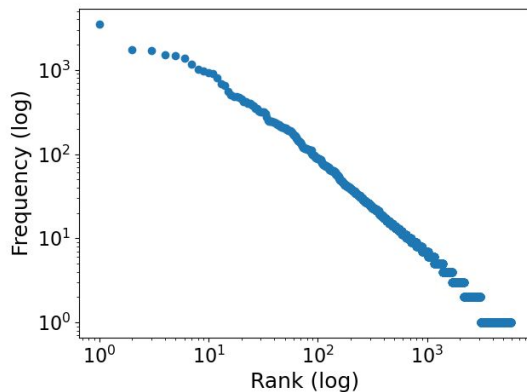
On the Origin of Species (212k+ words)
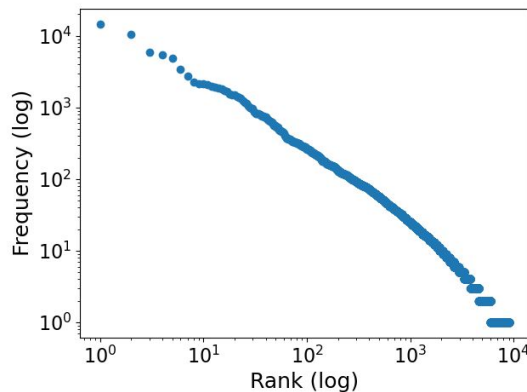
100MB Wikipedia dump (14.4M+ words)
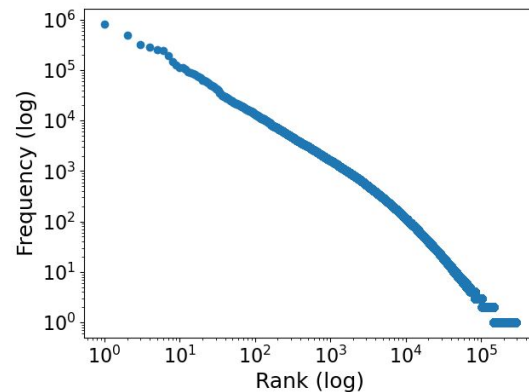
➜ Regardless of size and domain of corpus, there will be a lot of infrequent words!

Scale

- ~6.500 languages and ~150 language families

- Number of words (e.g., in English)
    - Dictionary: ~470,000

    - Web corpus: > 1,000,000

Unmodeled Representation

- 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**?
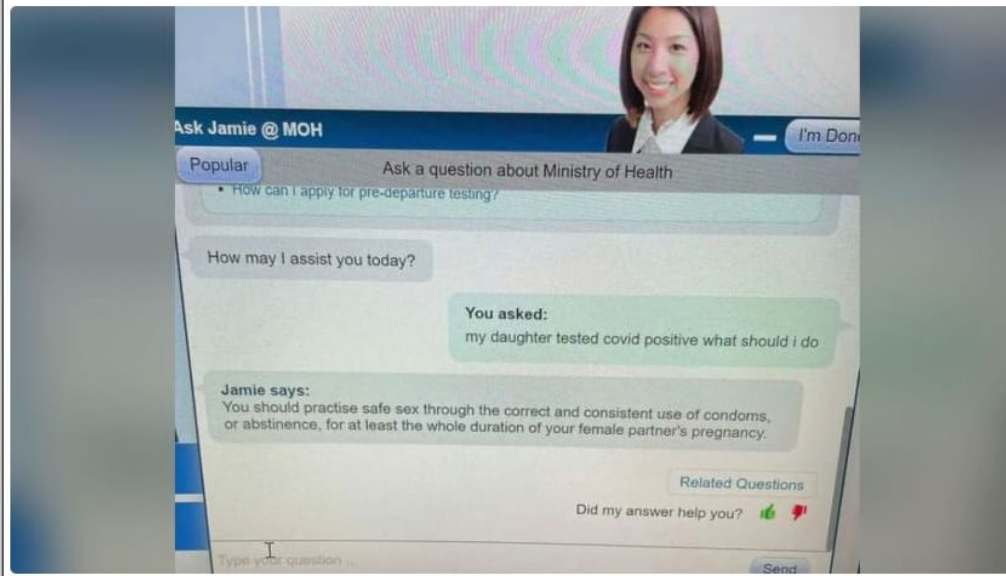
"I slipped and fell hard on the floor."

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

Outline

- ## What is NLP?
  - Basic definition
  - Prominent applications
  - Core building blocks
  - Fundamental tasks

- ## **Why is NLP so hard?**
  - Characteristics of language
  - **When NLP goes wrong**

- ## The Big Picture
  - NLP as a research field
  - Topics covered by CS4248

NLP in the Press — For the Wrong Reasons

NLP in the Press — For the Wrong Reasons

College Kid's Fake, AI-Generated Blog Fooled Tens of Thousands

Microsoft terminates its Tay AI chatbot after she turns into a Nazi

OpenAI Shuts Down GPT-3 Bot Used To Emulate Dead Fiancée

Not spam: estimated cost of 'false positive' junk mail amounts to more than €19.4 billion in Europe alone

Artificial intelligence has a problem with grammar

Why chatbots still suck in 2021

AI tools that companies use to scan resumes are stopping 27 million people finding new jobs, a Harvard report says

Facebook's data on you runs deeper than your therapist's notes

Our computers are sexist towards male and female politicians

AI Wrote Better Phishing Emails Than Humans in a Recent Test

Outline

- **What is NLP?**
  - Basic definition
  - Prominent applications
  - Core building blocks
  - Fundamental tasks

- **Why is NLP so hard?**
  - Characteristics of language
  - When NLP goes wrong

- **The Big Picture**
  - NLP as a research field
  - Topics covered by CS4248

What is NLP? — The Bigger Picture

Human Language
- Speech
- Writing

Linguistics

Computer Science
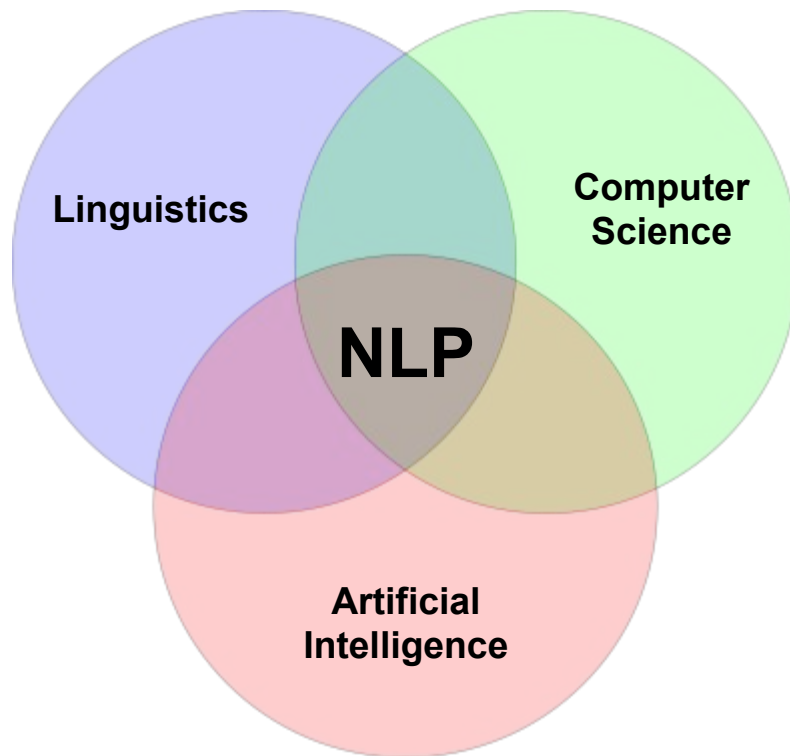
Algorithms, e.g.:
- Indexing / search
- Pattern matching

NLP

Artificial Intelligence

Machine Learning
Deep Learning

🏃‍♀️ 🏃 🏃‍♂️ What is Computational Linguistics

(5 minutes)

How does it differ from NLP?

Is it the same?

What is NLP? — The Bigger Picture

- ## NLP as machine learning
  - Symbolic, probabilistic, and connectionist ML have found their way into NLP
  - Good ML needs bias and assumptions ➜ NLP: linguistic theory & representations

- ## NLP as linguistics
  - NLP must contend with NL data as found in the world
  - NLP ≈ computational linguistics
  - Linguistics now use tools originating in NLP!

What is NLP? — The Bigger Picture

- **Fields with Connections to NLP**
  - Cognitive Science
  - Information Theory
  - Data Science
  - Political Science
  - Psychology
  - Economics
  - Education
  - Ethics

"Language shapes the way we think, and determines what we can think about."

Benjamin Lee Whorf

"Knowledge of languages is the doorway to wisdom."

Roger Bacon

"Language is the road map of a culture. It tells you where its people come from and where they are going."

Rita Mae Brown

"We should learn languages because language is the only thing worth knowing even poorly."

Kató Lomb

Desiderata of NLP Models

- What makes good NLP?
  - Sensitivity to a wide range of phenomena and constraints in language

  - Generality across languages, modalities, genres, styles

  - Strong formal guarantees (e.g., convergence, statistical efficiency, consistency)

  - High accuracy when judged against expert annotations or test data

  - Computational efficiency during training and testing (construction and production)

  - Explainable to human users ➜ transparency

  - Ethical considerations

In practice, often conflicting goals
(e.g., accuracy vs explainability)

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

- Advances in machine learning
  - Continuously growing model zoo (LSTM/GRU, CNN, VAE, Transformers, LLMs, etc.)

- Advances in understanding of language in social context

Course Meta Topics

- Linguistic Issues
  - What are the range of language phenomena?
  - What are the knowledge sources that let us disambiguate?
  - What representations are appropriate?
  - How do you know what to model and what not to model?

- Statistical Modeling Methods
  - Increasingly complex model structures
  - Learning and parameter estimation
  - Efficient inference: dynamic programming, search
  - Deep neural networks for NLP: RNNs, CNNs, Transformers

Summary

- ## Questions covered
  - What is NLP?
  - Why do we care about NLP?
  - Why is it challenging (for machines)?

- ## This week's main takeaways
  - NLP is everywhere
  - Language is complex, ambiguous, subjective, ever-changing, multifaceted
  - Human communication = language + shared context/understanding (e.g., world knowledge)

- ## Outlook for next lecture
  - Capturing strings and words
  - Text preprocessing / cleaning
  - Error/typo handling

  ➔ Getting your text ready for analysis
    (otherwise: "garbage in, garbage out")

Outlook for Next Week

Byte Pair Encoding Data Compression Example

aaabdaaabac

aaabdaaabac   Replace Z = aa

ZabdZabac   Replace Y = ab

ZYdZYac   Replace X = ZY

XdXac   Final compressed string

Replacement Table

Byte pair	Replacement
X	ZY
ab	Y
aa	Z

Pre-Lecture Activity for Next Week

- ### Assigned Task (due before class; do this individually in your own words — no 🤖 & not as a team)
  - ■ Post a 1–2 sentence answer to the following question into the common Discussion Forum
    (it's at Canvas > Discussion > [Pre-Lecture] Week 02)

"What is the relationship between a Finite State Machine
and Regular Expressions?"

Side notes:
- This task is meant as a warm-up to provide some context for the next lecture
- No worries if you get lost; we will talk about this in the next lecture
- You can just copy-&-paste others' answers but this won't help you learn better