NUS | Computing
National University of Singapore

Natural Language Processing: Foundations

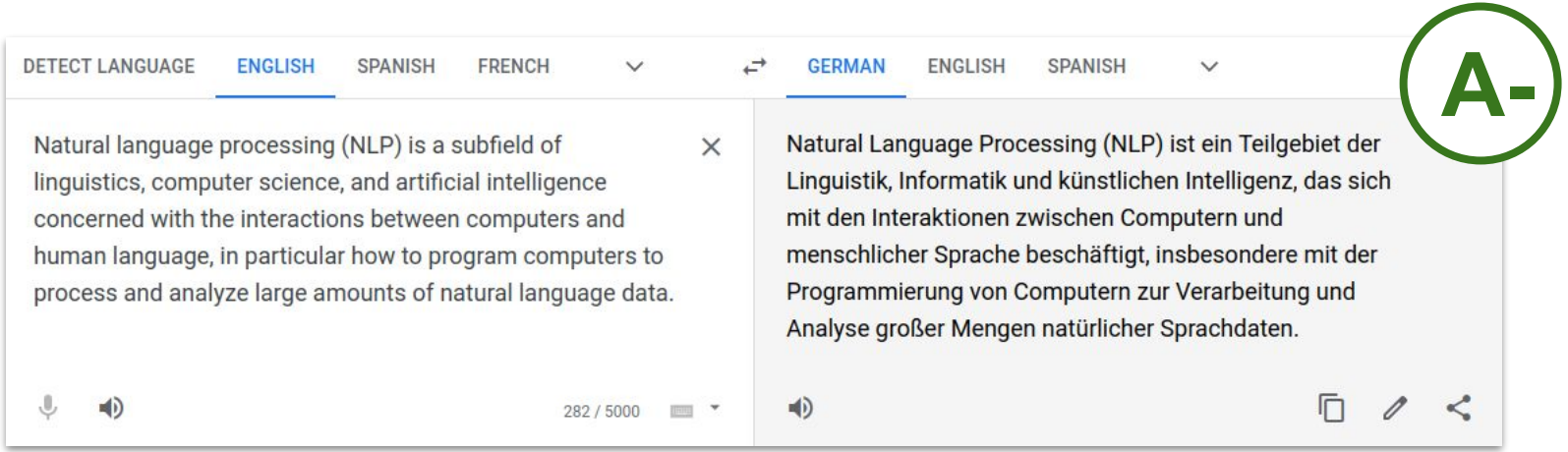
Section 1 — NLP in Everyday Life

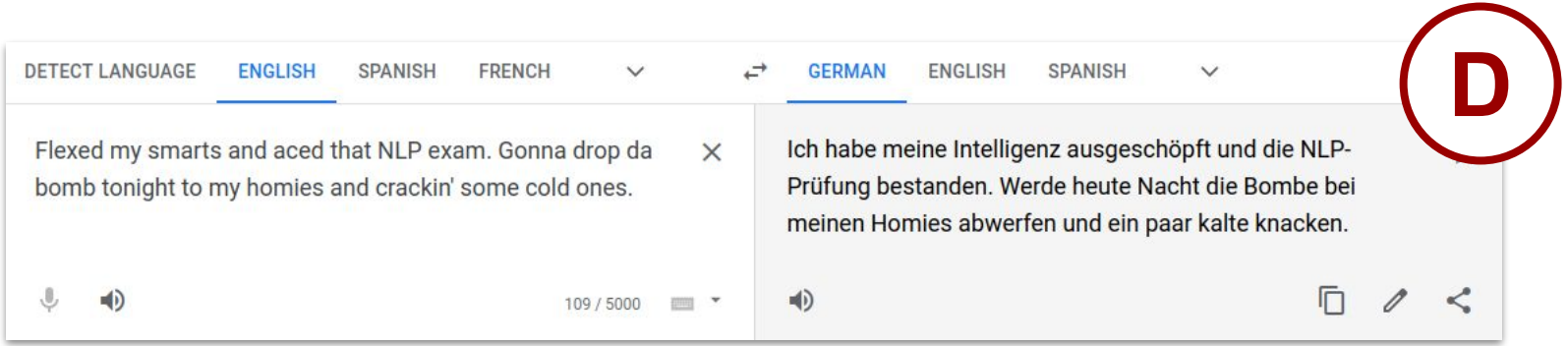
NLP Applications

- NLP application in almost daily use
  - Machine translation
  - Conversational agents (e.g., chat bots)
  - Text summarization
  - Text generation (e.g., autocomplete)

- Applications powered by NLP
  - Social media
  - Search engines
  - Writing assistants (e.g., grammar checking)

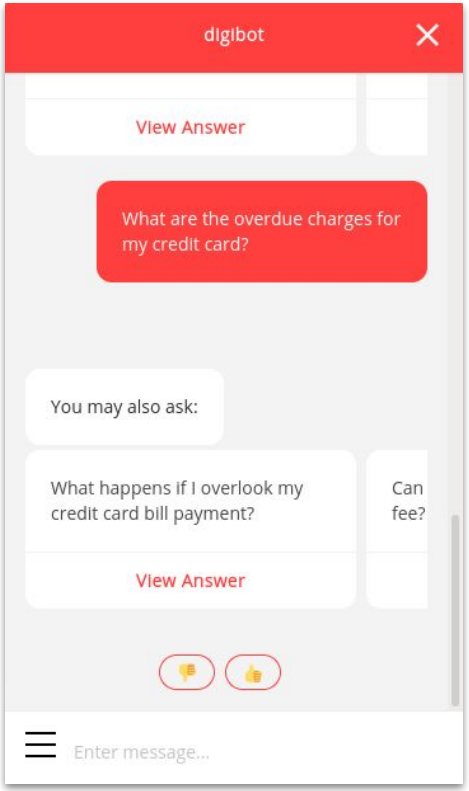
Machine Translation

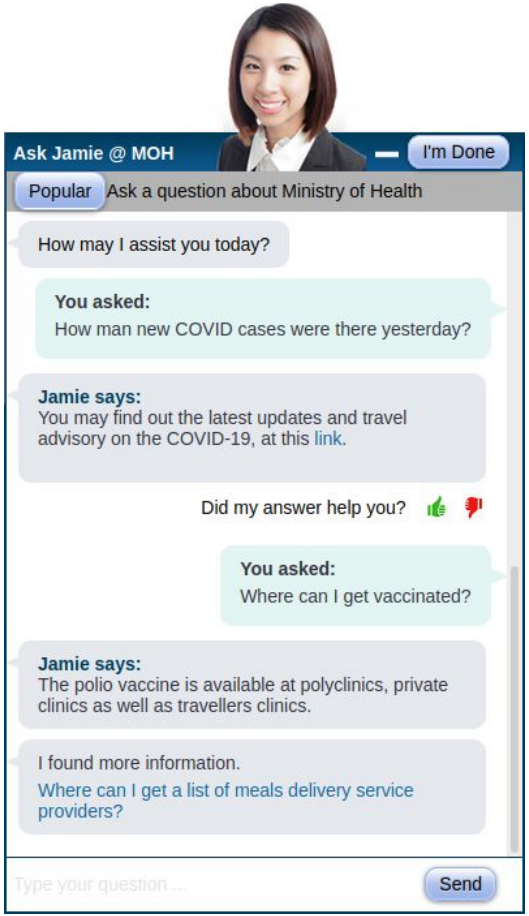




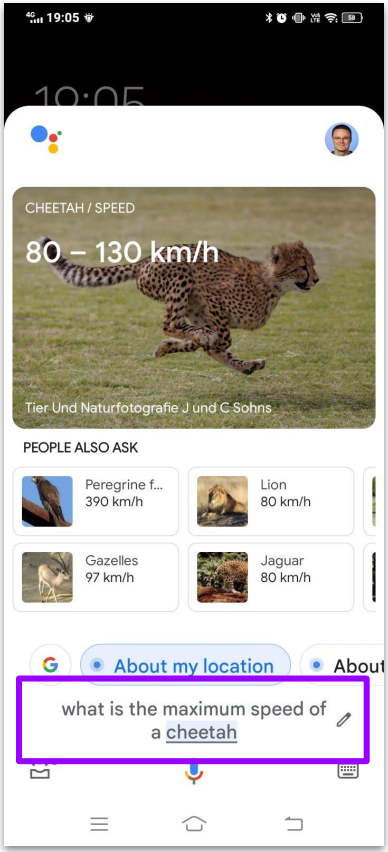
Conversational Agents

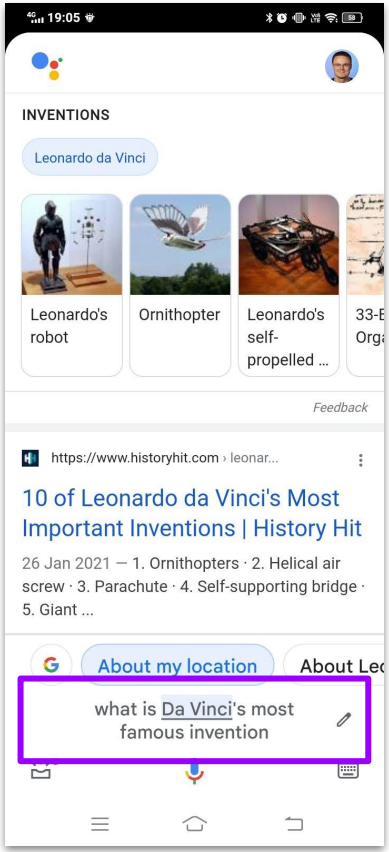
- Conversational agents — core components
  - Speech recognition
  - Language analysis
  - Dialogue processing
  - Information retrieval
  - Text-to-Speech

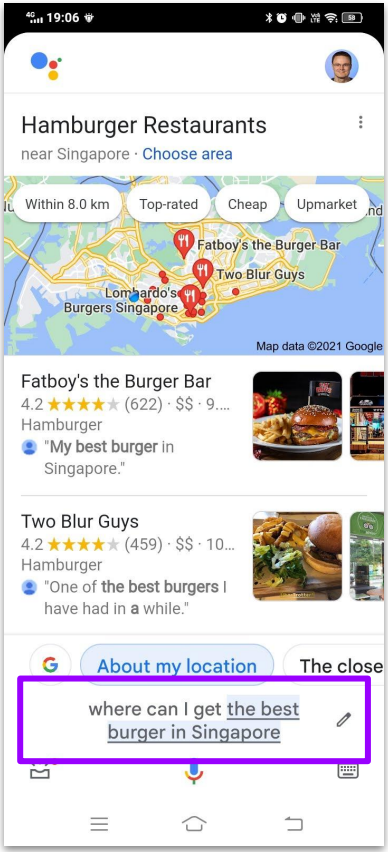


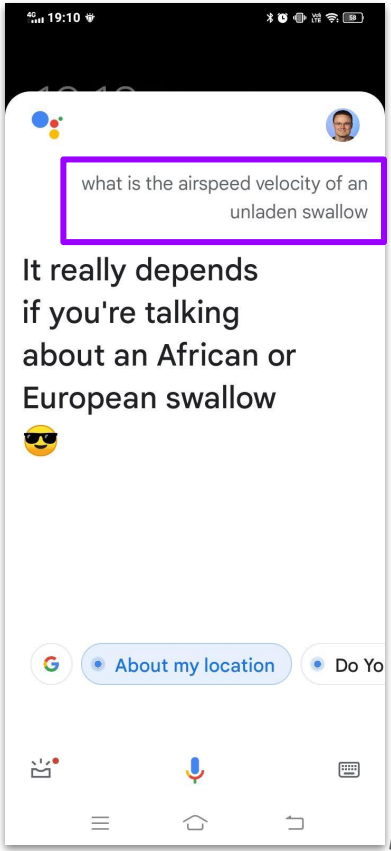


Conversational Agents — Question Answering

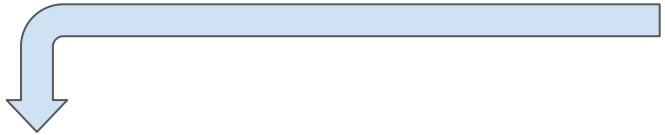




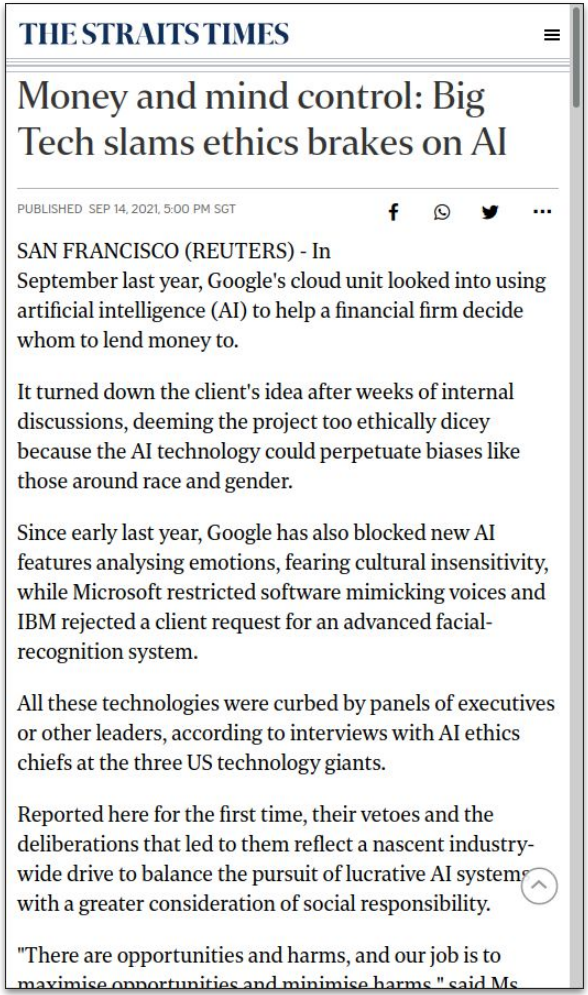




Text Summarization



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.

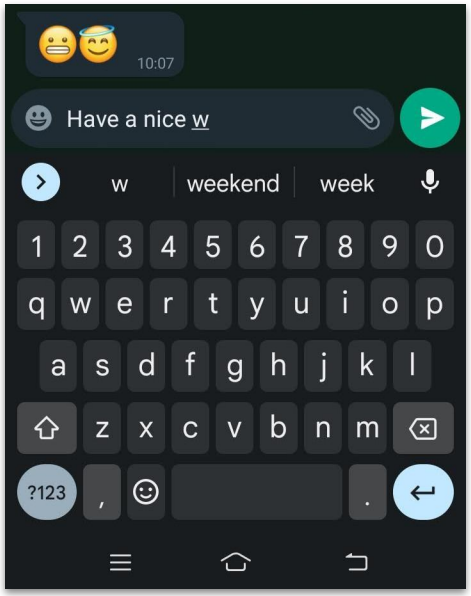
All these technologies were curbed by panels of executives or other leaders, according to interviews with AI ethics chiefs at the three US technology giants.

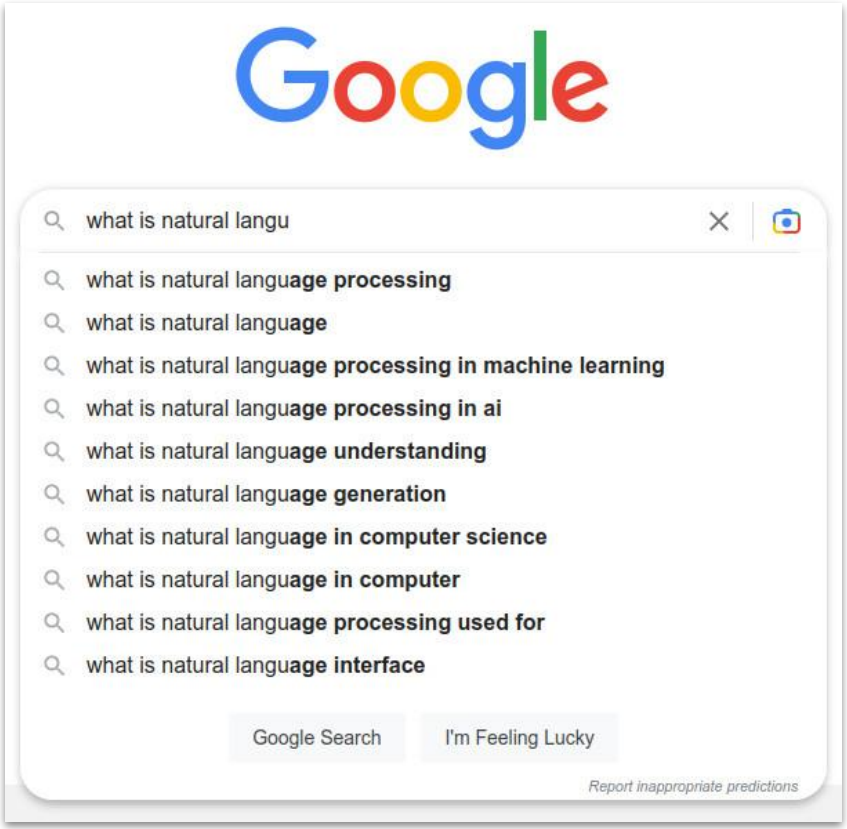
Reported here for the first time, their vetoes and the deliberations that led to them reflect a nascent industry-wide drive to balance the pursuit of lucrative AI systems with a greater consideration of social responsibility.

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

Text Generation

- Example: Autocomplete
  - Given the first words of a sentence, predict the next most likely word





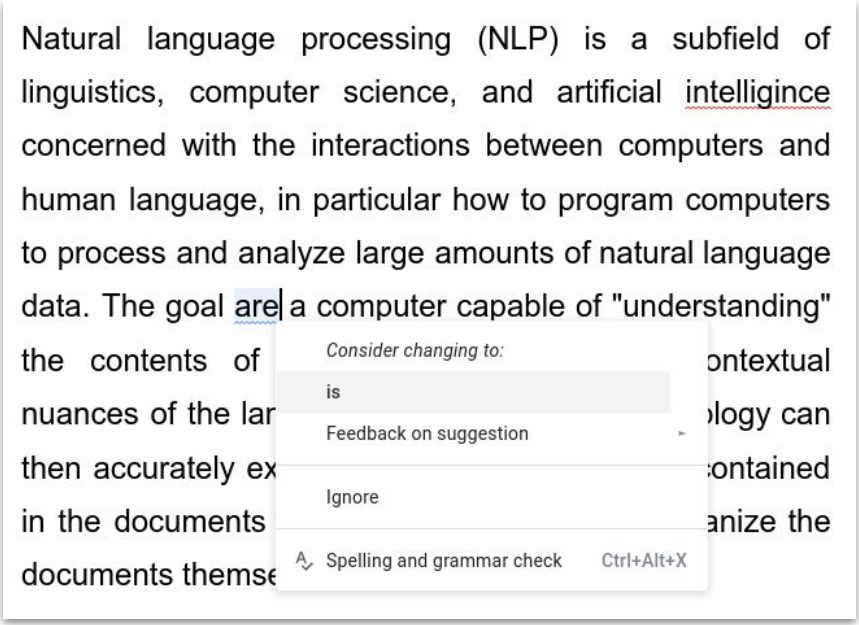
Text Generation

- 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



NUS | Computing
National University of Singapore

Natural Language Processing: Foundations

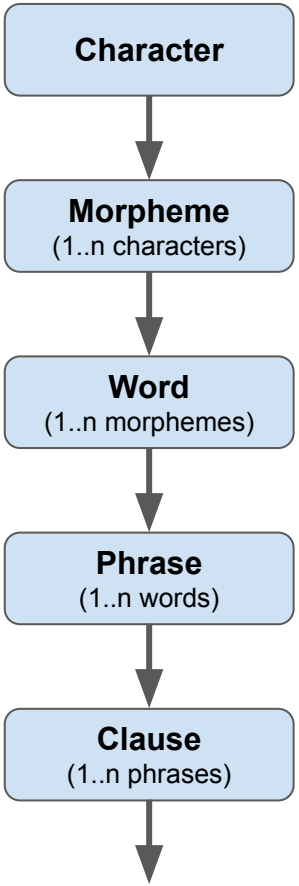
Section 1 — What is Language? What is NLP?

What is Natural Language?

- Natural Language
  - Means of communicating thoughts, feelings, opinions, ideas, etc.
  - Not formal, yet systematic: rules can emerge which were not previously defined
    (this includes that many rules can be bent until reaching a breaking point)
  - Characteristics: ambiguous, redundant, changing, unbounded, imprecise, etc.

- Text / Writing
  - Visual representation of verbal communication (i.e., Natural Language)
  - Writing system: agreed meaning behind the sets of characters that make up a text
    (most importantly: letters, digits, punctuation, white space characters: spaces, tabs, new lines, etc.)

Core Building Blocks of (Written) Language



- Basic symbol of written language
  (letter, numeral, punctuation mark, etc.)

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

- Smallest meaning-bearing unit in a language

re-act-ion

- Single independent unit of language that can be represented

reaction

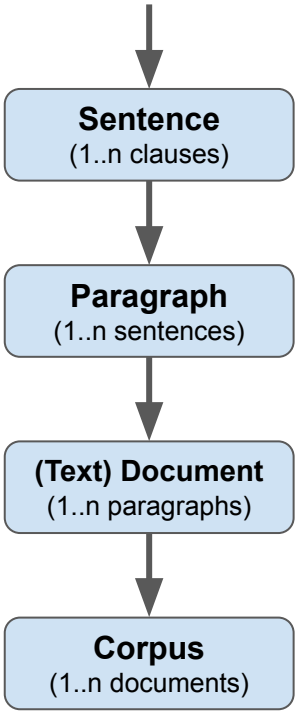
- Group of words expressing a particular idea or meaning

his quick reaction

- Phrase with a subject and verb

his quick reaction saved him

Core Building Blocks of (Written) Language



- 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)

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.

Morphemes

- 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

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

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

Examples

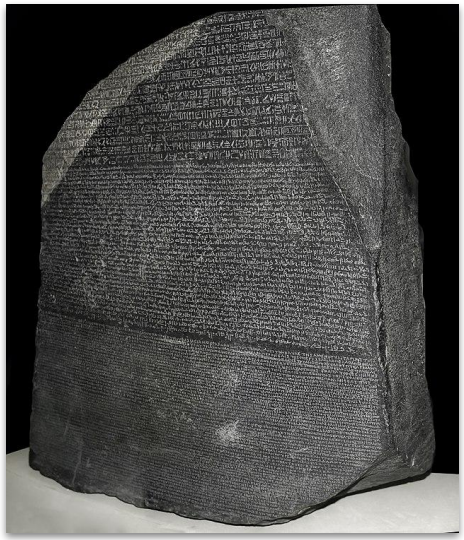
	Prefix	Prefix	Stem	Suffix	Suffix	Suffix
dogs			***dog***	*-s*		
walked			***walk***	*-ed*		
imperfection		*im-*	***perfect***	*-ion*		
hopelessness			***hope***	*-less*	*-ness*	
undesirability		*un-*	***desire***	*-able*	*-ity*	
unpremeditated	*un-*	*pre-*	***mediate***	*-ed*		
antidisestablishmentarianism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

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

Using Language Before ~1950

- Verbal communication between ***people***
  - Day-to-day conversations
  - Oral history

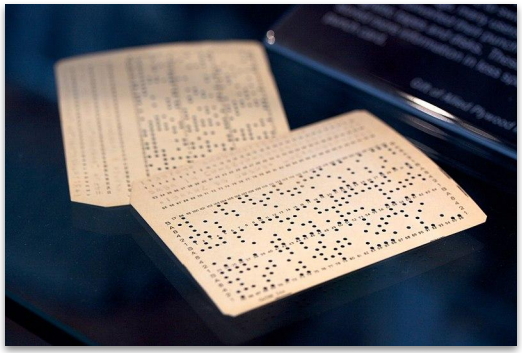
- Written communication for ***people***
  - Stone tablets, scrolls, books, etc.
  - Permanent record of written language



Source: Wiki Commons (CC BY-SA 4.0): Rosetta Stone

Since 1950: Communication with Machines

~50s-70s



Basic symbolic languages

(e.g., punch cards)

~80s

Formal languages

(e.g., programming languages)

Today

Natural language

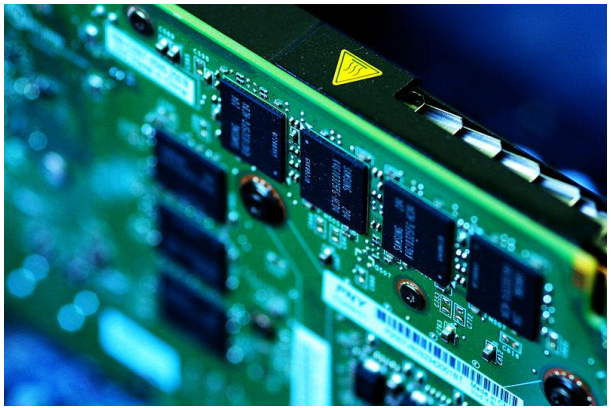
(e.g., conversational agents / chatbots)

Source: Wiki Commons (CC BY-SA 4.0): punch cards, programming

Communication with Machines

Humans

Machines



Natural Language

Analysis

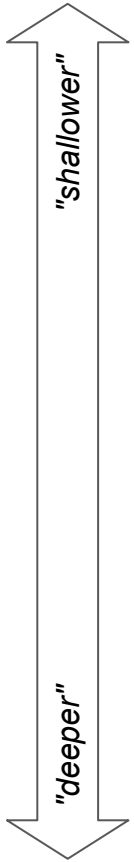
Generation

$\mathcal{R}$

Some abstract internal representation / model of language and the world

Source: Wiki Commons (CC BY-SA 4.0): gpu

NLP in One Slide



characters
morphemes
words

Lexical Analysis
(understanding structure & meaning of words)

- Tokenization
- Stemming
- Normalization
- Lemmatization

phrases
clauses
sentences

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

paragraphs
documents

Discourse Analysis
(meaning of sentences in documents)

- Coreference / anaphora resolution
- Ellipsis resolution
- Stance detection

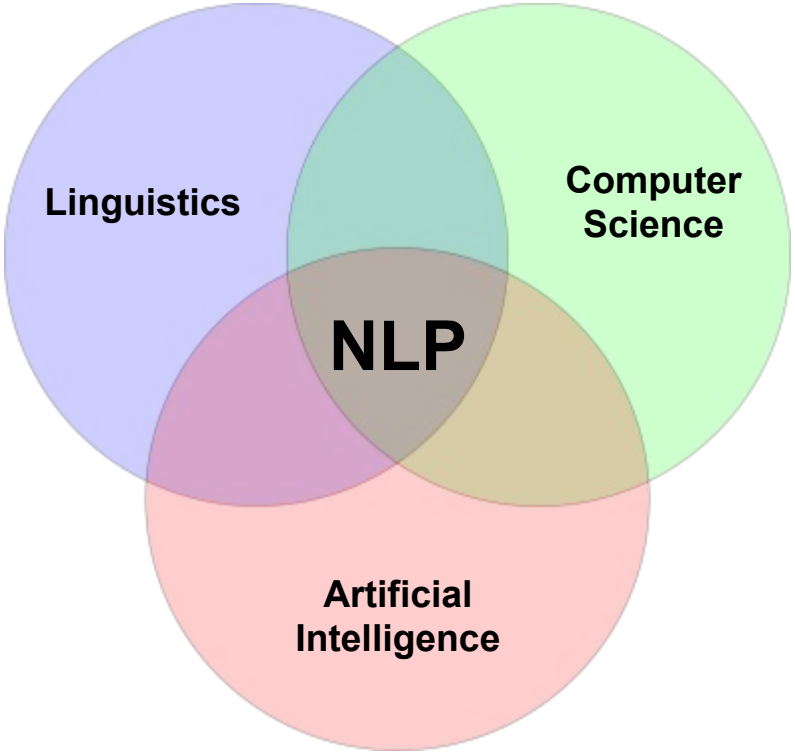
world knowledge
common sense

Pragmatic Analysis
(understanding & interpreting language in context)

- Textual Entailment
- Intent recognition

What is NLP? — The Big Picture

Human Language
- Speech
- Writing



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

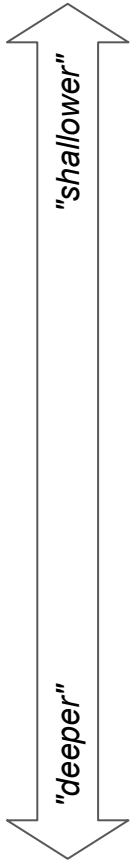
Machine Learning
Deep Learning

NUS | Computing
National University of Singapore

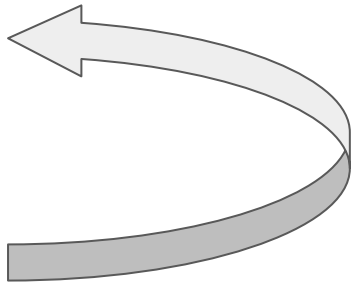
Natural Language Processing: Foundations

Section 1 — Important NLP Tasks

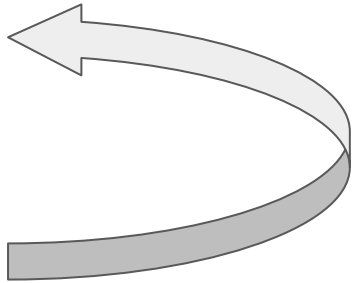
Important NLP Tasks



- 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



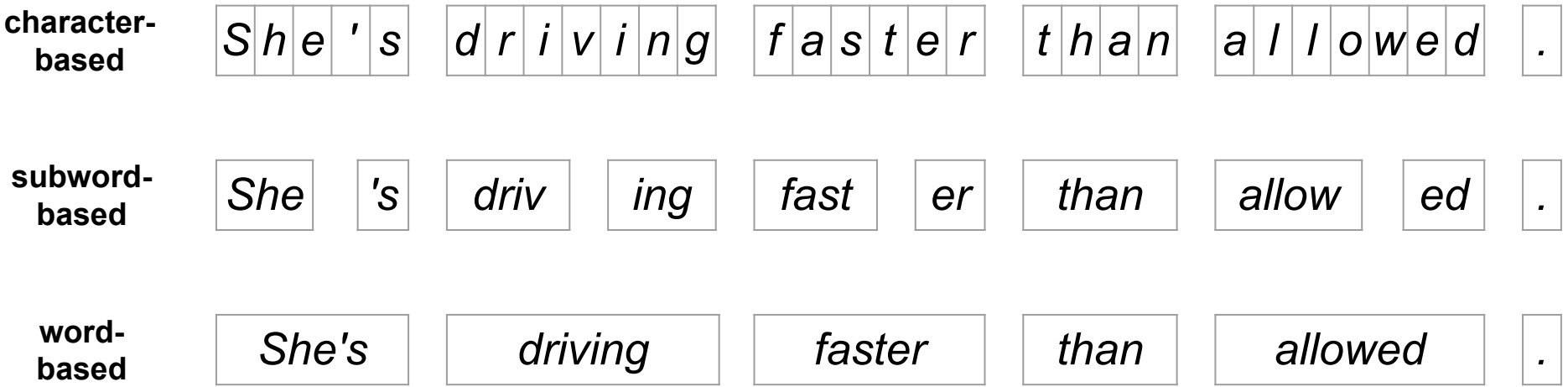
rely on



rely on

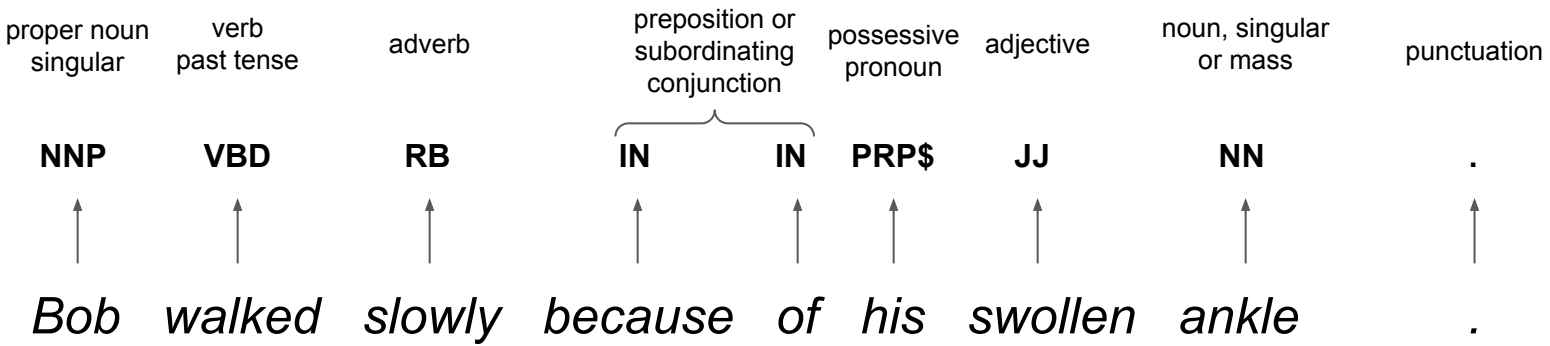
Lexical Analysis — Tokenization

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



Syntactic Analysis — Part-of-Speech Tagging

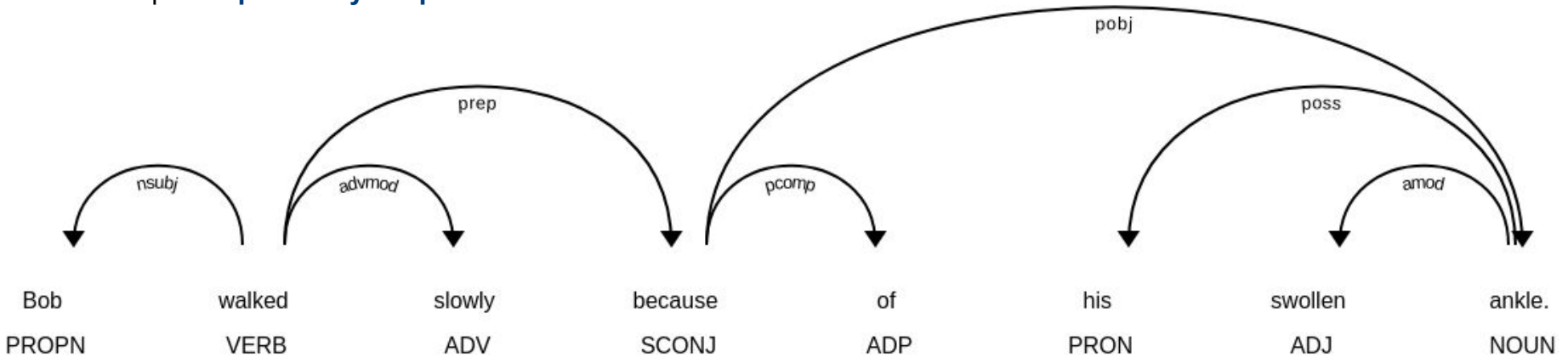
- Part-of-Speech (POS) tagging
  - Labeling each word in a text corresponding to a part of speech
  - Basic POS tags: noun, verb, article, adjective, preposition, pronoun, adverb, conjunction, interjection



Syntactic Analysis — Syntactic Parsing

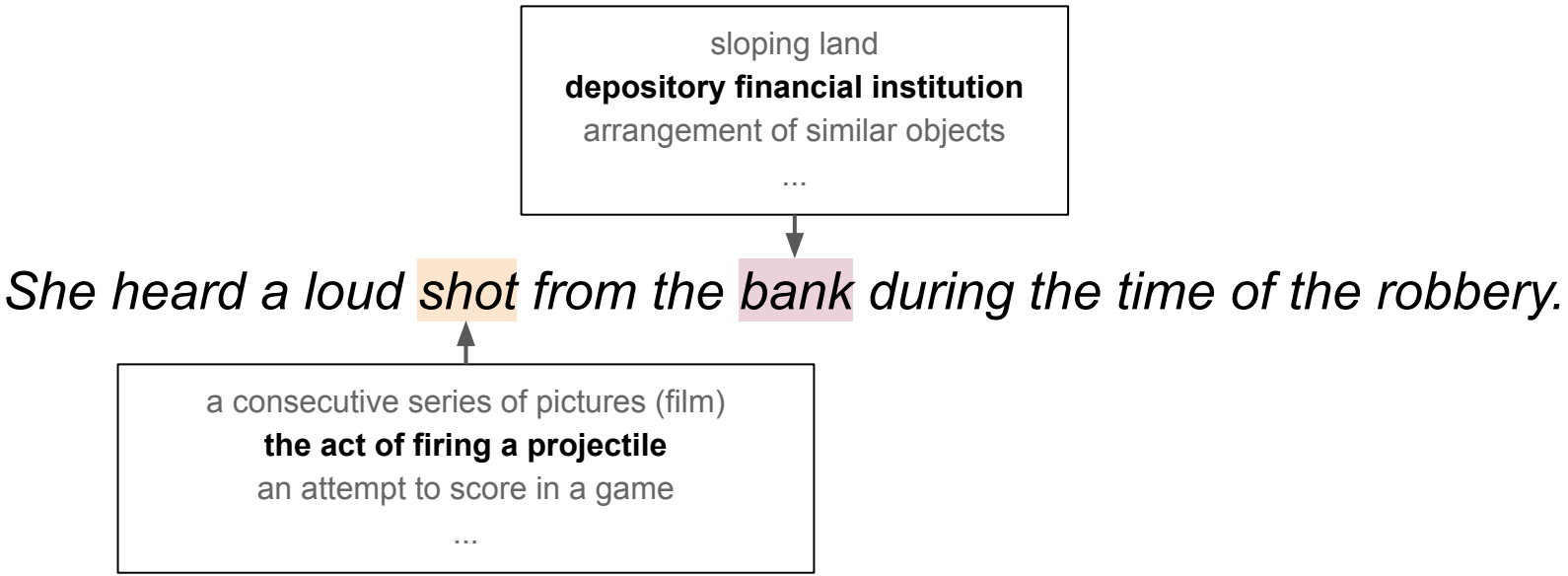
- Dependency parsing
  - Analyze the grammatical structure in a sentence
  - Find related words & the type of the relationship between them

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)

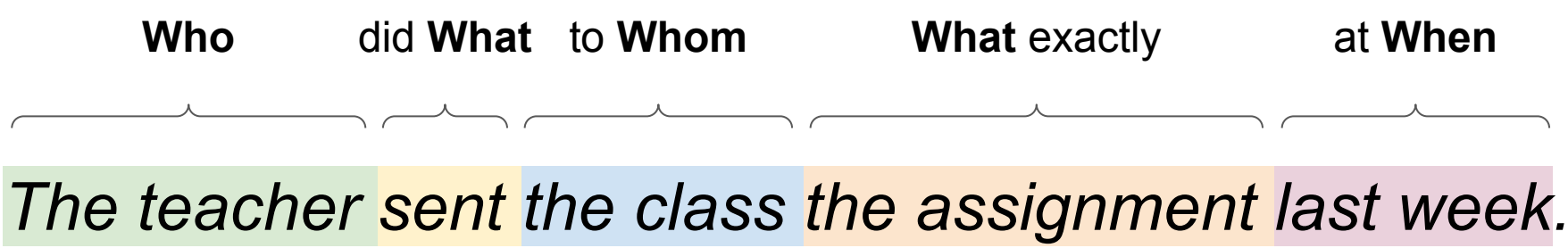


Semantic Analysis — Named Entity Recognition

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

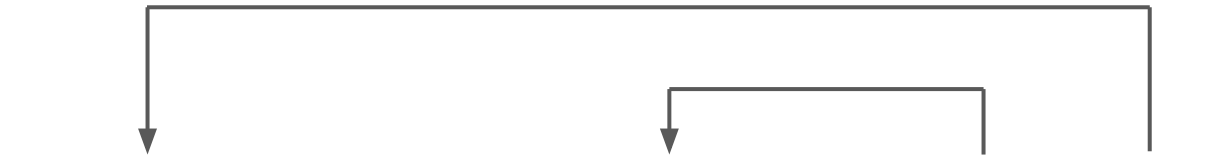
Semantic Analysis — Semantic Role Labeling

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



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.

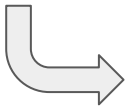


Mr Smith didn't see the car. Then the car hit Mr Smith.

Discourse Analysis — Ellipsis Resolution

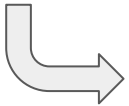
- Ellipsis Resolution
  - Inference of ellipses using the surrounding context
  - **Ellipsis**: omission of a word or phrases in sentence

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 chatbots

"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.

NUS | Computing
National University of Singapore

Natural Language Processing: Foundations

Section 1 — What makes NLP so Hard?

What Makes NLP so Hard?

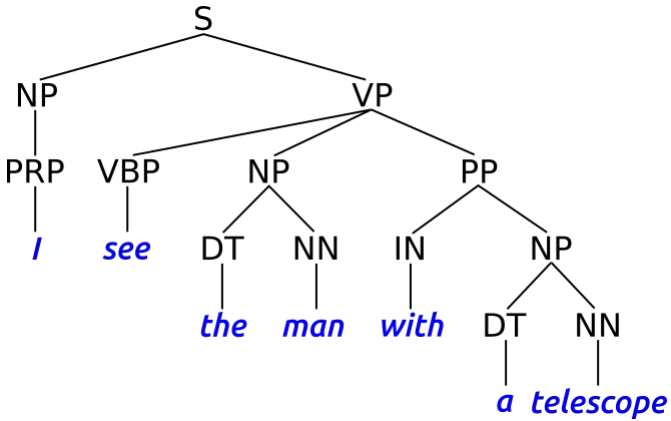
Challenges of Language ➜ Challenges of NLP

- ambiguous
- redundant
- changing
- unbounded
- imprecise

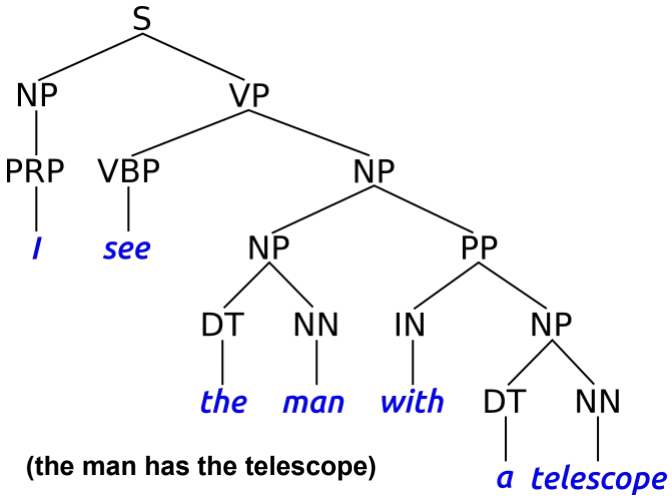
- expressive
- sparse
- variable
- "large-scale"
- …

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 semantics!



(I have the telescope)



(the man has the telescope)

Ambiguity

- 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

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

???

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

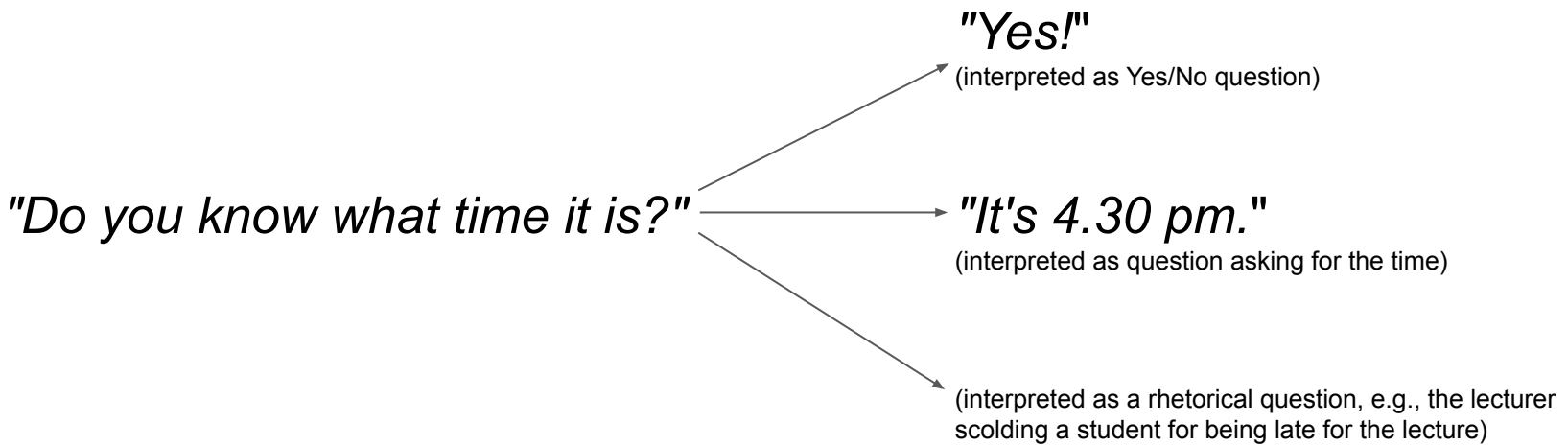
vs.

???

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

Ambiguity

- Pragmatic Ambiguity
  - Unclear semantics if context is unknown



Expressivity

- In general, the same meaning can be expressed with 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

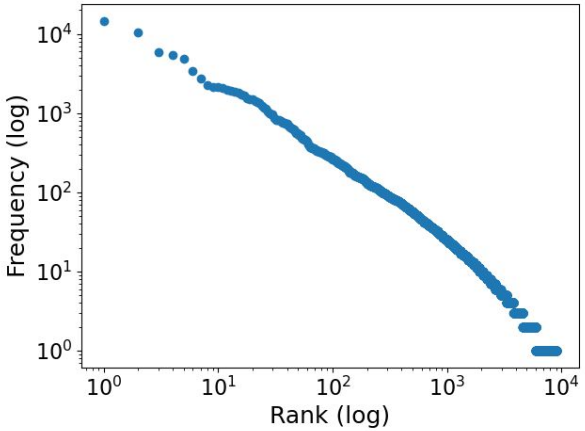
"Oh yeah...studying NLP 24/7 is reeeally my favorite way to spend a weekend!"

Variation

- No one-size-fits-all NLP solutions
  - Difference in underlying task
    (tokenizing, stemming, syntax parsing, part-of-speech tagging, 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

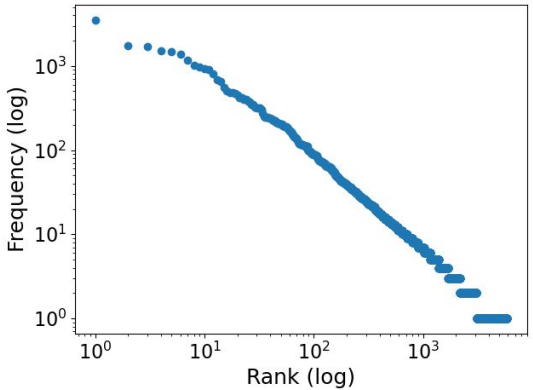
- Sparsity in text corpora
  - Word frequencies inversely proportional to their rank ➜ Zipf's Law
  - Example: *"On the Origin of Species"*
    (Charles Darwin, 1859; 212k+ words)



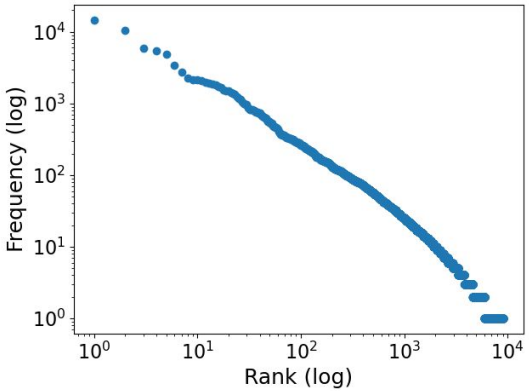
Rank	Word	Freq.
1	*the*	14,767
2	*of*	10567
3	*and*	5920
4	*in*	5477
5	*to*	4837
6	*a*	3460
7	*that*	2764
8	*as*	2242
9	*have*	2121
10	*be*	2116
...	...	...
101	*mr*	263
102	*parts*	260
103	*often*	260
104	*period*	259
105	*common*	256
...	...	...
1001	*increasing*	25
1002	*expected*	25
1003	*egg*	25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

Sparsity

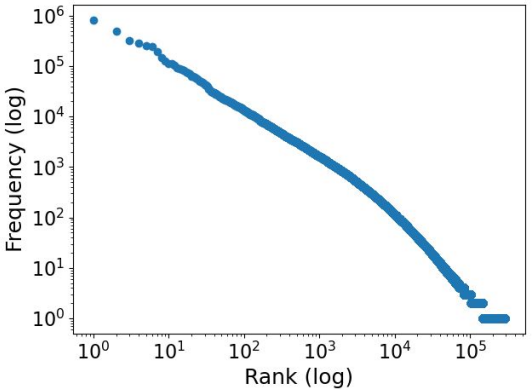
The Hound of the Baskervilles (63k+ words)



On the Origin of Species (212k+ words)



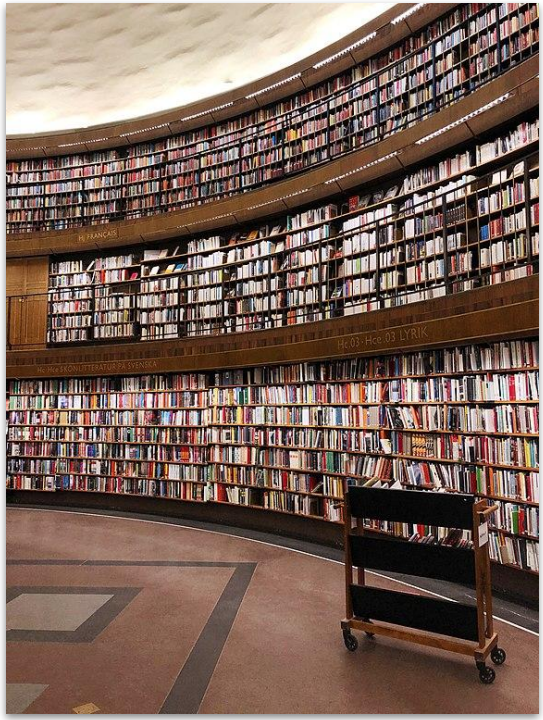
100MB Wikipedia dump (14.4M+ words)



➜ Regardless of size and domain of a 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



Source: Wiki Commons (CC0 1.0): Stockholm Public Library

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**?

NUS | Computing
National University of Singapore

Natural Language Processing: Foundations

Section 1 — When NLP Goes Wrong

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?*
  - ***Should*** *we organize users news feed based on their interests and likings to maximize user engagement?*
  - ***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)

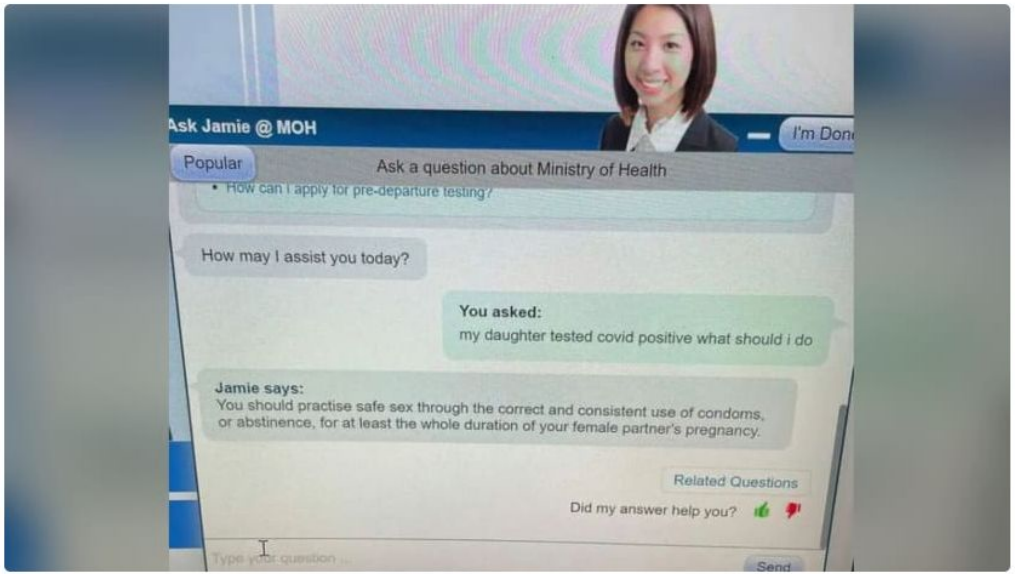
- Fundamental challenges in NLP
  - Most NLP techniques rely on statistical models
    (never 100% correct and often difficult to quantify errors)
  - Most NLP techniques learn from data
    (Is the data representative? Is the data biased?)

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

(Ian Malcolm; Jurassic Park, 1991)

NLP in the Press — For the Wrong Reasons

MOH temporarily disables Ask Jamie chatbot after 'misaligned replies'



Afifah Darke

05 Oct 2021 02:11PM
(Updated: 05 Oct 2021 02:17PM)

Source: https://www.channelnewsasia.com/singapore/moh-ask-jamie-covid-19-query-social-media-2222571

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

NUS | Computing
National University of Singapore

Natural Language Processing: Foundations

Section 1 — The Big Picture

What is NLP? — The Big 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
  - Linguists now use tools that originated from NLP!

What is NLP? — The Big 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, etc.)
- Advances in understanding of language in social context

Course Meta Topics

- Linguistic Issues
  - What is 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: LSTM, CNN, Seq2seq

Section 1 — Summary

- Main takeaway messages

NLP is important!

- Communication with machines and not only with other people
- NLP-powered applications more and more commonplace
- More on more written language will be generated by machines

NLP is challenging!

- NLP is challenging because Natural Language is challenging
- Natural Language is ambiguous, redundant, changing, unbounded, imprecise, expressive, etc.
- NLP methods need to be flexible, robust, efficient, accurate, etc.

NLP is interdisciplinary!

- NLP covers many fields
- Main fields: linguistics and computer science
- Many connected fields (e.g., ethics, psychology, education)