Natural Language Processing

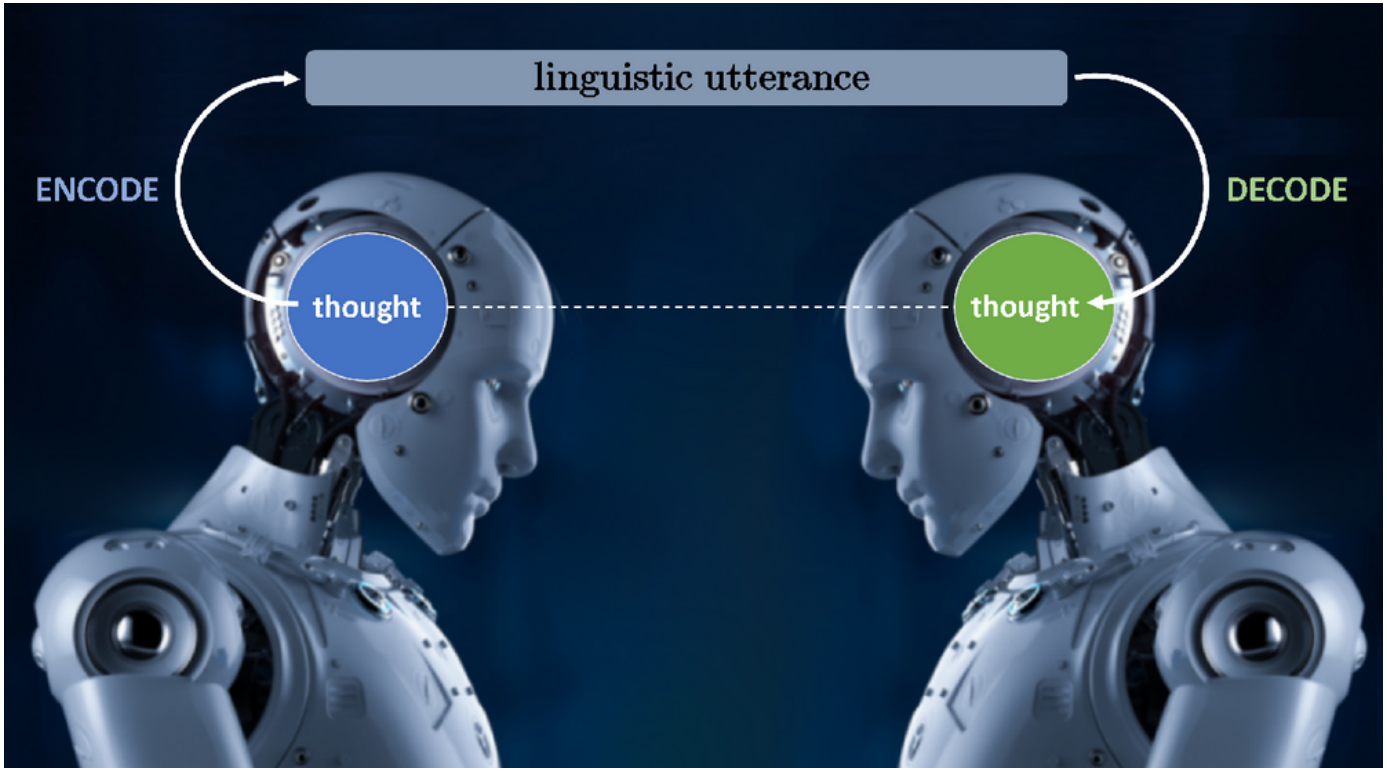
Lecture 1 : Introduction

Master Degree in Computer Engineering
University of Padua
Lecturer : Giorgio Satta

Natural language processing: An unexpected journey

©The Hobbit: An Unexpected Journey, 2012

What is natural language processing?



The gradient, Walid S. Saba

What is natural language processing?

There is an **impelling need** in our society to process extremely large and constantly growing amounts of text.

This is seen for instance in data analysis for

- business intelligence
- social media
- healthcare
- finance
- human resources
- advertising

The textual data people generate every day exceeds human processing powers. The solution, therefore, is to extract relevant information in some **automatic** way.

What is natural language processing?

Natural language processing (NLP) is a field of **artificial intelligence** (AI) that allows machines to read, derive meaning from text, and produce documents.

Terms 'natural language processing', 'computational linguistics' and 'human language technologies' may be thought of as essentially synonymous.

It works in the background of many services, from chatbots through virtual assistants to social media tracking.

Such language technologies are **already** showing major penetration into the information and communication industry.

What is natural language processing?

Some **well-known** end-to-end NLP applications

- chatbot

  ChatGPT, Bard, Bing Chat
- virtual assistant

  Siri, Alexa, Google Home
- machine translation

  Google Translate, DeepL
- sentiment analysis
- fake news detection

What is natural language processing?

NLP is also at the basis of several **generative AI** applications

- AlphaCode / Copilot (text to code)
- DALL-E / Midjourney (text to image)
- Pika / Lumiere / Sora (text to video)

What is natural language processing?

Source: Tracy Mayor, **MIT**

Why finance is deploying natural language processing

Case study: **Finance**

In finance, data that can help make timely decisions comes in text. Earnings reports are one example. A company will release its report in the morning, and it will say "Our earnings per share were a $1.12."

By the time that **unstructured** data makes its way into a database of a data provider where you can get it in a structured way, hours have passed and you've lost your edge.

NLP can deliver those transcriptions in minutes, giving analysts a competitive advantage.

What is natural language processing?

Source: Collins Ayuya, **Section**

Automated fake news detection

Case study: **Social networks**

Fake news refers to information content that is false, misleading or whose source cannot be verified. Automatic approaches to fake news detection involve NLP.

As an example, companies like Facebook, Twitter, TikTok, Google, Pinterest, Tencent, YouTube, and others are working with the World Health Organization to mitigate the COVID-19 driven infodemic.

What is natural language processing?

Source: **foresee medical**

Natural language processing in healthcare

Case study: **Health**

Huge volumes of unstructured patient data is input into **electronic health record systems**. 80% of healthcare documentation is unstructured text.

Healthcare NLP uses specialized engines capable of discovering previously missed or improperly coded patient conditions.

Very short history of natural language processing

Very short history of natural language processing

In summary:

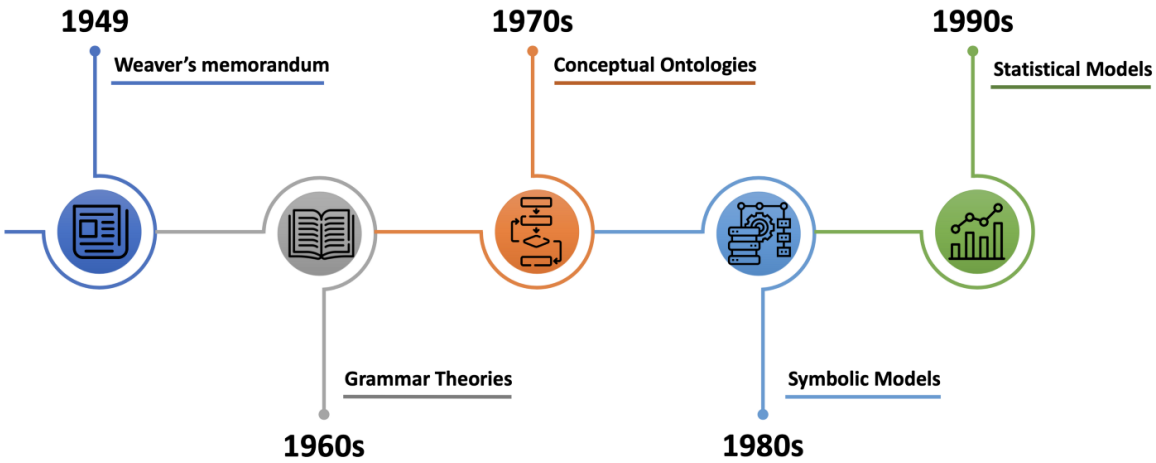
1950-1960: prehistory, scientific knowledge regarding artificial intelligence and linguistics extremely limited

1960-1990: **symbolic** models, rules handwritten by experts, very limited coverage

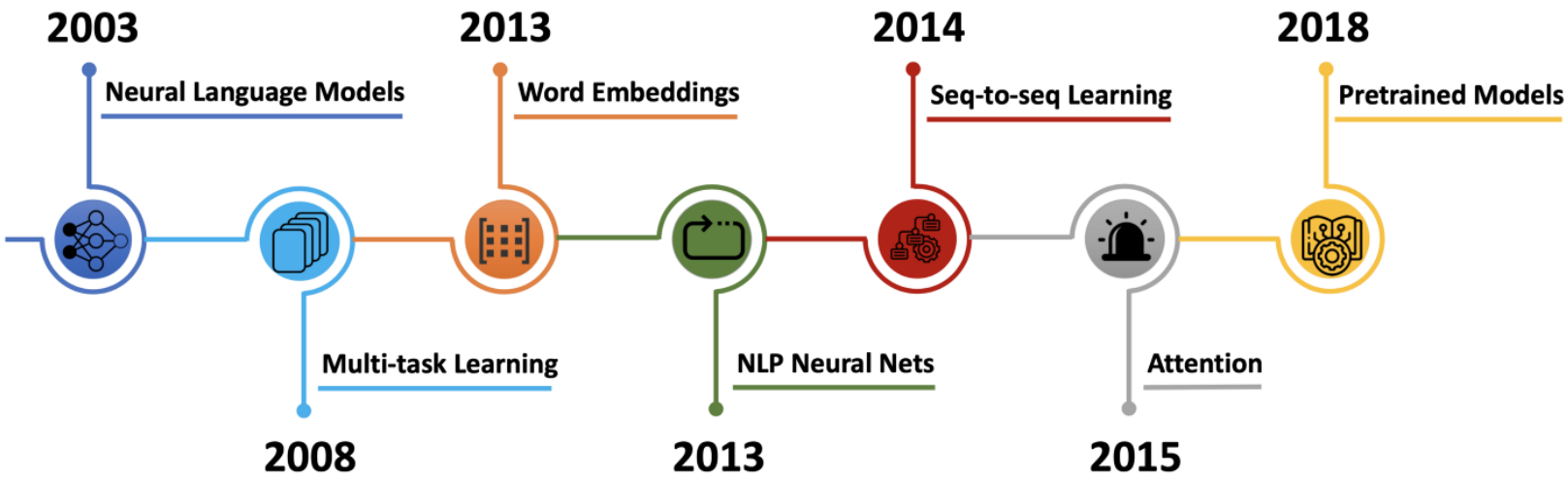
1990-2010: **statistical** models, machine learning on data annotated by experts, good coverage

2010-present: **neural** models, machine learning on non-annotated data, excellent coverage

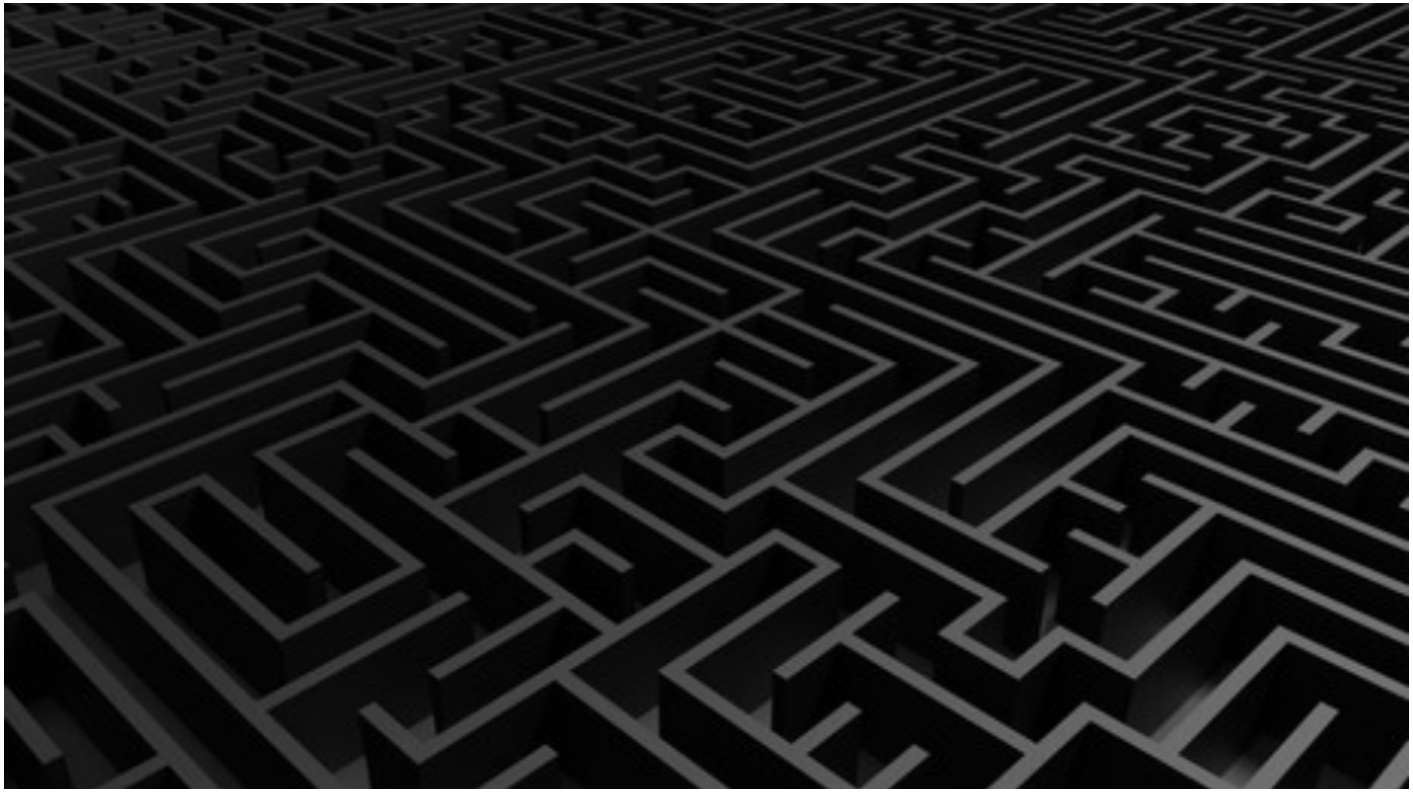
Very short history of natural language processing



Very short history of natural language processing



Why is natural language processing tricky?



©Shuttestock, Dark Maze

Why is natural language processing tricky?

NLP distinguishes itself from other AI application domains, as for instance computer vision or speech recognition.

Text data is fundamentally discrete. But **new words** can always be created.

Stan: an extremely enthusiastic and devoted fan (stalker-fan).

Nomophobia: anxiety caused by not having a working mobile phone.

Few words are very frequent, and there is a long tail of **rare words** (Zipf/Mandelbrot law).

Out-of-vocabulary words are always being discovered (Herdan/Heaps law).

More about the above two laws in next lectures.

Why is natural language processing tricky?

Language is **ambiguous**: units can have different meanings.

Language is **compositional**: meaning of a unit defined as a function of the meaning of its components.

Language is **recursive**: units can be repeatedly combined.

Language unveils **hidden structure**: local changes in a sentence might have global effects.

See next slides.

Ambiguity

Phonetic transcription **[ralt]** might mean write, right, rite

Word **can** belongs to several categories: noun, verb, or modal

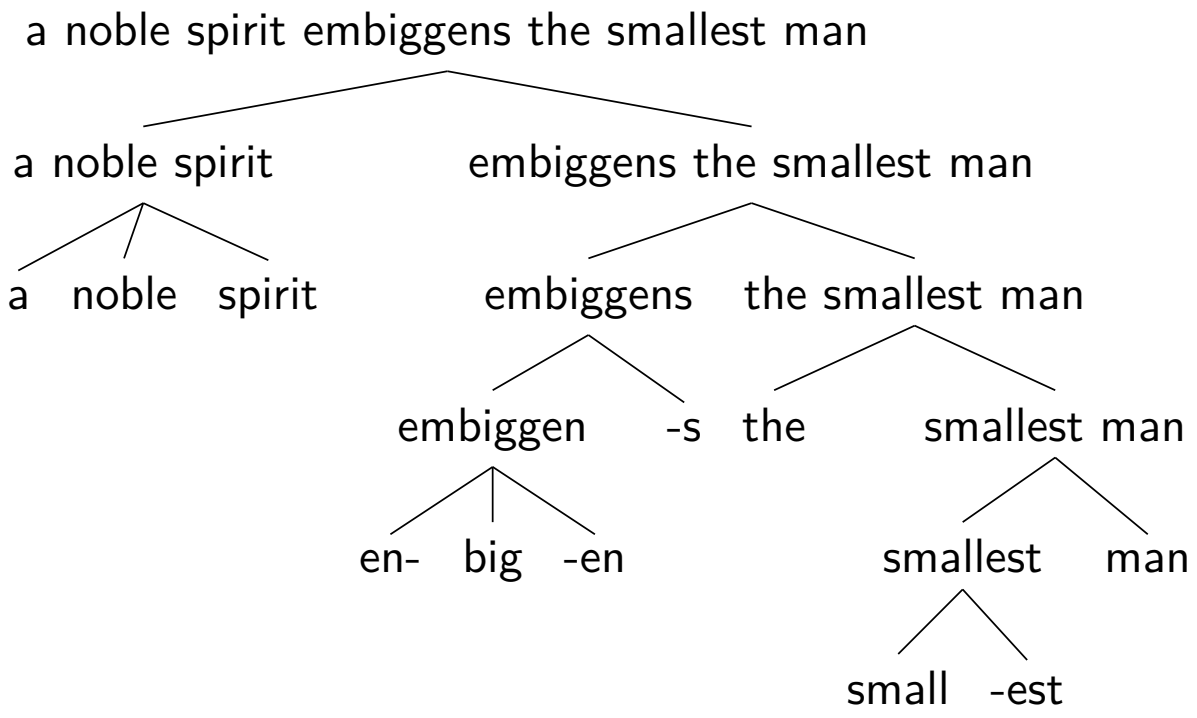
Word **bank** has different meanings: river bank or money bank

Morphological composition: word un-do-able is ambiguous between 'not doable' and 'can be undone'

Sentence 'I saw the man **with** the telescope' has two interpretations

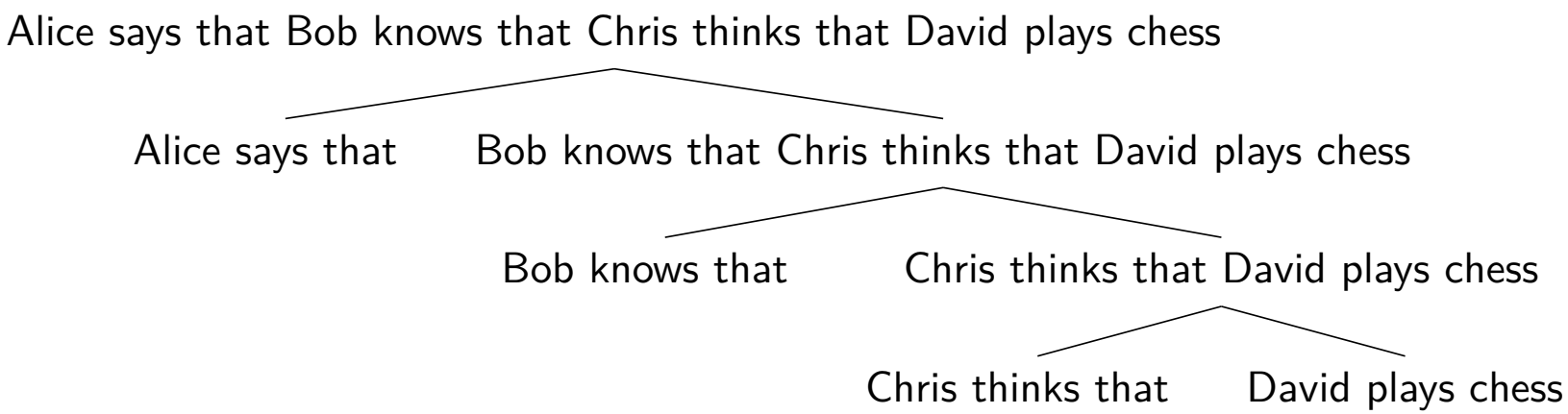
Two possible references for pronoun **him** in 'The son asked the father to drive him home'

Compositionality



At each level, **meaning** of a larger unit is provided by some function of the meaning of its immediate components and the way they are combined.

Recursion



The rules of the grammar can iterate to generate an infinite number of structures, each with its specific meaning.

Recursion is considered the main difference between human and other animals' languages.

Hidden structure

Local changes can **disrupt** the interpretation of a sentence. This suggests the existence of hidden structure.

Example : The trophy doesn't fit into the suitcase because **it** is too {small, large}.

Example of Winograd schema challenge, discussed later in this course.

How does natural language processing work?

How does natural language processing work?

NLP is an **interdisciplinary** area of research, based on several scientific fields (alphabetical order)

- cognitive science
- computer science
- linguistics
- machine learning
- mathematical logic
- statistics

How does natural language processing work?

NLP applications are based on ideas and tools such as (alphabetical order)

- deep learning & optimisation
- dynamic programming
- grammars & automata
- probability and information theory

How does natural language processing work?

NLP plays a significant role in these **neighbour** fields

- computational psycholinguistics
- computational social science
- digital humanities
- human-computer interaction
- information retrieval
- sociolinguistics
- speech processing/understanding
- text mining

Learning & knowledge

Raffaello, The School of Athens

Learning & knowledge

Rationalism:

> *A significant part of the knowledge in the human mind is not derived by the senses but is fixed in advance, presumably by genetic inheritance.*
>
> *Noam Chomsky*
> *Poverty of the stimulus, 1980*

Generative linguists have argued for the existence of a **language faculty** in all human beings, which encodes a set of abstractions specially designed to facilitate learning, understanding and production of language.

Learning & knowledge

Empiricism:

> *The view that there is no such thing as innate knowledge, and that knowledge is instead derived from experience, either sensed via the five senses or reasoned via the brain or mind.*

Originated in ancient Hindu and Greek philosophy

At the time of writing, many statistical NLP techniques work very well on texts, without the need to use special bias representing linguistic knowledge or mental representation of language.

Learning & knowledge

A recurring topic of debate in NLP is the relative importance of machine learning vs. linguistic knowledge

- 1950s: Empiricism I — information theory
- 1970s: Rationalism I — formal language theory and logic
- 1990s: Empiricism II — stochastic grammars
- 2010s: Empiricism III — deep learning

Source: K. Church and M. Liberman, The Future of Computational Linguistics: On Beyond Alchemy (2021).

Search & learning

Kir Simakov on Unsplash

Many natural language processing problems can be written mathematically in the form of **optimization**

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})}{\text{argmax}} \ \Psi(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$$

where

- $\boldsymbol{x}$ is the input, which is an element of a set $\mathcal{X}$;
- $\boldsymbol{y}$ is the output, which is an element of a set $\mathcal{Y}(\boldsymbol{x})$;
- Ψ is a scoring function, also called the **model**, which maps from the set $\mathcal{X} \times \mathcal{Y}$ to the real numbers;
- $\boldsymbol{\theta}$ is a vector of **parameters** for Ψ;
- $\hat{\boldsymbol{y}}$ is the predicted output, which is chosen to maximize the scoring function.

The **search** module is responsible for finding the candidate output $\hat{\boldsymbol{y}}$ with the highest score relative to the input $\boldsymbol{x}$.

This requires efficient algorithms.

The **learning** module is responsible for finding the model parameters $\boldsymbol{\theta}$ that maximizes the predictive performance.

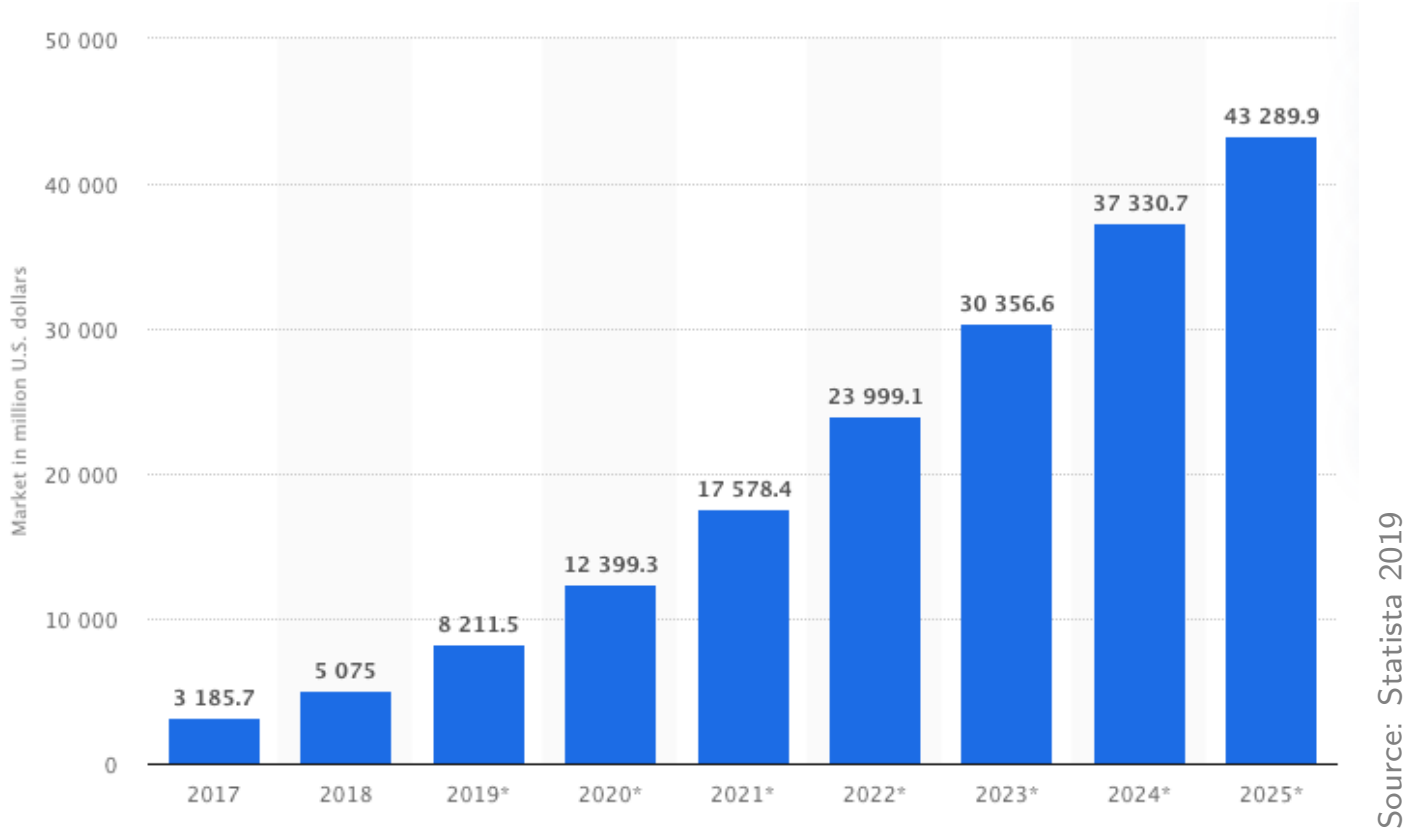
This requires machine learning.

Structured prediction is an umbrella term for supervised machine learning techniques that involves predicting structured objects, rather than scalar discrete or real values.

Miscellanea

Kristine Rosenblatt, Kristine's Kitchen

Market



The NLP market is predicted to be almost 14 times larger in 2025 than it was in 2017, increasing from around three billion U.S. dollars to over 43 billion.

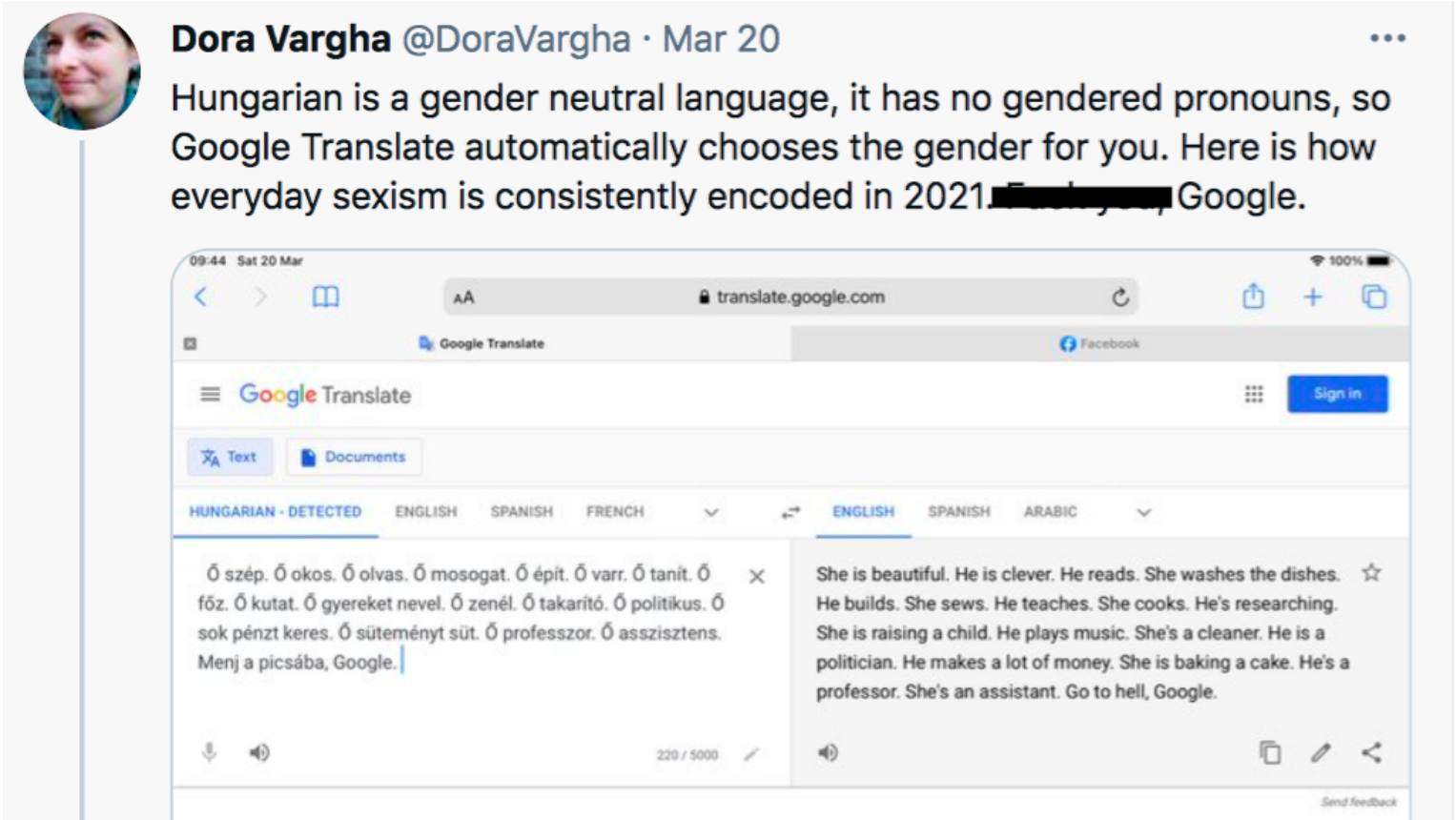
Environment

Consumption	**CO_2e (lbs)**
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Strubell et al., 2019
https://www.aclweb.org/anthology/P19-1355/

Model training incurs a substantial cost to the environment due to the energy required to power this hardware for weeks or months at a time.

Ethics



Dora Vargha @DoraVargha · Mar 20

Hungarian is a gender neutral language, it has no gendered pronouns, so Google Translate automatically chooses the gender for you. Here is how everyday sexism is consistently encoded in 2021 [redacted] Google.

https://twitter.com/DoraVargha/status/1373211762108076034?s=20

Ethics

THE WALL STREET JOURNAL.

English Edition | Print Edition | Video | Podcasts | Latest Headlines

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

By Catherine Stupp

Updated Aug. 30, 2019 12:52 pm ET

PRINT TEXT

Criminals used artificial intelligence-based software to impersonate a chief executive's voice and demand a fraudulent transfer of €220,000 ($243,000) in March in what cybercrime experts described as an unusual case of artificial intelligence being used in hacking.

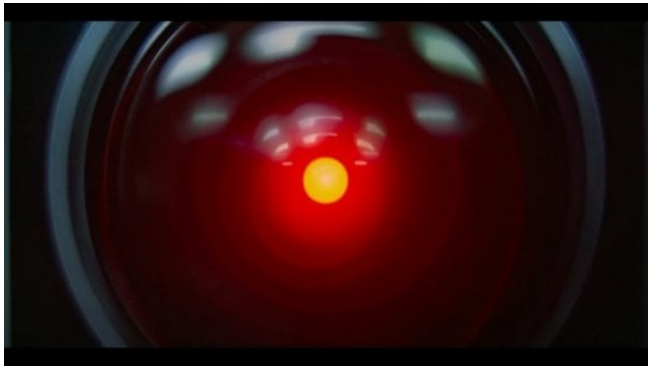
The CEO of a U.K.-based energy firm thought he was speaking on the phone with his boss, the chief executive of the firm's German parent company, who asked him to send the funds to a Hungarian supplier. The caller said the request was urgent, directing the executive to pay within an hour, according to the company's insurance firm, Euler Hermes Group SA.

Euler Hermes declined to name the victim companies.

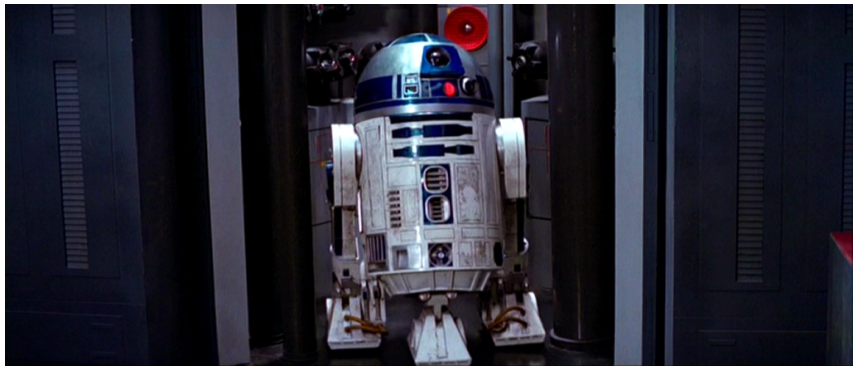
https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

NLP Legacy

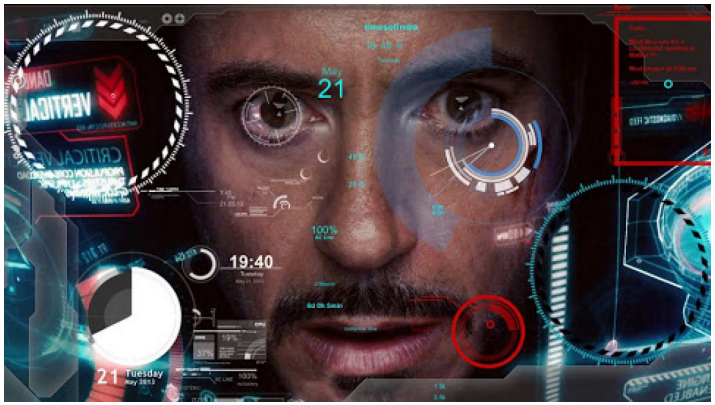
The field of natural language processing has had a recurring impact on popular culture.



HAL 9000 in 2001: A Space Odyssey (1968)



R2D2 in Star Wars (1977)



J.A.R.V.I.S. in Iron Man (2008)

Samantha virtual assistant in Her (2013)

NLP Legacy (cont'd)

Alien language in Arrival (2016)