UNIVERSITÉ DU LUXEMBOURG

Natural Language Processing

Lecture 1: Introduction

Salima Lamsiyah
University Luxembourg, FSTM, DCS, MINE Research Group
Salima.lamsiyah@uni.lu

Welcome to the NLP Course!

- **Instructor: Dr. Salima LAMSIYAH**
  - Postdoctoral Researcher in NLP and ML, University of Luxembourg
  - Office: Maison du Nombre, 4th floor, MNO-E04-0435080
  - Email: salima.lamsiyah@uni.lu

- **Research & Teaching Interests:**
  - Natural Language Processing (NLP)
  - Machine Learning & Deep Learning
  - Reinforcement Learning
  - AI for Education

✨ *Looking forward to exploring NLP together!*

Syllabus (Part 1)

- **Introduction**
  - Overview of NLP as a field
- **Linguistic Structure & Analysis**
  - Words & Morphological Analysis
  - Sequences: Part-of-Speech Tagging (POS), Named Entity Recognition (NER)
  - Syntactic Parsing: Phrase Structure, Dependencies
- **Modeling**
  - Text Classification
  - Language Modeling: n-gram, Neural LMs
  - Representation Learning: Word Vectors, Contextualized Embeddings

Syllabus (Part 2)

- **Applications**
  - Sentiment Analysis
  - Sentence Auto-Completion
  - Machine Translation
  - Chatbots & Conversational Agents
  - Information Retrieval & Summarization

- **Final Project (Team-based)**
  - Build an end-to-end NLP application of your choice using LLMs
- ***Goal**: Apply theory + modeling to solve a real-world NLP problem*

Prerequisites

- Python Skills
- Basics Probability and Statistics
- Background in Linear Algebra and Calculus
- Familiarity with ML is helpful but not assumed

Course Lecture Plan

- **Lecture 1:** Introduction to NLP
- **Lecture 2:** Text Preprocessing
- **Lecture 3:** Text Representation
- **Lecture 4:** Text Classification
- **Lecture 5:** Language Modeling
- **Lecture 6:** Logistic Regression for NLP
- **Lecture 7:** Neural Networks for NLP
- **Lecture 8:** Word Vectors & Embeddings
- **Lecture 9:** RNNs and Transformers
- **Lecture 10:** Large Language Models (LLMs) & RAG

Course Practical Assignments

- **Assignment 1:** Text Preprocessing & Cleaning
- **Assignment 2:** Text Classification (traditional & neural models)
- **Assignment 3:** Word Embeddings & Representation Learning
- **Assignment 4:** Sentence Auto-Completion with Language Models
- **Assignment 5:** Machine Translation (small domain)
- **Assignment 6:** Transformer-Based Chatbot
- **Assignment 7:** Applications of Large Language Models (LLMs)

Learning Goals

- Understand the main subfields and applications of **Natural Language Processing (NLP)**
- Apply **text preprocessing** and **representation methods** for linguistic data
- Implement **classification models** for NLP tasks (e.g., sentiment analysis, spam detection)
- Train and evaluate **language models** (n-gram, neural LMs)
- Use **word embeddings** and contextualized representations for downstream tasks
- Build and train **neural networks and Transformers** for NLP
- Develop end-to-end applications: **LLM-based Applications**
- Conduct a **final NLP project** demonstrating integration of theory and practice

Grading

- **Final Written Exam:** 30%

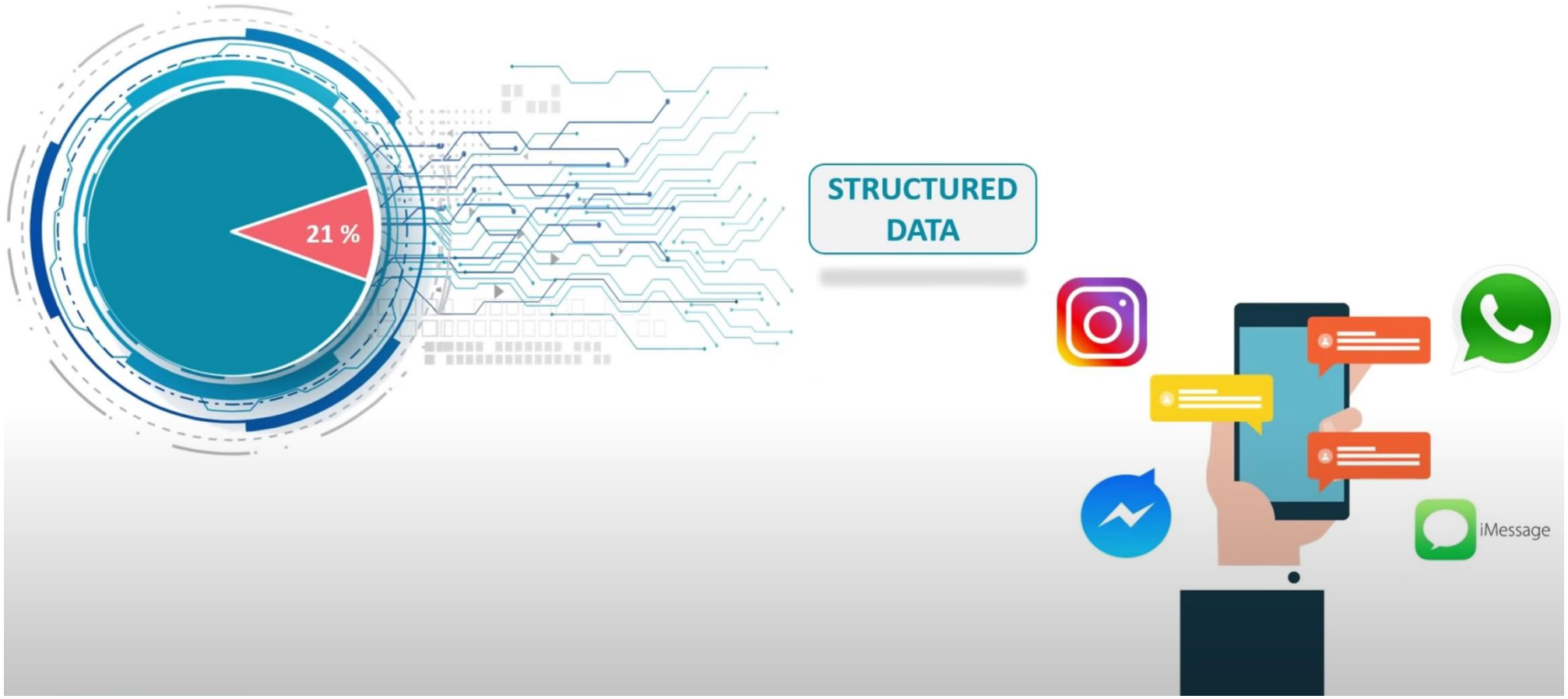
- **Final Project:** 30%
  - Project Implementation: 20%
  - Presentation: 10%

- **Practical Assignments:** 30%
- **Class Participation , Presentation, Attendance and Presence:** 10%

Plan for Today

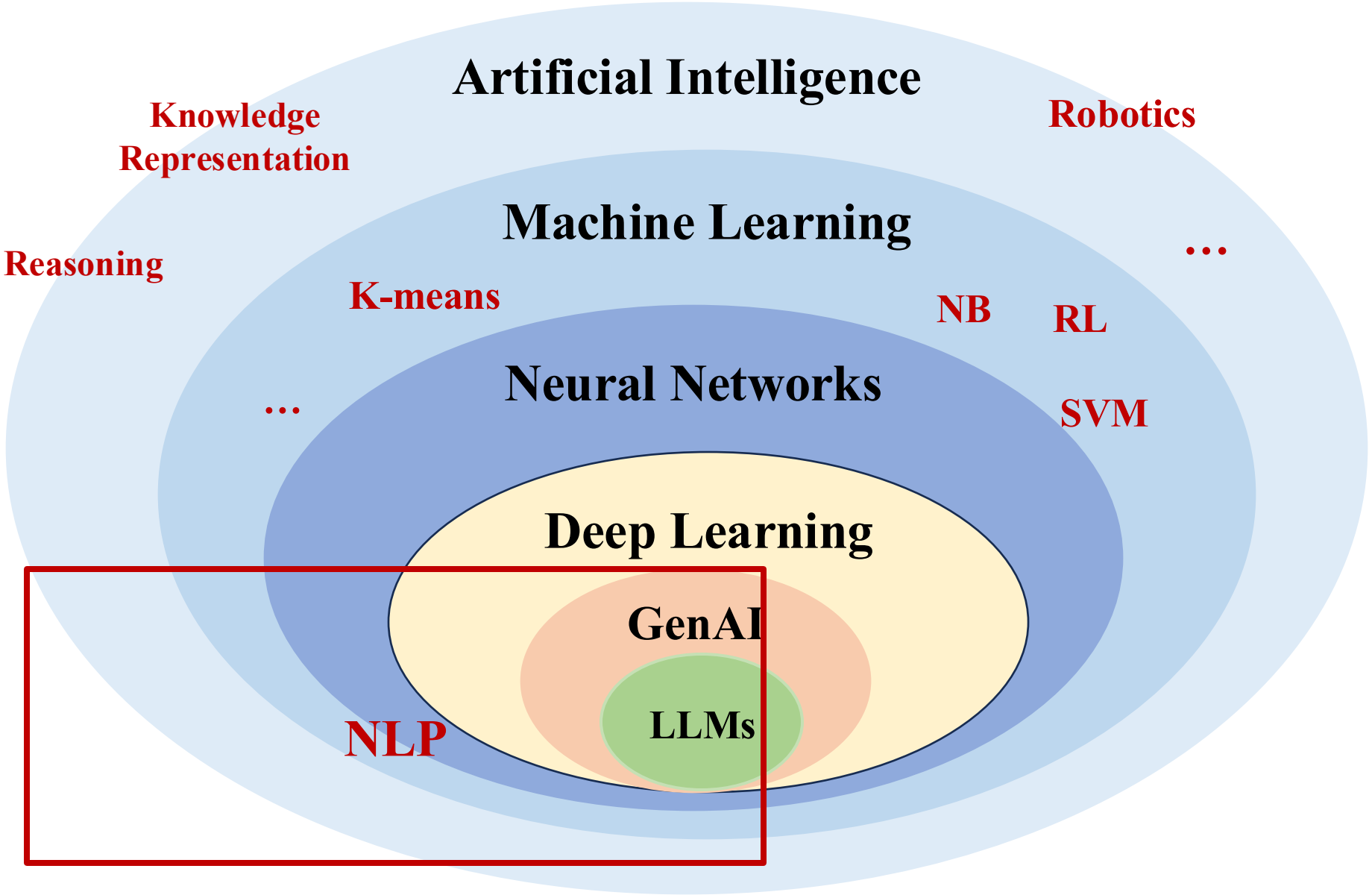
1. What is NLP?
2. Today's NLP Application
3. Why is NLP Hard?
4. NLP Components
5. Steps involved in solving an NLP problem
6. Fields with Connections to NLP

21^{st} Century: *We Are Living In The Age of Information Overload*



https://www.g2.com/articles/structured-vs-unstructured-data

AI Terminology: Artificial Intelligence



Natural Language Processing

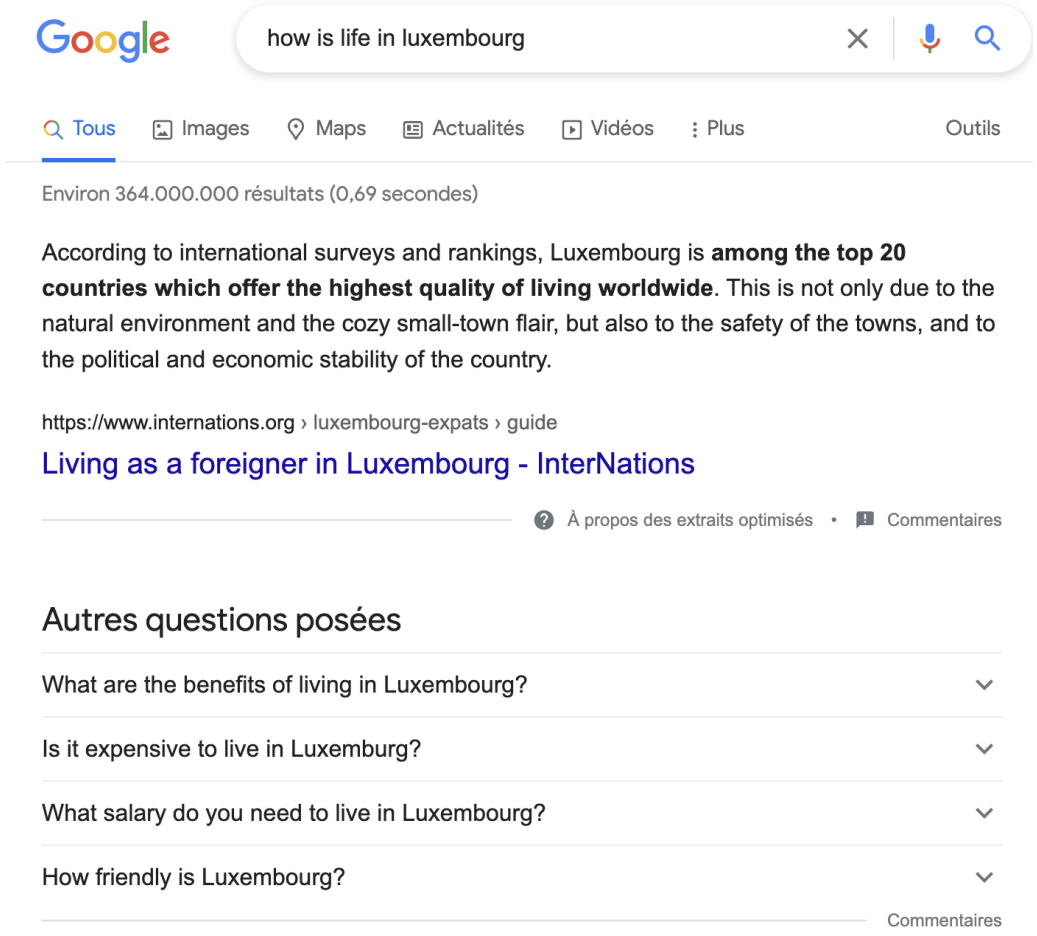
- **Natural Language Processing (NLP)** is a field at the intersection of:
  - Computer Science
  - Artificial Intelligence
  - And Linguistics.

- **Goal**: for computers to process or "Understand" natural language in order to perform tasks:
  - Translation, Question Answering, Siri, Google Assistant, …

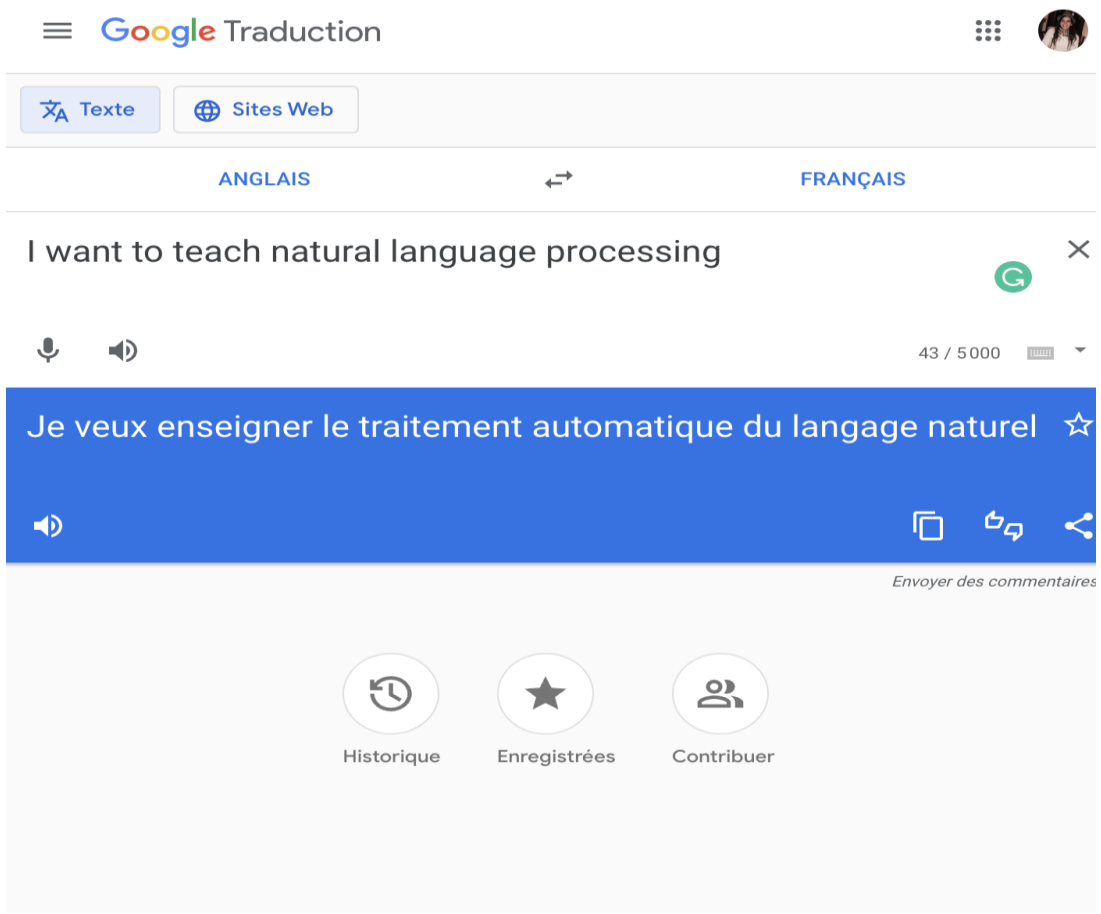
- Fully **understanding** and **representating** the **meaning** of a language is a difficult goal.
  - Perfect language understanding is AI-complete (AI-hard)

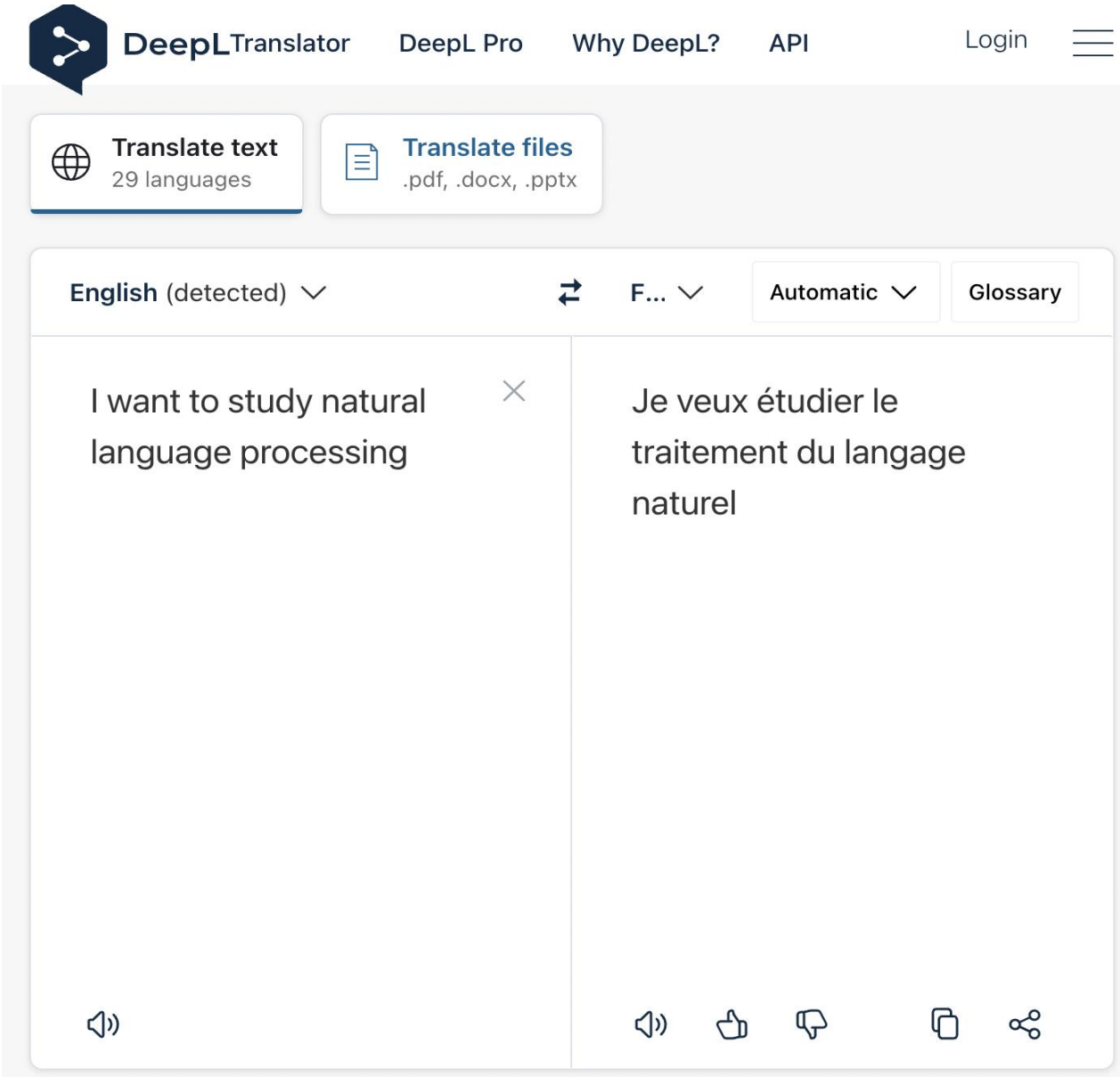
NLP Applications

Question Answering and Chatbot: **ChatGPT**



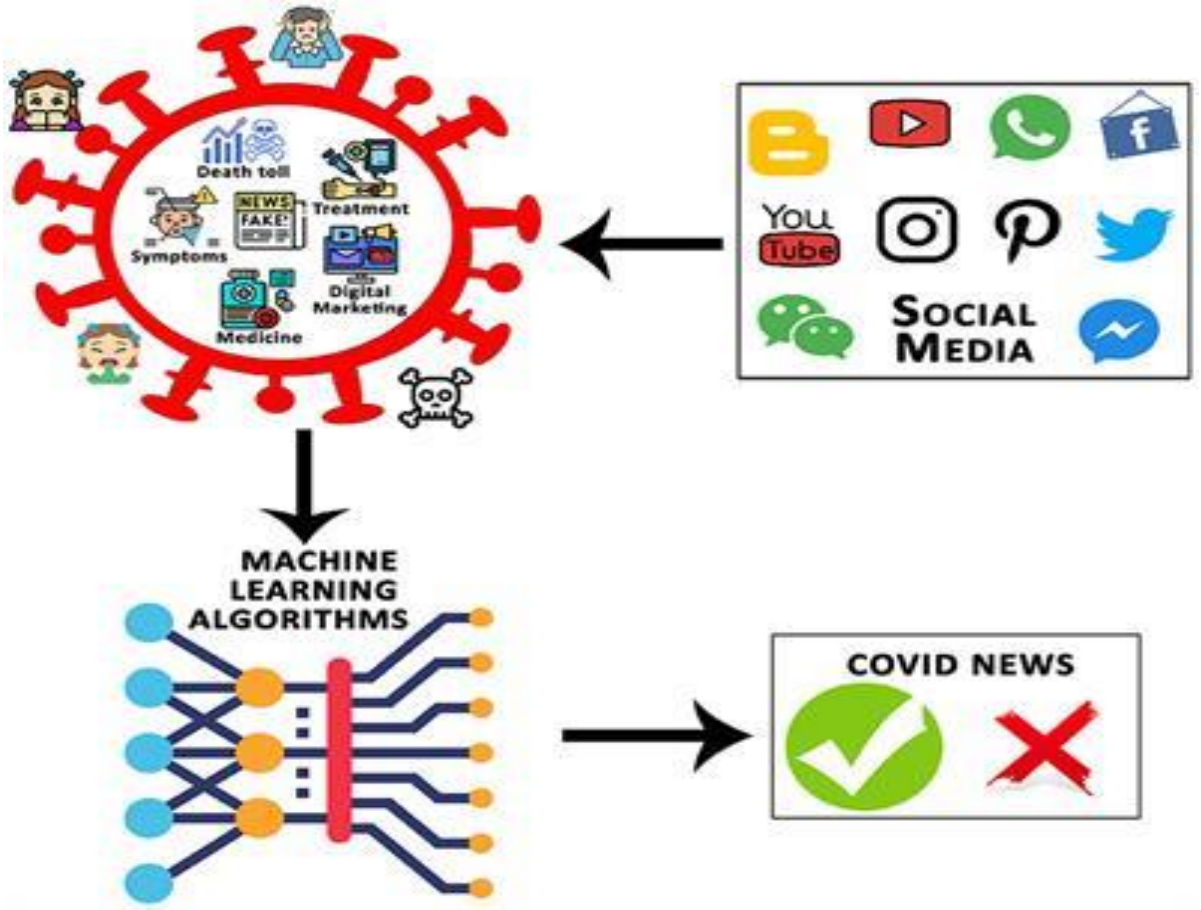
Machine Translation



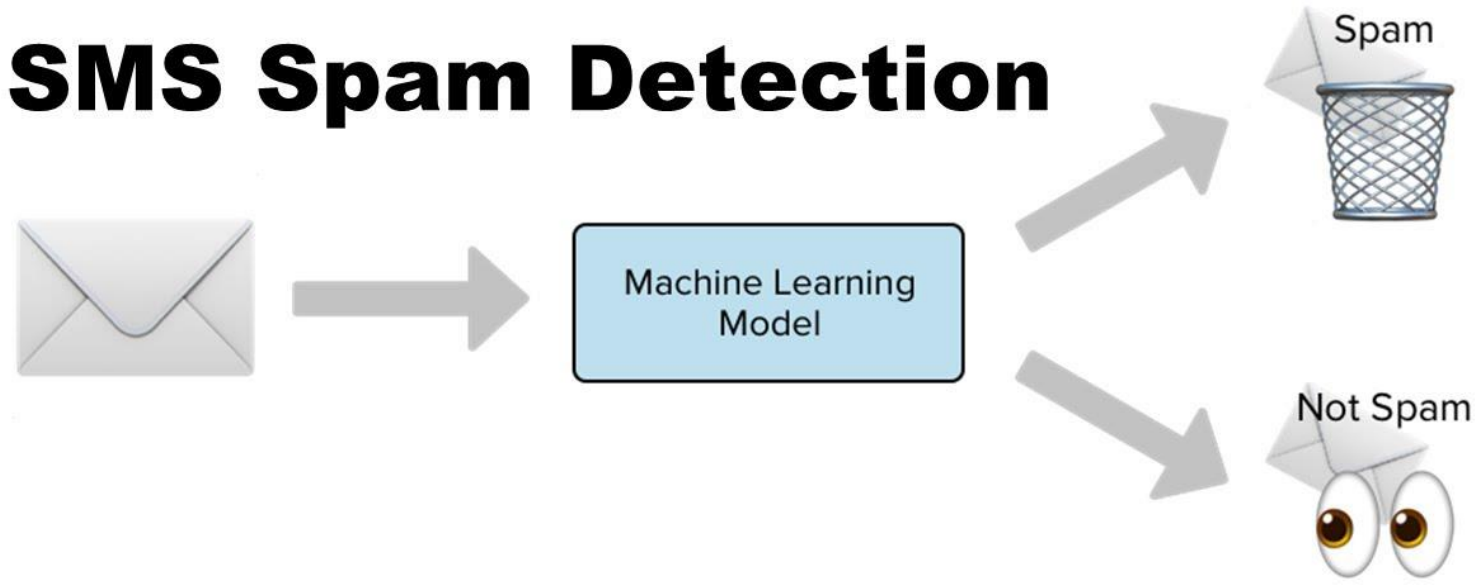


Fact verification: trustworthy or fake?

Have you
Covid19?
Drink Alcohol!



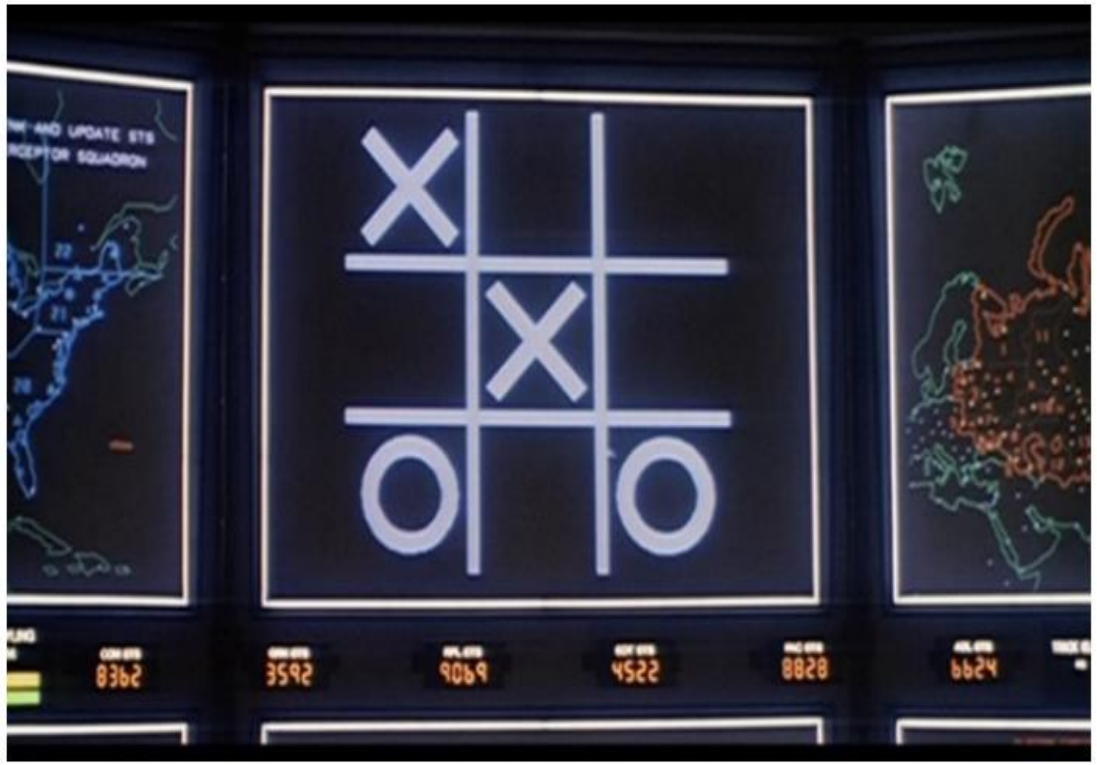
Spam Detection



Sentiment Analysis

- Sentiment analysis is used to determine whether a data is postive, negative or neutral.
- It is often performed to help businesses monitor brand and product sentiment in customner feedback, and understand customer needs.
- https://monkeylearn.com/sentiment-analysis/

Games



MIT researchers just got a computer to accomplish yet another task that most humans are incapable of doing: It learned how to play a game by reading the instruction manual.

The MIT Computer Science and Artificial Intelligence lab has a computer that now plays Civilization all by itself — and it wins nearly 80% of the time. Those are better stats than most of us could brag about, but the real win here is the fact that instruction manuals don't explain how to win a game, just how to *play* it.

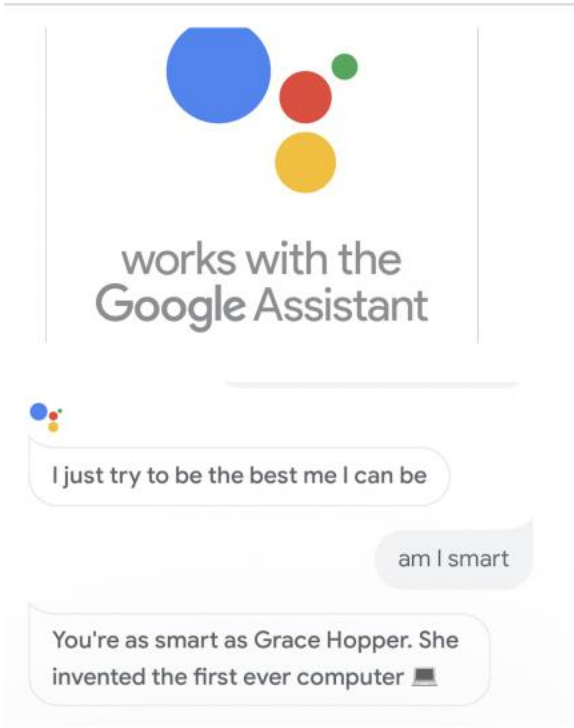
Conversational agents

works with the Google Assistant

alexa

Conversational agents

- A conversational agent contains
  1. Speech recognition
  2. Language analysis
  3. Dialog processing
  4. Text to speech

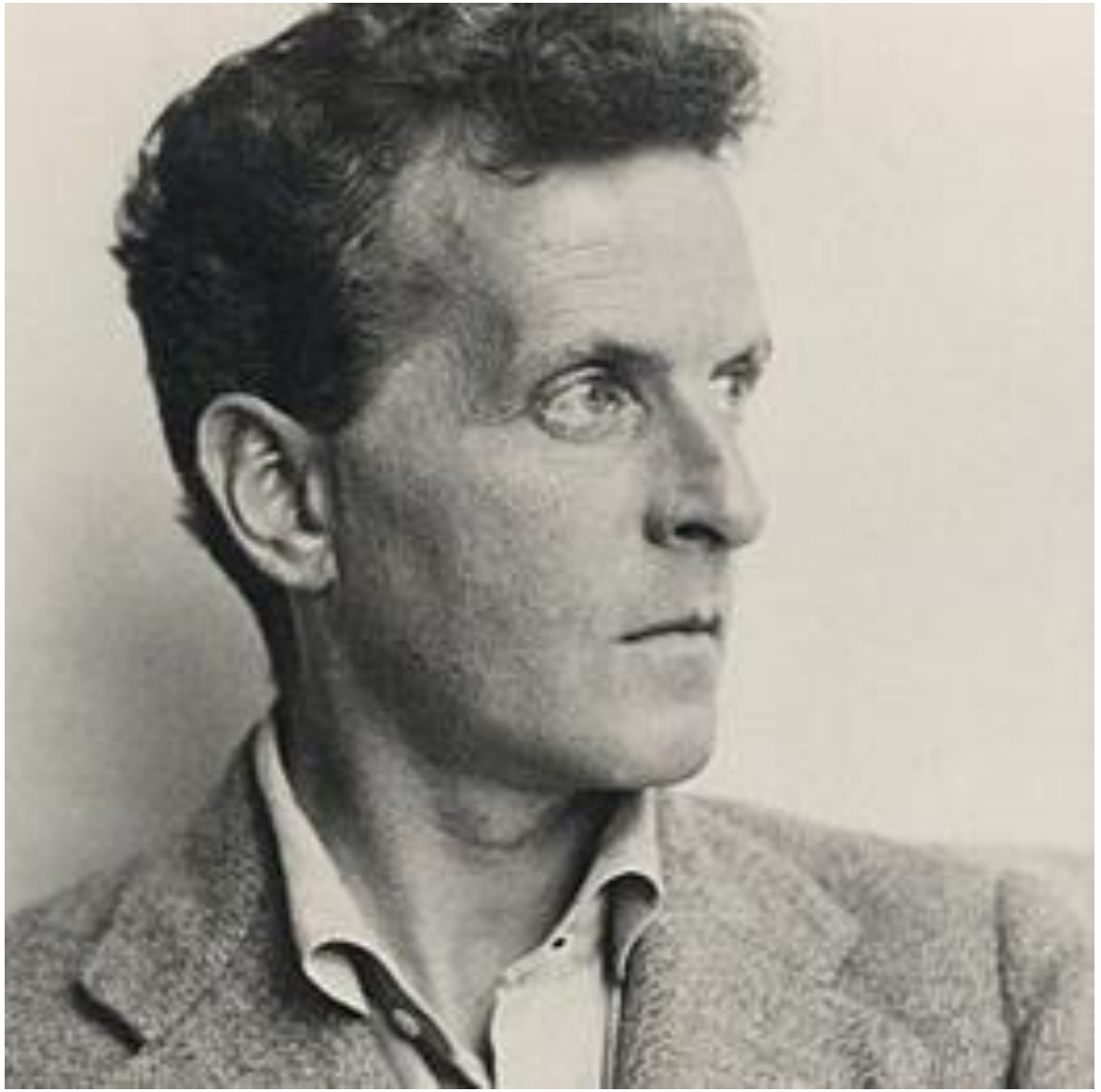


Why NLP is hard ?

Why NLP is hard?

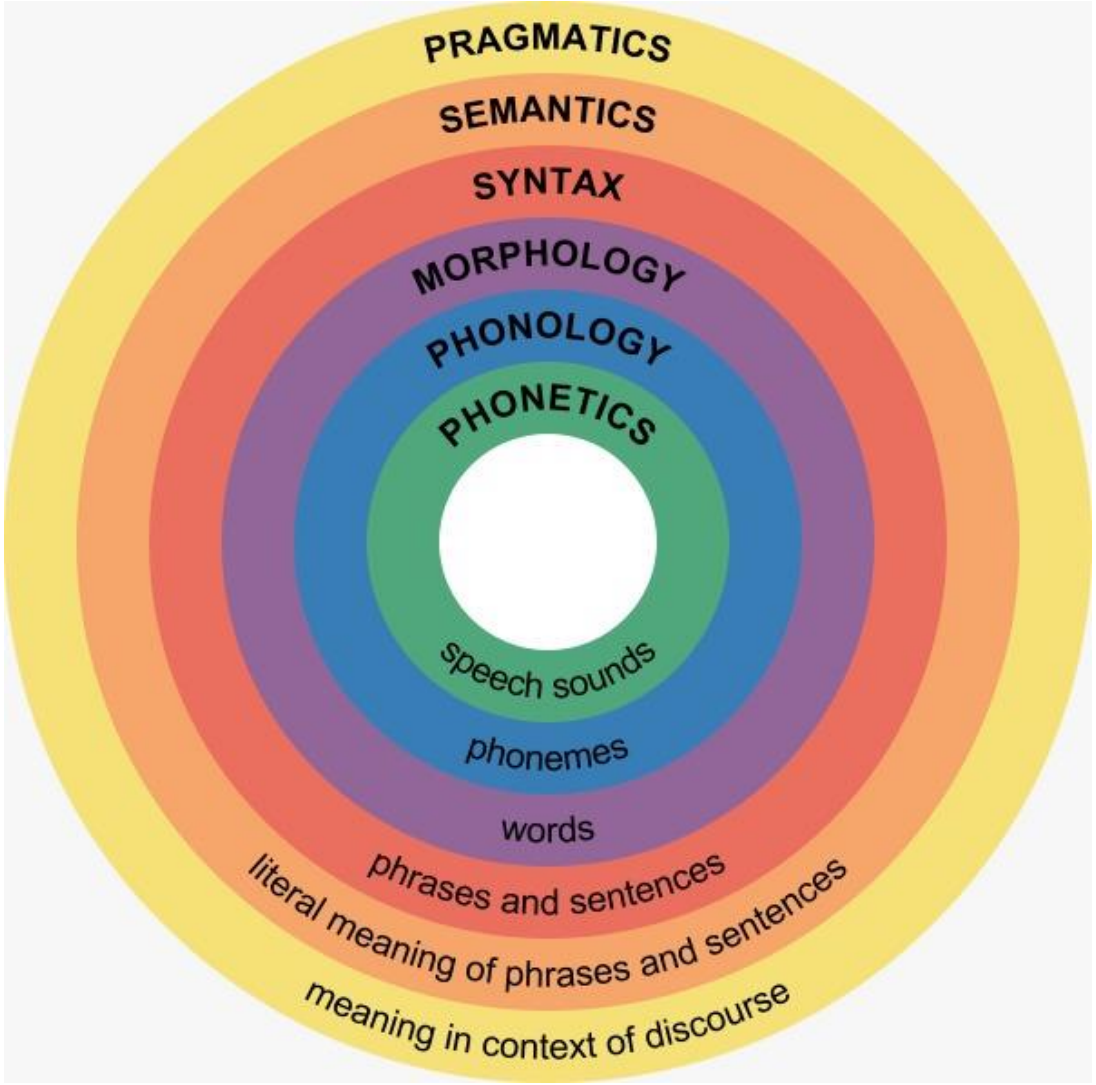
“The limits of my language are the limits of my world”

Ludwig Wittgenstein



Why NLP is hard?

- Language consists of **many levels of linguistic knowledge.**
- Humans fluently integrate all of these to produce and understand language
- Ideally, so would a computer!



Why is NLP hard?

- Ambiguity
- Scale
- Variation
- Expressivity
- Unmodeled Variables
- Unknown representation

Ambiguity

- Ambiguity at multiple levels:
  - ***Lexical Ambiguity***: The fisherman go the **bank** (finance or river?)
  - ***Syntactic ambiguity*: I can see a man with a telescop**
  - ***Semantic ambiguity*: I gave a present to the children or The chicken is ready to eat**
  - ***Referential ambiguity*: John met David after he finished work. (he = John or David ?)**

Ambiguity – Semantics

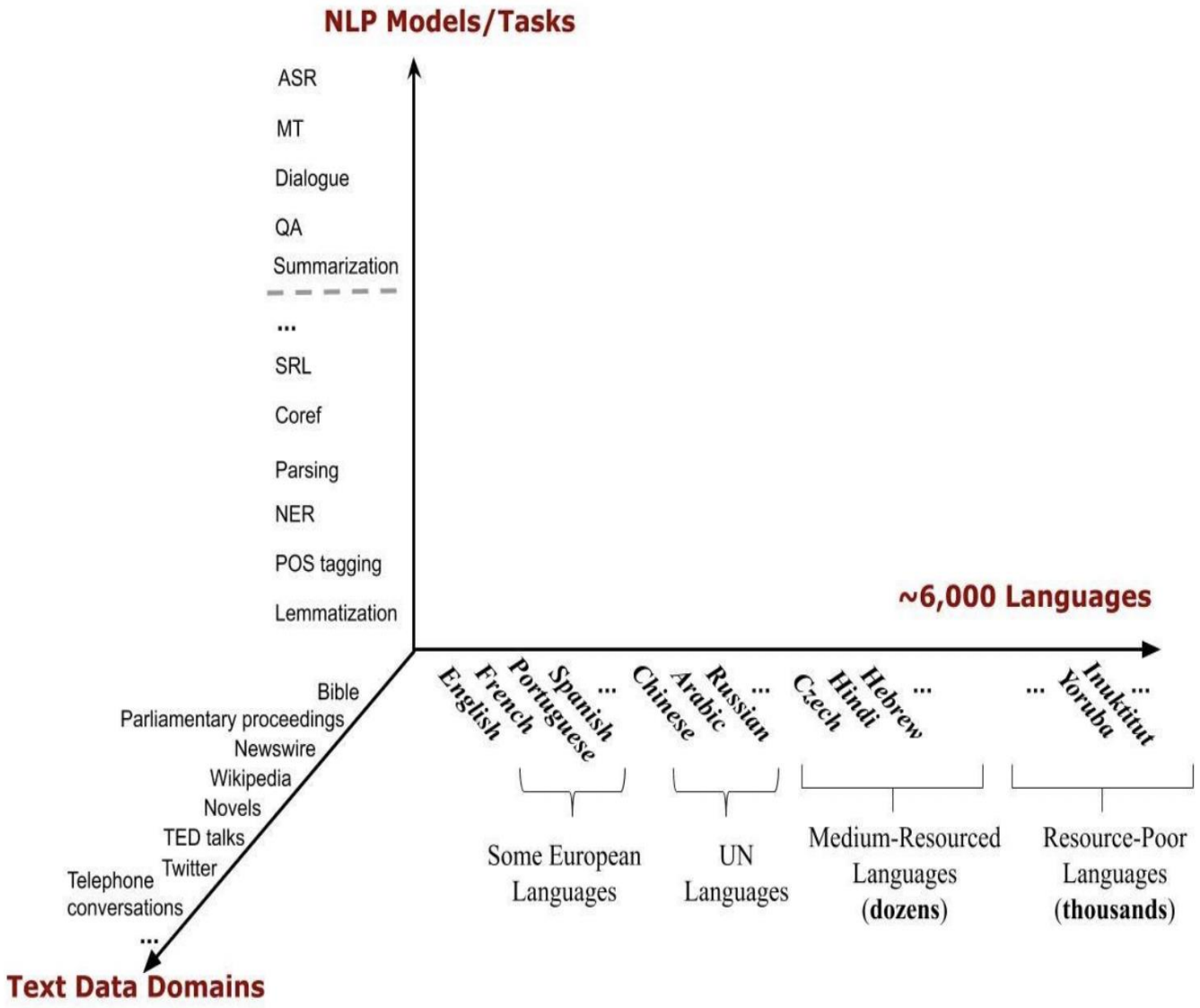
- Every fifteen minutes a woman in this country gives birth.

- Our job is to find this woman, and stop her!:

– Groucho Marx

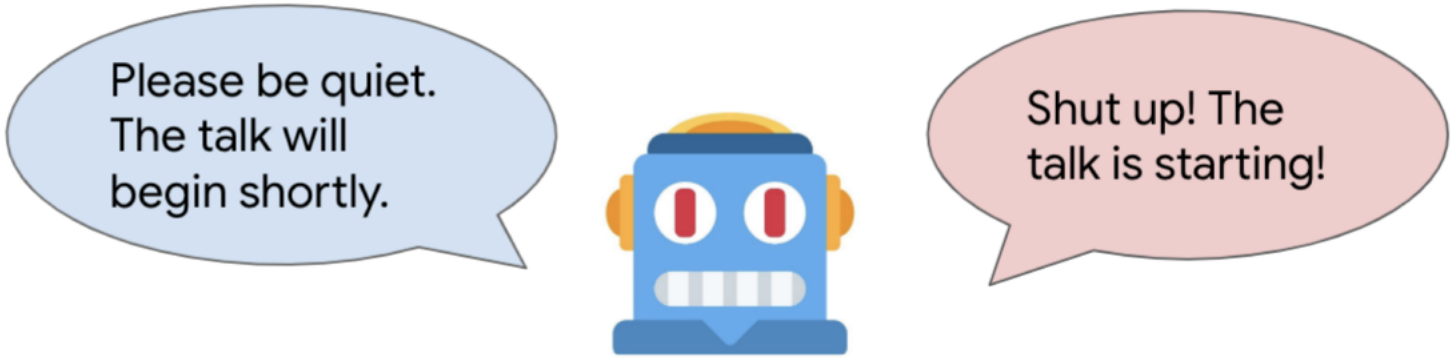
Scale and Variation

- **~7K languages**
- **Thousands of language varieties**
- **Variation of domains (news, biomedical, historical, …)**



Expressivity

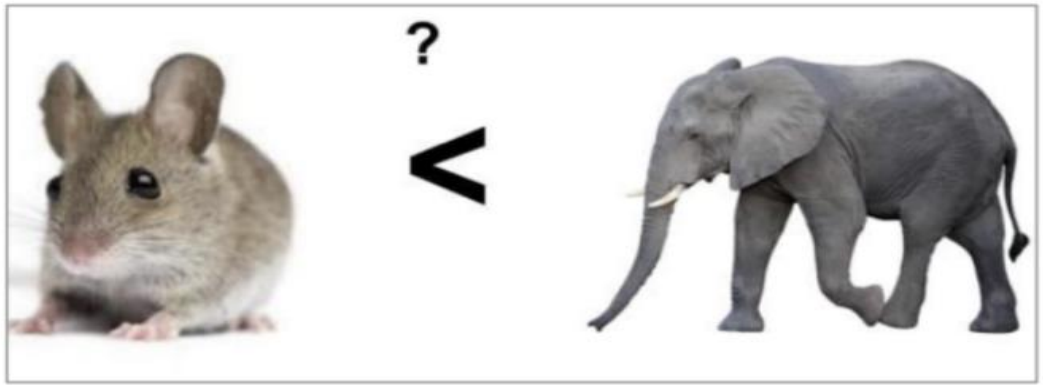
- Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:
  - **She gave the book to Aria** vs. **She gave Aria the book**
  - **Is that door still open?** vs. **Please close the door**



Unmodeled Variables

- Language often depends on **world knowledge** and variables that aren't explicitly present in the text.
- World knowledge
  - I dropped the glass on the floor and it broke
  - I dropped the hammer on the glass and it broke

"Drink this milk"



Unknown Representation

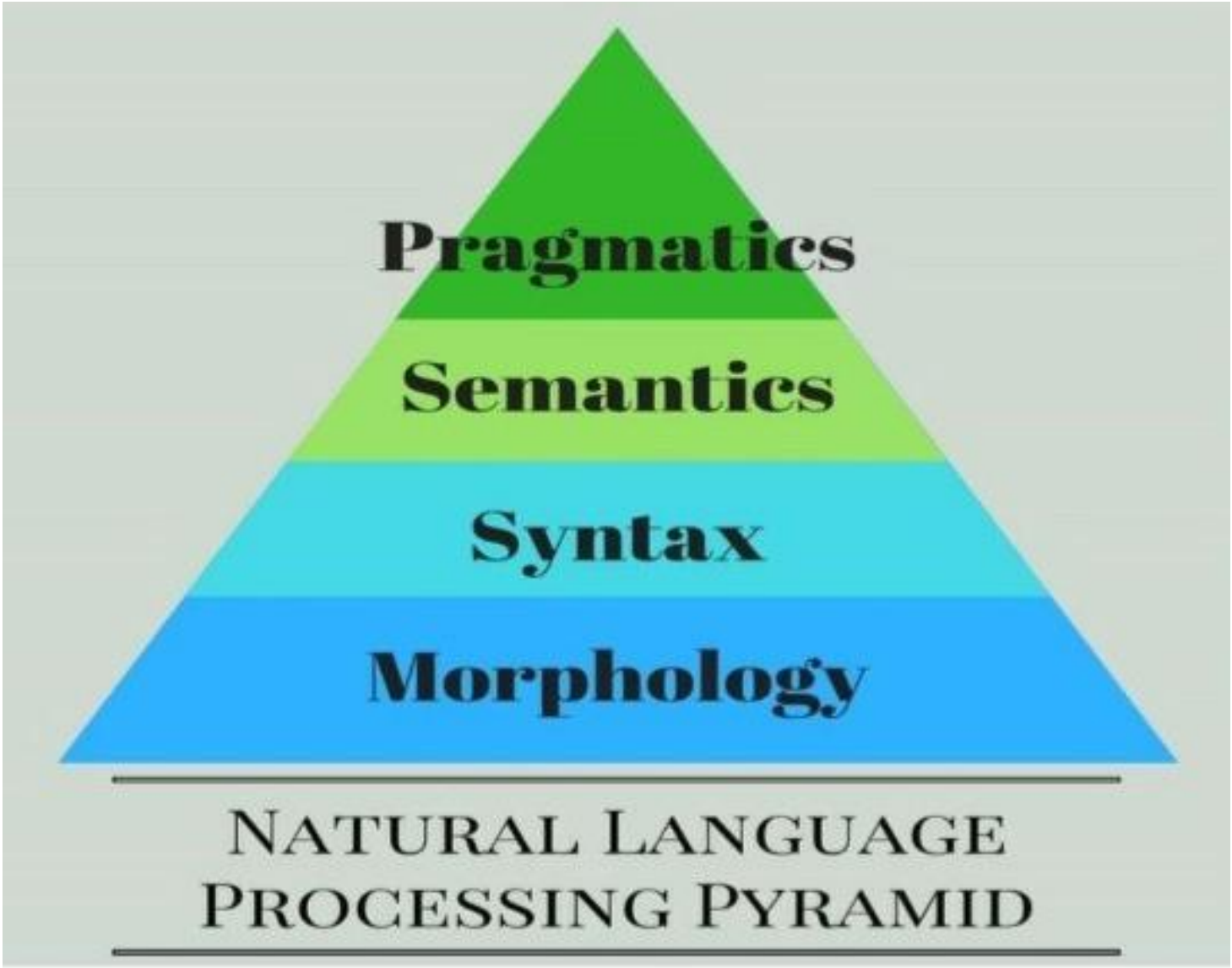
- Very difficult to capture what is the representation of the text or speech, since we don't even know how to represent the knowledge a human has/needs:
  - What is the "*meaning*" of a word or sentence?
  - How to model context?
  - Other general knowledge?

The Challenges of "Words"

- Segmenting text into words
- Morphological variation
- Words with multiple meanings: bank, mean
- Domain-specific meanings: latex
- Multiword expressions: make a decision, take out, make up

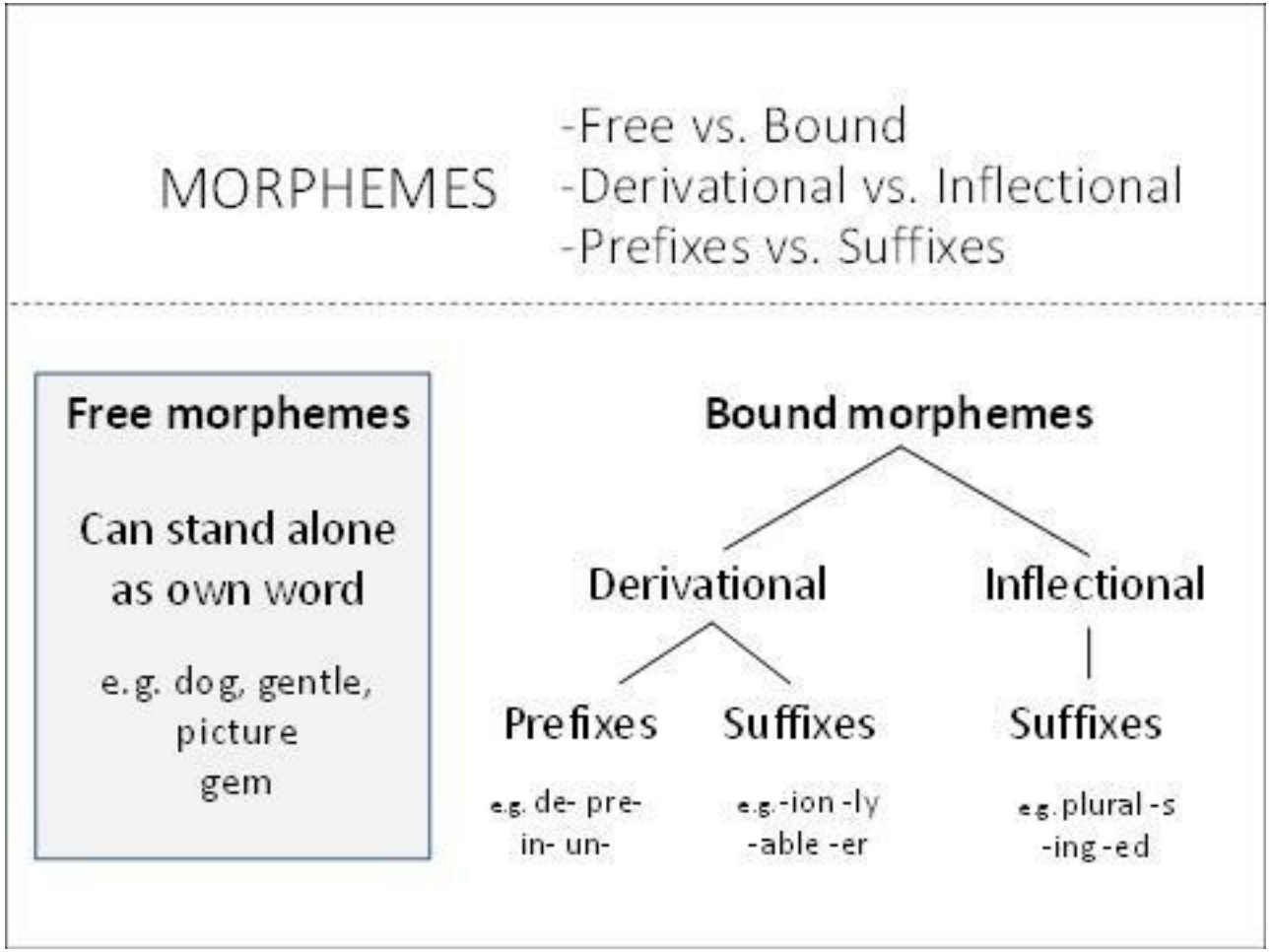
How to deal with Ambiguity?

1. Morphological Analysis
2. Syntactic Analysis
3. Semantic Analysis
4. Pragmatics Analysis



Morphological Analysis

- **Morphological Analysis** is the field of linguistics that studies the structure of words.
- It identifies how a word is produced through the use of morphemes.
- **Morpheme**: is the smallest element of a word that has a grammatical function and meaning.



Morphological Analysis

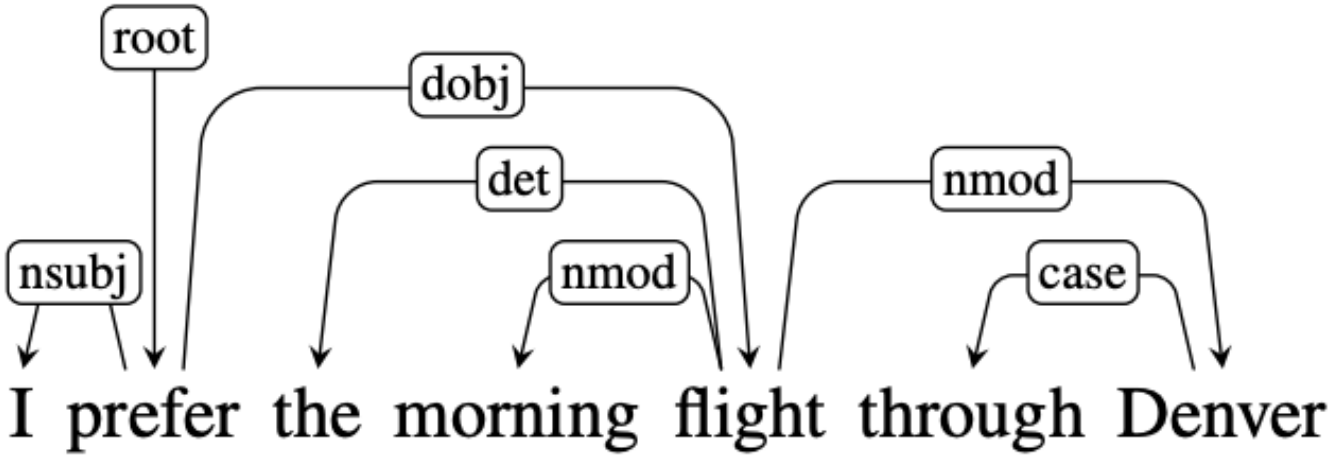
- **Tokenization** consists of splitting a sentence into an ordered list of words, usually referred to as tokens.

- **Stemming** is the process of removing affixes (prefix, suffix, infix) from a word in order to obtain a word **stem (root)**.

- **Lemmatization** returns the canonical form, dictionnary form of a word. The output we will get after lemmatization is called ‘lemma’.

Syntactic Analysis

- **Syntactic Analysis** studies the syntactic relationships between words in a sentence based on formal grammar rules;



Dependency parsing [Source]

Syntactic Analysis

- **Part-of-speech (POS) tagging** assigns to each token its corresponding part-of-speech tag i.e. its syntactic word category (verb, adverb, noun, …)

Why	not	tell	someone	?
adverb	adverb	verb	noun	punctuation mark, sentence closer

Semantic Analysis

- **Semantic Analysis** aims to understand the meaning and interpretation of words and sentence structure.

  - **Named entity recognition:** seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages**,**

  - **Word sense disambiguation** is defined as the ability to determine which meaning of word is activated by the use of word in a particular context.

  - **Semantic role labelling** is the process that assigns labels to words or phrases in a sentence that indicates their semantic role in the sentence, such as that of an agent, goal, or result.

Pragramatic Analysis

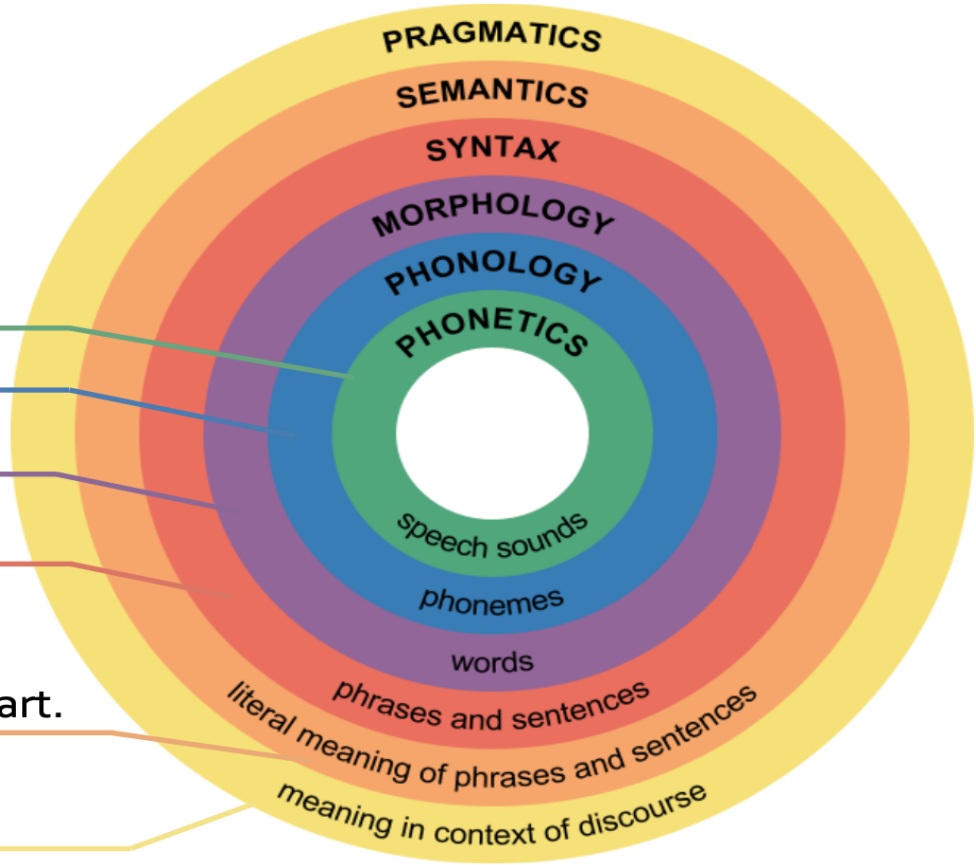
- **Pragramtic Analysis** aims to study the language and the context of use; that is, interpreting the sentence according to the general world knowledge and the communication situation.

***Alice**: Will you come to the ball tonight?*

***Caroline**: I heard that **Bernard** will be there!*

Large Language Models (LLMs)

- Level of linguistic knowledge



Not applicable

Not applicable

Great. LLMs are robust to typos, coinage, cacography

Great. Dependency parsing and coreference resolution are almost solved. Syntactic information is captured in attention (Clark et al., 2019).

Great. Most of current NLP benchmarks focus on this part.

It depends! LLMs still get confused when they meet unique contexts or special users (e.g., those in underrepresented groups)

Large Language Models (LLMs)

- **Addressed Key Challenges:**
  - Solved many issues for NLP, particularly in high-resource languages.

- **Ongoing Struggles:**
  - Still face difficulties with low-resource languages.

- **Further work required for specific domains and low resource languages:**
  - Cybersecurity, Education, Biomedical, Historical, …

Natural Language Processing Tasks

Natural Language Understanding Tasks:
sentiment, classification, NLI, semantic, social knowledge, etc

Natural Language Generation Tasks:
summarization, dialogue, translation, QA, style transfer, writing, etc

Reasoning:
Logical reasoning, commonsense reasoning, etc

Multilingual:
English + low-resource/non-Latin languages

Factuality/Faithfulness

Natural Language Processing

- **Core Technologies**
  - Text represention
  - Part-of-speech tagging
  - Language modeling
  - Syntactic Parsing
  - Named entity recognition
  - Word sense disambiguiation
  - Sematinc role labeling
  - ...
- **Applications**
  - Chatbots
  - Machine Translation
  - Question Answering
  - Dialogue Systems
  - Information Extraction
  - Summarization
  - Sentiment Analysis
  - …

Steps to solve an NLP problem

1. Collect Data
2. Text Preprocessing
   - Tokenization
   - Stemming
   - POS …
3. Text Representation
4. Machine Learning Model Selectin and Training
5. Evaluation of the Selected Model
6. Deploy the Model

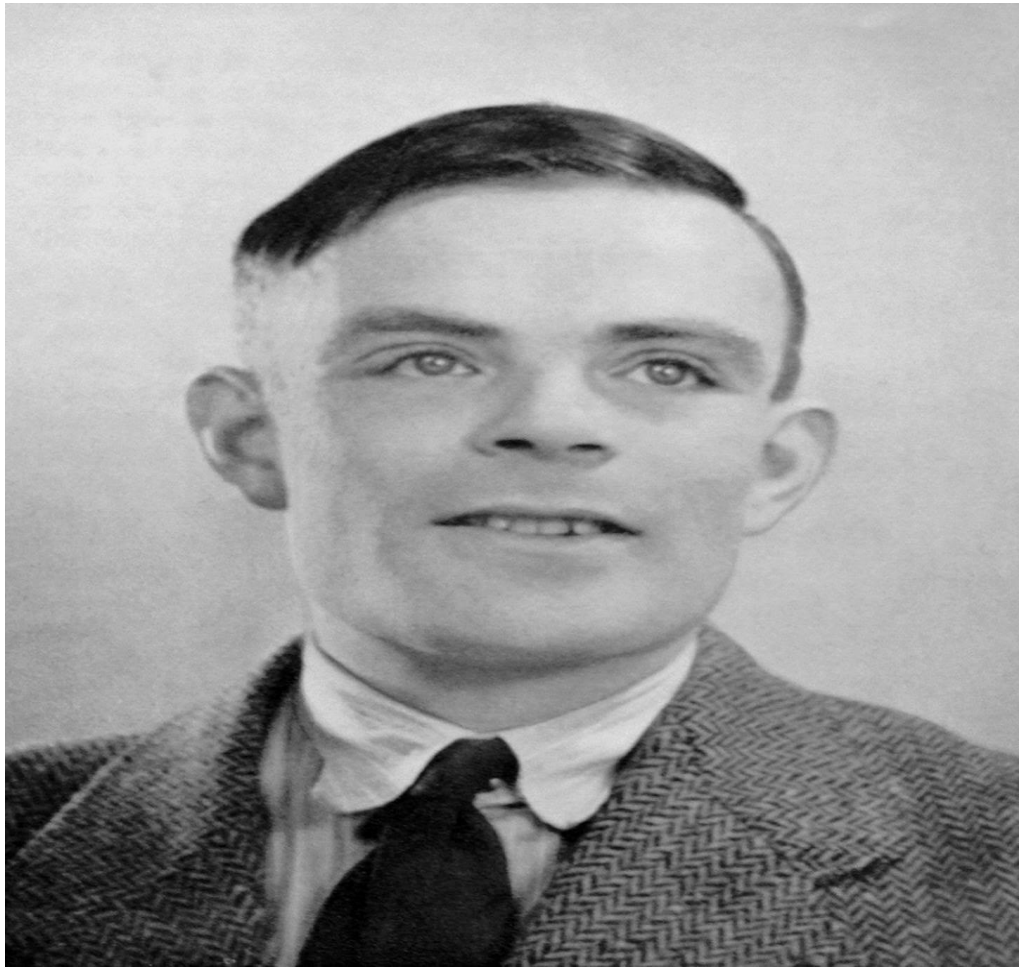
Field with Connections to NLP

- Machine learning
- Linguistics
- Cognitive science
- Information theory
- Logic, Data Science
- Political science
- Psychology, Education
- Economics
- Education
- …

Turing Test

Distinguishing human vs computer only through written language

A method to empirically determine whether a computer has achieved intelligence



`Alan Turing

https://www.youtube.com/watch?v=3wLqsRLvV-c&ab_channel=CambridgeUniversity

Reading

- Please refer to this document : Chapter 1

  - Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed. draft)
  - Jacob Eisenstein. Natural Language Processing
  - Yoav Goldberg. A Primer on Neural Network Models for Natural Language Processing
  - Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning

Relevant Scientific Conferences

- Association for Computational Linguistics (ACL)
- North American Association for Computational Linguistics (NAACL)
- International Conference on Computational Linguistics (COLING)
- Empirical Methods in Natural Language Processing (EMNLP)
- Conference on Computational Natural Language Learning (CoNLL)

Next Class

- Text Preprocessing