

Generated by DALL-E



VOYAGE, VOYAGE

Luísa Coheur

OVERVIEW

- Learning objectives
- Topics
 - Concepts
 - Main challenges
 - Historical overview
- Key takeaways

LEARNING OBJECTIVES

LEARNING OBJECTIVES

- After this class, students should be able to:
 - Define concepts such as Natural Language, paraphrases, ellipsis, co-reference, etc.
 - Explain some of the main challenges in NLP
 - Explain the concepts related with the linguistic approach
 - Explain how NLP move forwards over the last 80 years
 - Explain some NLP tasks

TOPICS

OVERVIEW

- Learning objectives
- Topics
 - Concepts
 - Main challenges
 - Historical overview
- Key takeaways

NATURAL LANGUAGE

- Language
 - Grammatical system
 - Own rules
 - Used to communicate
- Natural
 - Natural evolution due to people communication
 - New words everyday!
 - Yet, some control in adding them to dictionaries...

ACTIVE LEARNING MOMENT: THINK-PAIR-SHARE



EXERCISE

Can you identify some new words in your native language (or in English)?

Can you identify words that were removed from our vocabulary?

EXAMPLES OF NATURAL LANGUAGES

- Portuguese, English

- Portuguese Sign Language

- ...

- What about Esperanto?

NATURAL LANGUAGES

- Esperanto is a language designed Dr Ludwig Lazarus Zamenhof (1887) to be easy to learn and promote communication
 - Example (morphology):
 - All nouns end in “o”
 - noun -> corresponding adjective: replace the “o” with an “a”
 - To make a plural, add a “j” to the end

EXERCISE



The Esperanto alphabet

a b c ĉ d e f g ĝ h ĥ i j ĵ k l m n o p r s ĝ t u ī v z

ESPERANTO

Saluton
Ĝis
Dankon
Nomo
Amiko
Hundo
Kato

ESPERANTO

Jaro
Patro
Patrino
Viro
Virino
Doktoro
Tago

ESPERANTO

Tempo
Trinku
Mangî
Dormu
Hejmo
Banejo
Restoracio

Vi estas welcome You're welcome De Nada

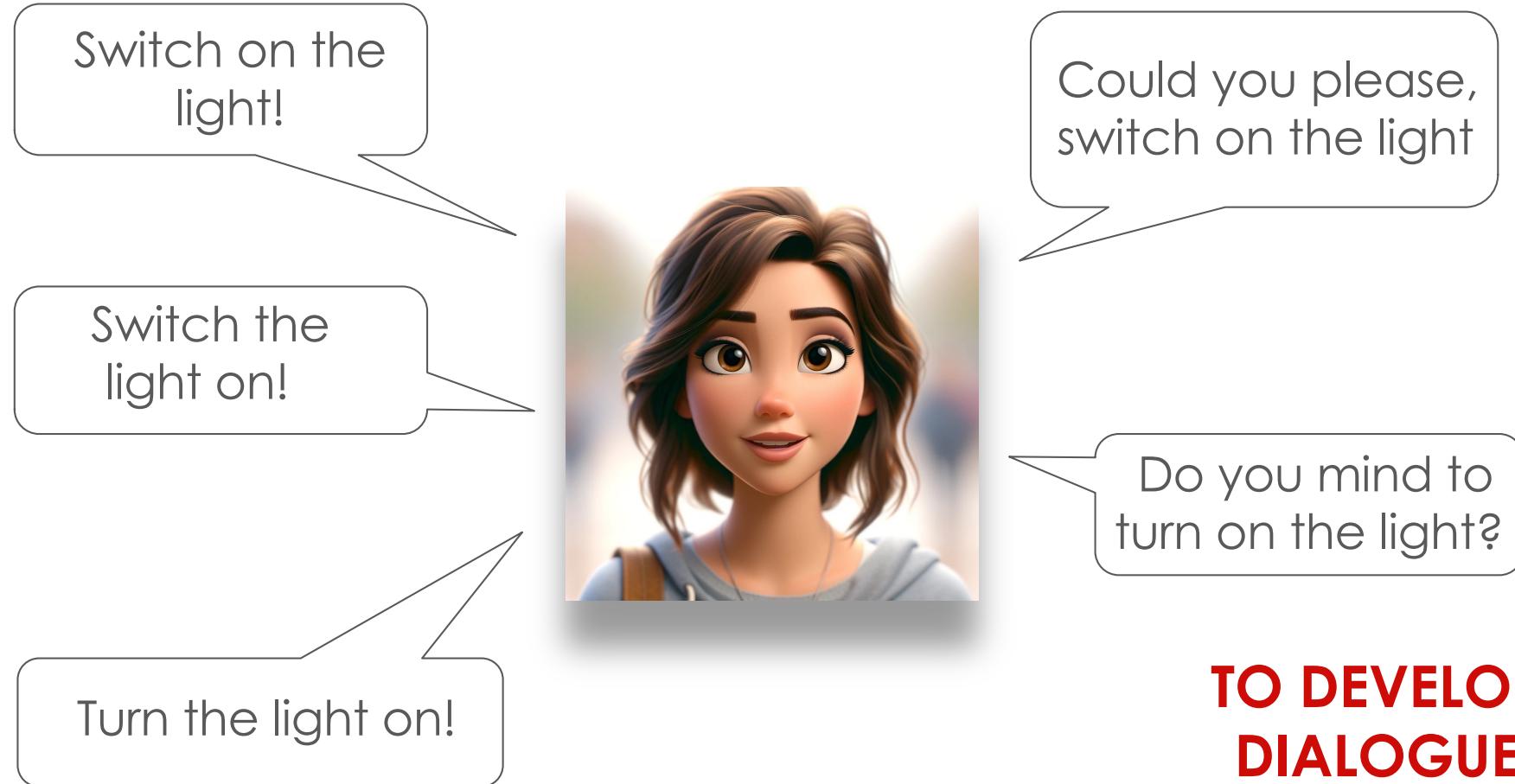
NATURAL LANGUAGE PROCESSING

- Natural Language Processing (NLP)
 - Field of Artificial Intelligence and Linguistics concerned with the interactions between computers and human (natural) languages

OVERVIEW

- Learning objectives
- Topics
 - Concepts
 - Main challenges
 - Historical overview
- Key takeaways

MAIN CHALLENGES



LANGUAGE
VARIABILITY

**TO DEVELOP
DIALOGUE
SYSTEMS is a
main
NLP TASK!**

MAIN CHALLENGES

It is so dark...



“Open” the light!

Lumos!

LANGUAGE
VARIABILITY

BY THE WAY

- Concept: two sentences with the same meaning are called **paraphrases**

MAIN CHALLENGES

Can you
turn on the
light,
please?



Yes, I can!

AMBIGUITY

MAIN CHALLENGES

Can you
turn on the
light,
please?



AMBIGUITY/VAGUENESS

BY THE WAY

- As we have seen, a single sentence can have different meanings
- The source of those different meanings can vary
 - There are many different types of ambiguity

BY THE WAY

- Some words have multiple meanings (concept: lexical ambiguity)



And if your dog's name is Step?

BY THE WAY

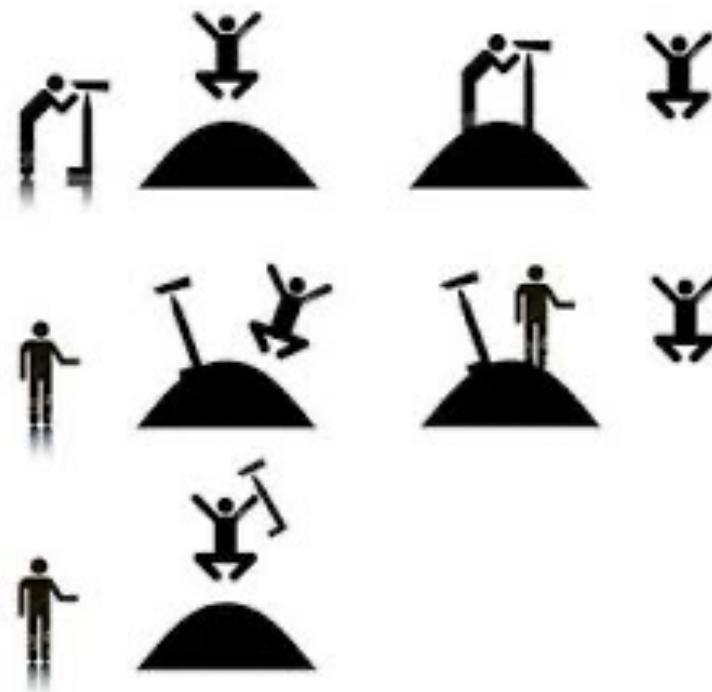
- My favorite ambiguity moment



I know a little German

BY THE WAY

- A sentence has multiple parse trees (concept: syntactic ambiguity)
 - I saw the man on the hill with a telescope



BY THE WAY

- All ambiguities “are” semantic, but some are not related with lexical and syntactic ambiguity
 - John and Mary are married.
 - One with the other? With other people?
 - John kissed his wife and Peter too
 - Well...
- And there are many other types...

**Word sense
disambiguation
is an
NLP TASK!**

MAIN CHALLENGES

Can you
turn on the
light,
please?



Now the radio!

ELLIPSIS

Now turn
it off!

CO-REFERENCE

Co-reference
resolution is an
NLP TASK!

MAIN CHALLENGES

Can you
turn on the
light,
please?



Now the radio!

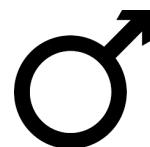
ELLIPSIS

Agora
desliga-a!

CO-REFERENCE

MAIN CHALLENGES

Can you
turn on the
light,
please?



Now the radio!

ELLIPSIS

Agora
desliga-o!

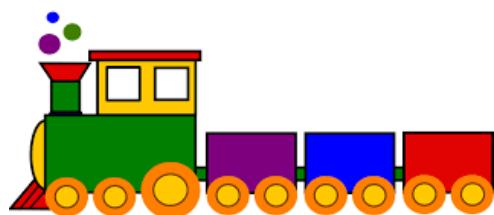
CO-REFERENCE

BY THE WAY

- Concepts:
 - Ellipsis: the omission of a word or words that are able to be understood from contextual clues
 - Co-reference: elements of a text that have the same reference

MAIN CHALLENGES

Can you
turn on the
light,
please?



**There is a course
focused on
Speech in P4**



NOISE

MAIN CHALLENGES

Can you
turn on the
light,
please?



DIFFERENT
ACCENTS

ENGLISH

From <https://www.worldaccent.com/>

MAIN CHALLENGES

DIFFERENT AGES

Can you turn on
the light, please?



Nha nha nhã?



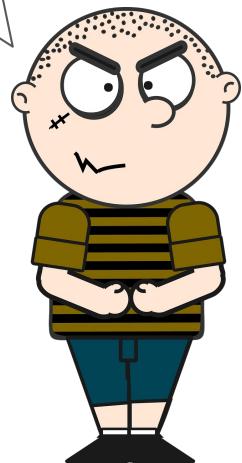
MAIN CHALLENGES

At 4pm I will be waiting
for you outside school!



MAIN CHALLENGES

At 4pm I will be waiting
for you outside school!



CONTEXT

We had a
project about
Cyberbullying
detection!

MAIN CHALLENGES

“Those who believe in telekinetics, raise my hand.” – Kurt Vonnegut

SARCASM AND IRONY

Generated by DALL-E



OVERVIEW

- Learning objectives
- Topics
 - Concepts
 - Main challenges
 - [Historical overview](#)
- Key takeaways

40s

- Machine Translation
 - Warren Weaver (1949) suggested using cryptographic techniques and statistical methods to translate languages



We will have a class dedicated to Machine Translation, a main NLP task!

RULE-BASED APPLICATIONS

60s



- Daniel G. Bobrow developed Student (PhD. Thesis, 1964)
 - Student solves algebra problems (implemented in LISP)
 - Example:
 - User: “The price of a radio is 69.7 dollars. If the price is 15 percent less than the marked price, find the marked price.”
 - STUDENT: The marked price is 82 dollars

RULE-BASED APPLICATION

60s

- Joseph Weizenbaum developed Eliza (1966)

```
Welcome to
      EEEEEE  LL      IIII  ZZZZZZ  AAAAAA
      EE      LL      II      ZZ  AA  AA
      EEEEEE  LL      II      ZZZ  AAAAAAAA
      EE      LL      II      ZZ  AA  AA
      EEEEEE  LLLLLL  IIII  ZZZZZZ  AA  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU: Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU: They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU: Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU: He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU: It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU: ■
```



ALSO: RULE-BASED APPLICATION

60s

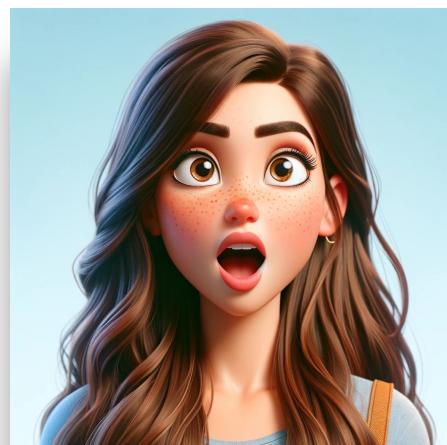
- How did ELIZA work?

USER: I think my cat has powers.

USER: I think my <anything>

ELIZA: Why do you think your <anything>

ELIZA: Why do you think your cat has powers.



80's

RULE-BASED

CHIT-CHAT AND “PERSONA” BOTS

Talk with Captain Kirk



QUESTION/ANSWERING AND
TASK ORIENTED DIALOGUE
SYSTEMS



Who was Fernando
Pessoa?

- I need a ticket to Braga
- ...

Question/Answer and the development of
Dialogue System are NLP TASKs!

90's

- Release of linguistic and computational resources



**We will have a
class dedicated
to datasets
(corpora) and
we will use
several
computational
NLP resources
along this course**

90's

- International evaluation fora/campaigns/shared tasks



Image from <https://www.publicdomainpictures.net/>

These campaigns put some order in the evaluation process. They also release new datasets and evaluation metrics.

Besides the class dedicated to datasets, we will have a class dedicated to (evaluation) metrics

90's, 00's

- Advances in Machine Learning



Emergence of the feature-engineering approach!
Ex: SVM (1995)

We will dedicate some time to the (adorable) feature engineering approach!

Since 00's

- More and more computational resources become available
 - Some are more recent, but NLTK, for instance, started to be developed in 2001



NLTK

...



We will use NLTK, spaCy
and Stanza in the labs!

2011

- WATSON (IBM) wins Jeopardy!

Photo: IBM



“AI wow” (in a long time)

SINCE 2011

- Virtual Assistants started to emerge



Hi, I'm Cortana

2012

- Meet Edgar Smith at the Monserrate Palace (by INESC-ID). It wasn't a global event, but it was really cool!



SINCE 2013

- Deep Learning
 - neural word embeddings
 - Transfer Learning
 - Multi-task Learning
 - ...

We will dedicate many classes to this!



2016

- Chatbots
(again)



Zuckerberg's Facebook Messenger launches 'chat bots' platform

Jessica Guynn | USA TODAY
Published 5:02 PM EDT Apr 12, 2016

SAN FRANCISCO — You already chat with friends on Facebook Messenger. Now Messenger wants you to chat with businesses and get updates from them, too.



Image from <https://eu.usatoday.com/story/tech/news/2016/04/12/facebook-messenger-f8-chat-bots/82919056/>

SINCE 2016 (MAYBE BEFORE)

- Several tools to build chatbots become available
 - Although they called themselves “natural language understanding” frameworks, they “only” take a sentence as input and:
 - Find the intention behind it (**classification**)
 - **Extract named entities**
 - Do something with this information.
 - Some have a module to define the **dialogue flow** (usually handcrafted)



Amazon Lex



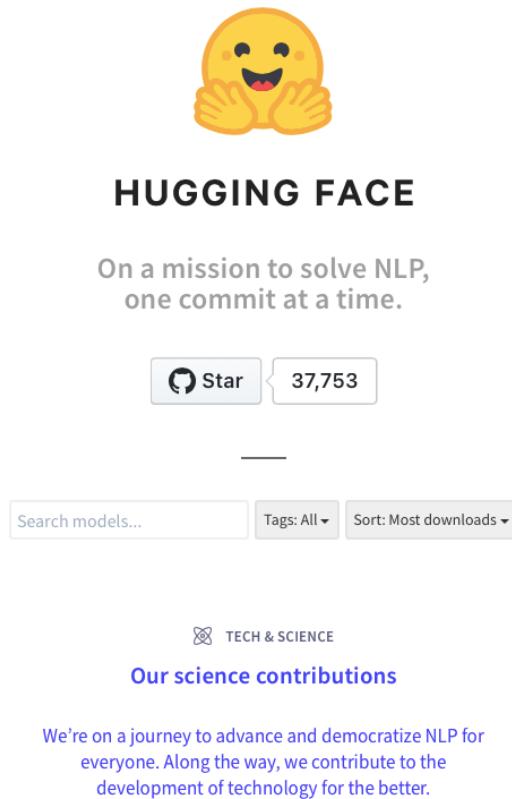
Dialogflow



We will study several classification tasks in NLP and also Named Entity Recognition, which is an NLP task

2016

- More computational stuff (mainly deep learning models)
 - Hugging Face is an NLP-focused startup with a large open-source community



SINCE 2016

- Shared Tasks/Prizes
 - [...] new benchmark [...] requiring a single system to perform ten disparate natural language tasks [...]



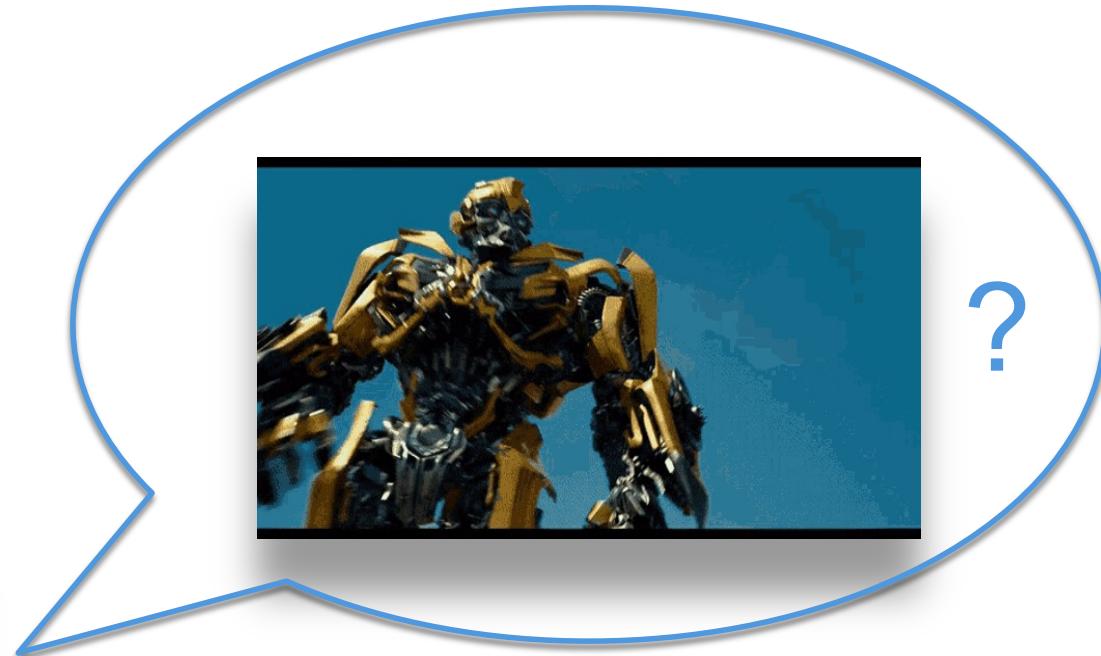
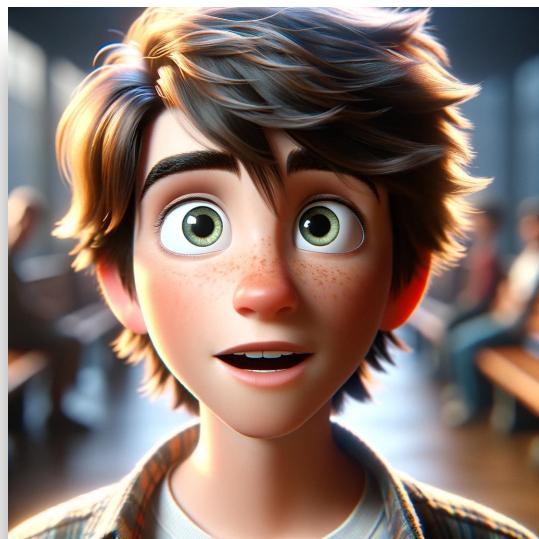
- \$500,000 prize will be awarded to the team that creates the best socialbot. The second- and third-place [...] \$100,000 and \$50,000, respectively.



A Portuguese team from NOVA got the first places twice in the Alexa competitions!

2017

- Transformers



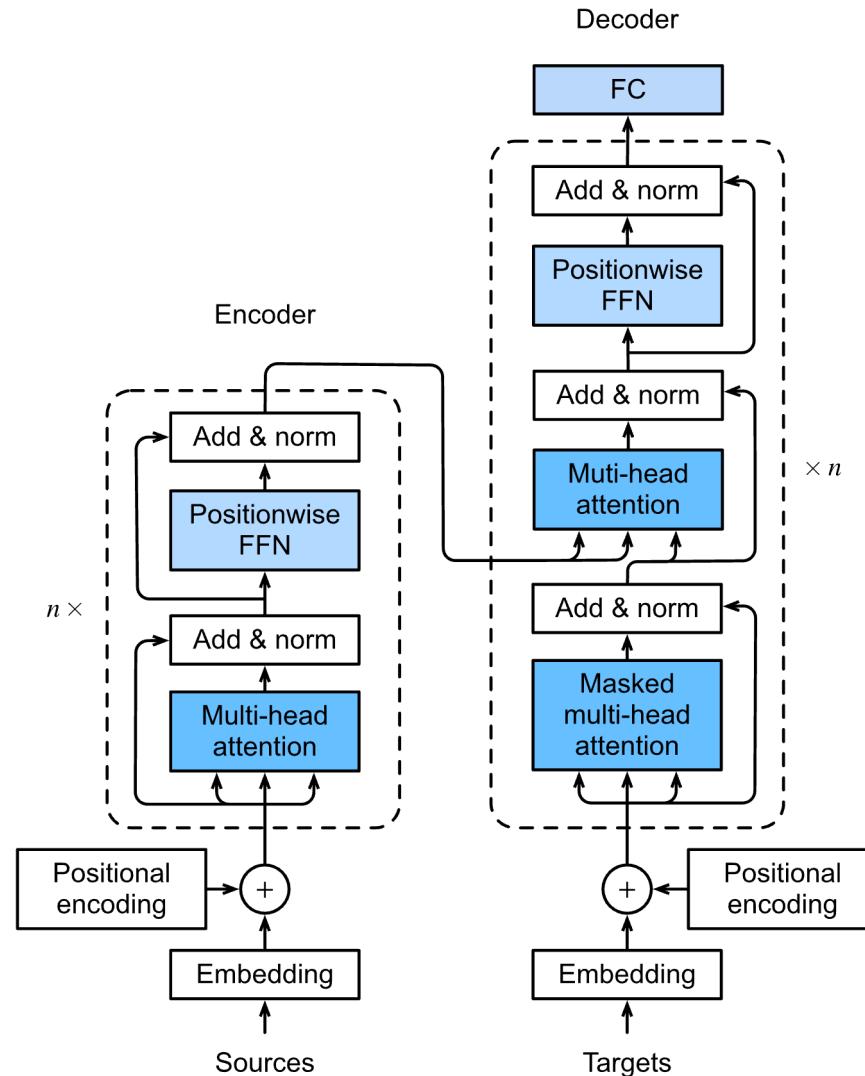
?

NOPE!
Not those transformers...

2017

- Transformers are a successful deep learning architecture
- Transformers are behind most of the nowadays Large Language Models (LLMs)

We will study this architecture



IN THE MEANWHILE



CRAZY CONFERENCE NUMBERS

CRAZY CONFERENCE NUMBERS

LONG PAPERS ACL

Year	Submitted	Accepted	Rate
1997	264	83	24%
1998 (w/COLING)	550	137	25%
1999	320	80	25%
2000	267	70	26.2%
2001	260	69	27%
2002	256	66	26%
2003	360	71	20%
2004	348	88	25%
2005	423	77	18%
2006 (w/COLING)	630	147	23%
2007	588	131	22.3%
2008	470	119	25%
2009	569	121	21%
2010	638	160	25%
2011	634	164	26%
2012	571	111	19%
2013	664	174	26%
2014	572	146	26.2%
2015	692	173	25.0%
2016	825	231	28.0%
2017	751	195	25%
2018	1045	256	24.5%
2019	1740	447	25.7%
2020	2244	779	22.7%

2021	3350	710 (long and short) for main conference + 457 in "Findings of ACL"	21.2% main / 37.5% findings
2022	3378	701 (long and short) for main conference + 361 in "Findings of ACL"	20.8% main / 31.4% findings

2024	4835	940 + 975 (findings)
2025	> 8300	1700 + 1392 (findings)

CRAZY COMPUTATIONAL COSTS

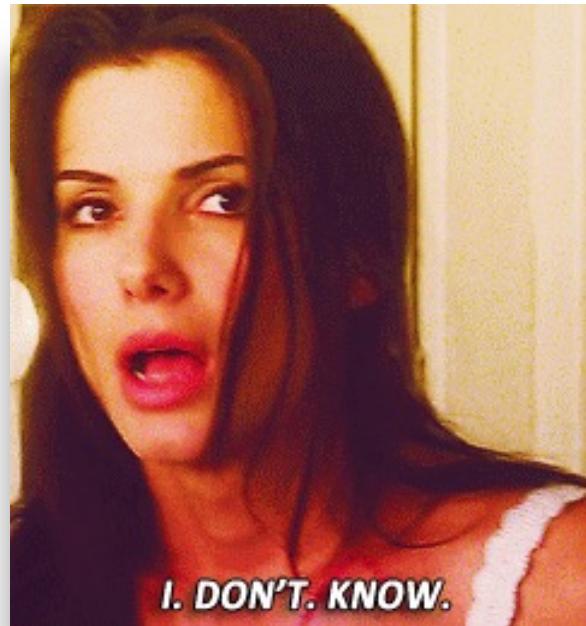
Google
Facebook
OpenAI
...



Rest of the world

NEW MODELS POPPING EVERYDAY

- In an online Transformers class:
chat: what about **decepticons?**



From wifflegif

2019



OpenAI

- GPT-2 (Open-AI)
 - Um modelo que usa transformers
 - O modelo não foi logo libertado por razões éticas

Legolas and Gimli advanced on the orcs, raising their weapons with a harrowing war cry.

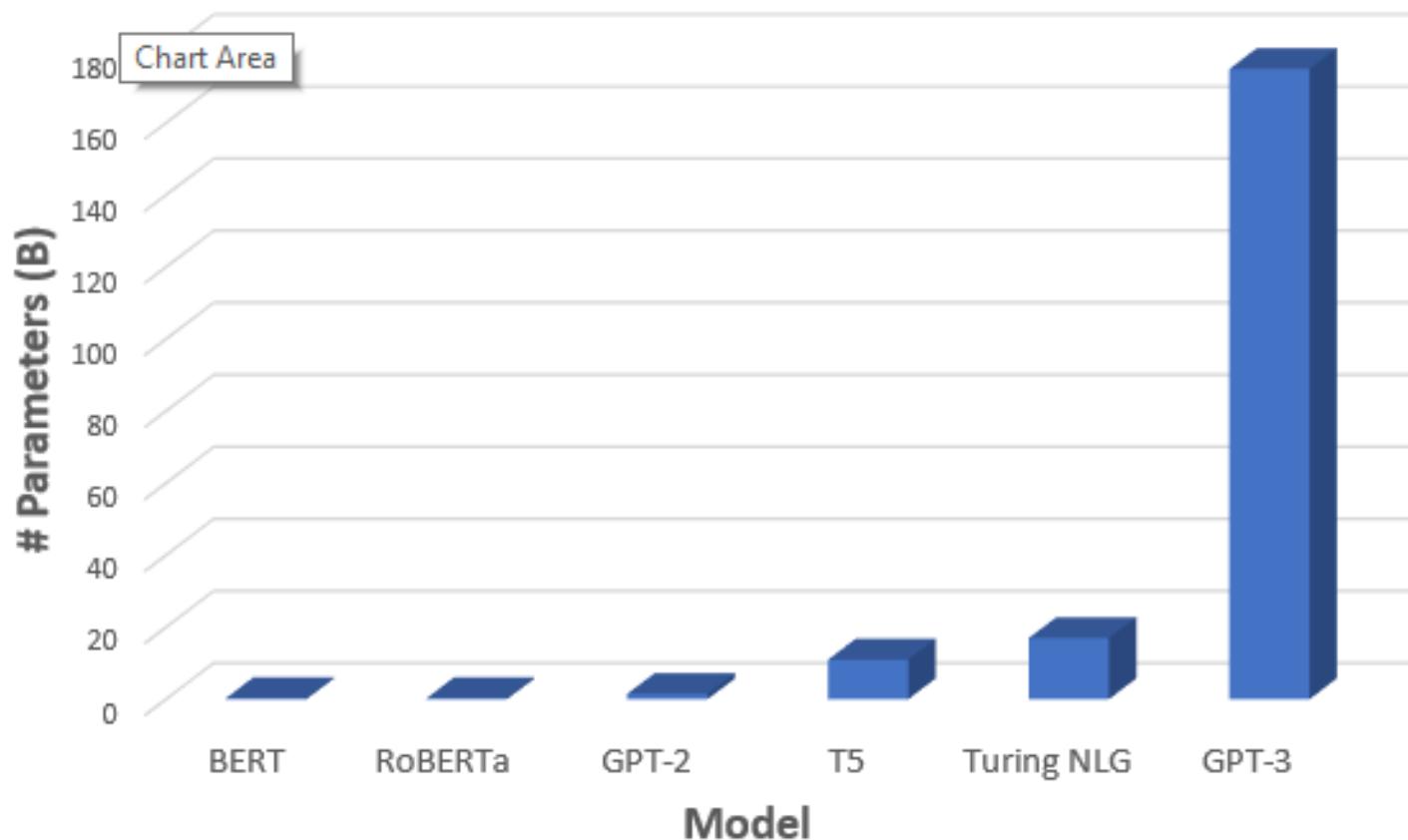
EXAMPLE

2020



OpenAI

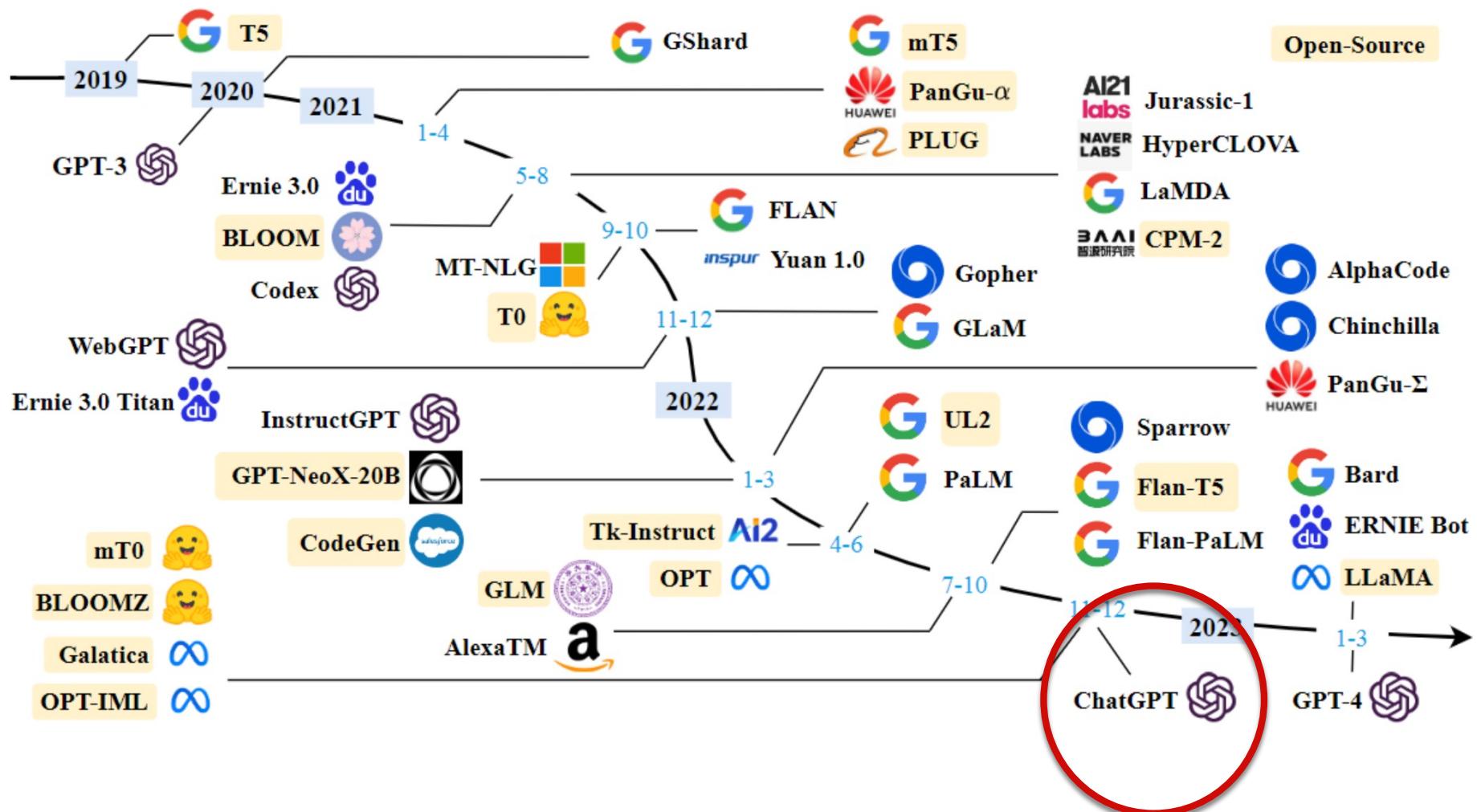
- GPT3
 - 175 mil milhões de parâmetros



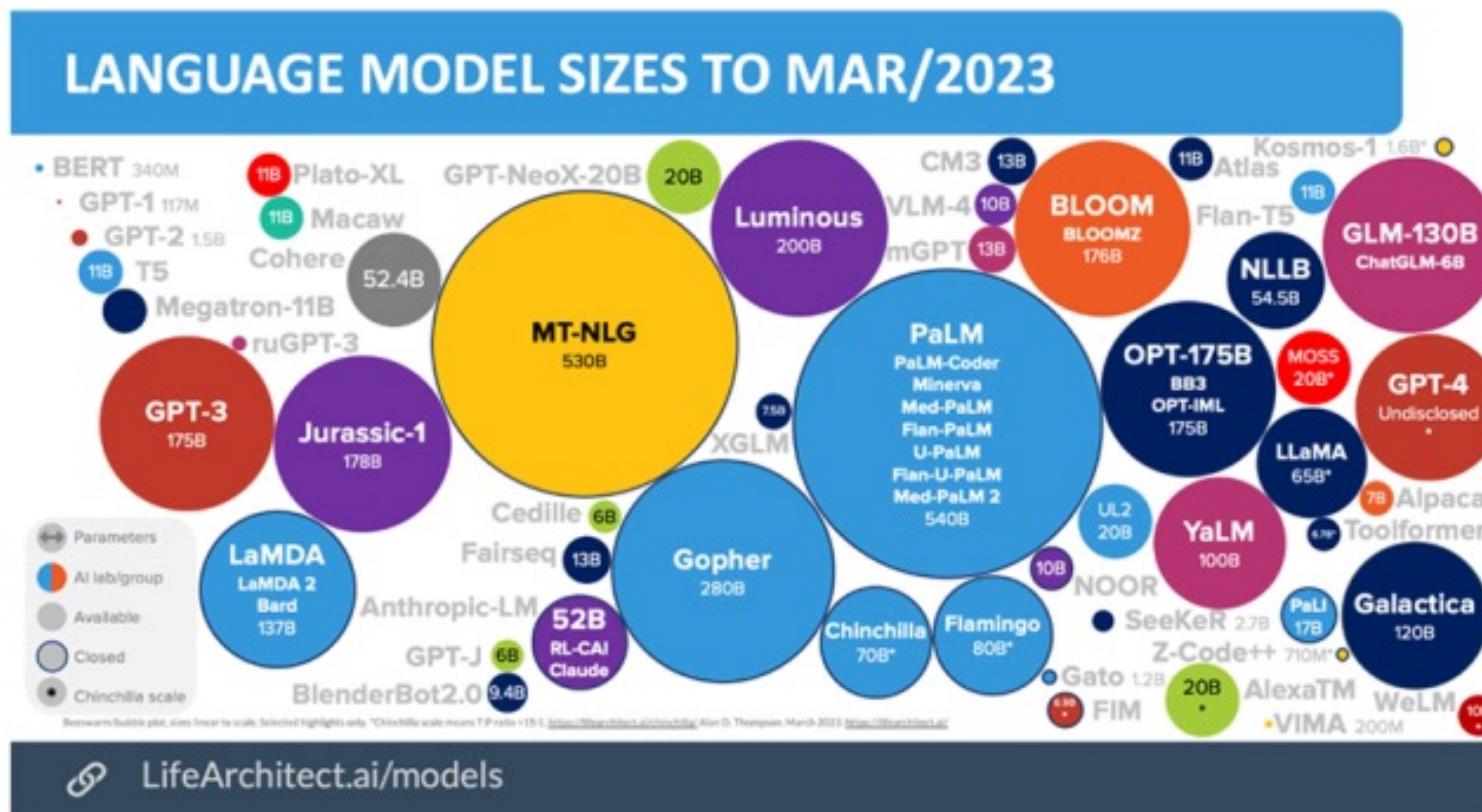
ABOUT GPT-3

- The cost of AI is increasing exponentially. A single training cycle for the 175 Billion parameter model takes about 355 years on a single V100 GPU, or around \$4,600,000 using on-demand Lambda Cloud GPU instances.
- (from <https://lambdalabs.com/blog/demystifying-gpt-3/>)

2022



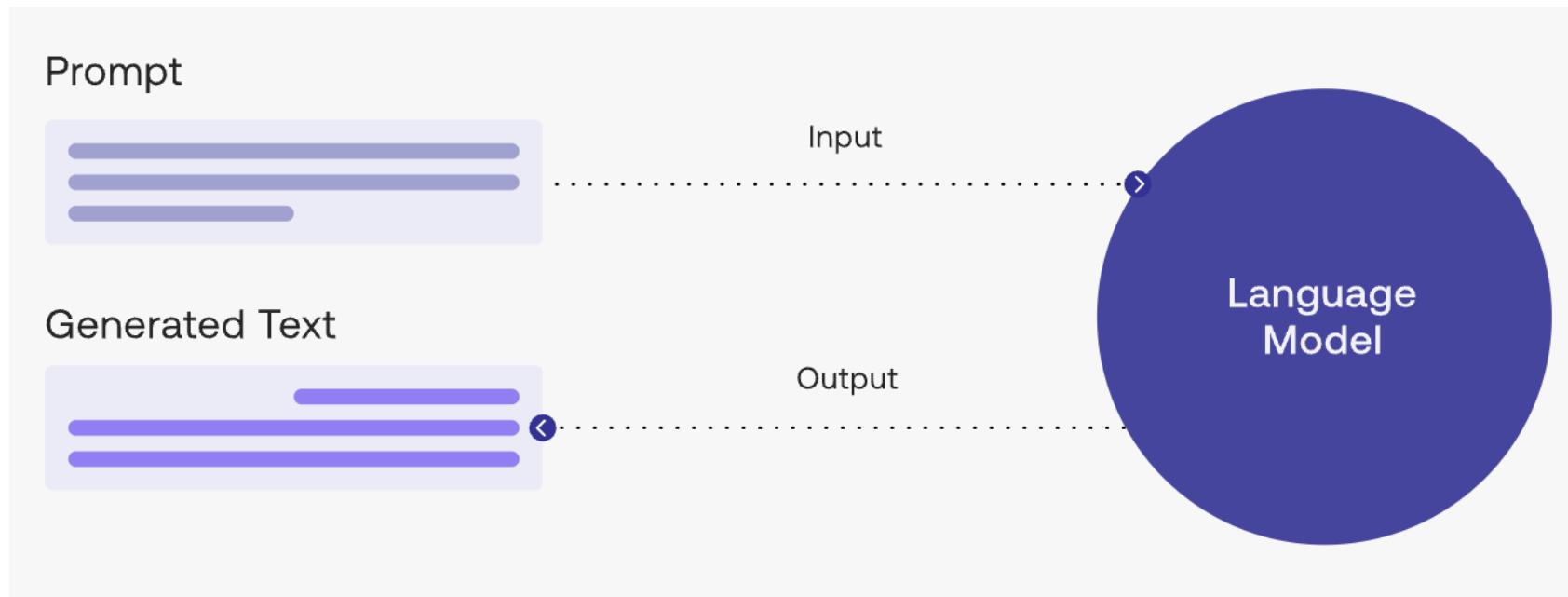
2023



2023

We will play with ChatGPT
and will talk about this

- A new job role has emerged: Prompt Engineer
- The job is dead some months after
 - LLM are also very good prompt engineers



2025

- This is how I see it!



KEY TAKEAWAYS

KEY TAKEAWAYS

- Many challenges, many different approaches (some taking advantage of linguistic knowledge), many different applications, many concepts to understand (see learning goals)
- Long process before the current achievements
- Everything is moving fast nowadays, and we are still trying to understand what we have done

MINI-MAP 1

