Distributional Semantics in the wild II

ESERCITAZIONI DI LINGUISTICA APPLICATA A.A. 2024/2025

Mattia Proietti

Topics of the lab

- **Vector semantics**
  - Static Embeddings
  - Contextual Embedding
- **Word2Vec**
- **Transformers Language Models**
  - **BERT**
  - **Generative Language Models** (GPT family)

Topic of the day

- **Contextual Embeddings**
  - **Transformers** language models

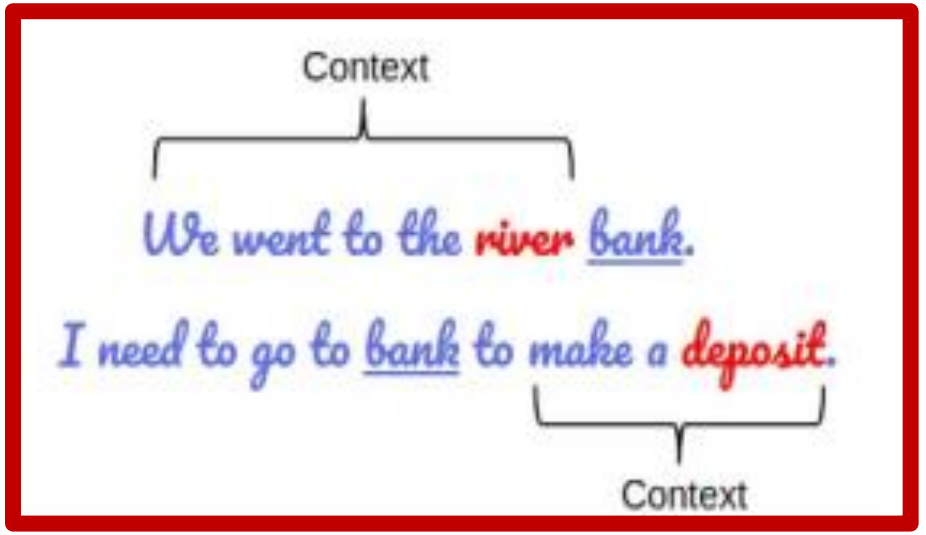
- **BERT**
  - What is it?
  - What can be used used for?
  - Masked Language Modelling

- **GPT**
  - What is it?
  - What can be used for?
  - Causal Language Modelling

Contextual Embeddings (recap)

- Discriminate different contexts
- Can assign different meanings to words based on the surrounding
- Take into account the position
- Help to disambiguate meaning

- Mostly generated by **Transformers**
  - Using the **Attention mechanism**



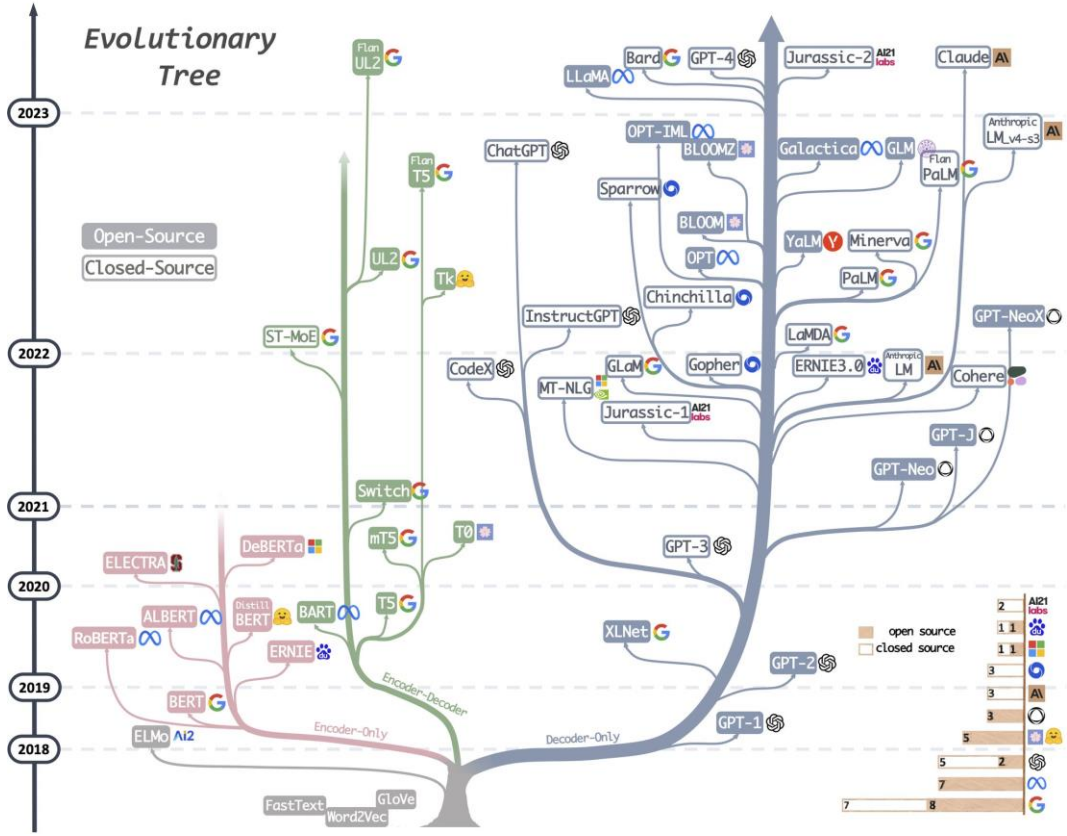
Transformers Language Models

Three branches:

- Encoder-only (BERT)
- Encoder-Decoder (T5)
- Decoder-only (GPTs)

- **Pre-Trained** on large corpora
  - (but with different objectives!)

- Can be **Fine-Tuned** for down-stream tasks (e.g. summarization, QA, translation etc.)



Pre-Training vs **Fine-Tuning**

Pre-training

- Large corpus
- Semi-supervised
- Use a training objective:
  - **Masked Language Modeling** (BERT)
  - **Causal language modelling** (GPT)

Pre-Training vs Fine-Tuning

Pre-training

- Large corpus
- Semi-supervised
- Use a training objective:
  - **Masked Language Modeling** (BERT)
  - **Causal language modelling** (GPT)

Fine-Tuning

- Small curated corpus
- Supervied training
- Specific tasks: QA, summarization, NER

BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers)

- TLM based on **encoders**
- **Bidirectional** contextual embeddings
- Trained on Wikipedia (2.5B) and Book Corpus (800M)
- With **Masked Language Modeling** (and Next Sentence Prediction)

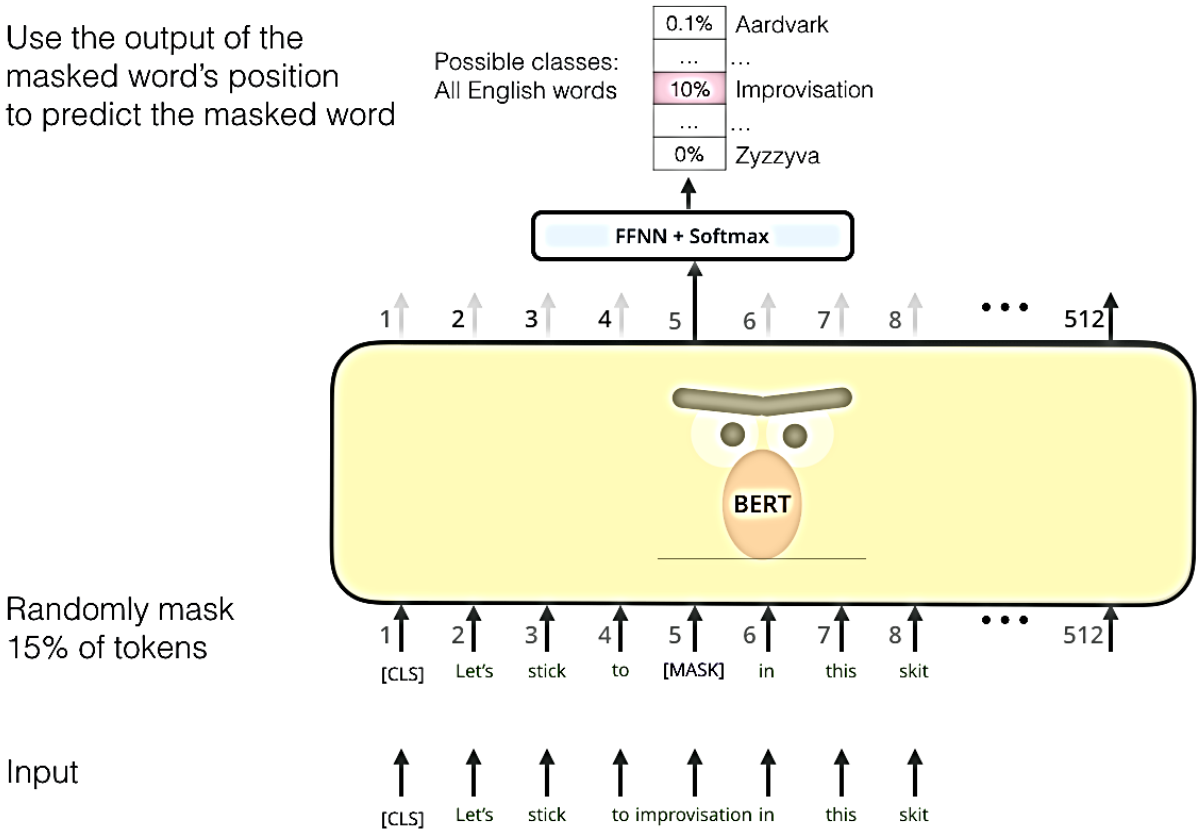
What can we do with it?

Extract vector representations of text

Fine-tune it on a given task
Use a specialized model "off the shelf"

BERT (Masked Language Modeling)

- Mask 15% of the word at random
- Gain some knowledge of the language



BERT

Let's try it on Colab!

What we'll see:

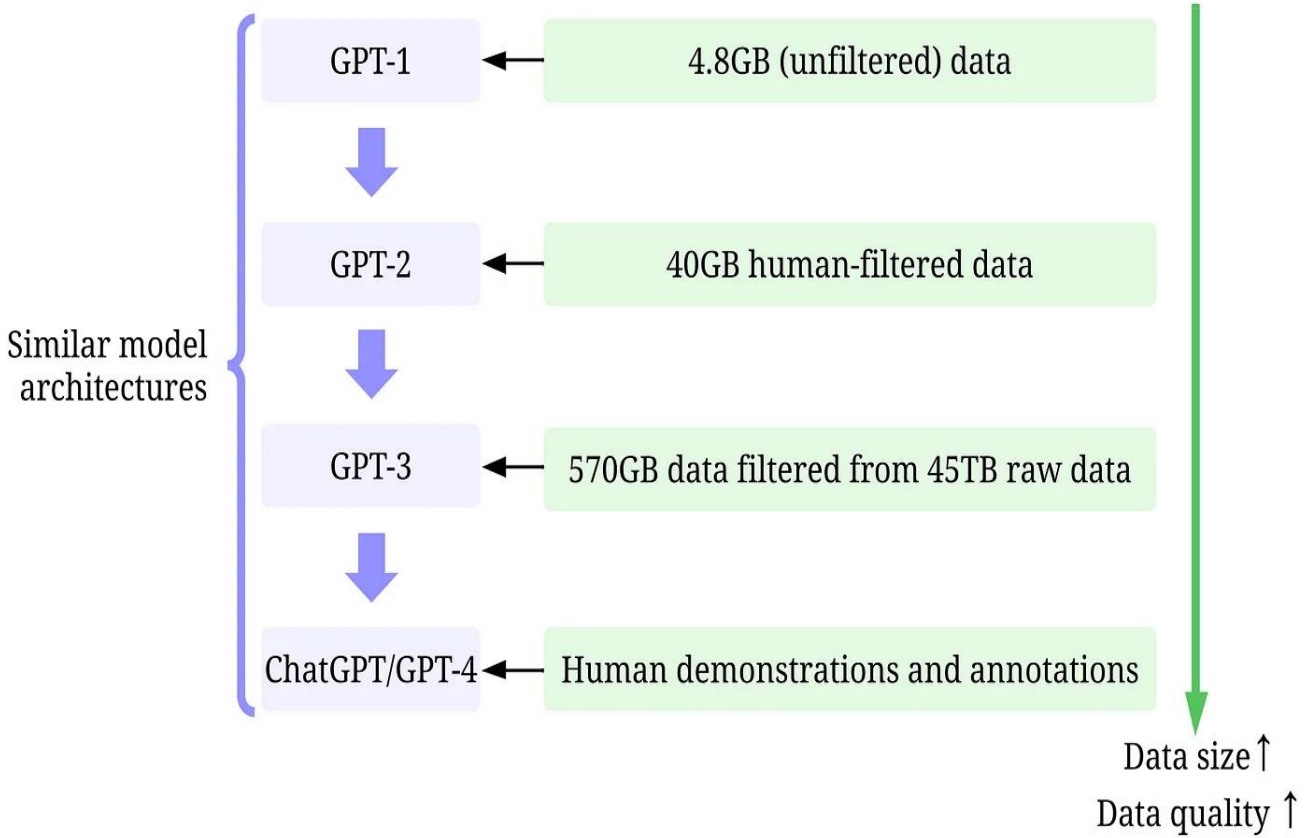
- A closer look at contextual embeddings
- Masked Language Modelling examples
- Question Answering
- Classification using BERT as feature extractor

GPT (**G**enerative **P**re-Training for **T**ransformers)

- Family of decoder-only models
- Different sizes
- Open source till GPT-2
- **Unidirectional**

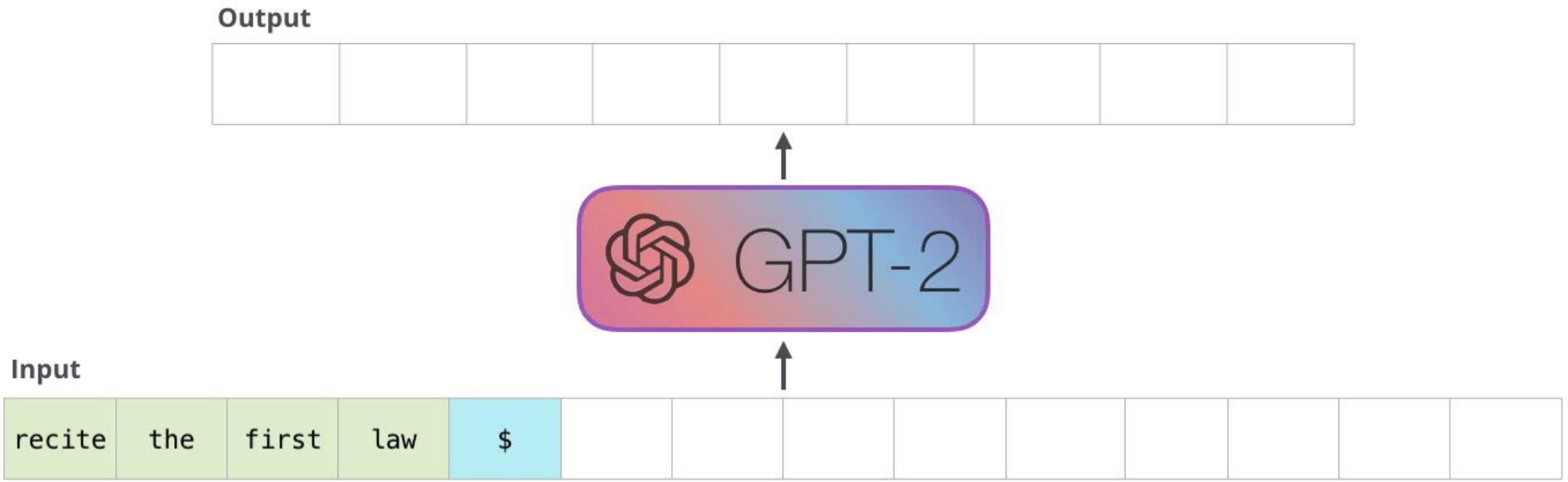
- Trained on the **Causal Language Modeling**
  - A.k.a. Next Word Prediction

- **GPT ≠ ChatGPT**
  - The first is a language model!
  - The second is a chat model based on the first!



GPT (Causal Language Modeling)

Causal Language modelling is done by masking the last word of a sequence recursively till the end of the sequence.



BERT vs **GPT**

- Encoder Architecture
- Bidirectional context
- Masked Language Modelling

Feature Extraction
Task-Specific Fine-tuning

Best used for

- Decoder-only architecture
- Unidirectional (left-to-right) context
- Causal Language Modelling

Text Generation
Task-Specific Fine-tuning

To know more on transformers

- Devlin et al.(2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- Radford et al. (2018), Improving Language Understanding with generative pre-training
- Vaswani et al (2017) Attention is all you need
- Jay Alammar, The illustrated Transformer
- McCormick blog post on input formatting