Course: Intelligent Systems

Unit 4: Language Technologies

Language technologies
Part 4

Mariano Rico
2022
Technical University of Madrid

NLP at a glance

- Session 1 (**29th Nov**)
  - Encodings
  - Corpus
  - Normalization
  - Hands-on 1
- Session 2 (in 2 weeks, **Tue 13 Dec**)
  - Part of Speech
  - Sparse Vector models
  - TF-IDF
  - Sentiment analysis
  - Hands-on 2
- Session 3 (in 3 weeks, **Tue 20 Dec**)
  - Document classification
  - Information extraction
  - Hands-on 3
- Session 4 (after Xmas, **Today**)
  - The neural revolution
  - Language Models 4 NLP tasks
  - Hands-on 4

Table of Contents

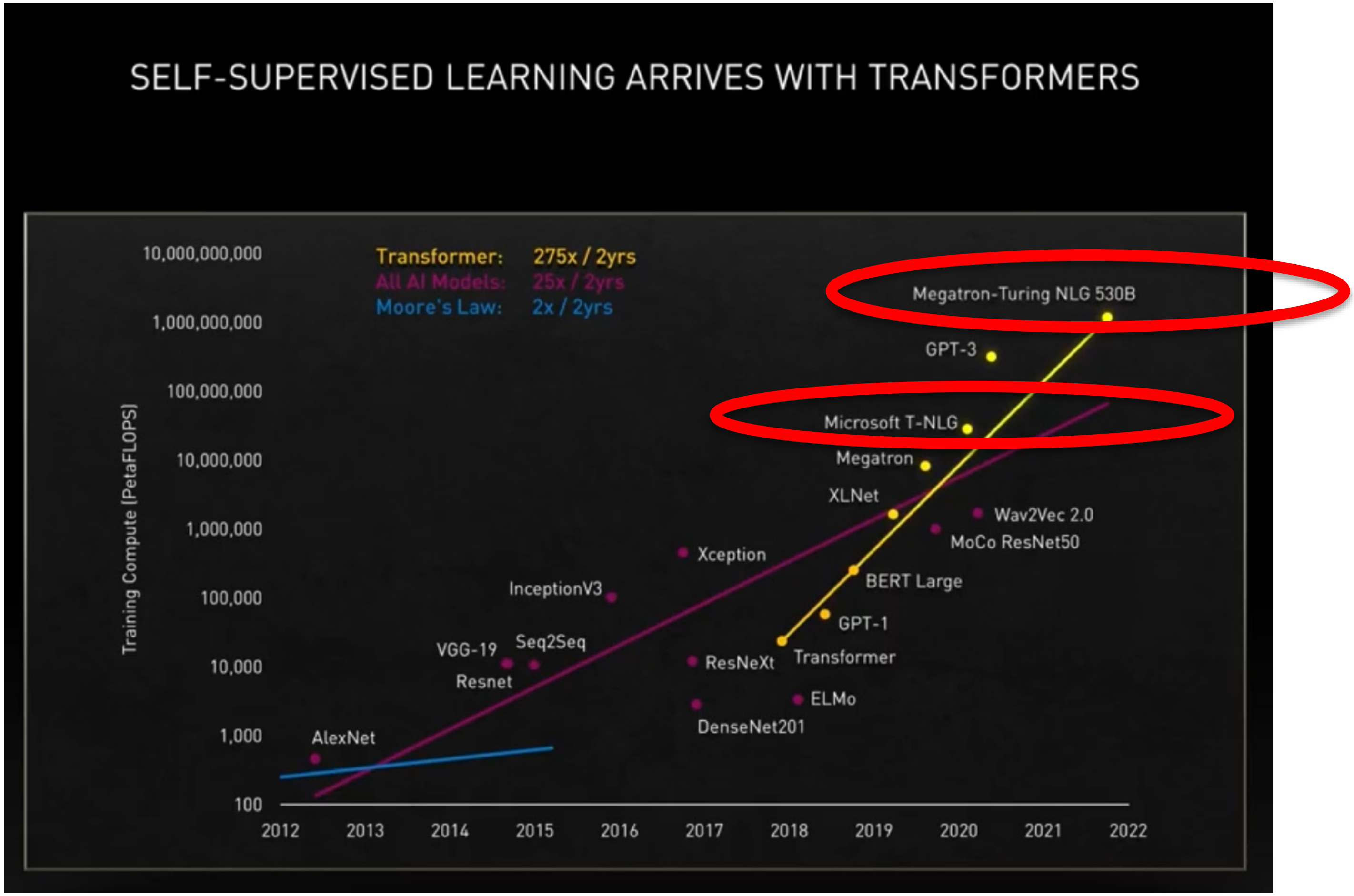
Acknowledgements

- Thanks to Pablo Calleja
  - Many slides in this presentation were made by him

THE NEURAL (R)EVOLUTION

A technological race

- Nov. 2021



NVIDIA [Nov. 2021 GTC Keynote]() by Jensen Huang

A technological race

- Dec. 2021 (less than 1 month later)

Microsoft Research Blog

Efficiently and effectively scaling up language model pretraining for best language representation model on GLUE and SuperGLUE

Published December 2, 2021

By Jianfeng Gao, Distinguished Scientist & Vice President; Saurabh Tiwary, Vice President & Distinguished Engineer

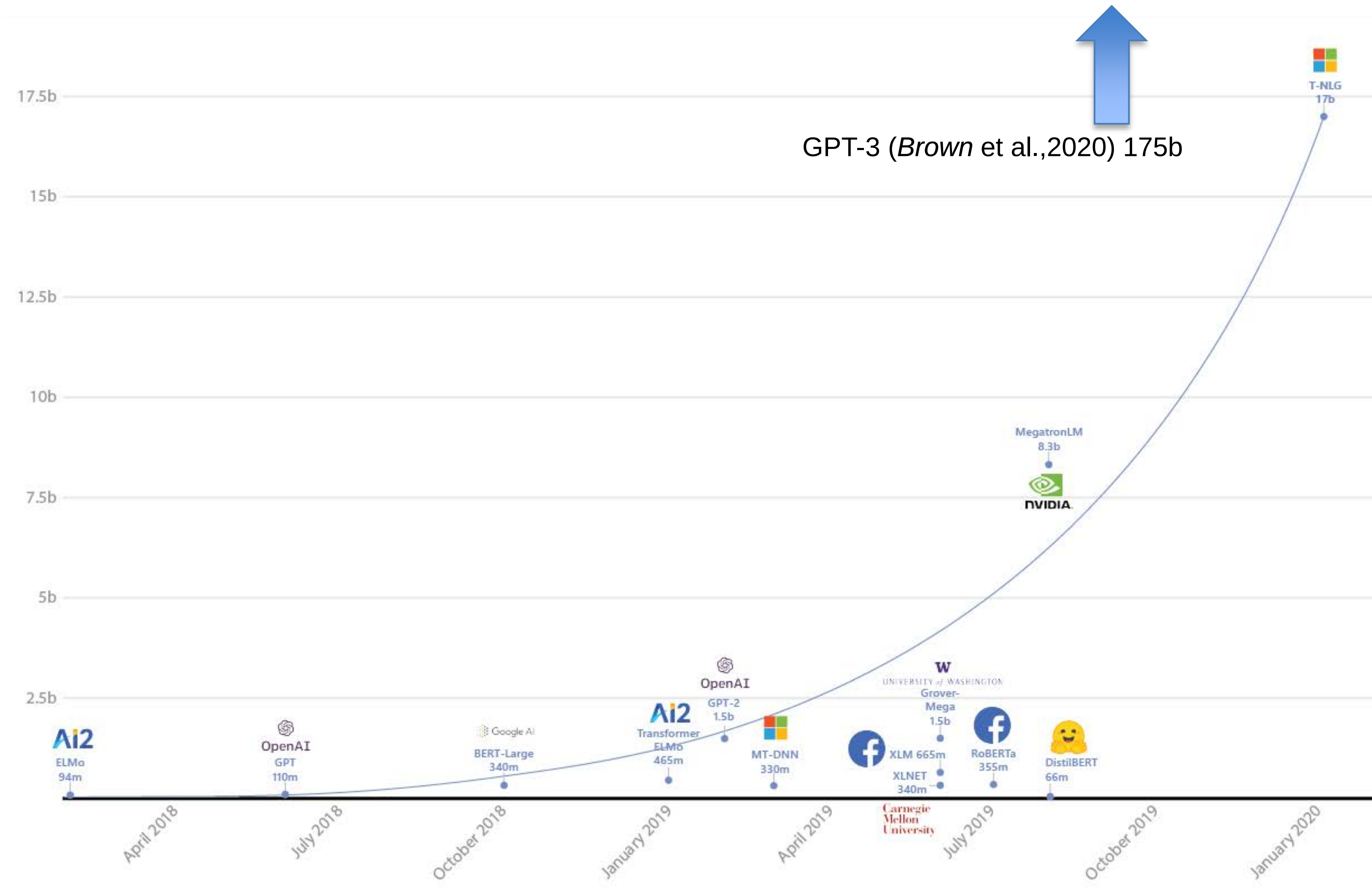
Research Area

Artificial intelligence

MS Research blog Dec. 2021

A technological race

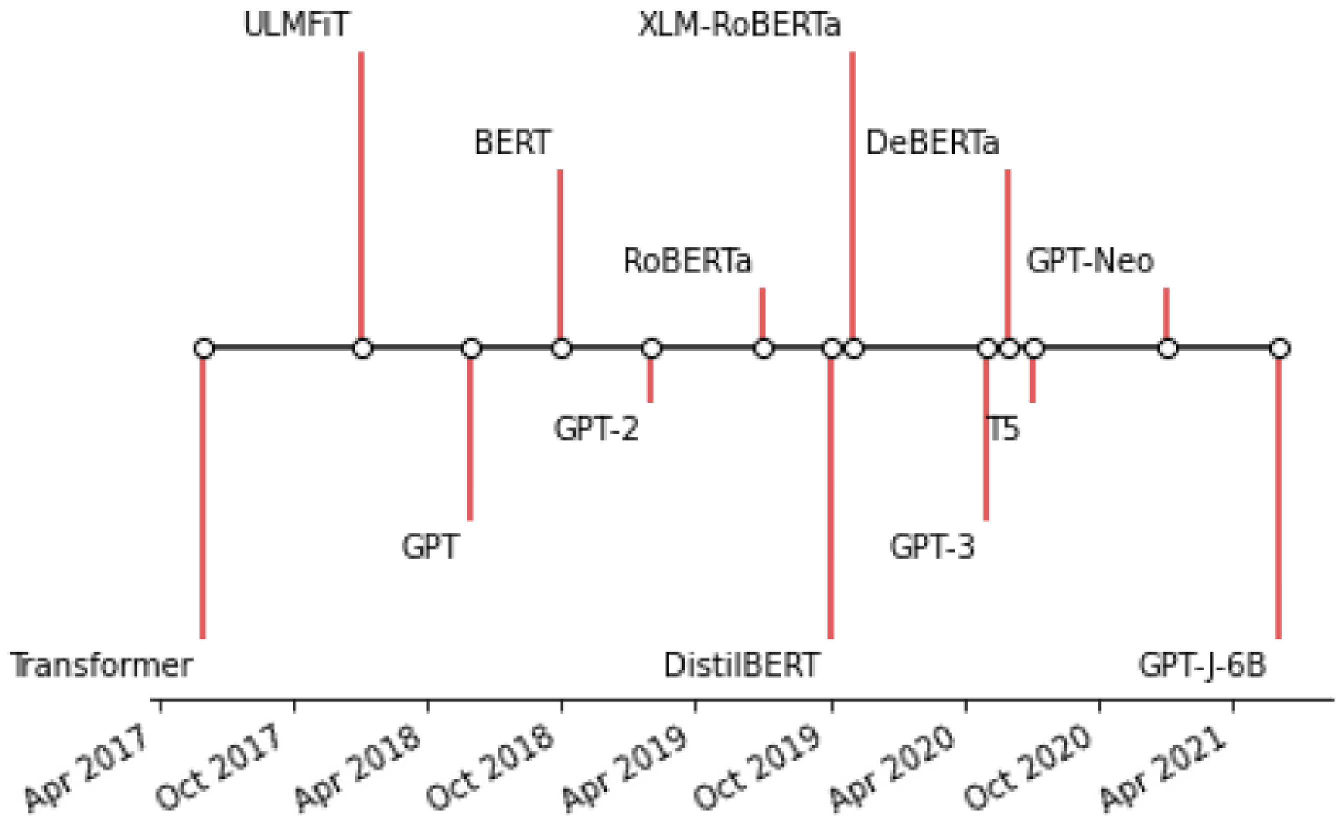
- Evolution: number of parameters and actors



MS Research blog [Feb. 2020]

A technological race

- Evolution: evolution of transformers



Tunstall *et al.* 2021 (O'Reilly media)

A technological race

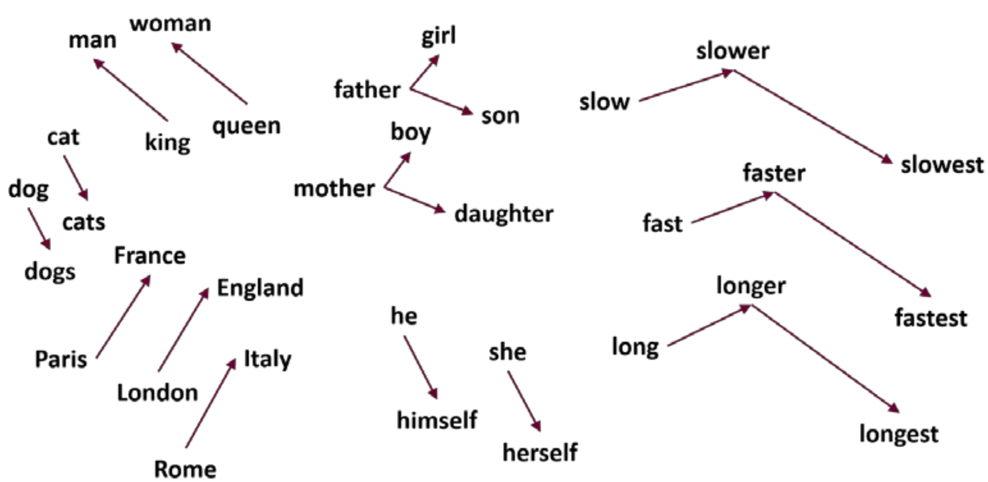
- (R)evolution: things to come
  - Explainable AI
    - Can you trust current AI?. Beyond a black-box model for neural systems
  - Reduction of hardware dependency
    - Do you have hardware to create a neural model?
    - What is the carbon fingerprint of creating a huge model?
    - I am a minority language. How can I get a model for my language?

All started with embeddings

- Distributional Hypothesis (Harris, 1954)

  Words with similar meanings tend to occur in similar contexts

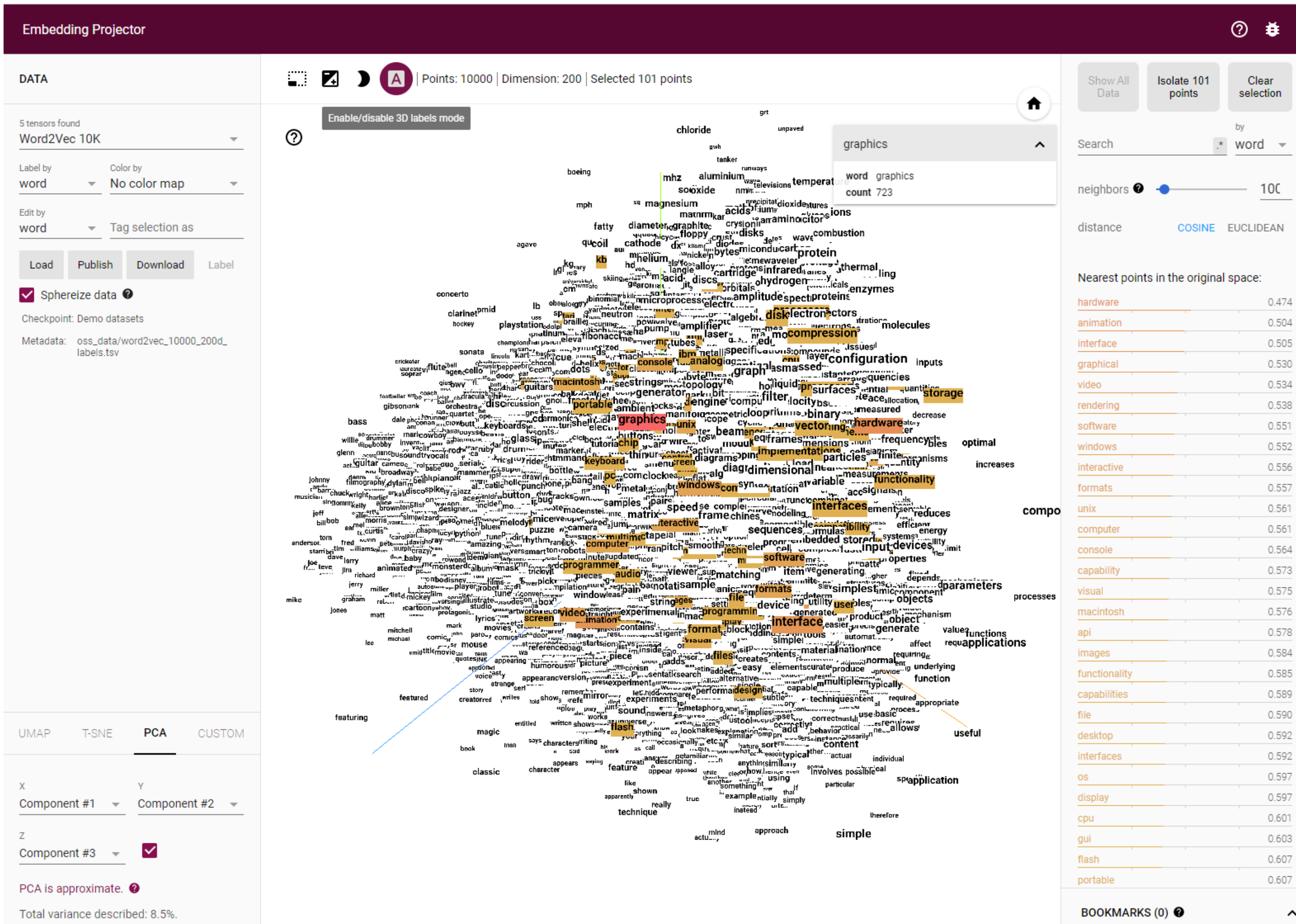
- Word2Vec ([Mikolov 2013]())
  – Also for relations!! ➔ semantic similarity!!



[Source]()

All started with embeddings

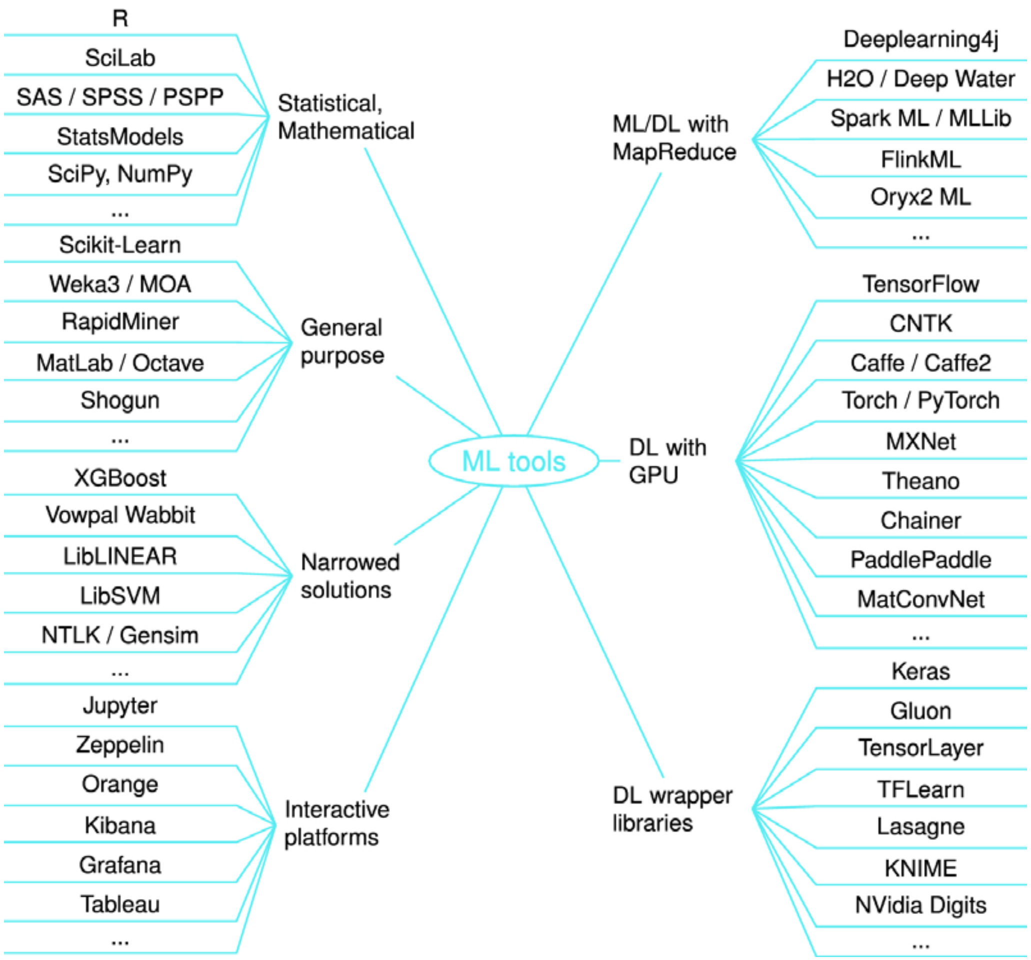
- Play with them [here](https://projector.tensorflow.org/)



Source: https://projector.tensorflow.org/

Development environments

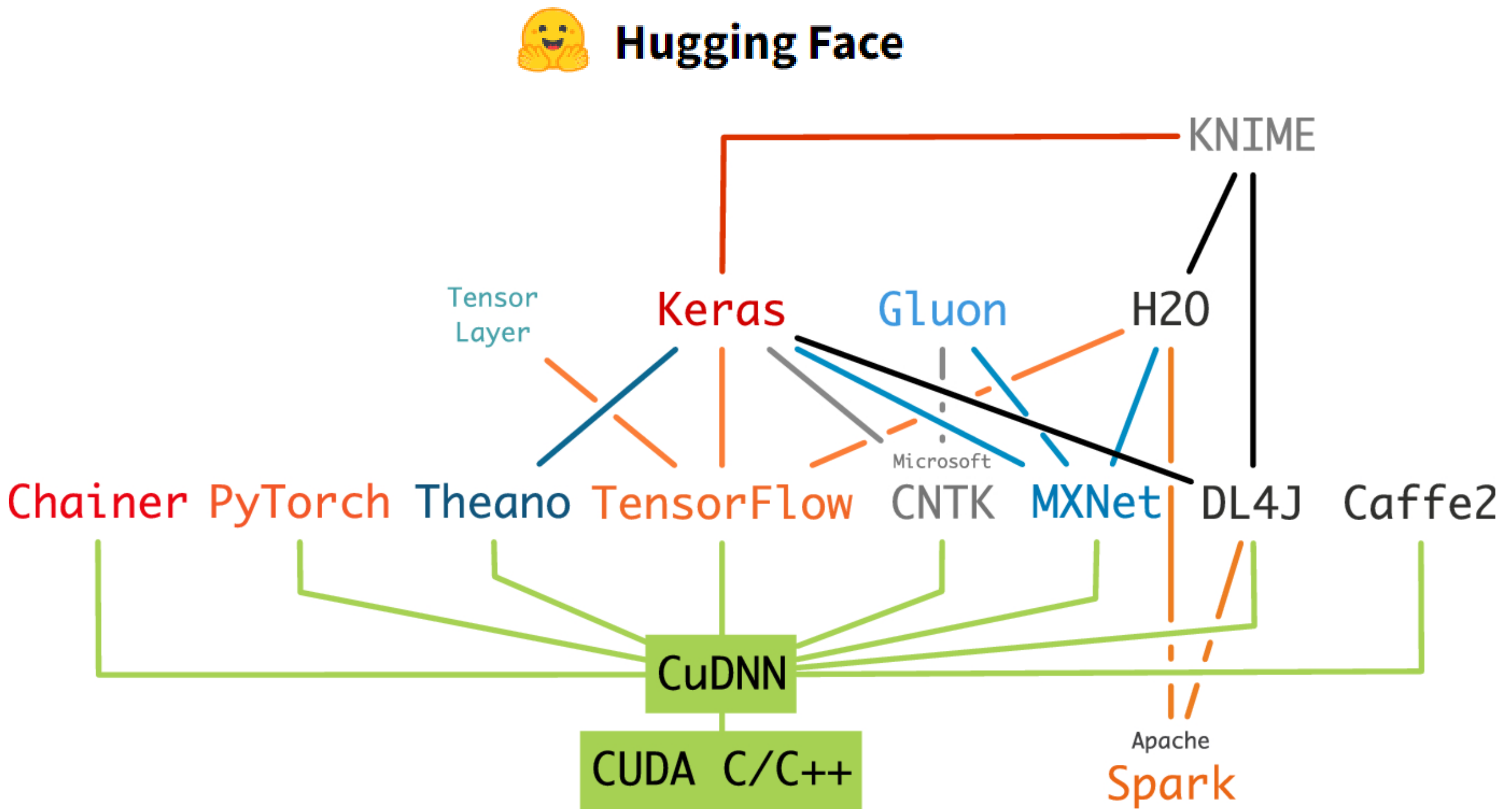
- ML frameworks and libraries



[Springer, Jan. 2019]

Development environments

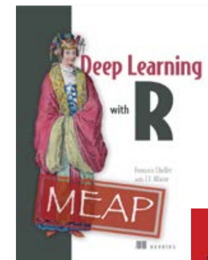
- DL frameworks and libraries



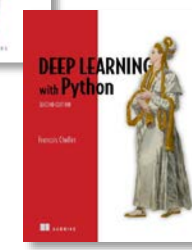
[Springer, Jan. 2019]

Deep Learning using R

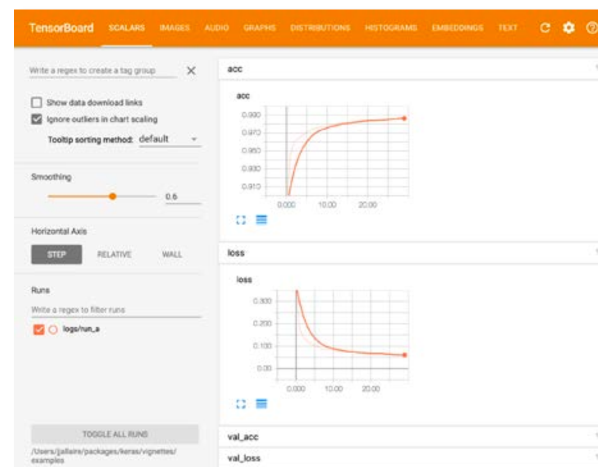
- Although most code is Python there are options for R:
  - Keras from Rstudio (keras.rstudio.com)
    - Cheatsheet (keras 2.1.2, 2017, before [TF2](TF2))
      - A Spanish version by Carlos Ortega (R Users Madrid)
    - TensorBord: visualizaing the state of the neural net
    - TFruns: track and visualize training runs (integrated with Rstudio's addins)

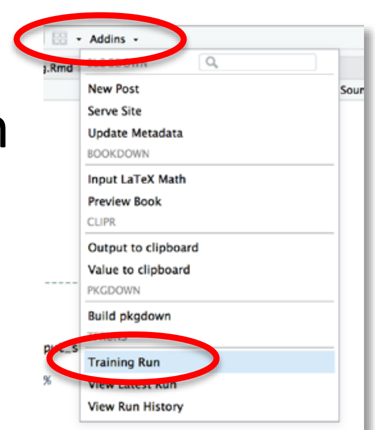


2018 (TF1)



2021 (TF2)

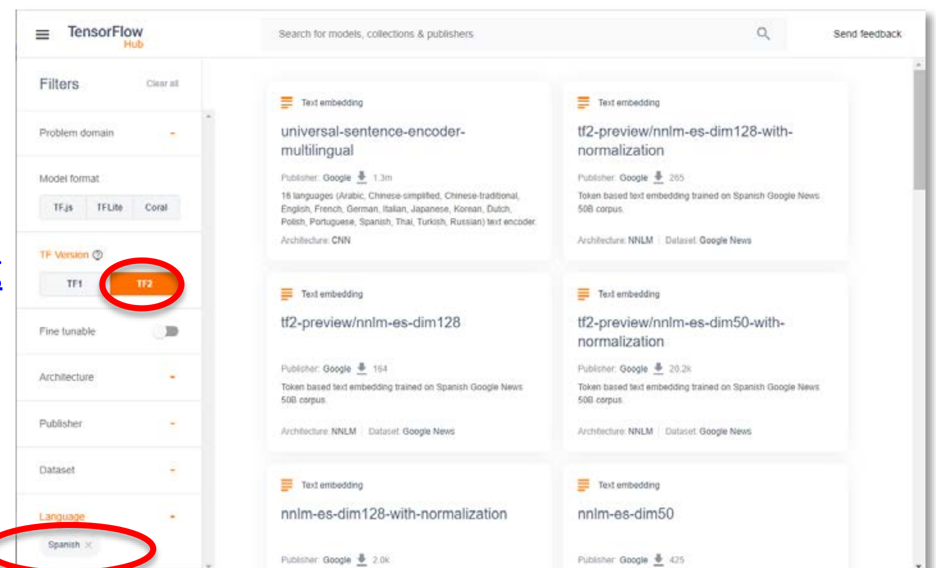




Deep Learning using R

- Tensorflow (TF1 y TF2)
  - You can use local GPUs (only NVIDIA) but also cloud GPUs like
    - Google CloudML
    - Cloud Server (Amazon EC2, Google Compute Engine)
    - Paperspace Cloud Desktop (only TF1?)
  - Packge `tfhub`: using models from Tensorflow Hub as a keras layer
    - TF1 and TF2
    - 19 TF2 Spanish models for the Spanish language ☺
    - Many examples
      - » Simple transfer learning
      - » Text classification
      - » Attention (seq2seq almost Transformer)

2019
(pre Transformer)



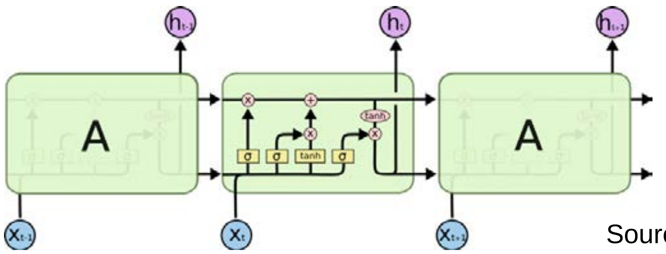
Deep Learning using R

- Using Hugging Face
  - Package `reticulate` can load any Python code
    - Even PyTorch
  - See examples in the hands-on

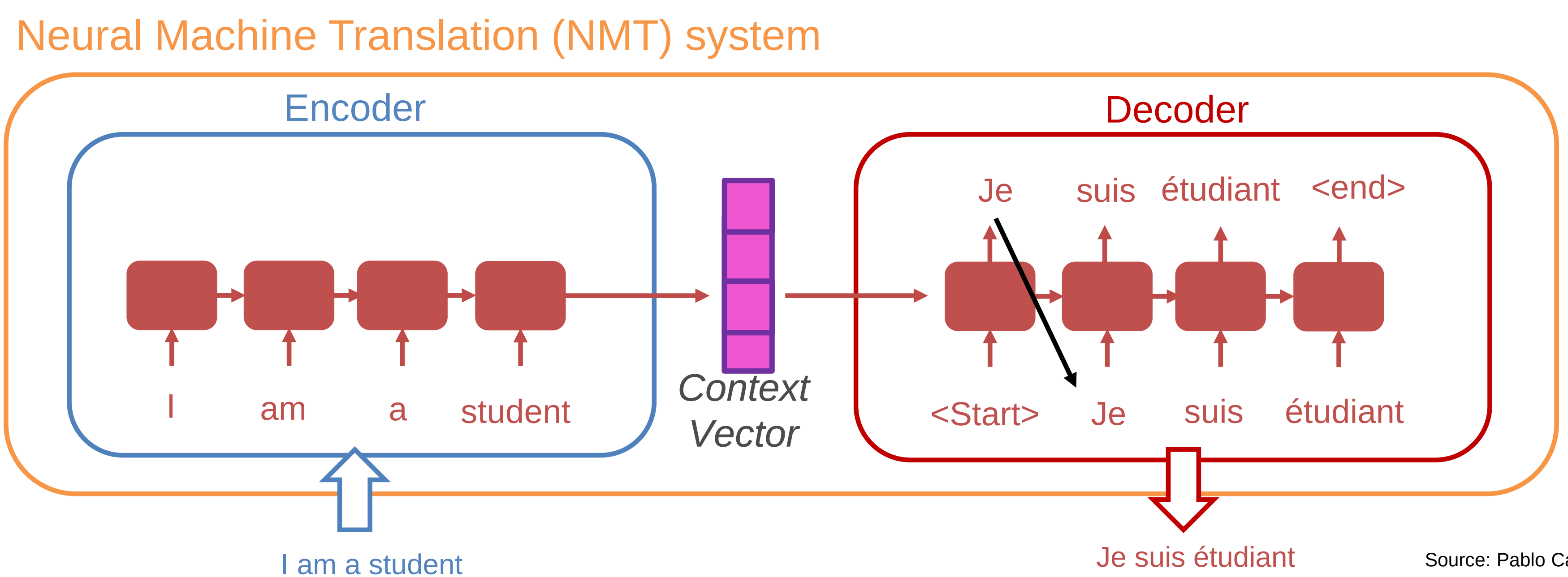
TRANSFORMERS

Why transformers

- Evolution of Recurrent Neural Networks (RNN)
  - LSTM
    - Relevant words are lost in long sentences (attention focused on nearby words)



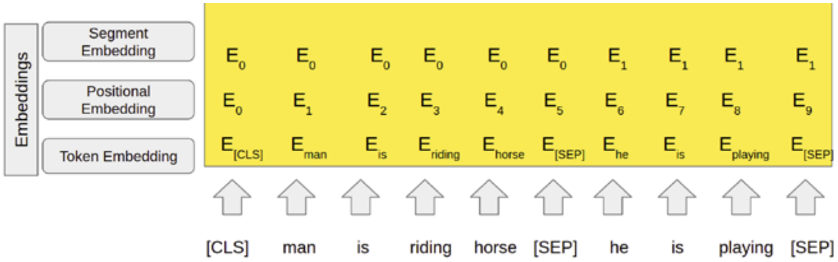
Source: here



Source: Pablo Calleja

Why transformers

- Enhances the capture of context information
  - How?: Attention (is all you need)
    - Attention mechanism: an alignment score function to quantify the relevance of each token to another token
      - There are several types of attention mechanisms. Transformers use the *scaled dot-product attention*
  - Instead of processing word by Word (as RNNs do), the whole sentence in processed **in parallel**
  - Instead of 1 encoder and 1 decoder (as RNNs do), we have many of them
  - Uses positional embeddings for each token, as well as segment embeddings to separate sentences



- More info (barely math)
  - Illustrated transformer

BERT / DISTIL**BERT** /RO**BERT**A

BERT

- BERT: Bidirectional Encoder Representations from Transformers
  - Encoder: the model uses the encoder part of the transformer
  - Bidirectional means:
    - Pay attention both forward and backwards tokens (transformers only backwards)
    - Achieved with a novel technique named Masked Language Model (MLM)

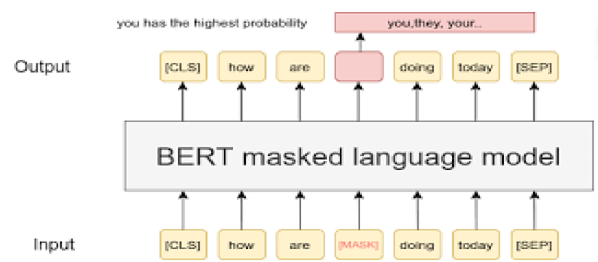
- The paper (v1 Oct. 2018, v2 May 2019)

BERT

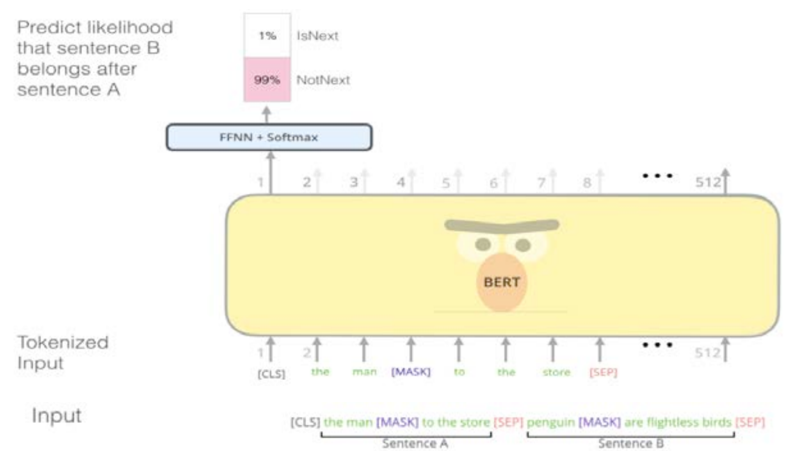
- BERT: Bidirectional Encoder Representations from Transformers
  - Designed to be used as a pre-trained model that can be [fine-tuned]
    - This pre-trained model can be slightly modified (typically by adding output neural layers) to perform NLP tasks such as:
      - Question answering
      - Sentiment analysis
      - Named entity recognition
      - Text summarization

BERT

- Model training for two different tasks:
  - Masked Language Model (MLM)
    - 15% of tokens in the input are masked
      - 80% replaced with [MASK]
      - 10% with a random Word
      - 10% with the original Word



  - Next Sentence Prediction (NSP)
    - BERT is trained with pairs of sentences and predicts if the second is the subsequent
      - 50% are subsequent pairs and 50% are random
      - Uses special tokens for the classification. [CLS] at the beginning, and [SEP] at the end of each sentence. [CLS] token is used to predict IsNext/NotNext



DistilBERT

- Created by 🤗 Hugging Face ([paper]() 2020)
- It is a *distilled* BERT
  - 40% smaller
  - 60% faster
  - Retains 97% of the language understanding capabilities
- Methodology
  - BERT is the “teacher” model. DistillBERT is a “student” model with
    - half number of layers (but keeping layer sizes)
    - Without token-type embeddings
    - Without pooling

RoBERTa

- Created by Facebook (paper 2019)
- It is a "Robustly optimized" BERT approach
  - Modifications to the BERT pre-training process:
    - Longer model training times
      - Larger batches and more data
    - Removed one of the two BERT tasks:
      - The *Next Sentence Prediction* (NSP) task
    - Longer sequences for training
    - Changes in the method used for masking the training data

Comparison of BERT-based models

	BERT	RoBERTa	DistilBERT	XLNet
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base**: 66	**Base**: ~110 **Large**: ~340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base**: 8 x V100 x 3.5 days; 4 times less than BERT.	**Large:** 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	3% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	**Base**: 16 GB BERT data **Large**: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation	Bidirectional Transformer with Permutation based modeling

What about non English languages?

- Like Spanish
  - MarIA (by RAE+BSC)
    - Github repo with
      - Models (links to 🤗)
        - » RoBERTa (b & L)
        - » GPT2 (b & L)
      - Fine-tuned models for
        - » POS (Part of Speech)
        - » NER (Named Entity Recognition)
        - » QA (Question-Answering)
      - Evaluation results
      - Usage examples (Python)

First massive Artificial Intelligence system in the Spanish language, MarIA, begins to summarize and generate texts

11 November 2021

Launched five months ago, the system expands its capabilities to use the language. Creative and business applications and those related to the digitization of Public Administration increase.

Evaluation ✅

Dataset	Metric	RoBERTa-b	RoBERTa-l	BETO*	mBERT	BERTIN**	Electricidad***
UD-POS	F1	0.9907	0.9898	0.9900	0.9886	0.9898	0.9818
Conll-NER	F1	0.8851	0.8772	0.8759	0.8691	0.8835	0.7954
Capitel-POS	F1	0.9846	0.9851	0.9836	0.9839	0.9847	0.9816
Capitel-NER	F1	0.8960	0.8998	0.8772	0.8810	0.8856	0.8035
STS	Combined	0.8533	0.8353	0.8159	0.8164	0.7945	0.8063
MLDoc	Accuracy	0.9623	0.9675	0.9663	0.9550	0.9673	0.9493
PAWS-X	F1	0.9000	0.9060	0.9000	0.8955	0.8990	0.9025
XNLI	Accuracy	0.8016	0.7958	0.8130	0.7876	0.7890	0.7878
SQAC	F1	0.7923	0.7993	0.7923	0.7562	0.7678	0.7383

** A model based on BERT architecture.*

*** A model based on RoBERTa architecture.*

**** A model based on Electra architecture.*

For the RoBERTa-base

```python
from transformers import AutoModelForMaskedLM
from transformers import AutoTokenizer, FillMaskPipeline
from pprint import pprint
tokenizer_hf = AutoTokenizer.from_pretrained('PlanTL-GOB-ES/roberta-base-bne')
model = AutoModelForMaskedLM.from_pretrained('PlanTL-GOB-ES/roberta-base-bne')
model.eval()
pipeline = FillMaskPipeline(model, tokenizer_hf)
text = f"¡Hola <mask>!"
res_hf = pipeline(text)
pprint([r['token_str'] for r in res_hf])
```

What about non English languages?

- Like Spanish
  - flairNLP (Humbold Univ.)
    - NER models (links to 🤗) for several languages
      - English, German, Dutch, Spanish
      - Top performance
    - Also other models for POS
    - It is a development framework (Python + PyTorch)
      - With tutorials and an enthusiastic community

State-of-the-Art Models

Flair ships with state-of-the-art models for a range of NLP tasks. For instance, check out our latest NER models:

Language	Dataset	Flair	Best published	Model card & demo
English	Conll-03 (4-class)	94.09	*94.3 (Yamada et al., 2020)*	Flair English 4-class NER demo
English	Ontonotes (18-class)	90.93	*91.3 (Yu et al., 2020)*	Flair English 18-class NER demo
German	Conll-03 (4-class)	92.31	*90.3 (Yu et al., 2020)*	Flair German 4-class NER demo
Dutch	Conll-03 (4-class)	95.25	*93.7 (Yu et al., 2020)*	Flair Dutch 4-class NER demo
Spanish	Conll-03 (4-class)	90.54	*90.3 (Yu et al., 2020)*	Flair Spanish 4-class NER demo

The state of the art in NER: here

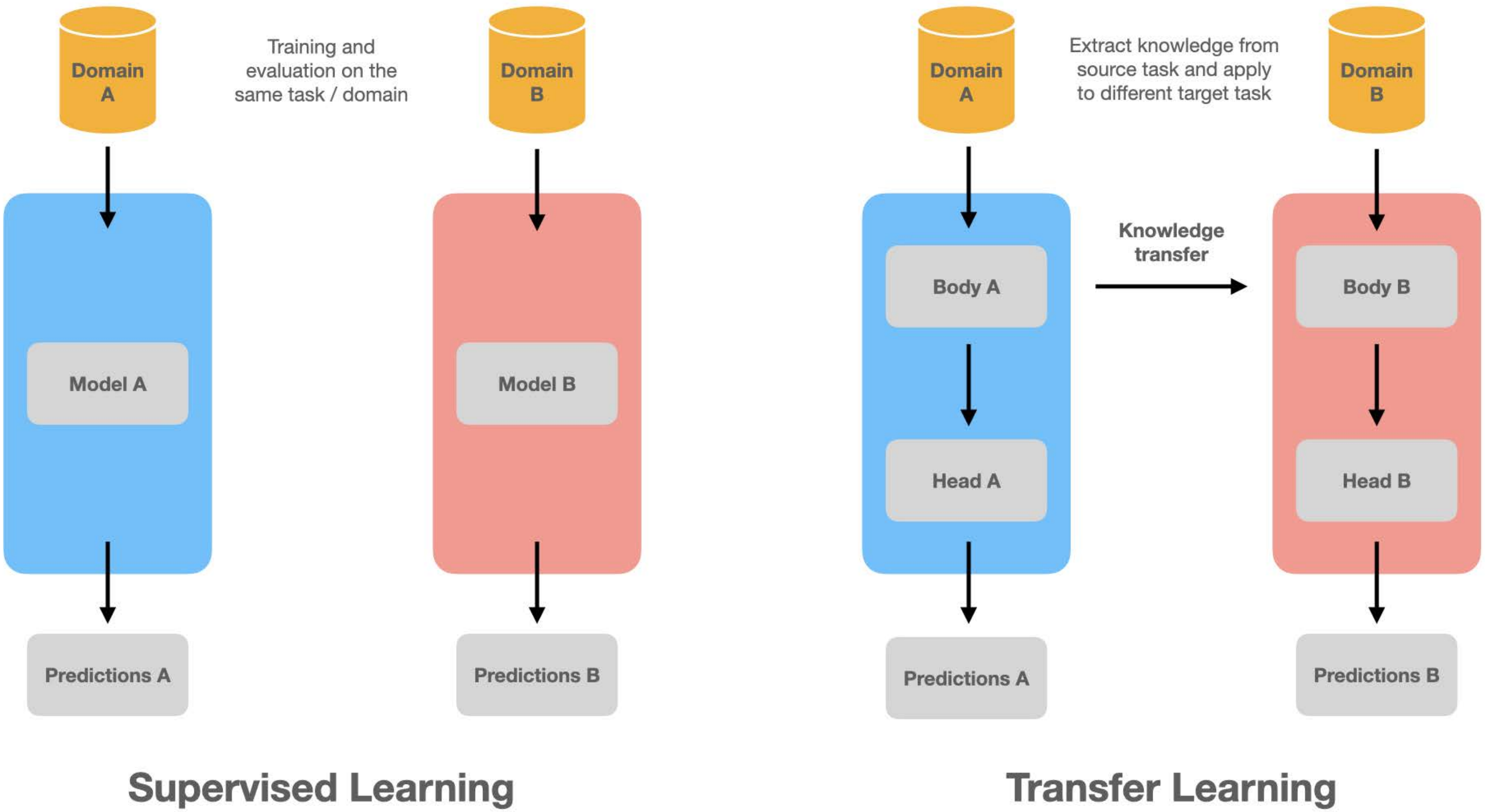
What about non English languages?

- Like Spanish
  - David vs. Goliath
    - RigoBERTa (IIC) They claim better results in 10 in 13 tasks

	Dataset	BETO	BERTIN	MarIA	RigoBERTa
NER	CANTEMISTNER	89.9%	79.5%	92.3%	**93.3%**
NER	CAPITEL	87.0%	86.5%	**87.8%**	87.4%
NER	CONLL2002	89.6%	**90.1%**	89.9%	89.5%
NER	MEDDOCAN	84.7%	72.2%	84.1%	**85.0%**
NER	MEDDOPROF1	80.5%	71.0%	80.7%	**83.1%**
NER	MEDDOPROF2	81.8%	44.2%	78.5%	**86.4%**
Class	MLDOC	95.4%	94.4%	**95.6%**	**95.6%**
Class	PAWS-X	89.7%	90.1%	88.9%	**91.0%**
NER	PHARMACONER	61.4%	47.1%	57.1%	**70.0%**
QA	SQAC	76.2%	75.0%	86.6%	**89.7%**
QA	SQUADES	75.6%	70.0%	81.8%	**85.4%**
Class	TASS2020	46.1%	46.1%	**47.3%**	46.7%
Class	XNLI	81.7%	79.4%	81.6%	**83.4%**
	TOTALES	76.5%	69.6% 1	77.3% 3	**79.8%** 10

LANGUAGE MODELS 4 NLP TASKS

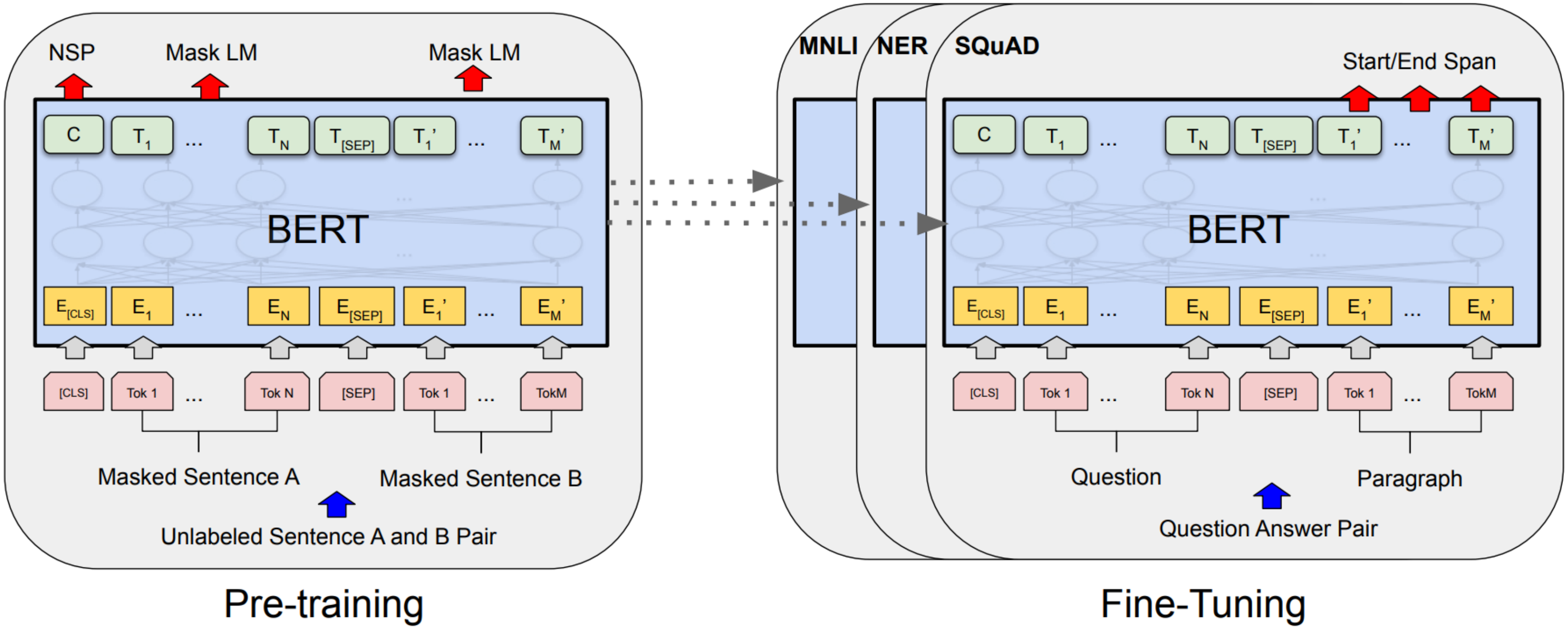
Transfer Learning
The main concept



Tunstall *et al*. 2021 (O'Reilly media)

Fine-tuning BERT

Standing on the shoulders of giants

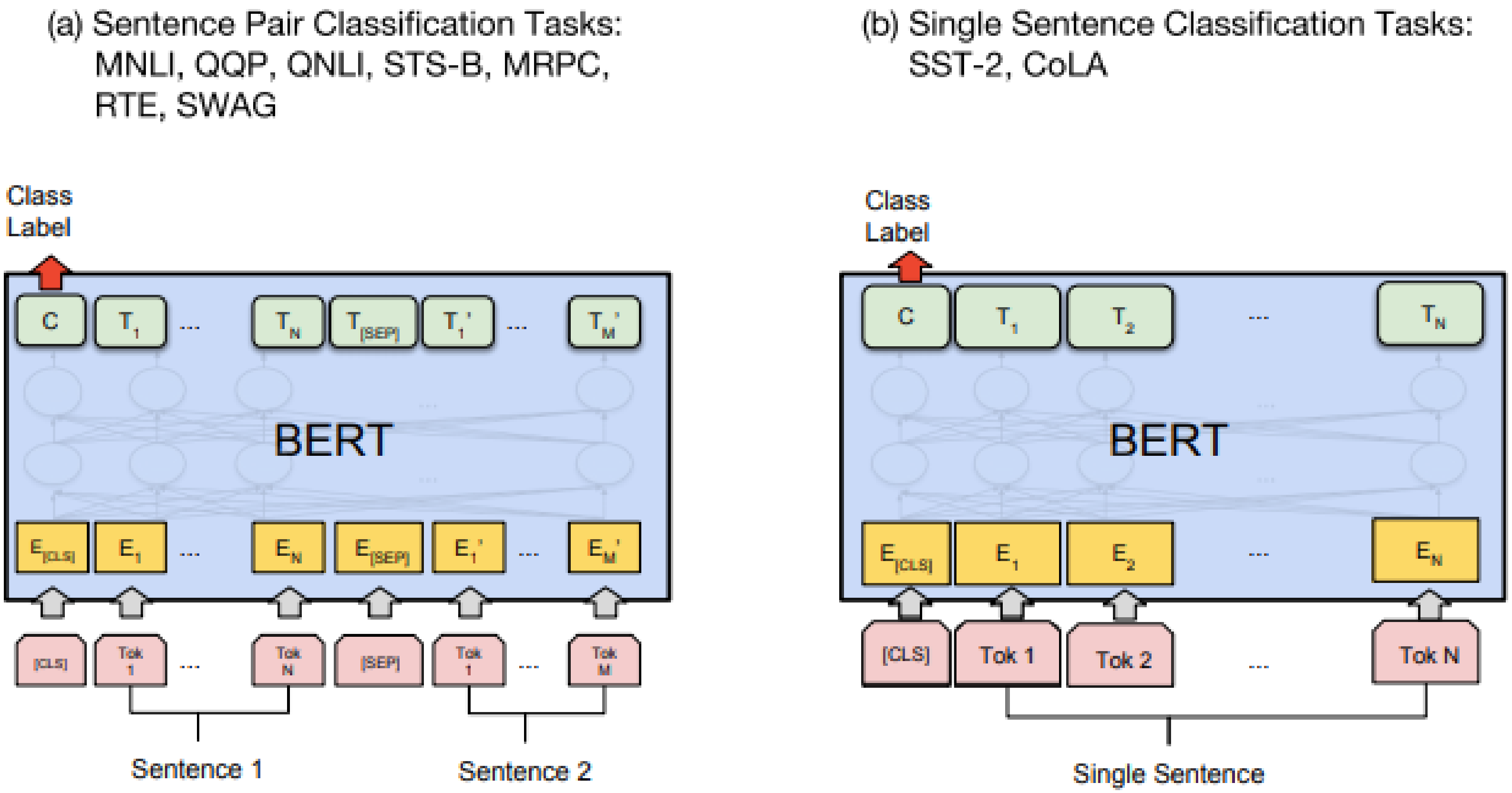


There is also training, but less than the computational effort to create the pre-training model

Source: [BERT paper (2018-9)]

Fine-tuning BERT

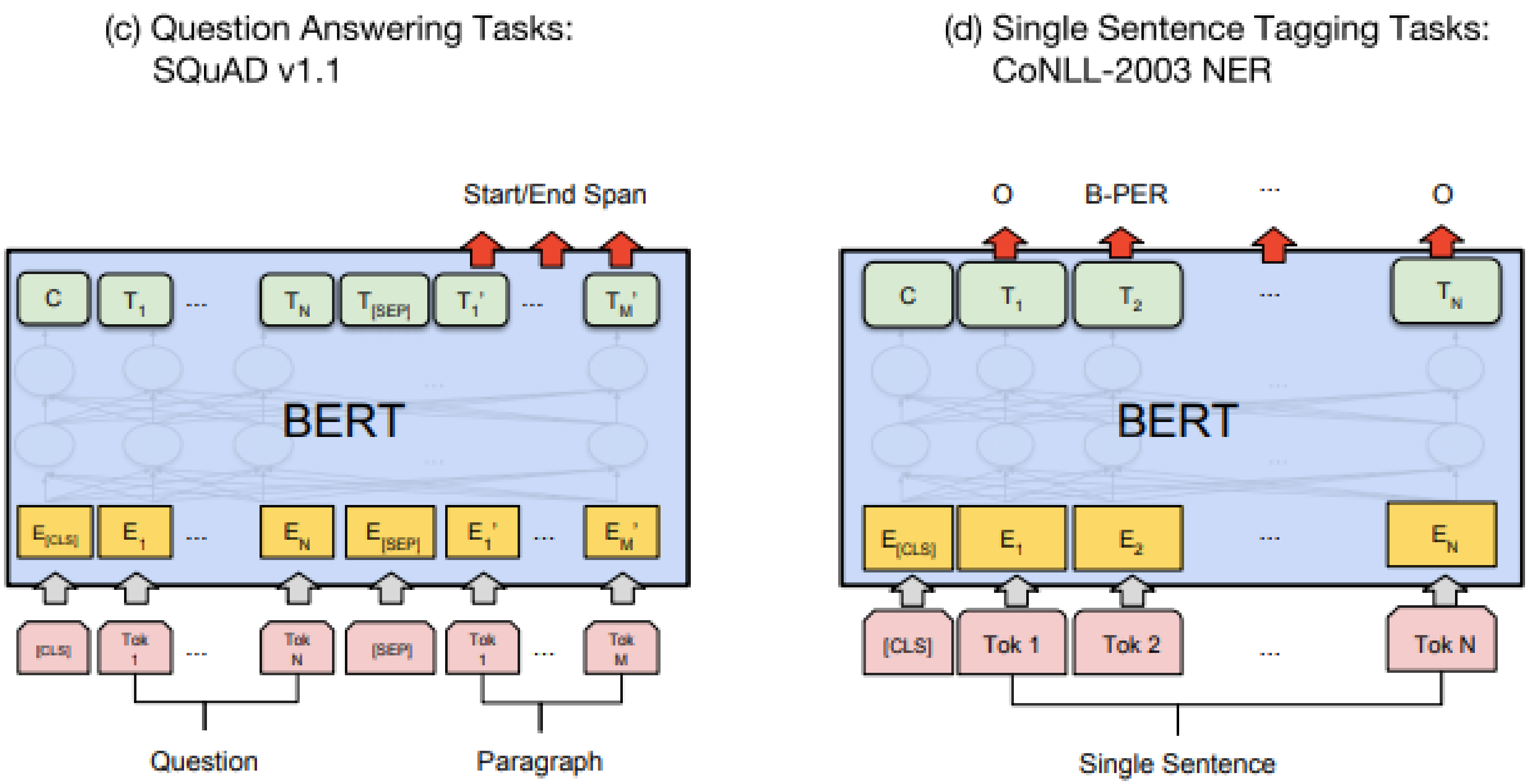
- BERT can be adapted for NLP tasks such as



Source: [BERT paper (2018-9)]

Fine-tuning BERT

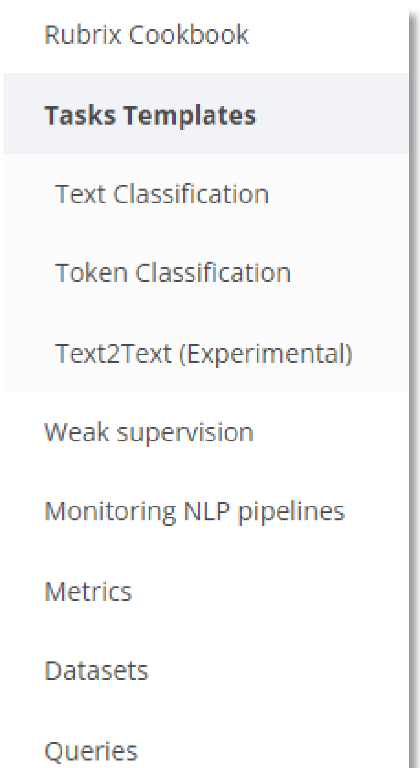
- BERT can be adapted for NLP tasks such as



Source: [BERT paper (2018-9)]

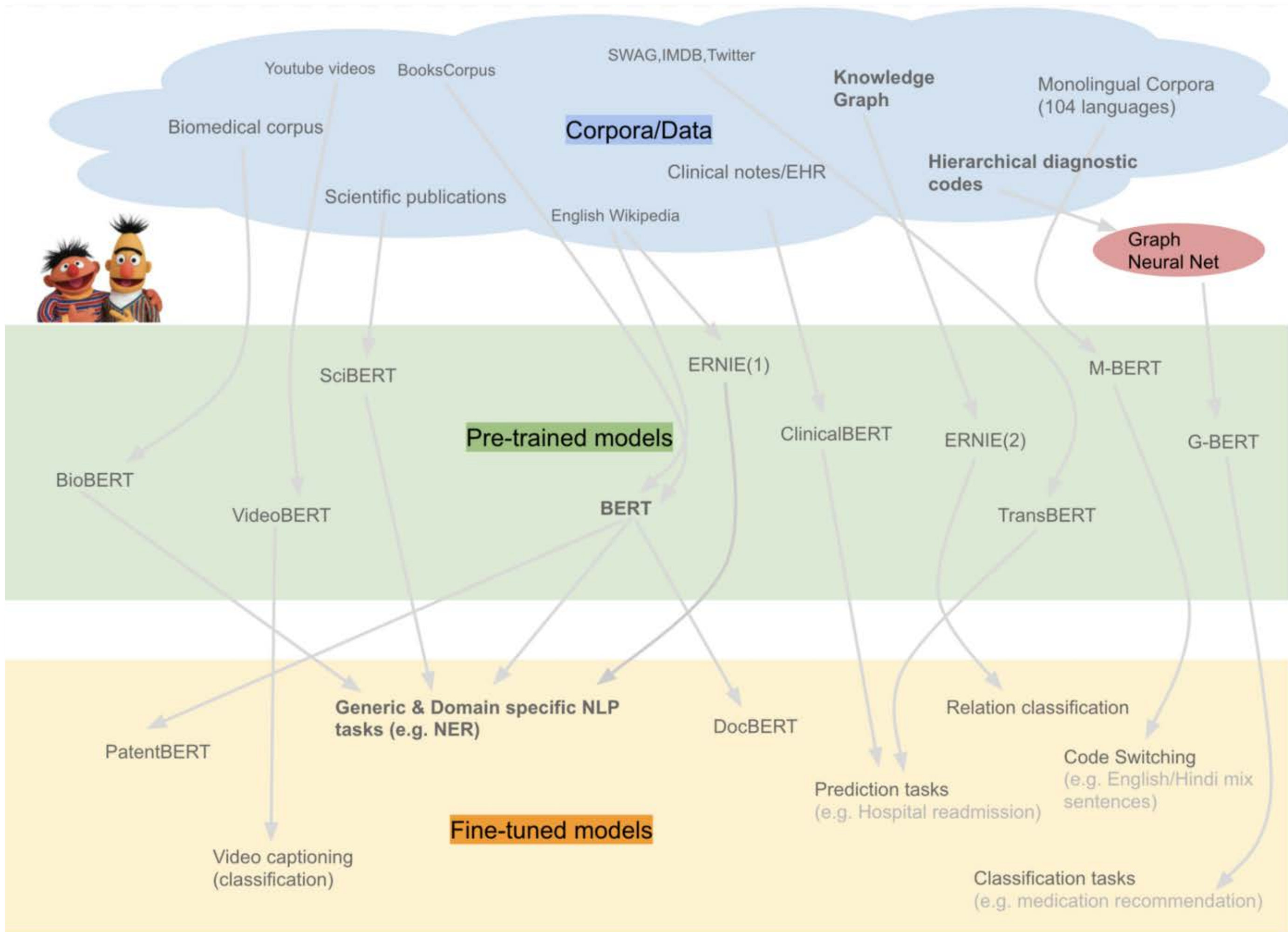
Examples (Python) with Rubrix

- Rubrix is an open-source framework
  - Created by Daniel Vila (OEG ‘s PhD)
- Examples here



Fine-tuning BERT

- Evolution and dependencies



Source: [here]

At last!! ☺

HANDS-ON 4

Questions?

Course: Intelligent Systems

Unit 4: Language Technologies

Language technologies
Part 4

Mariano Rico
2022
Technical University of Madrid