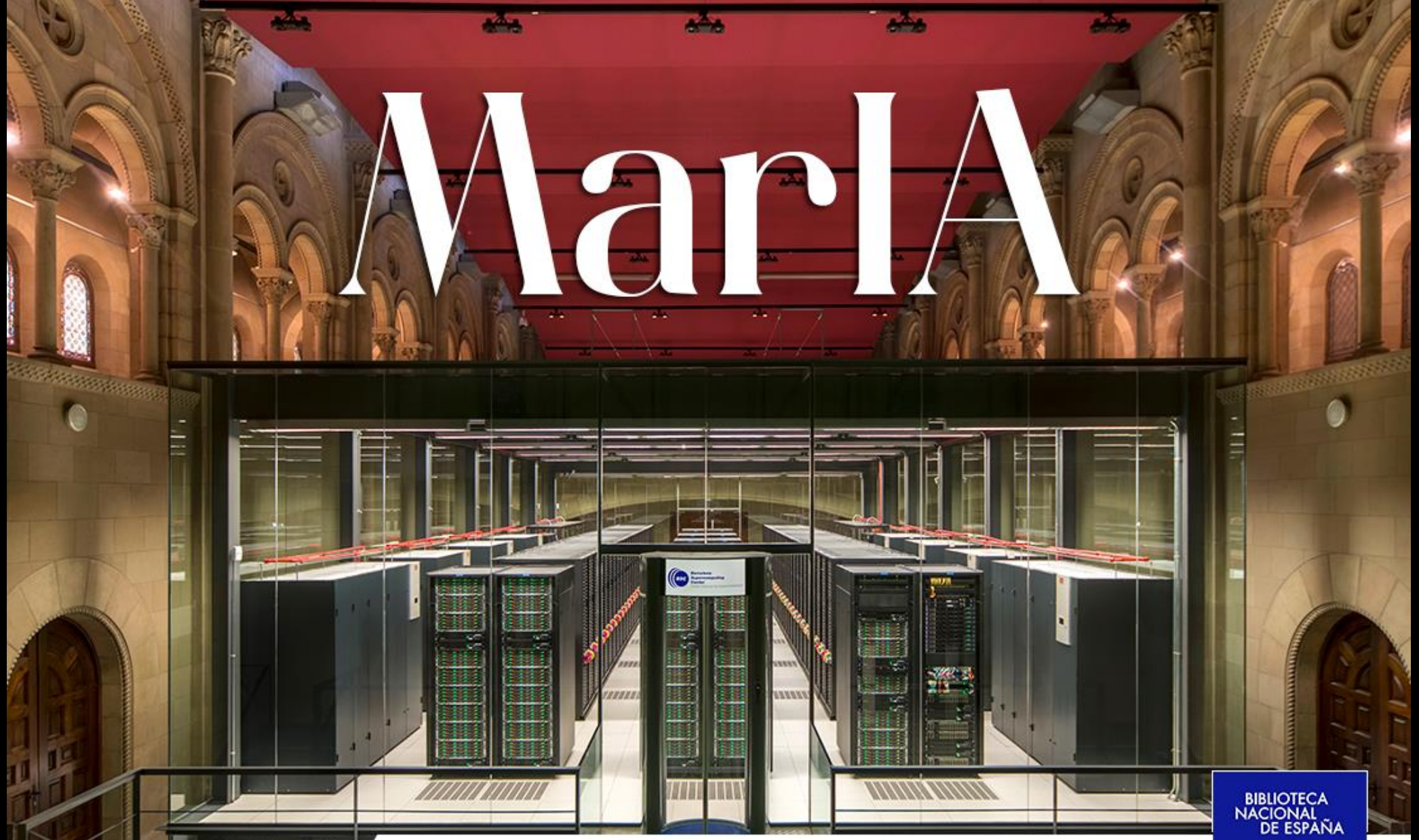
MarIA



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

BIBLIOTECA NACIONAL DE ESPAÑA
BNE

The impact of large scale language modelling. Reflections and proposals for the future.

Aitor Gonzalez-Agirre

(Slides by Marta Villegas)

Index

- What do we do?
- NLP
- DNN revolution

WHAT WE DO

PlanTL..... Promotion of Language Technologies (LT) in Spain
AINA..... Promotion of LT for Catalan

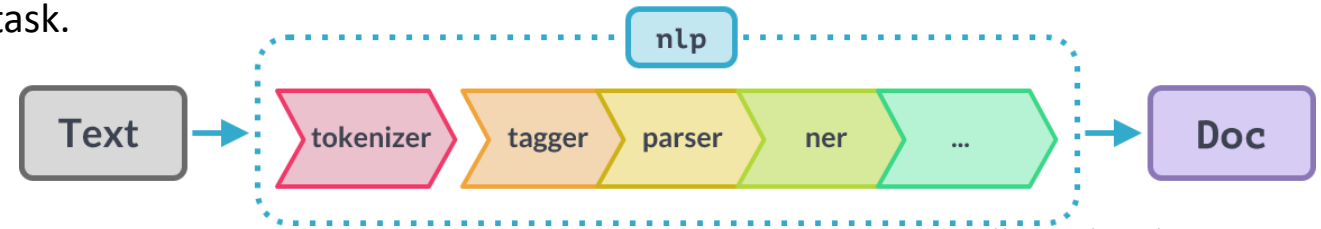
ICTUSnet / MARATO..... Information extraction in clinical reports
INTELCOMP..... Applying LT assist STI policy makers
NextProcurement..... Apply LT in public procurement (coordinators)
IBERIFIER..... Iberian hub againts disinformation

ELG / ELE..... EU initiatives about language infrastructures.

Ajuntament BCN..... Applying LT in citizen participation services.
PRESTO..... Chatbot to detect psichiatric problems in medical staff

NLP

- Natural language processing (NLP) is the subfield of artificial intelligence concerned with language and has the goal of giving computers the ability to understand and generate human language in the same way human beings can.
- Until very recently, NLP applications were implemented as pipelines of different components, each performing a specific subtask.

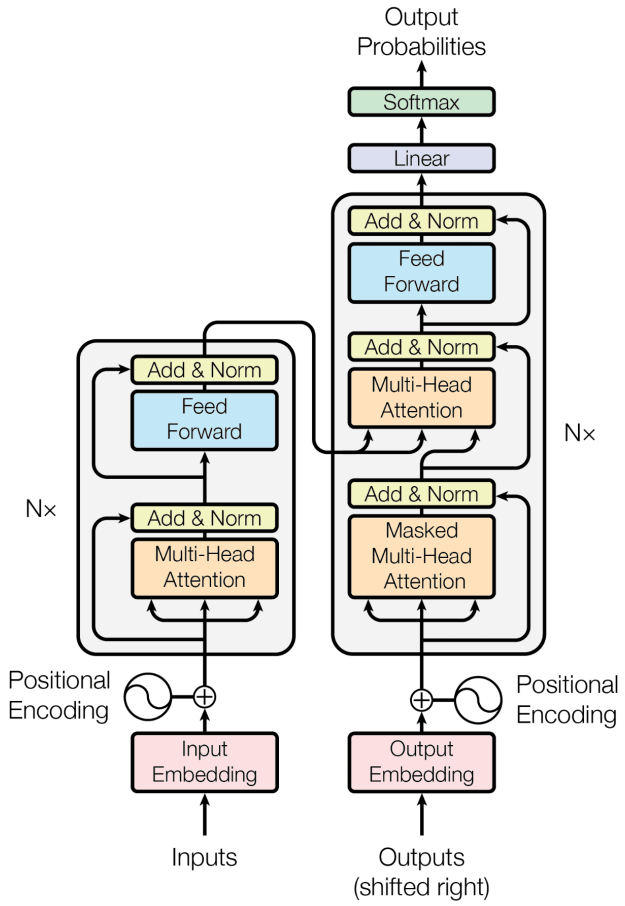


https://spacy.io/usage/processing-pipelines

- Normally, one choose one framework from the ones available (there were many: UIMA, GATE, NLTK, Stanford, OpenNLP ...) and build a processing pipeline, ideally reusing or adapting existing components.
- Machine learning was essentially based in feature engineering.
- When you work with pipelines you deal with workflow orchestration, interoperability, type system,…
- The main objective of the Plan-TL was the development of a platform on which components could be shared and executed.

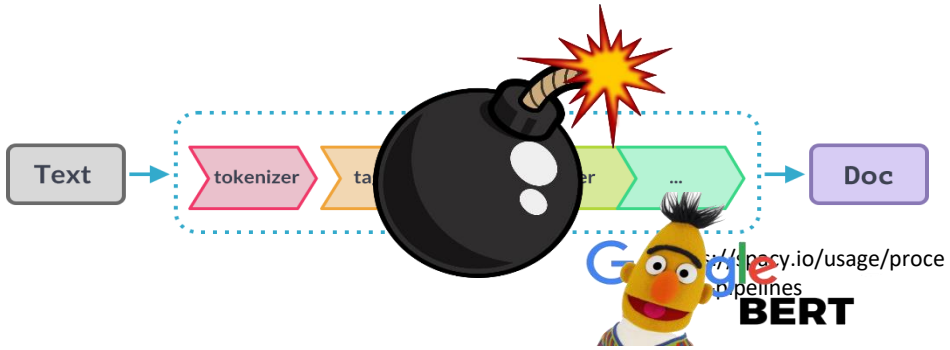
Attention Is All You Need

Then, in 2017, a new network architecture is presented:

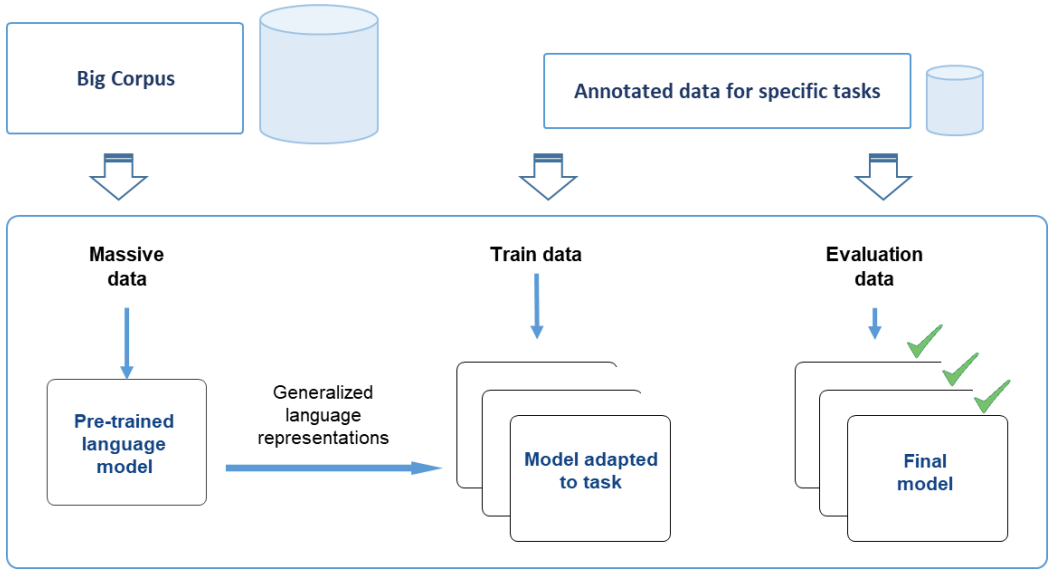


DNN revolution

In 2018, **Google** published a **transformer** model (**BERT**) and everything changed in the NLP domain.



The use of DNN (GPUs) and deep learning in NLP were like an earthquake, we stopped talking about pipelines, components and interoperability to talk about pre-trained models, self-supervised learning, fine-tuning, end-to-end learning, transfer learning.



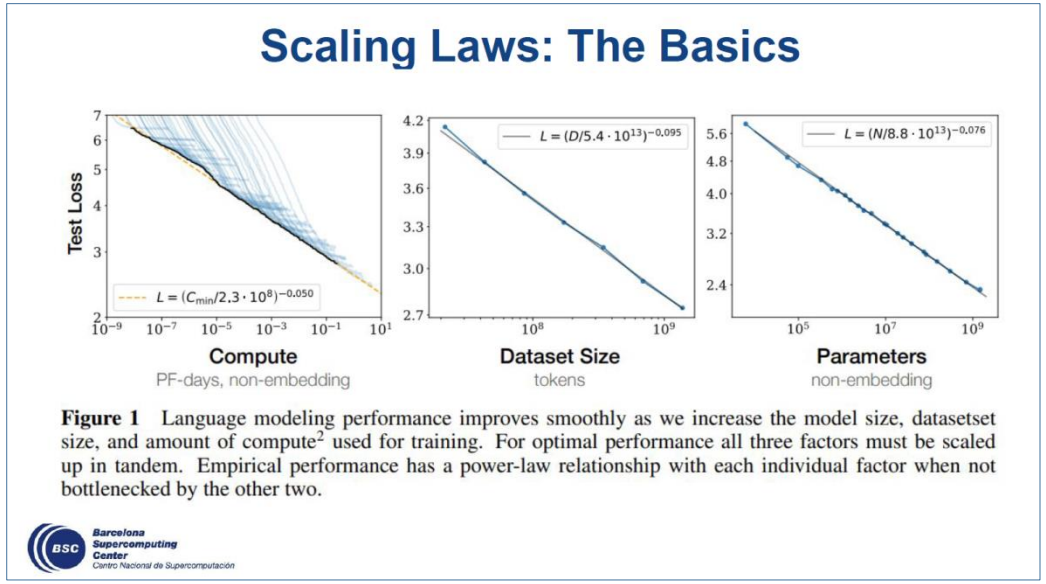
DNN revolution

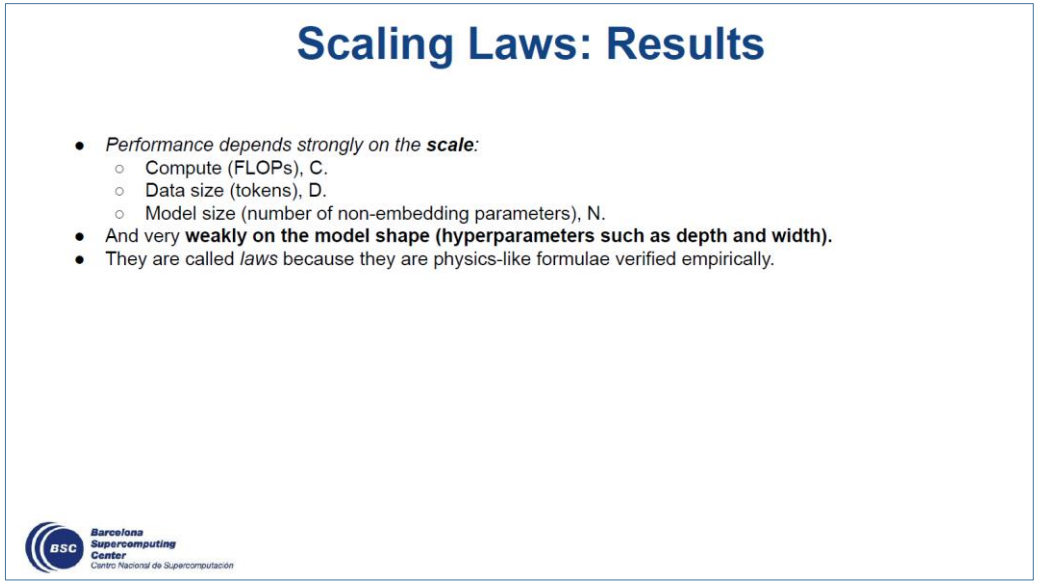
- Transformers allow for **scale** and **homogenization**.
- Transformers **outperform any previous models** and pipelines that were developed to solve a specific tasks and, in addition, allow for **transfer learning.**
- **Transformers Are Not Only for NLP** (https://medium.com/swlh/transformers-are-not-only-for-nlp-cd837c9f175) Vision, Biology, Chemistry and other Sciences.
    - DeepMind has developed **AlphaFold** to solve the task for predicting Protein 3D structure

DNN revolution

In **2019, OpenAI** announced its **GPT3** model: huge model trained on a massive data to simply predict the next word in a text.

GPT3 was an experiment in scale, that demonstrated that major gains are achieved by just scaling up the model size, the amount of data and the training time.





Jordi Armengol's talk about Language Models at Scale

DNN revolution

We move from fine-tuning to few-shot learning

[GPT-3: Language Models are Few-Shot Learners]

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

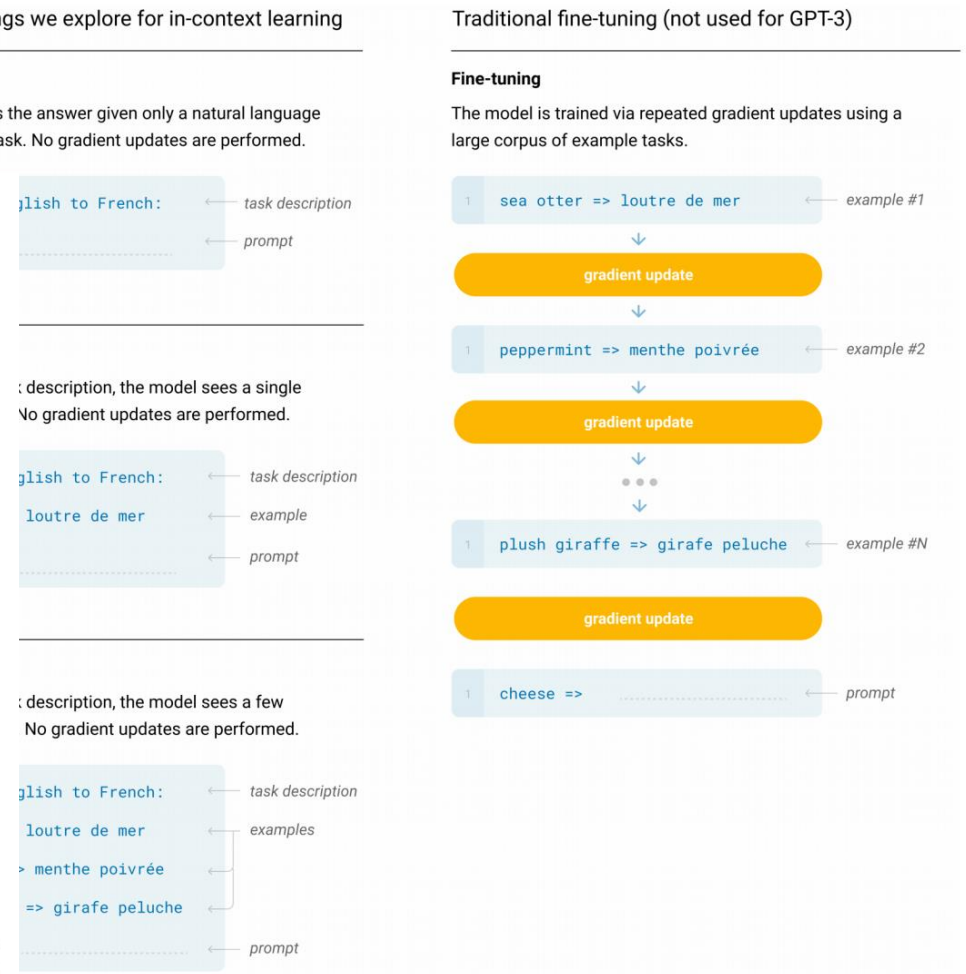
To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.
A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:
In our garage we have a Burringo that my father drives to work every day.

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.

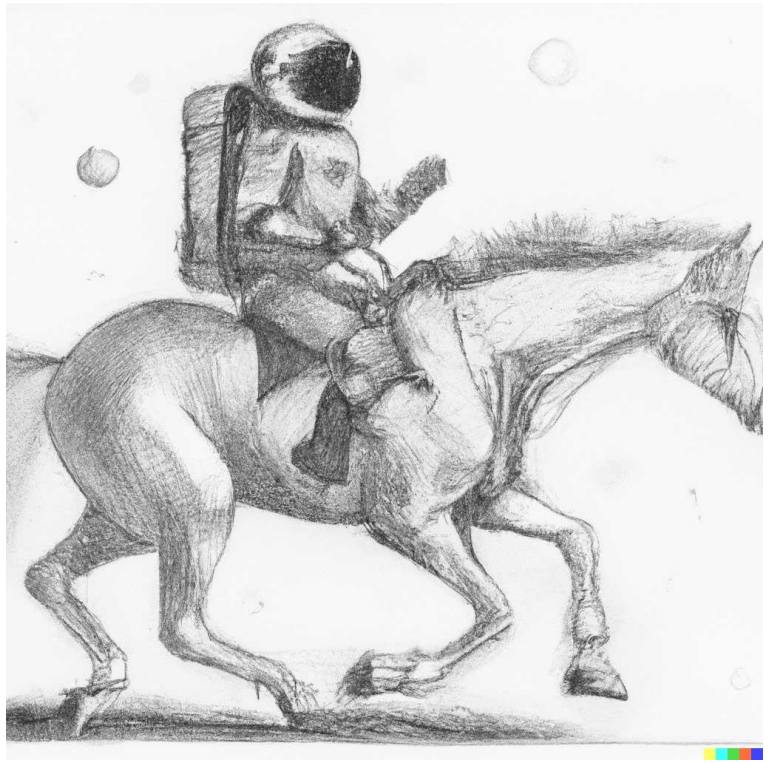
Figure 3.16: Representative GPT-3 completions for the few-shot task of using a new word in a sentence. Boldface is GPT-3's completions, plain text is human prompts. In the first example both the prompt and the completion are provided by a human; this then serves as conditioning for subsequent examples where GPT-3 receives successive additional prompts and provides the completions. Nothing task-specific is provided to GPT-3 other than the conditioning shown here.



https://blog.inten.to/gpt-3-language-models-are-few-shot-learners-a13d1ae8b1f9

OpenAI

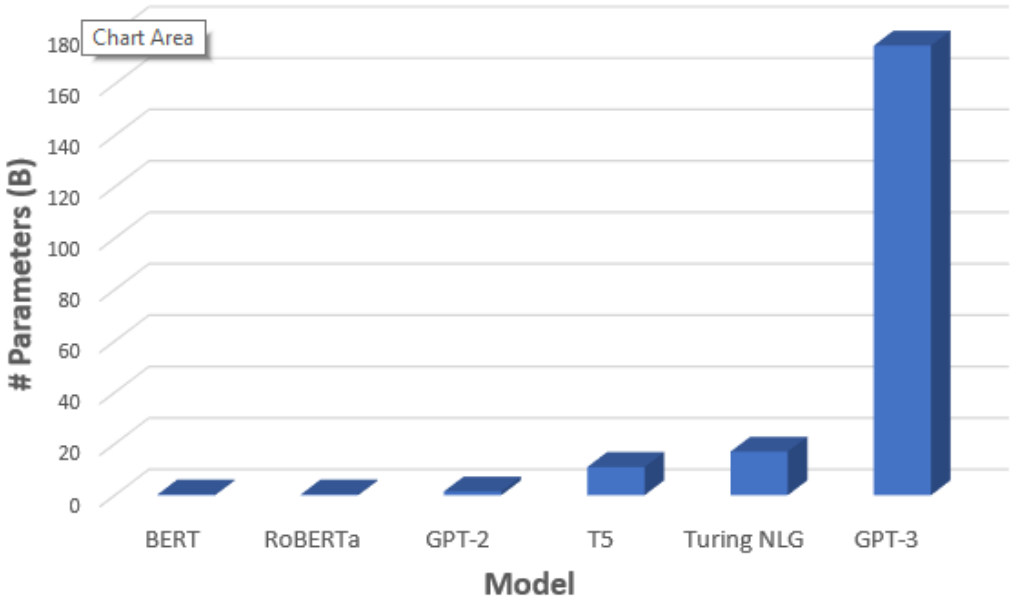
An astronaut riding a horse in a photorealistic style



An astronaut riding a horse as a pencil drawing

DNN revolution

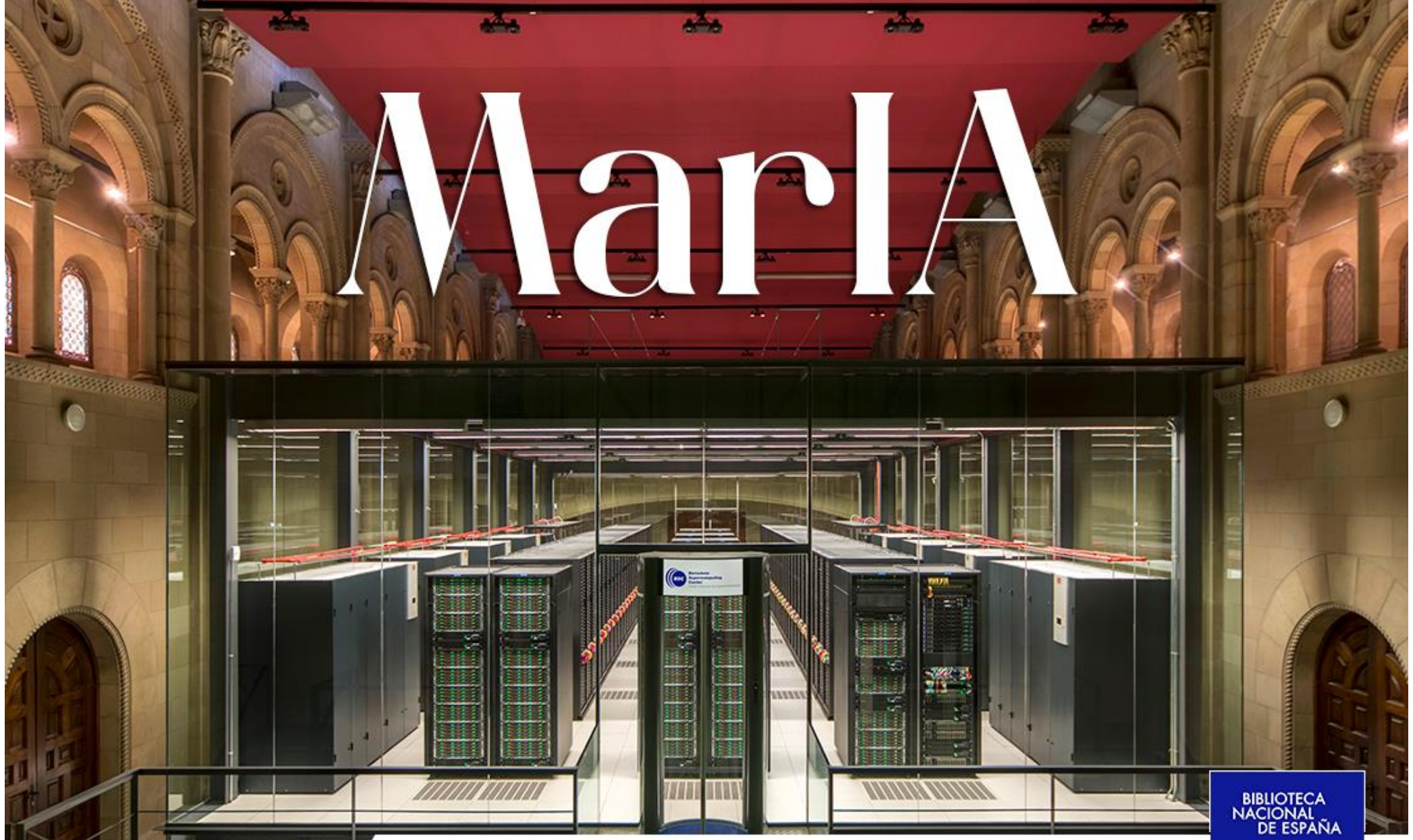
Parameters



Trainig data

Model	Year	Training data
BERT	2018	16 GB
RoBERTa	2019	161 GB
GPT2	2019	40 GB
GPT3	2020	950 GB

MarIA



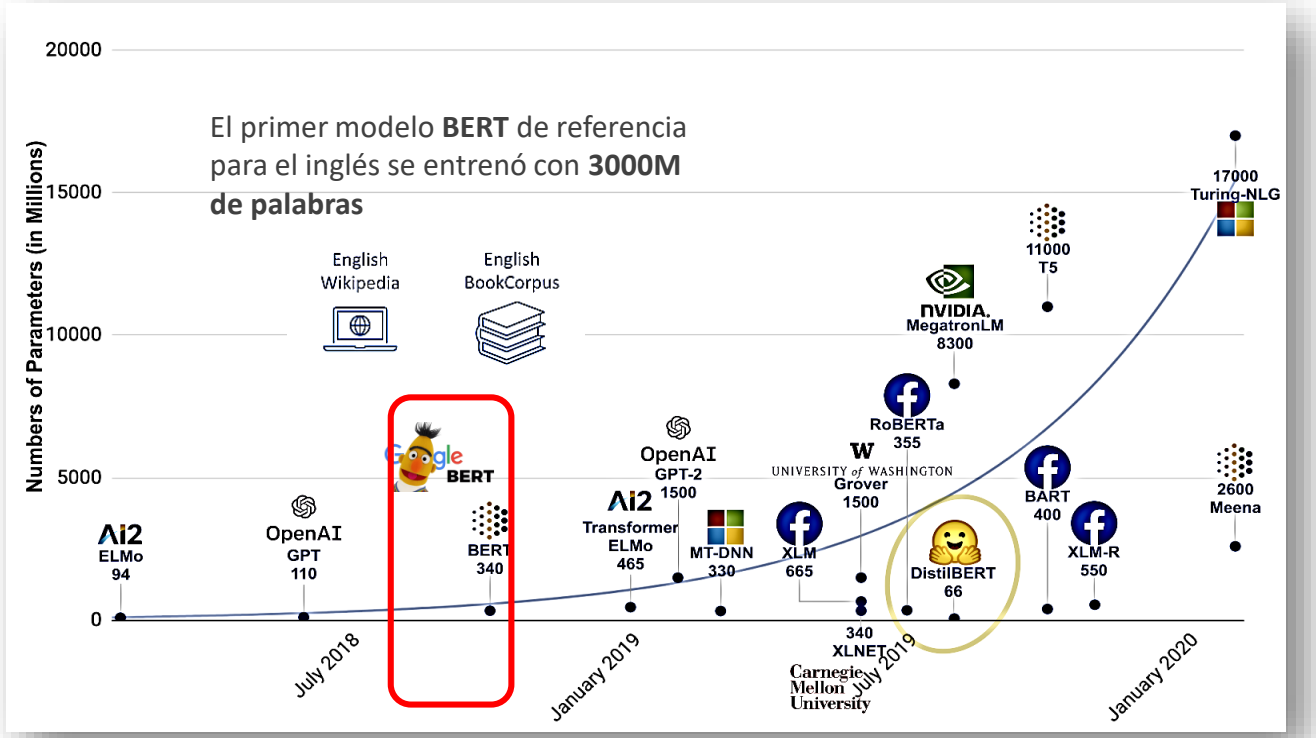
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

BIBLIOTECA NACIONAL DE ESPAÑA
BNE

Redes neuronales, un cambio de paradigma en las TL

- La arquitectura neuronal multicapa ha supuesto un cambio de paradigma y una **evolución disruptiva** en la IA y las TL
- Desde primer BERT de Google a finales del 2018, ha habido una **explosión de modelos y arquitecturas**.
- Estos modelos permiten generar aplicaciones de **mayor calidad** y a **menor coste**.
- Para generar estos models se necesitan **datos masivos y de calidad**.

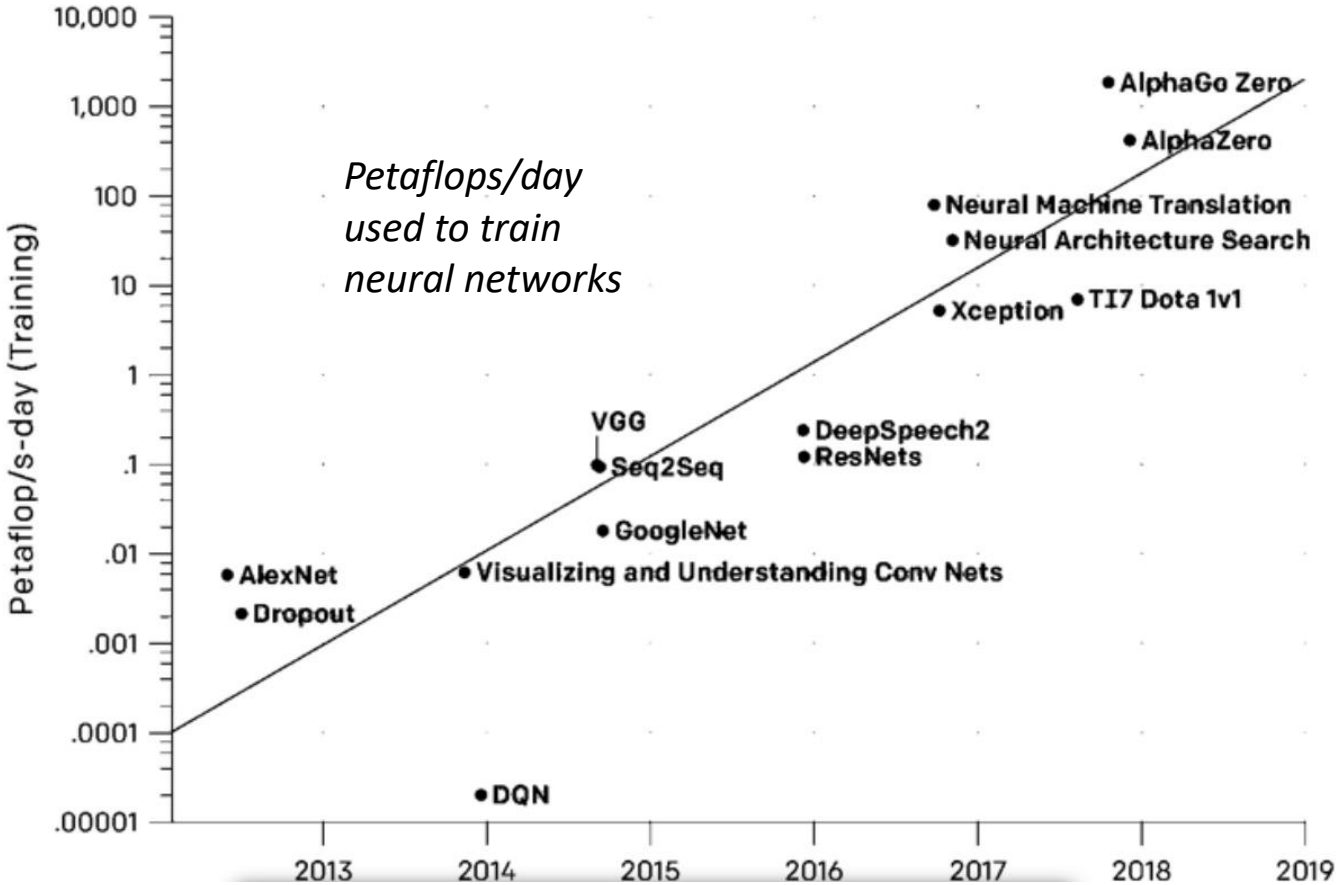


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Redes neuronales, un cambio de paradigma en las TL

Entrenar sistemas de IA es muy costoso

Since GPUs were first used in AI (2012), **computing power** available to generate AI models has increased exponentially – and improvements in computing power has been key for **AI progress**.



OpenAI https://blog.openai.com/ai-and-compute/

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

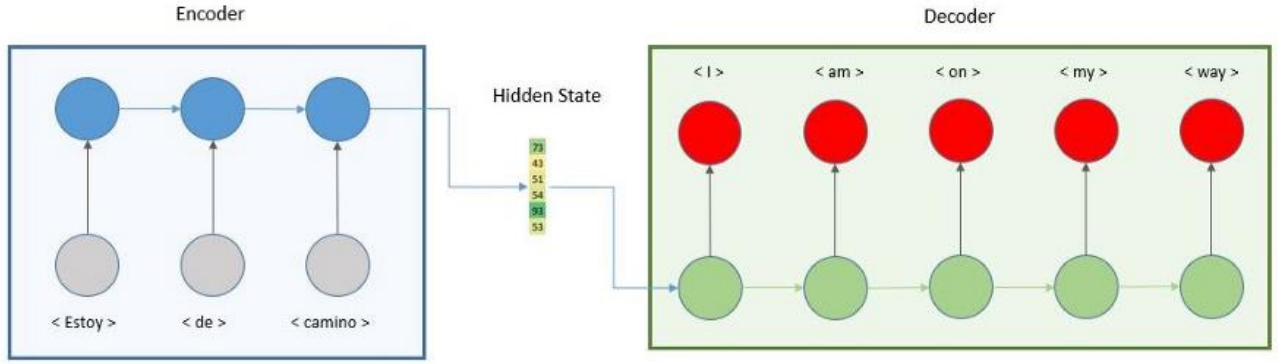
Modelo del lenguaje

Los **modelos del lenguaje** se han convertido en un componente clave en el ámbito de la IA y el Procesamiento del Lenguaje Natural.

Un **modelo de la lengua** es la distribución de la probabilidad de las palabras, de los caracteres e incluso de las oraciones de una lengua y lo usamos cada día en todo tipo de aplicaciones y servicios, desde los más básicos, como correctores ortográficos o predictores de la escritura, hasta los asistentes de voz

Desde el lanzamiento de **BERT** de Google en octubre de 2018, los modelos basados en **Transformers** han revolucionado el estado del arte.

La arquitectura **Transformer** se diseñó originalmente para la traducción siguiendo el concepto "sequence to sequence" (seq2seq). Constan de un **Encoder**, una **representación intermedia** y un **Decoder**.



https://towardsdatascience.com/what-is-an-encoder-decoder-model-86b3d57c5e1a

Del modelo Seq2Seq surgieron los modelos

- Encoders: BERT, RoBERTa, utilizados en clasificación de textos, NER, QA,…
- Decoders: GPT, utilizados en la generación de texto
- Encoder-decoder: BART, T5 utilizados en traducción automática, resumen automático,…

Modelo del lenguaje

Los **modelos del lenguaje** se han convertido en un componente clave en el ámbito de la IA y el Procesamiento del Lenguaje Natural.

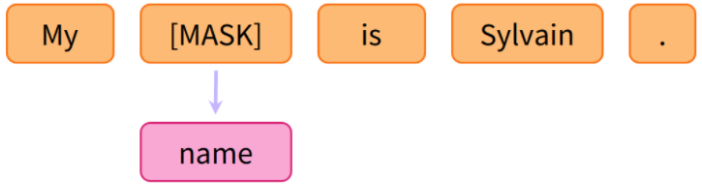
Desde el lanzamiento de BERT de Google en octubre de 2018, los modelos basados en **Tansformers** han revolucionado el estado del arte.

La arquitectura **Transformer** se diseñó originalmente para la traducción.

- La mayor evolución en el modelo de Transformers es la capacidad de paralelizar el proceso de entrenamiento. En los modelos anteriores, para entrenar un modelo para mapear una oración A a una oración B se procesaban todas las palabras de A y B secuencialmente con redes recurrentes.
- El segundo punto fuerte es la capacidad de obtener modelos pre-entrenados con muchos datos y, luego, adaptarlos a tareas específicas con ‘pocos’ datos.. (**Aprendizaje por transferencia o transfer learning**)

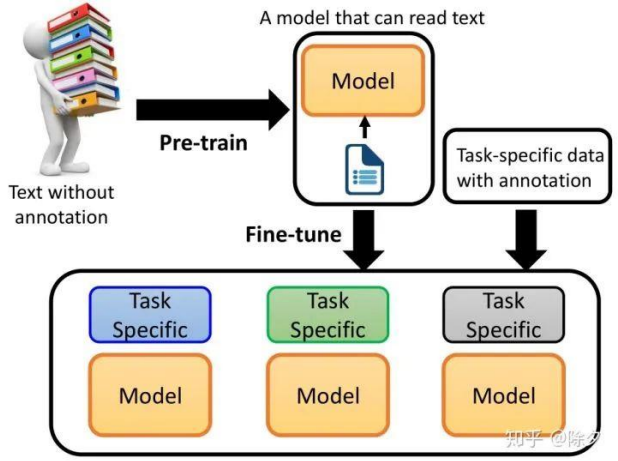
Modelo del lenguaje

Los modelos se entrenan de manera **no supervisada** utilizando datos masivos y la técnica de **masking**, el sistema tiene que adivinar la palabra 'enmascarada' utilizando el contexto.



La mayor evolución en el modelo de Transformers es la **capacidad de paralelizar** el proceso de entrenamiento. En las arquitecturas anteriores, para entrenar un modelo se procesaban las palabras secuencialmente con redes recurrentes. Los Transformes permiten procesar secuencias de palabras a la vez.

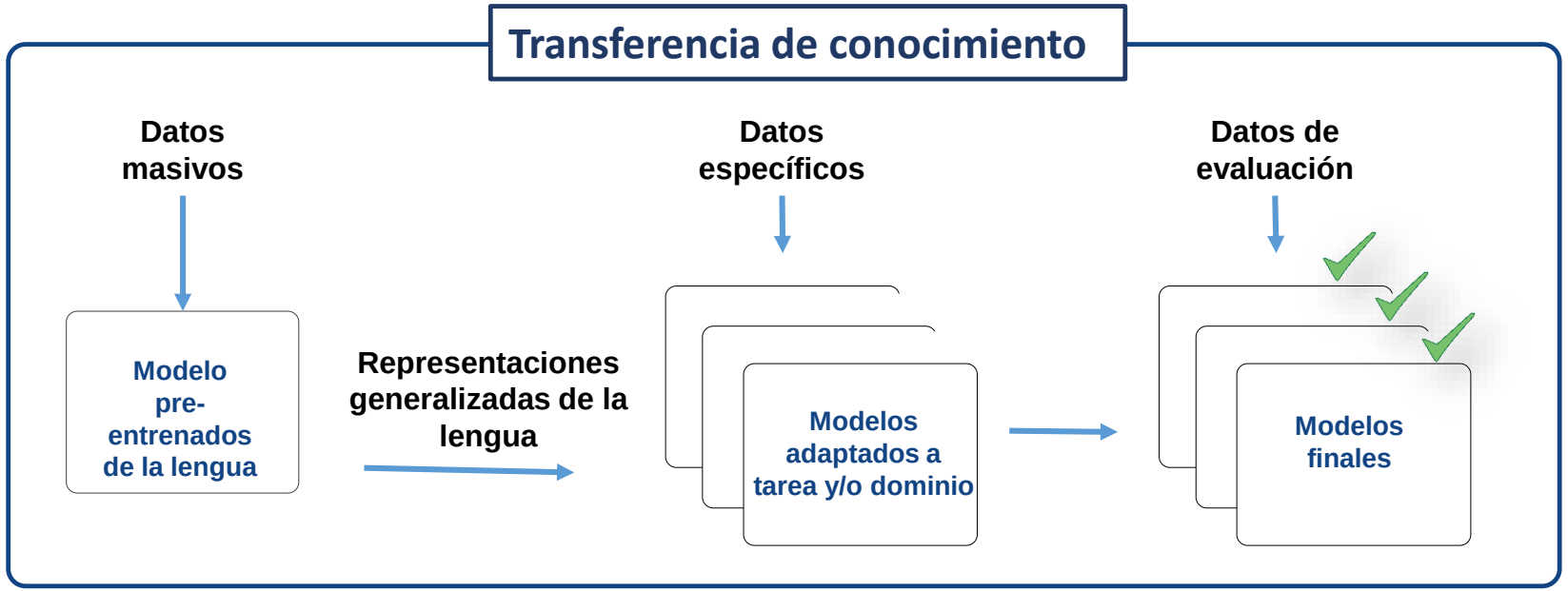
Otra característica es la capacidad de obtener modelos pre-entrenados con muchos datos y, luego, adaptarlos a tareas específicas con 'pocos' datos. (**Aprendizaje por transferencia** o transfer learning)



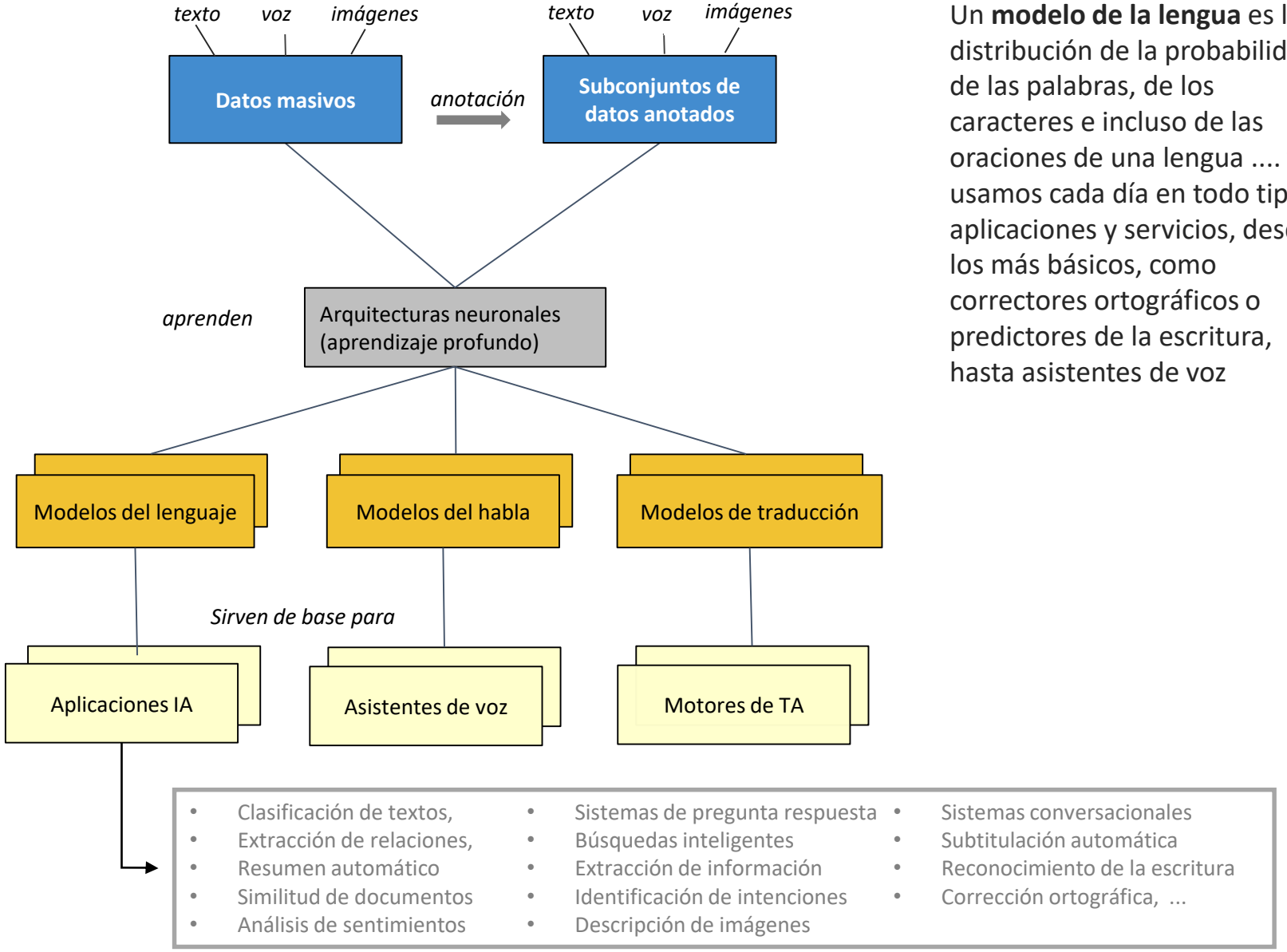
https://www.programmersought.com/article/17834625815/

Modelo del lenguaje

- Obtener datos masivos de calidad es **costoso en tiempo y recursos**.
- Afortunadamente, disponemos de métodos de **transferencia del conocimiento** que permiten reutilizar los modelos pre-entrenados con datos masivos para adaptarlos a nuevas tareas o dominios; evitando tener que reentrenar un modelo desde cero.
- Esto sopone un **ahorro en recursos, tiempo y energía** y la **reutilización** de modelos.



¿Un modelo de la lengua?



Un **modelo de la lengua** es la distribución de la probabilidad de las palabras, de los caracteres e incluso de las oraciones de una lengua .... y lo usamos cada día en todo tipo de aplicaciones y servicios, desde los más básicos, como correctores ortográficos o predictores de la escritura, hasta asistentes de voz

Corpus

- **Corpus Biblioteca Nacional de España**. 60% de los datos están ya limpios: 120B (billones americanos) de tokens. Este será, de largo, el mayor corpus del Español: Beto (3B tokens), OSCAR (26B tokens)
- **Corpus Biomédico**, compilado, limpio y deduplicado (972M tokens). Modelo ya generado utilizando la arquitectura RoBERTa
- **Corpus del catalán**, compilado, limpio y deduplicado (1.770M tokens). Modelo ya generado utilizando la arquitectura RoBERTa
- Pipeline de preproceso-limpieza finalizada: https://github.com/temu-bsc/corpus-cleaner

- Compilación y limpieza de datos textuales.

  - Pre-procesados
  - Limpios
  - Deduplicados

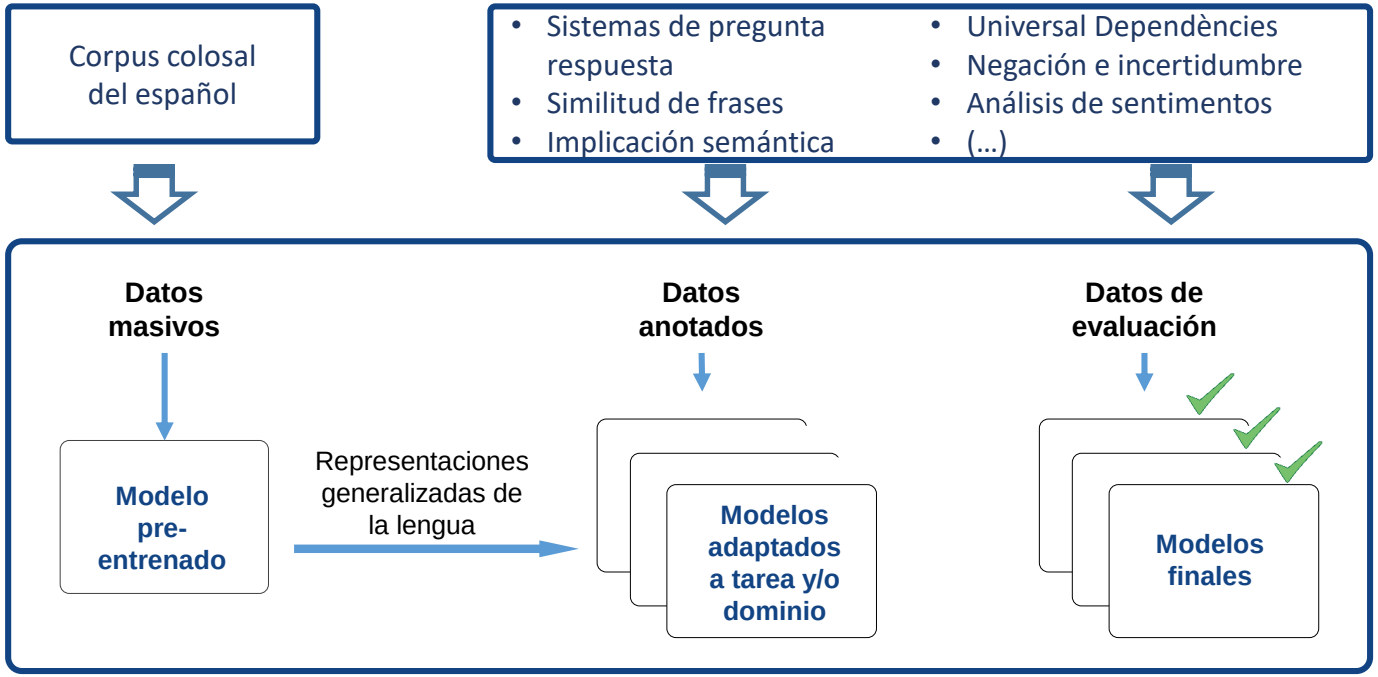
- La tecnología avanza rápidamente, con constantes innovaciones y mejoras.
- Disponer de datos suficientes garantiza la actualización: la tecnología cambia pero los datos son permanentes

Corpus biomédico

Corpus name	Text Size (GB)	Final size (GB)	Raw tokens	Cleaned tokens	Num. sentences
Clinical cases cardiology	0.0035	0,001	149,904.00	147,790.00	9,970.00
Clinical cases radiology	0.0066	0,001	177,366.00	170,997.00	9,948.00
libros_casos_clinicos	0.0083	0,007	1,137,555.00	1,024,797.00	68,833.00
Clinical cases COVID	0.0084	0,001	82,201.00	82,091.00	3,896.00
EMEA corpus	0.087	0,034	13,797,362.00	5,377,448.00	284,575.00
Patents	0.087	0,084	14,022,520.00	13,463,387.00	253,924.00
wikipedia_life_sciences	0.172	0,088	18,771,176.00	13,890,501.00	832,027.00
barr2_background	0.188	0,159	28,868,022.00	24,516,442.00	1,029,600.00
Pubmed	0.211	0,013	1,957,479.00	1,858,966.00	103,674.00
REEC (casos clínicos)	0.823	0,028	4,581,755.00	4,283,453.00	220,726.00
mespen_medline	1.2	0,38	6,864,901.00	4,166,077.00	322,619.00
pdfs_general	3.3		9,124,996.00	7,146,139.00	5,252,481.00
Scielo	3.891	0,631	61,837,972.00	60,007,289.00	2,668,231.00
Medical crawler	606	4,5	?	746,368,185.00	32,766,976.00
TOTAL	**615.9858**	**5,927**	**261,373,209.00**	**972,503,562.00**	**43,827,480.00**

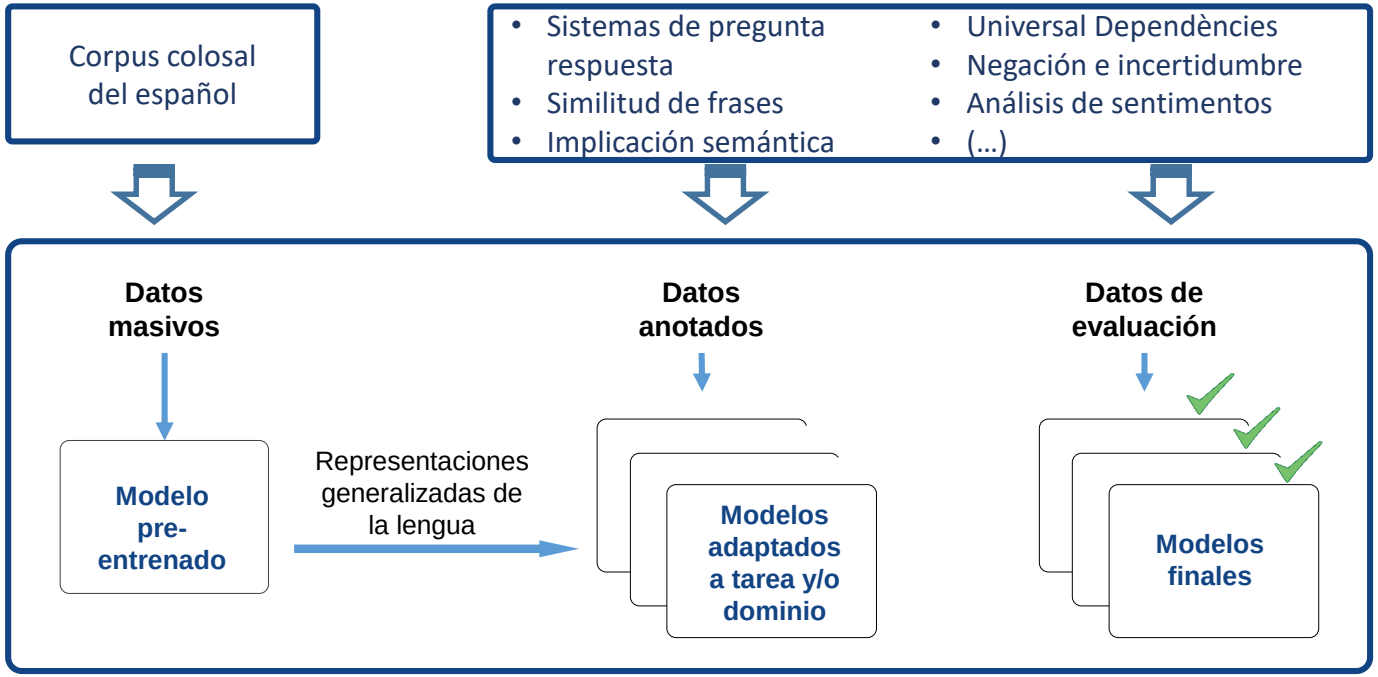
Datos anotados

- Se generan datos anotados, suficientes y de calidad, para poder entrenar y evaluar sistemas en las tareas más habituales y críticas.
- El objetivo es poner al alcance de la industria y la investigación los recursos y modelos de lengua necesarios para que puedan desarrollar aplicaciones inteligentes.



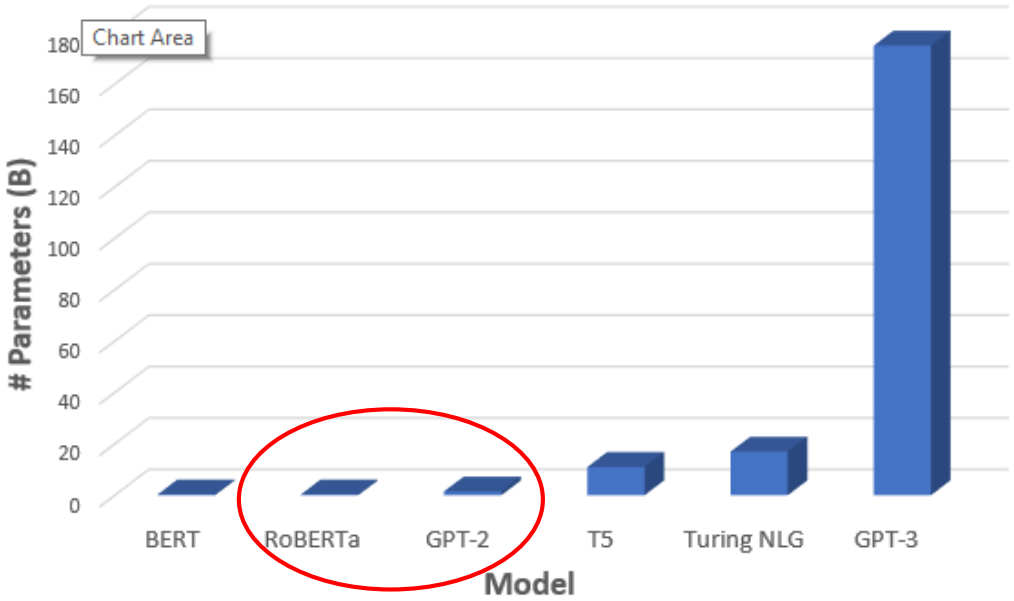
Datos anotados

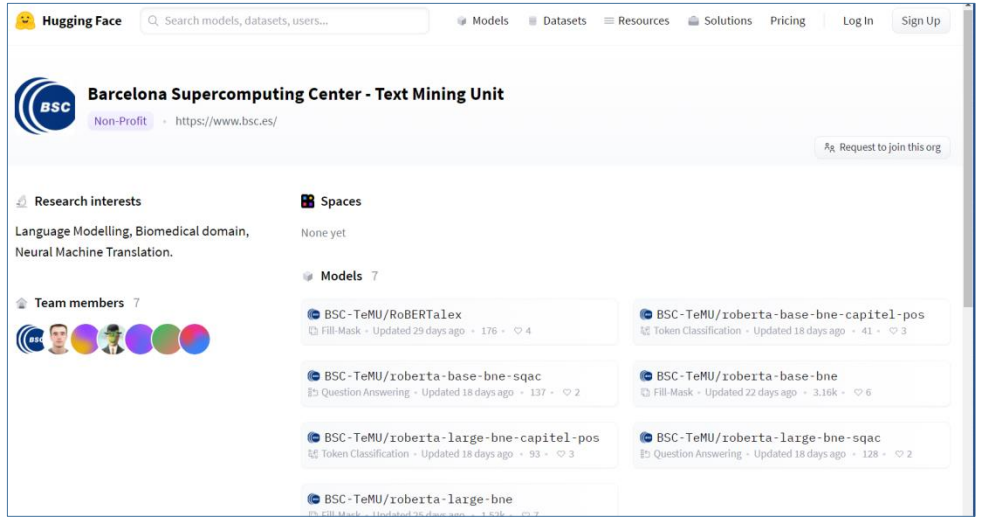
- Se generan datos anotados, suficientes y de calidad, para poder entrenar y evaluar sistemas en las tareas más habituales y críticas.
- El objetivo es poner al alcance de la industria y la investigación los recursos y modelos de lengua necesarios para que puedan desarrollar aplicaciones inteligentes.



Models at TeMU

- General Spanish models: RoBERTA base and RoBERTa large and GPT2 trained with 570GB of data from BNE.
- General Catalan models: RoBERTa base
- Domain specific models: RoBERTa base for biomedical, clinical and legal domains.



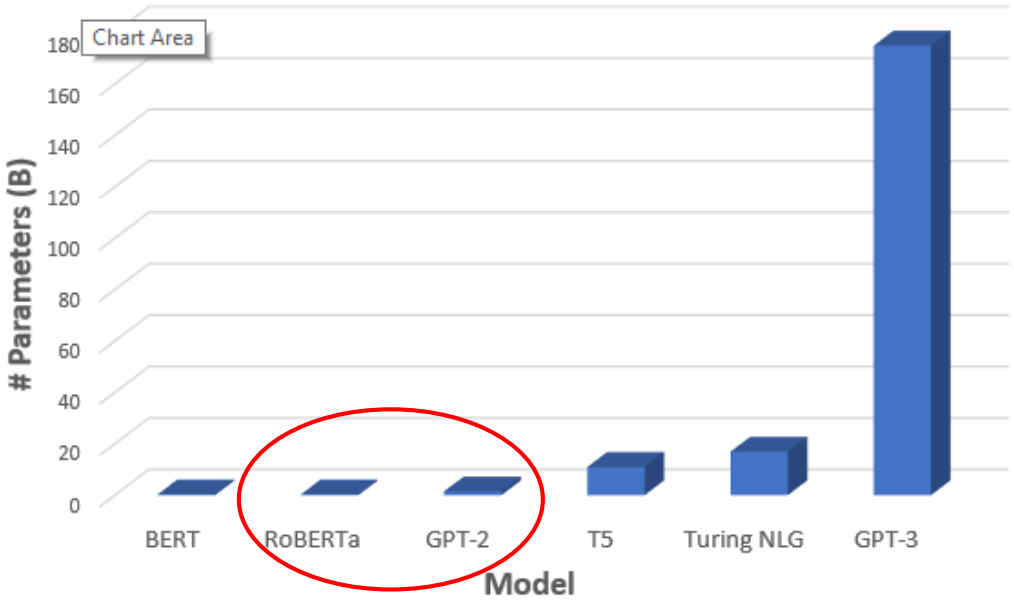


https://huggingface.co/BSC-TeMU

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Models at TeMU

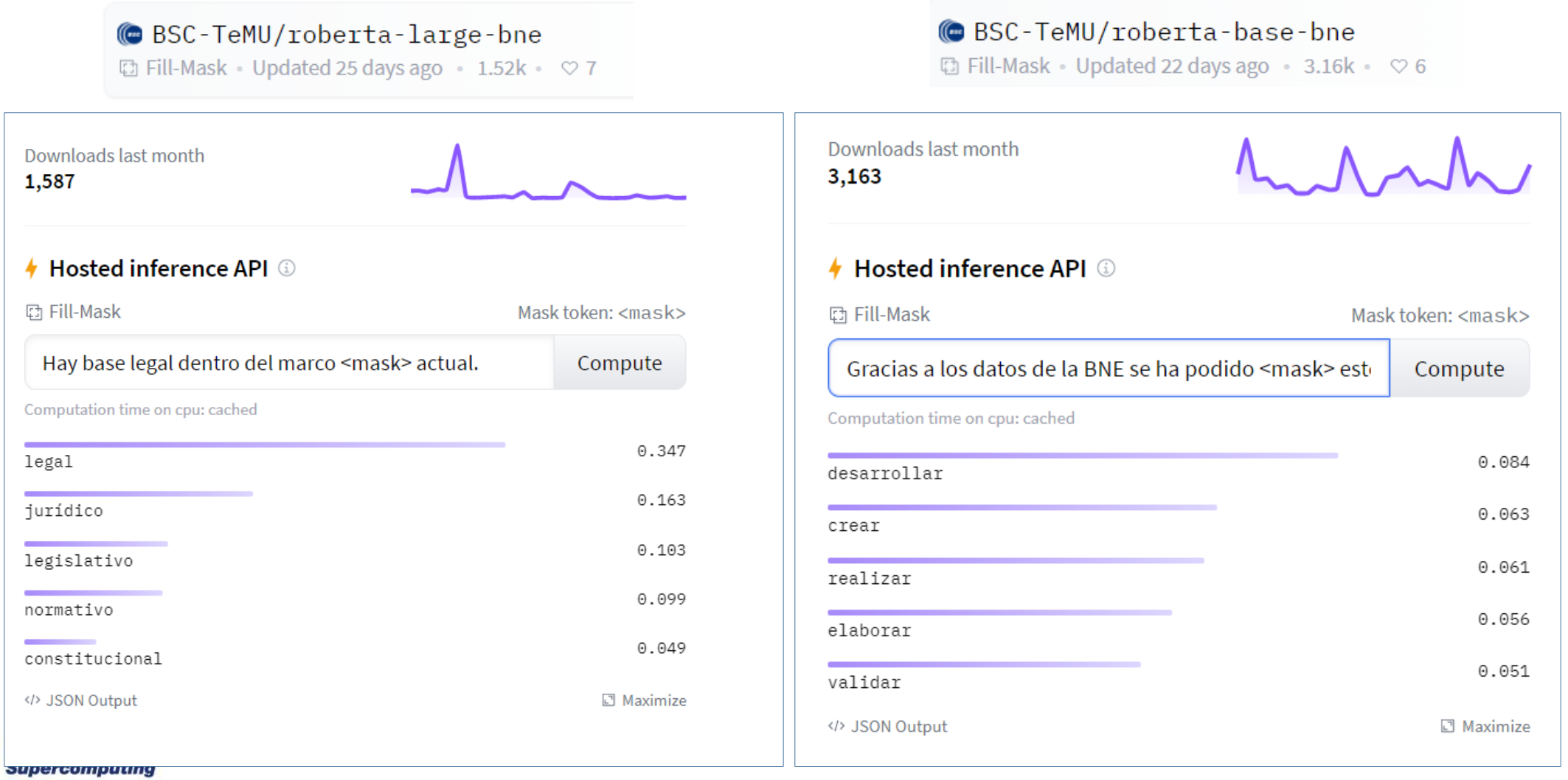
- General Spanish models: RoBERTA base and RoBERTa large and GPT2 trained with 570GB of data from BNE.
- General Catalan models: RoBERTa base
- Domain specific models: RoBERTa base for biomedical, clinical and legal domains.



Resources used:

- **Cleaning the BNE crawler (59TB)** we used 100 nodes, each with 48 CPU cores, of MareNostrum 4 during 96 hours, obtaining a total of 570GB of high quality data.
- **Training RoBERTa base**: 48 hours with 16 computing nodes each with 4 NVIDIA V100 GPUs of 16GB VRAM.
- **Training RoBERTa large**: 96 hours with 32 computing nodes each with 4 NVIDIA V100 GPUs of 16GB VRAM.
- **Training GPT2**: 10 days with 32 computing nodes each with 4 NVIDIA V100 GPUs of 16GB VRAM.

Models at TeMU



Supercomputing Center
Centro Nacional de Supercomputación

Models at TeMU

Fine-tuned models

BSC-TeMU/roberta-base-bne-capitel-pos

Token Classification • Updated 18 days ago • 41 • ♡ 3

Downloads last month

41

Hosted inference API

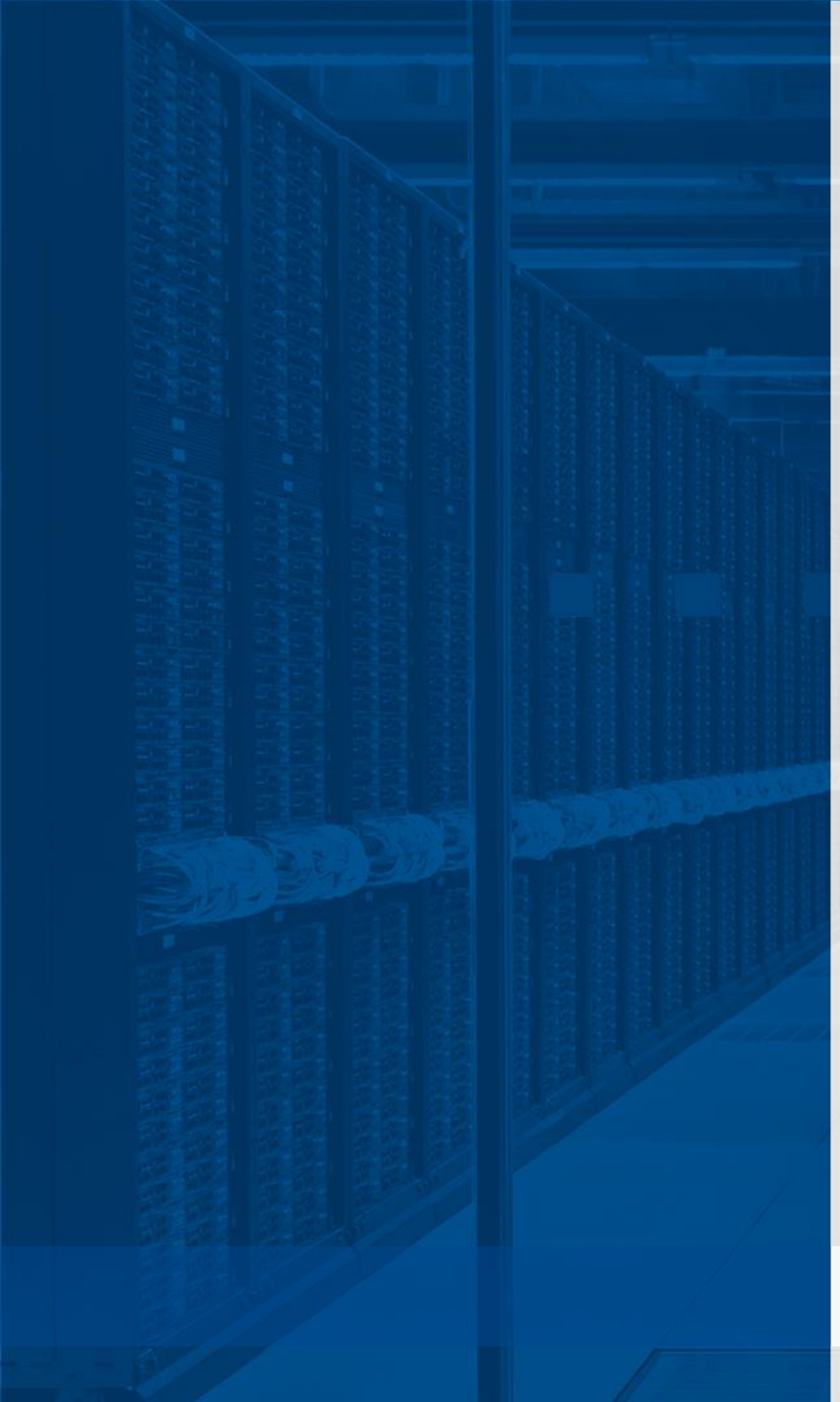
Token Classification

El Tribunal Superior de Justicia se pronunció ayer: "Ha | Compute

Computation time on cpu: cached

El DET Tribunal NOUN Superior ADJ de ADP Justicia NOUN se PRON pronunció VERB ayer ADV : PUNCT " PUNCT Hay VERB base NOUN legal ADJ dentro ADV del ADP marco NOUN jurídico ADJ actual ADJ ". PUNCT

Barcelona Supercomputing Center
Centro Nacional de Supercomputación



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Gracias!

Más información:

communication@bsc.es