Aplicaciones prácticas de NLP

Alejandro Vaca Serrano

Presentación

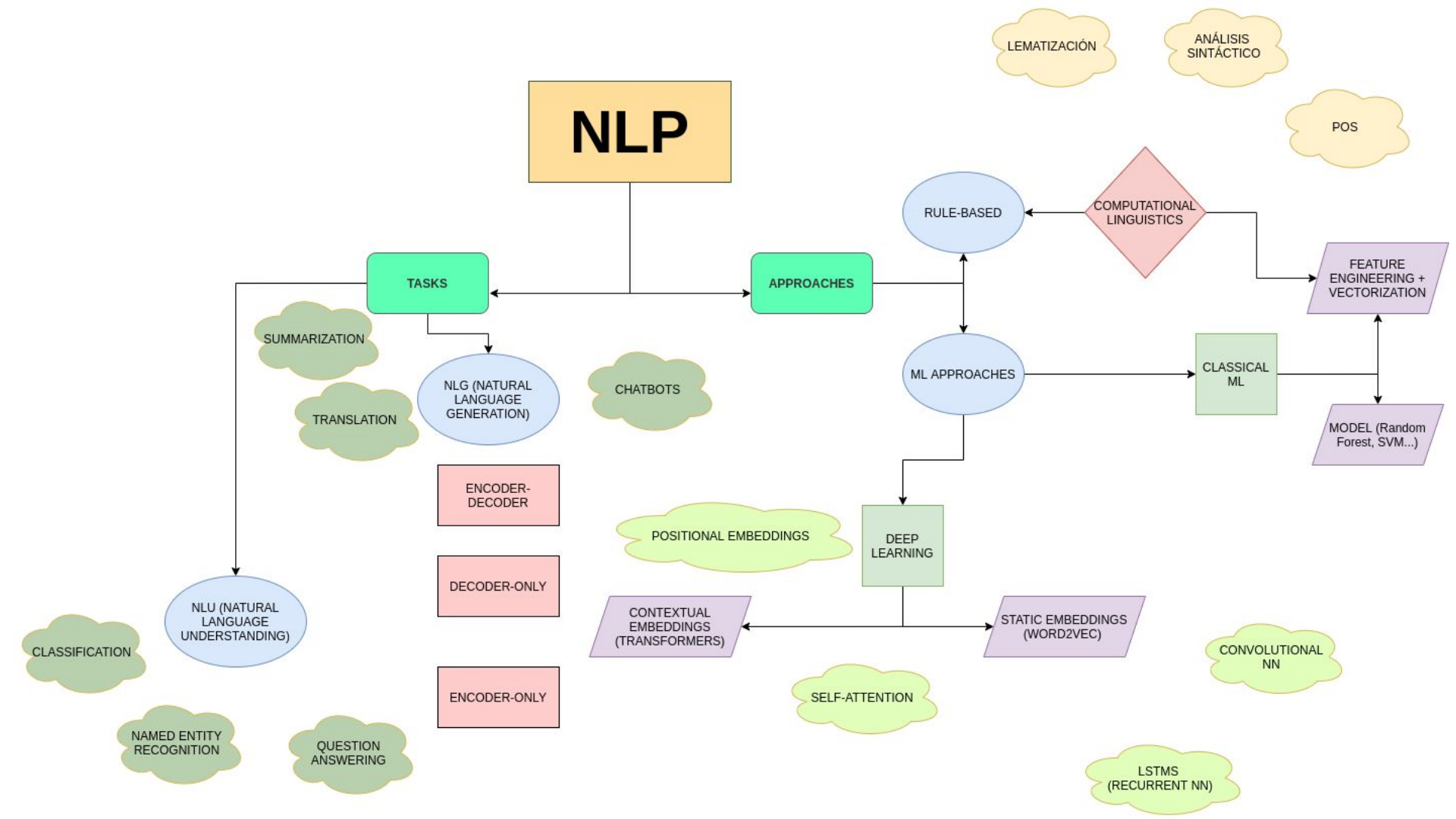
- ADE Bilingüe @ **CUNEF** 2018
- Máster en Data Science & Big Data @ **AFI** 2019
- Data Scientist en **Instituto de Ingeniería del Conocimiento** - Área de Inteligencia Artificial
- 👨‍🏫 NLP Professor - Master of Artificial Intelligence @ ESIC
- 👨‍🏫 NLP Professor - Bachelor of Data Science & Big Data @ IE University
- 👨‍🏫 Deep Learning Professor @ DataHack
- 🥇 1er Premio **Cajamar UniversityHack 2020** - reto Minsait Land Classification.
- 🏆 Premio Especial a **Mejor Data Scientist @ SpainAI Hackaton 2021**
- 🥇 1er Premio reto **Computer Vision @ SpainAI Hackaton 2021**
- 🥇 1er Premio reto **Prescriptive Analytics - Time Series @ SpainAI Hackaton 2021**
- 🥉 3er Premio reto **NLP @ SpainAI Hackaton 2021**
- 🥇 1er Premio **Hackaton SomosNLP 2022** por projecto BioMedIA
- 🏆 Premio Especial Proyecto Más Popular @ **Hackaton SomosNLP 2022**
- 🥇🥇 1er Premio Tareas 1 y 2 (2/2) del reto **EXIST2022 @ Iberlef2022**
- 🎤 Speaker @ **LREC 2022** - Marseille (France)
- 🎤 Speaker @ **NAACL 2022** (North American Association of Computational Linguistics) - Seattle (USA)

Objetivos de la Masterclass

1. Comprender mejor en qué consiste el área de NLP.
   a. Qué tareas se pueden hacer.
   b. Qué tipos de modelos hay y qué puede hacer cada uno.
2. Tomar perspectiva de la trayectoria de este campo.
3. **Comenzar a desarrollar la habilidad esencial de saber cómo unir piezas que resuelven pequeñas tareas para crear soluciones completas que resuelvan problemas complejos.**
4. Tener una visión amplia de los recursos disponibles en español (sobre todo modelos).
5. Que os guste tanto tanto el NLP, que me propongáis como profesor de esta materia al año que viene en el Máster 😎

Índice

1. Repaso rápido de modelos de lenguaje y de NLP.
2. Modelos de lenguaje en español: RigoBERTa.
3. Neuraculus: resolviendo dudas científicas sobre el COVID-19 con NLP.
4. BioMedIA: Una inteligencia artificial de voz a voz que genera respuestas a preguntas abiertas sobre biomedicina en español.


NLP
TASKS
APPROACHES
SUMMARIZATION
NLG (NATURAL LANGUAGE GENERATION)
TRANSLATION
CHATBOTS
NLU (NATURAL LANGUAGE UNDERSTANDING)
CLASSIFICATION
NAMED ENTITY RECOGNITION
QUESTION ANSWERING
ENCODER-DECODER
DECODER-ONLY
ENCODER-ONLY
RULE-BASED
COMPUTATIONAL LINGUISTICS
LEMATIZACIÓN
ANÁLISIS SINTÁCTICO
POS
FEATURE ENGINEERING + VECTORIZATION
ML APPROACHES
CLASSICAL ML
MODEL (Random Forest, SVM...)
DEEP LEARNING
POSITIONAL EMBEDDINGS
CONTEXTUAL EMBEDDINGS (TRANSFORMERS)
STATIC EMBEDDINGS (WORD2VEC)
SELF-ATTENTION
CONVOLUTIONAL NN
LSTMS (RECURRENT NN)

Revolución de The Transformer

Fig. Evolution of NLP Models

1. **BoW**: No tenemos en cuenta el orden ni la semántica.
2. **Embeddings estáticos + capas recurrentes / convolucionales**.
   a. Aprendemos vectores (embeddings) que contienen el significado de las palabras, e introducimos estos vectores como features de una red neuronal formada por estos tipos de capas.
   b. Aunque ya tenemos en cuenta el significado de cada palabra, este no se adapta en función del texto en el que aparece la palabra, los vectores son siempre iguales.
   c. Gracias a CNN y RNN, tenemos en cuenta el orden
3. **Embeddings contextuales (Transformer models)**:
   a. Los vectores de embedding se adaptan en función del contexto, el significado cambia con las palabras del contexto.
   b. No aprendemos únicamente una capa de embedding, sino todo un pipeline (una serie de capas) de contextualización.
   c. Añadimos positional encodings para tener en cuenta el orden.
   d. Introduce capas de Atención! (Acordaros: el Tinder para palabras 🔥)

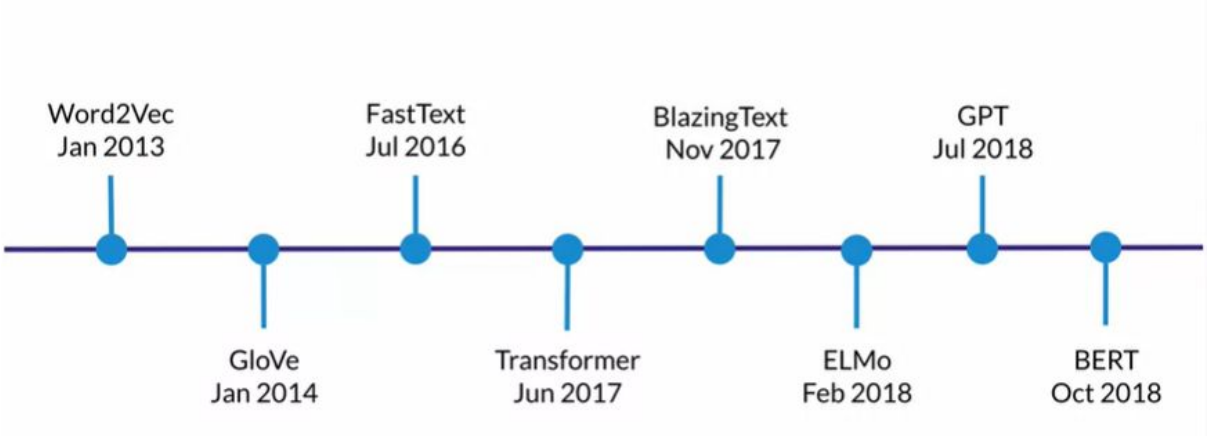
Afi Escuela

Revolución de The Transformer

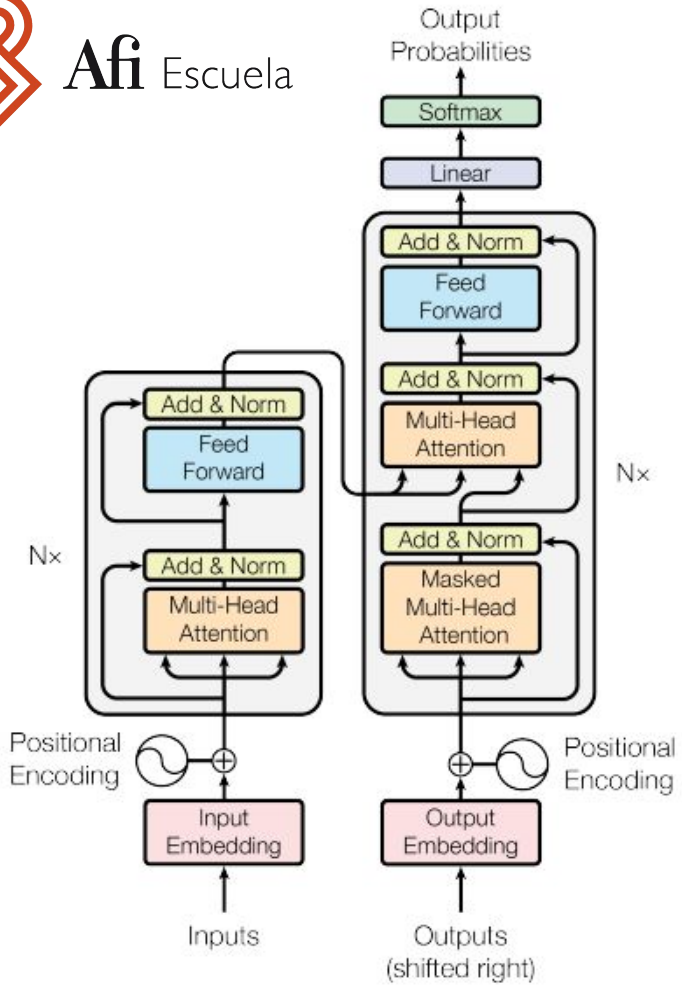
1950

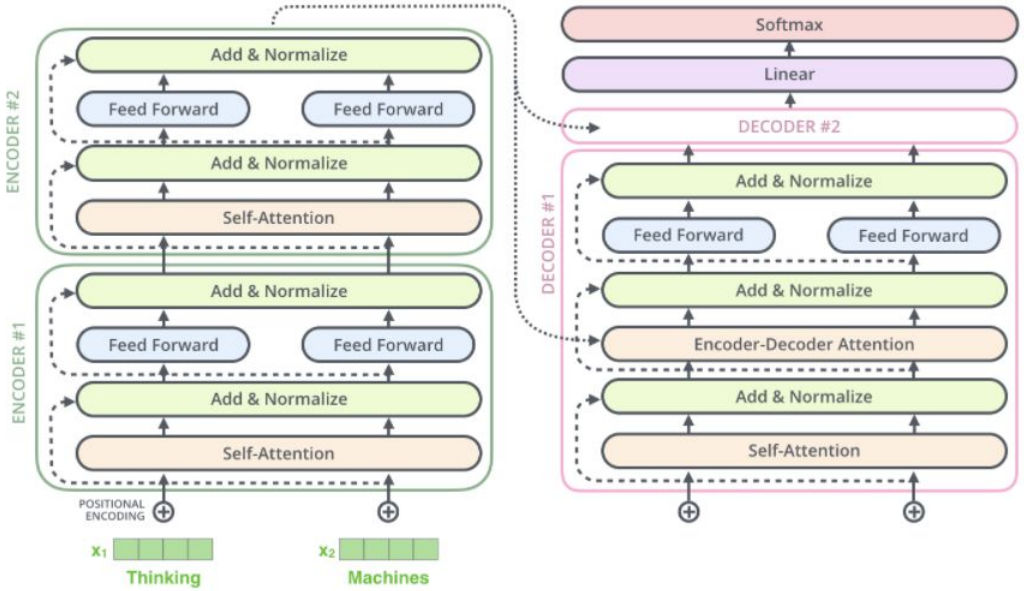
Fig. Evolution of NLP Models

2020



Afi Escuela





- En el problema original de The Transformer (traducción de inglés a francés), los **encoders** se encargan de codificar el texto en inglés, mientras que los **decoders**, a partir de esta codificación, se encargan de decodificar el texto en inglés al texto en francés, es decir, son los que generan el texto.
- En las capas de **encoder**, los tokens se atienden todos a todos; en las de **decoder**, solo a los anteriores.

¿Cuál es la gracia de The Transformer?

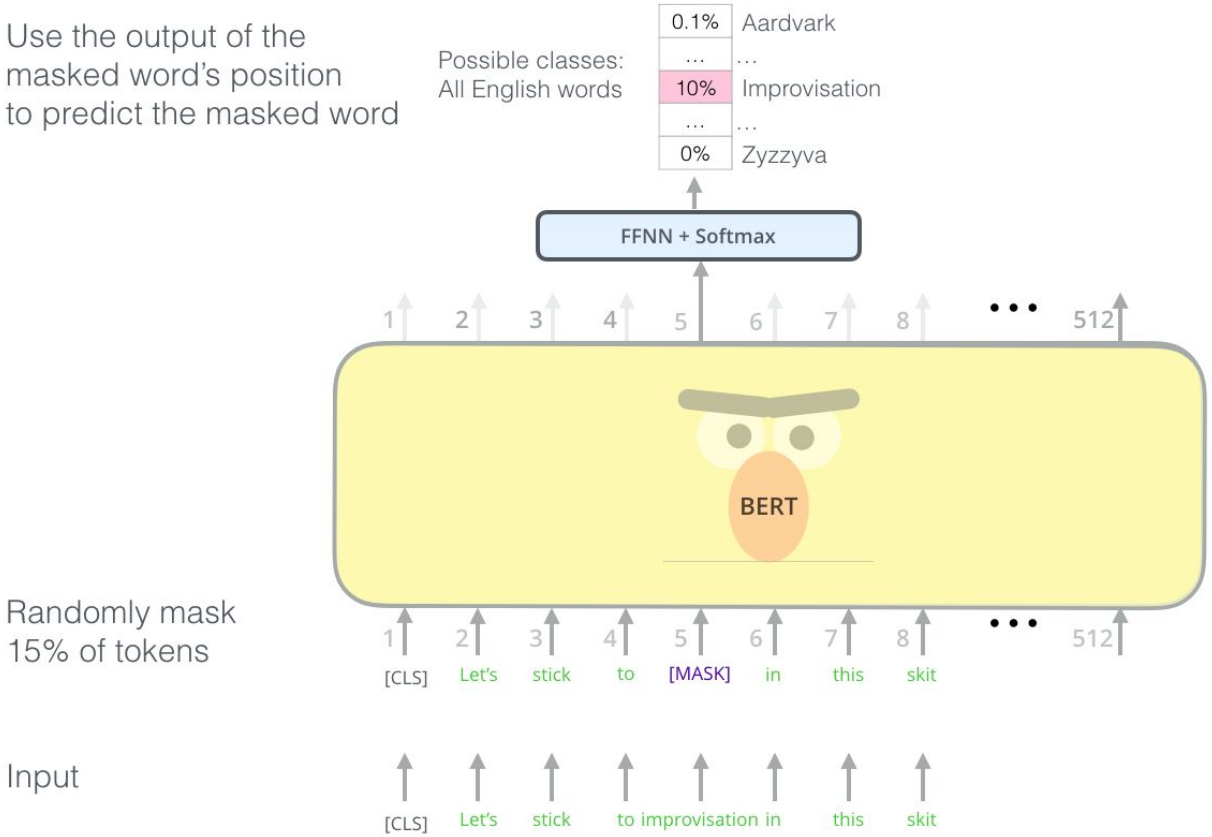
- Populariza las capas de atención.
-

Modelos de lenguaje Pre-Entrenados

- A partir de The Transformer, surge una nueva forma de trabajar con texto:
  - pre-entrenar primero modelos gigantes a base de encoders o decoders como los de la arquitectura Transformer (basados en la operación de atención).
  - a partir de estos modelos, hacer un ajuste fino a la tarea en concreto.
- De esta forma, tenemos modelos de lenguaje basados en encoders:
  - BERT, ROBERTA, DEBERTA…
  - Especialmente buenos en tareas de NLU (Natural Language Understanding), como: clasificación, detección de entidades, question-answering (QA) extractivo, etc.
- Modelos de lenguaje basados en decoders:
  - GPT-2, GPT-3.
  - Se usan sobre todo para generar texto, como noticias falsas, o para auto-completar código o texto.
- Modelos de lenguaje Encoder-Decoder (ambos):
  - PEGASUS, BART, T5, ProphetNet…
  - Son muy buenos en tareas de generación de texto condicionada o compleja; como en problemas de: traducción, resumen de textos, chatbots de dominio abierto, QA abstractivo…

¿Qué son los modelos de lenguaje?

- Modelos que “saben leer” en un idioma.
- Aprenden la estructura general del lenguaje.
- Normalmente se les entrena haciendo una tarea del tipo “fill in the blanks”, como la de los exámenes de inglés.
- Este “fill in the blanks”, se denomina Masked Language Modelling (MLM)



RECAPITULANDO MODELOS DE LENGUAJE…

A partir de una arquitectura basada en The Transformer…

Pre-entrenamiento: el modelo aprende a leer.

Fine-tuning: enseñamos al modelo a hacer tareas concretas.

Modelos de lenguaje en español

- Modelos disponibles en español:

dcc CIENCIAS DE LA COMPUTACIÓN UNIVERSIDAD DE CHILE

- ***BETO***
- BERT base.
- Corpus: SUC (18GB)
- > 700k steps

- ***BERTÍN***
- RoBERTa base.
- Corpus: MC4-es (sample - 142GB)
- approx. 250k steps

- ***MarIA***
- RoBERTa (base y large) + GPT-2
- Corpus: Preparado por la Biblioteca Nacional de España; 570GB limpios.
- approx. 500k steps

Un modelo de lenguaje del Estado del Arte en Español: RigoBERTa

Mayor corpus

Mayor calidad de corpus

Mejor hardware de entrenamiento

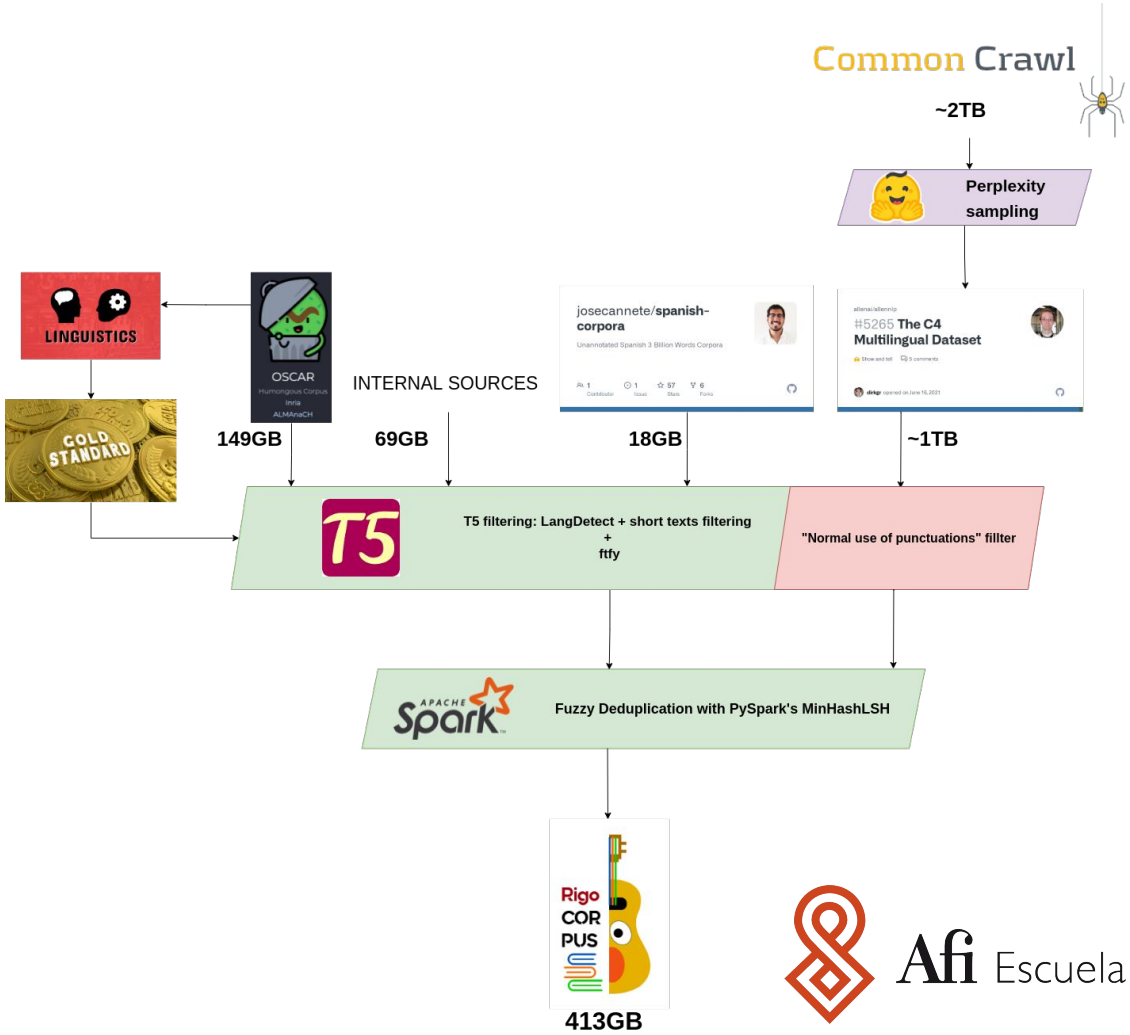
Mejor arquitectura neuronal

Adaptaciones a dominios

Afi Escuela

RigoCORPUS

- Limpieza basada en T5 y GPT-3 + reglas lingüísticas.
  - Detector de idioma.
  - Filtrado de textos cortos.
  - Corrector codificación
  - De-duplicado fuzzy
- Tamaño final comparable al del BSC + BNE (570GB vs 413GB)



Afi Escuela

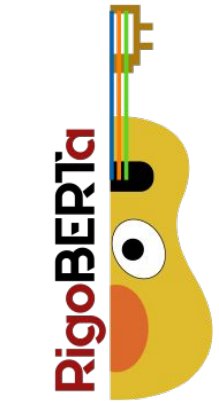
RigoBERTa – Modelo

SuperGLUE

Leaderboard Version: **2.0**

	Rank	Name	Model	Score
	1	ERNIE Team - Baidu	ERNIE 3.0	90.6
+	2	Zirui Wang	T5 + UDG, Single Model (Google Brain)	90.4
+	3	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	90.3
	4	SuperGLUE Human Baselines	SuperGLUE Human Baselines	89.8
+	5	T5 Team - Google	T5	89.3
+	6	Huawei Noah's Ark Lab	NEZHA-Plus	86.7
+	7	Alibaba PAI&ICBU	PAI Albert	86.1
+	8	Infosys : DAWN : AI Research	RoBERTa-iCETS	86.0
+	9	Tencent Jarvis Lab	RoBERTa (ensemble)	85.9
	10	Zhuiyi Technology	RoBERTa-mtl-adv	85.7

Arquitectura de RigoBERTa



Afi Escuela

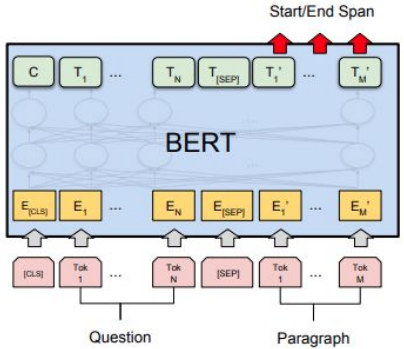
iic instituto de ingeniería del conocimiento

DeBERTa

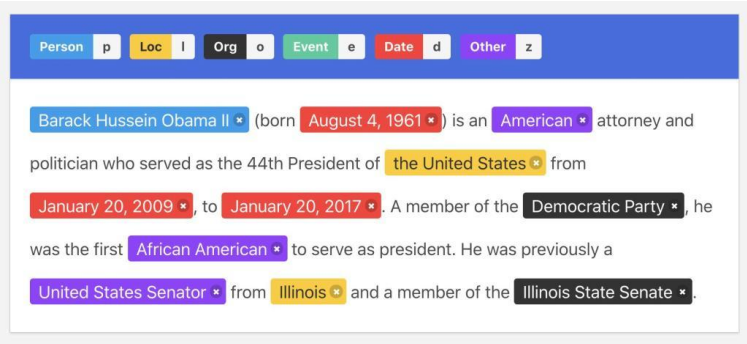
- Mejora a BERT y RoBERTa a igualdad de tamaño en la mayoría de tareas.
- Incluye embeddings posicionales relativos en cada capa ⇒ Mejor captación de información posicional ⇒ Muy útil para Question Answering especialmente.
- Disentangled Self-Attention: Cada palabra se representa usando 2 vectores, contenido y posición, los pesos de atención se calculan de forma separada en base a estos 2 vectores.
- Converge más rápido que RoBERTa.
- Enhanced Mask Decoder: Añadimos embeddings posicionales absolutos justo antes de la capa softmax que hace MLM.
- Hubo que hacer algunas modificaciones de bugs que tenía la implementación en la librería Transformers.

Evaluación de los Modelos

- Tareas:
  - Clasificación de textos.
  - NER (Named Entity Recognition): detección de entidades en textos. En la práctica, lo que hacemos es clasificar cada palabra de una frase.
  - Extractive Question Answering: lo resolvemos como un problema de clasificación; a partir de una pregunta y un texto, al que llamamos contexto, identificamos el token de inicio y el token del final de la respuesta.



QA



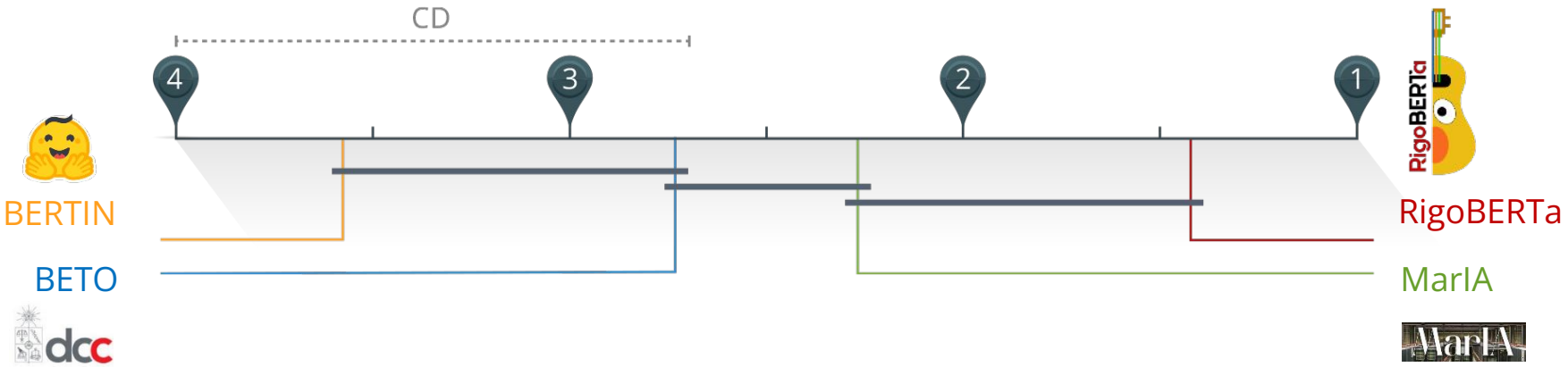
NER

Resultados

	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%
Anonymize	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%** ★
Classification	MLDOC	95.4%	94.4%	**95.6%** ★	**95.6%** ★
Paraphrasing	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%** ★
Sentiment	TASS2020	46.1%	46.1%	**47.3%** ★	46.7%
Entailment	XNLI	81.7%	79.4%	81.6%	**83.4%** ★
	TOTALS	76.5%	69.6% x1	77.3% x3	**79.8%** ★ x10

Afi Escuela

Test de Nemenyi

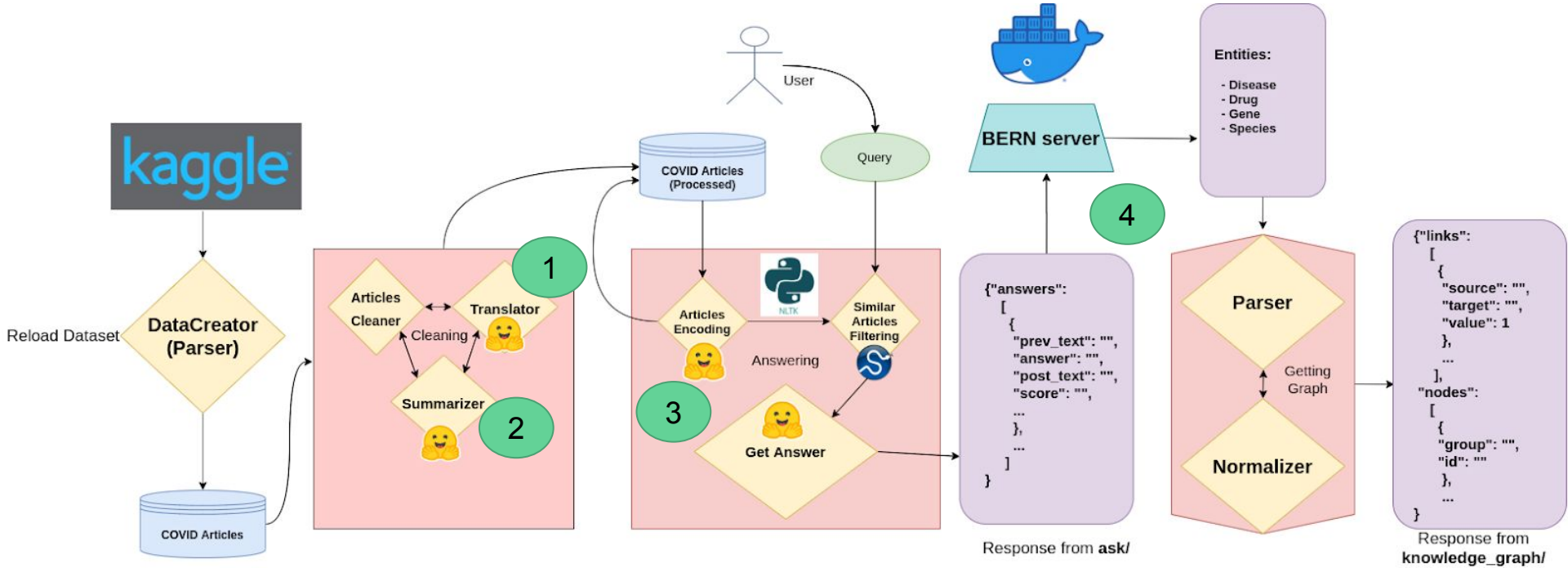


Paper de RigoBERTa

Neuraculus: Resolviendo dudas científicas sobre el COVID-19 con NLP

- Ya hemos visto cómo hacer un modelo que “lea” en un idioma mejor que ninguno (😎), ahora veamos para qué pueden servirnos estos modelos en la práctica.
- Contexto: habíamos estado ya unos cuantos meses de pandemia, se desarrolló entre mayo y octubre de 2020.

Afi Escuela



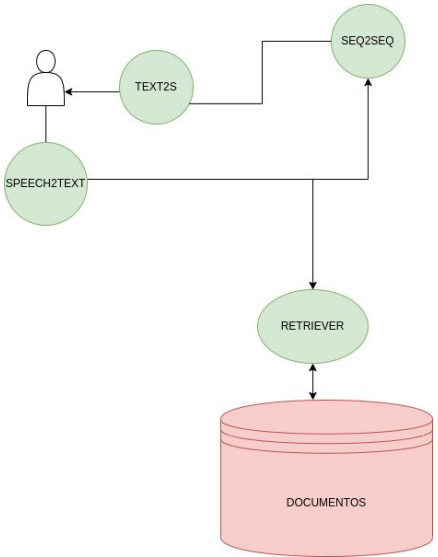
1. Summarization.
2. Translation.
3. Similar articles filtering + Extractive Question Answering.
4. NER + Normalization

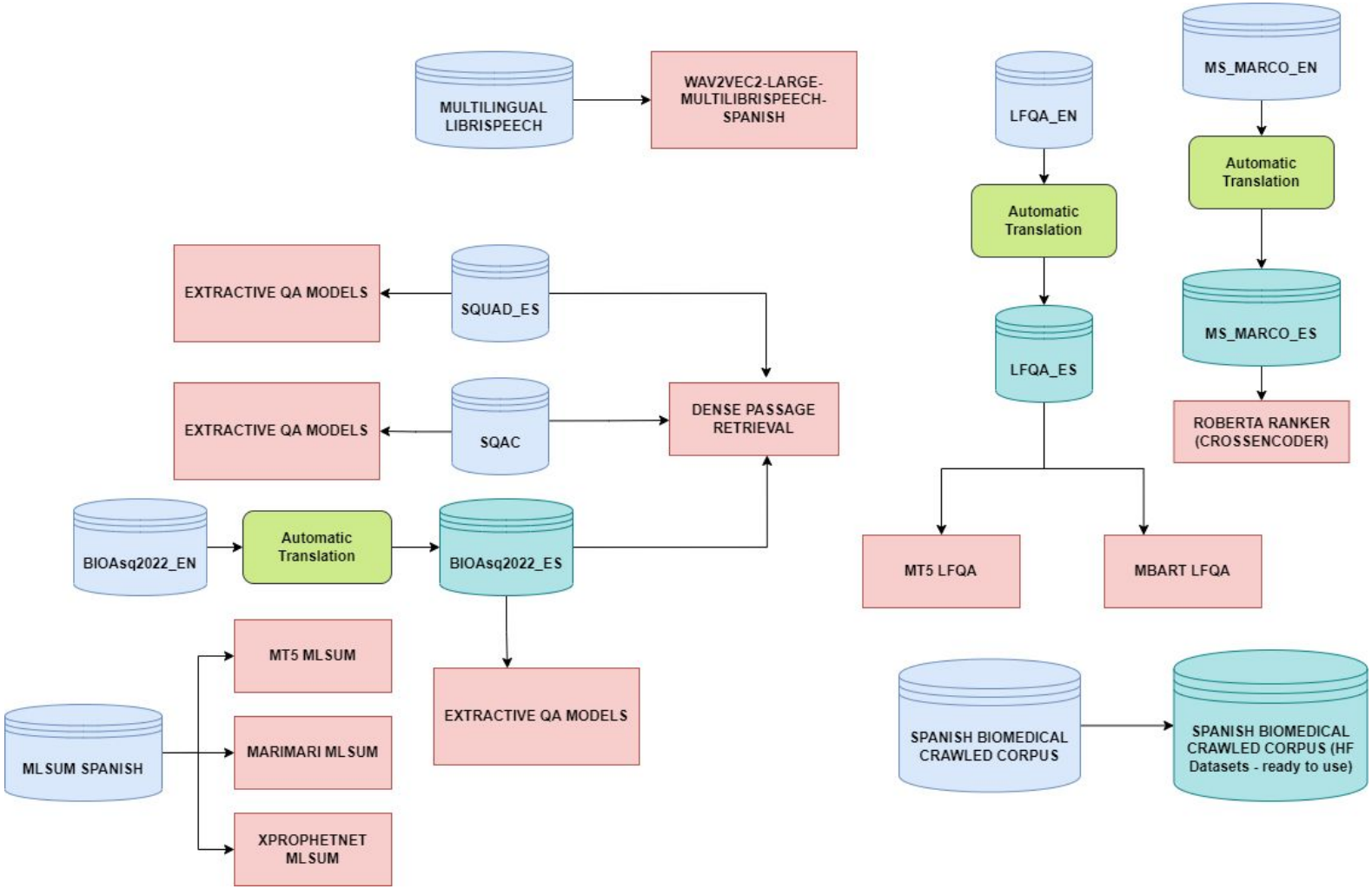
https://youtu.be/gv9QftCawnw?t=862
https://youtu.be/gv9QftCawnw?t=1949

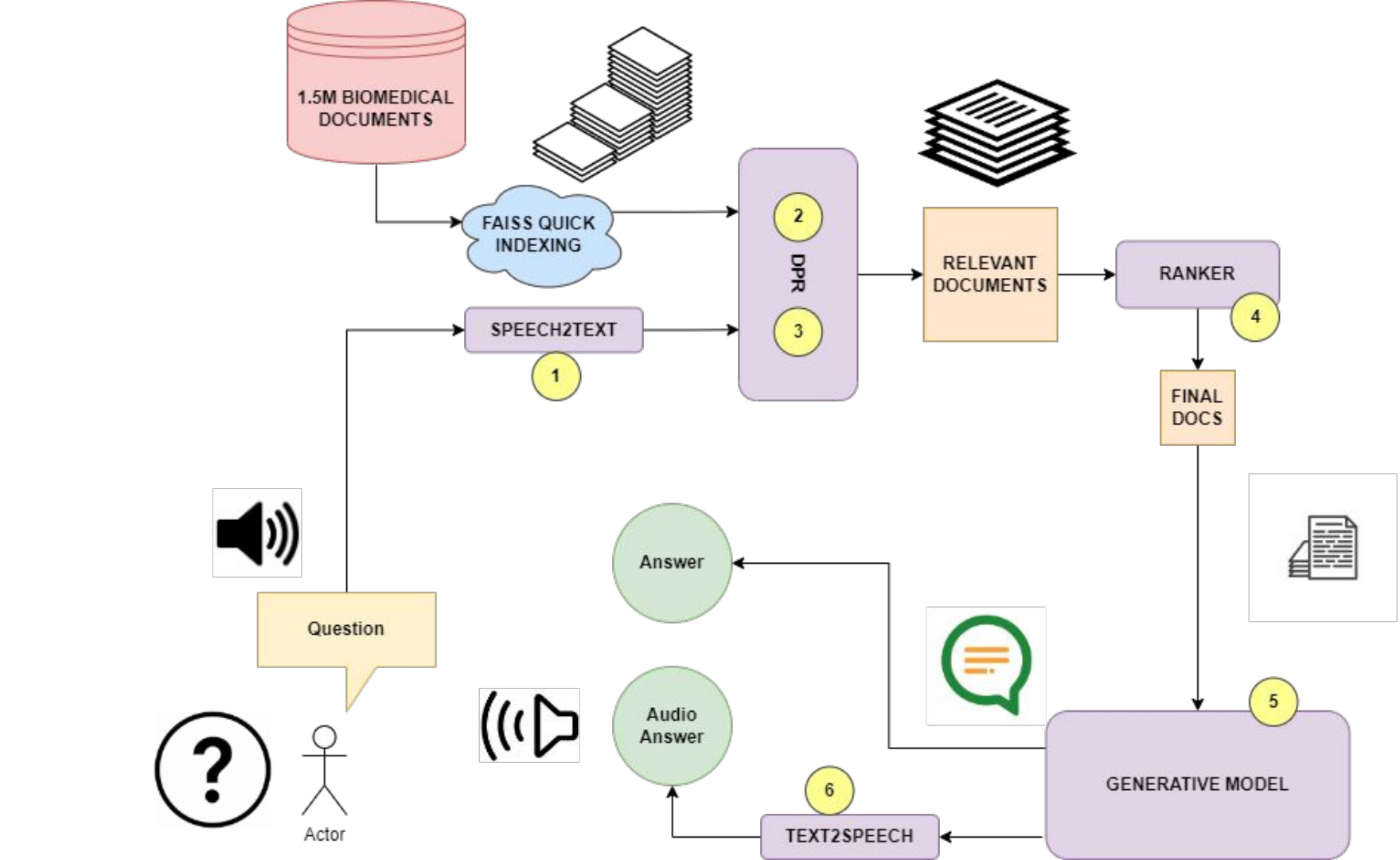
Afi Escuela

BioMedIA: Abstractive Question Answering for the BioMedical Domain in Spanish

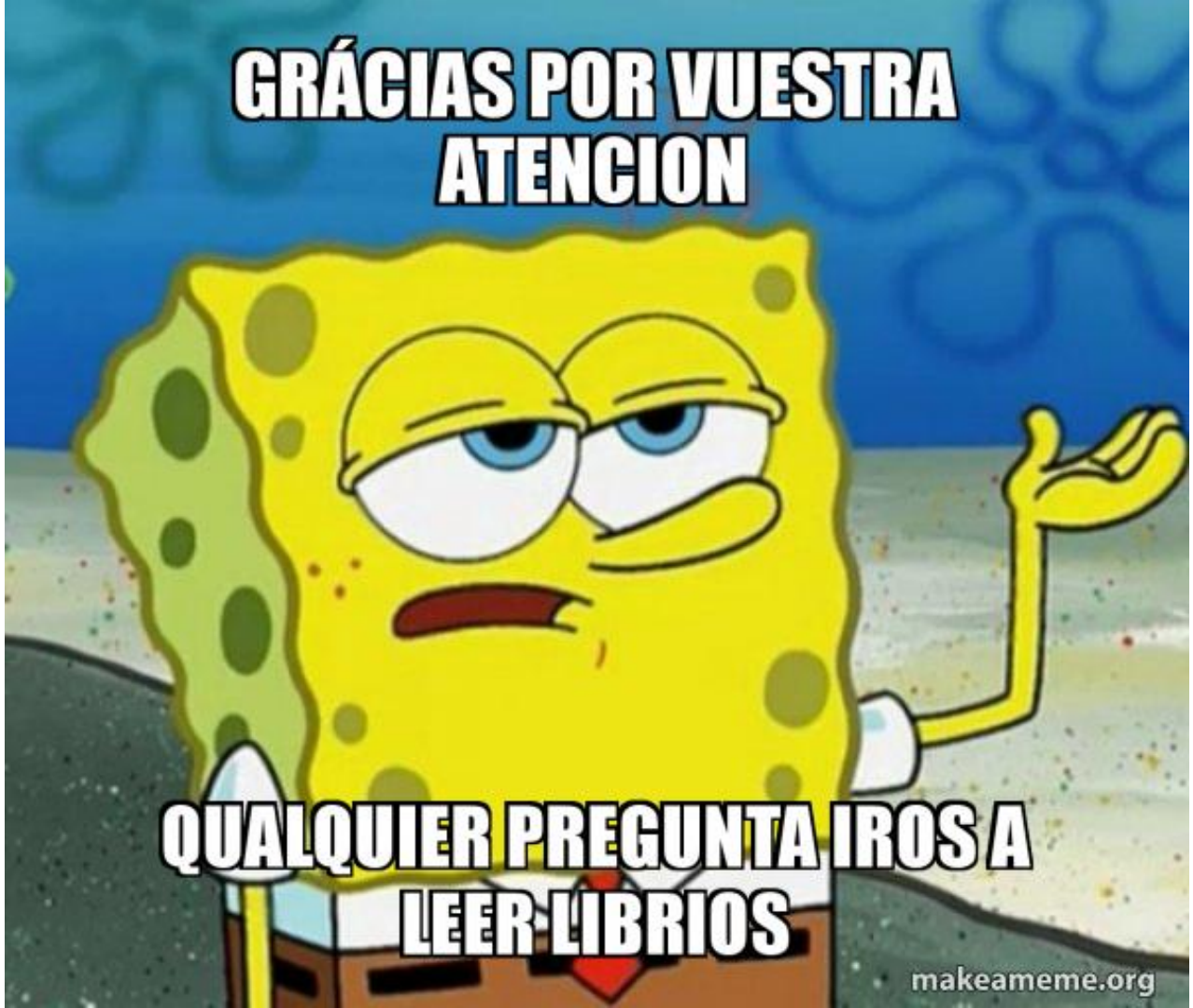
- Se desarrolla en el contexto del **Hackaton SomosNLP 2022**, el Hackaton de NLP en Español más grandes hasta la fecha. Del 14 al 31 de Marzo de 2022.
- Objetivo: potenciar el NLP en español.
- Idea original:




MULTILINGUAL LIBRISPEECH
WAV2VEC2-LARGE-MULTILIBRISPEECH-SPANISH
LFQA_EN
MS_MARCO_EN
Automatic Translation
Automatic Translation
EXTRACTIVE QA MODELS
SQUAD_ES
MS_MARCO_ES
LFQA_ES
EXTRACTIVE QA MODELS
SQAC
DENSE PASSAGE RETRIEVAL
ROBERTA RANKER (CROSSENCODER)
BIOAsq2022_EN
Automatic Translation
BIOAsq2022_ES
MT5 LFQA
MBART LFQA
MT5 MLSUM
EXTRACTIVE QA MODELS
MLSUM SPANISH
MARIMARI MLSUM
SPANISH BIOMEDICAL CRAWLED CORPUS
SPANISH BIOMEDICAL CRAWLED CORPUS (HF Datasets - ready to use)
XPROPHETNET MLSUM


1.5M BIOMEDICAL DOCUMENTS
FAISS QUICK INDEXING
2
DPR
3
RELEVANT DOCUMENTS
RANKER
4
SPEECH2TEXT
1
FINAL DOCS
Question
Answer
Audio Answer
5
GENERATIVE MODEL
6
TEXT2SPEECH
Actor


LIVE DEMO!
WHAT COULD GO WRONG?
memegenerator.net


GRÁCIAS POR VUESTRA ATENCION
QUALQUIER PREGUNTA IROS A LEER LIBRIOS
makeameme.org


GRACIAS POR VUESTRA ATENCIÓN
HACED PREGUNTAS QUE PUEDA RESPONDER
memegenerator.net