BioMedIA

A Complete Voice-to-Voice Generative Question Answering System for the Biomedical Domain in Spanish

Alejandro Vaca Serrano, David Betancur Sánchez, Alba Segurado, Guillem García Subies, Álvaro Barbero Jiménez

12 de febrero de 2023

instituto de ingeniería del conocimiento
www.iic.uam.es

Pioneros en inteligencia artificial desde 1989

Investigamos, innovamos y desarrollamos aplicaciones basadas en el conocimiento que nos proporcionan los datos.

NUESTROS ASOCIADOS

Santander

IBM

UAM Universidad Autónoma de Madrid

Realizamos análisis de datos en cualquier sector

Entorno Bancario

Entorno RRHH

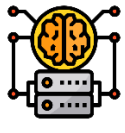
Entorno Digital

Entorno Energía

Entorno Salud

Entorno Inteligencia de Cliente

Líneas de investigación



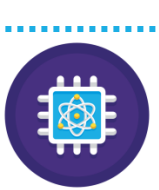
Machine learning + Deep Learning

Explicabilidad algorítmica

Deep Reinforcement Learning y Optimización

Machine Learning for NLP

Computer Vision



Quantum Computing

Encriptación homomórfica

Alejandro Vaca
Data Scientist @IIC

David Betancur
Data Scientist @IIC

Alba Segurado
Data Scientist @IIC

Guillem García
Data Scientist @IIC

Álvaro Barbero
Chief Data Scientist @IIC


OKAY
BUT WHY


whyyyy?

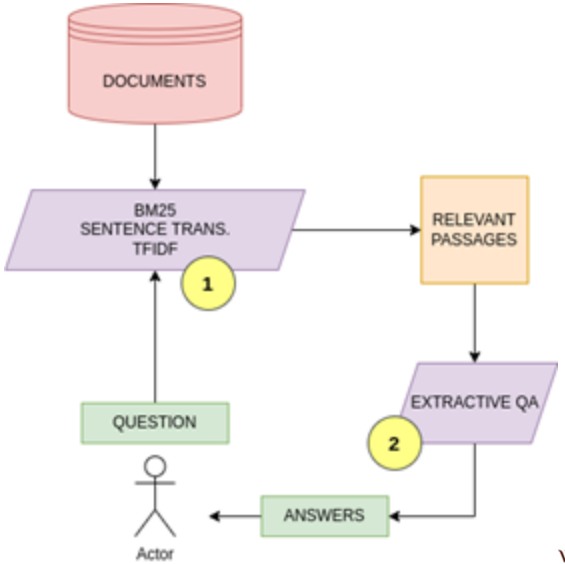

WHY WHY WHY

Objectives and Motivations

- Satisfacer la curiosidad.
- Importance of health for the population.
- Crear recursos de NLP en Español del Estado del Arte.

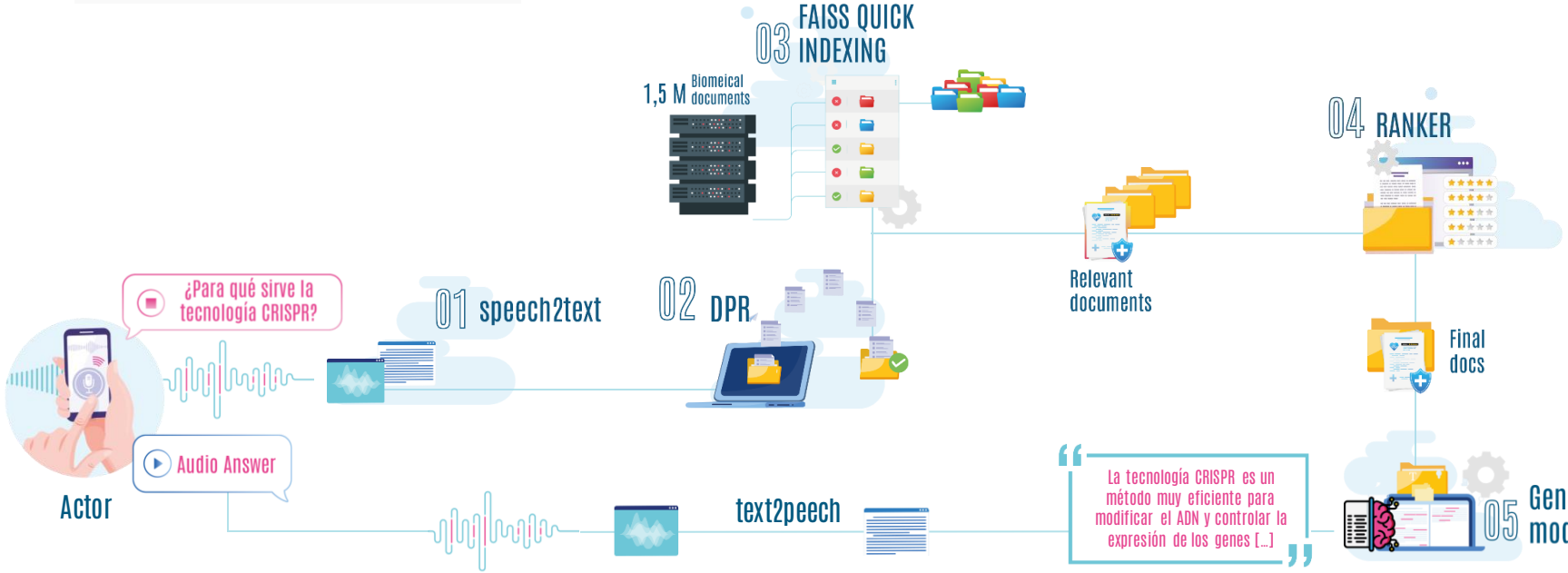
Sistema típico de QA

- Sentence Transformer / BM25 / TFIDF para obtener los textos similares a la pregunta.
- Modelo de QA extractivo para obtener las respuestas.
- Sólo 1 fuente de información a la vez.
- Sólo partes exactas del texto.



BioMedIA

instituto de ingeniería del conocimiento



01. Speech2Text

BioMedIA

instituto de ingeniería del conocimiento

Actor

- XLSR-53 large (Wav2Vec2)
- Configuración similar a la de los autores originales (Meta).
  - Congelamos el feature extractor.
  - Dropouts: Attention (0.1), hidden (0.1), feat_proj (0.0), mask_time (0.4), layerdrop (0.1).
- Multilingual Librispeech (Spanish Portion): unas 920 horas de audio.

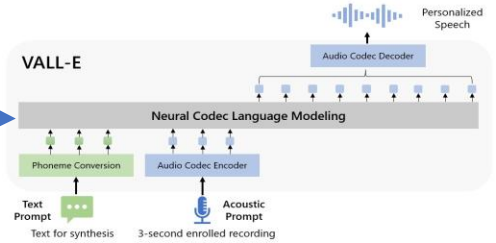
Model	WER
xlsr-53	11.5
ours	**7.3***

Table 1
Word Error Rate (WER) (Ali and Renals, 2018) for Speech to Text models on Multilingual Librispeech test split. Lower is better.


BioMedIA
iic
instituto
de ingeniería
del conocimiento


WHISPER


BioMedIA
iic
instituto
de ingeniería
del conocimiento

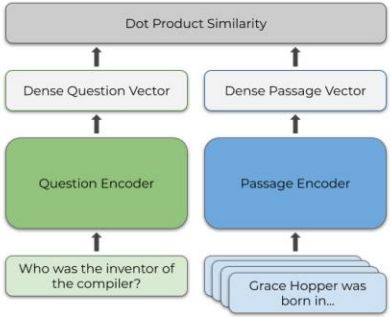

Personalized Speech
VALL-E
Audio Codec Decoder
Neural Codec Language Modeling
Phoneme Conversion
Audio Codec Encoder
Text Prompt
Acoustic Prompt
Text for synthesis
3-second enrolled recording

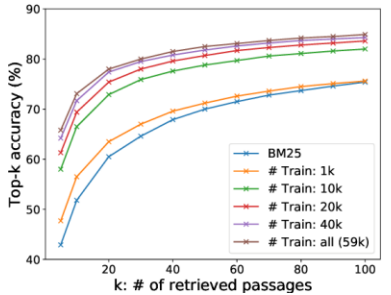
02. Dense Passage Retrieval (DPR)

BioMedIA

instituto de ingeniería del conocimiento

- Estado del arte para Passage Retrieval, entrenado por Facebook.
- La pregunta y el pasaje son codificados por dos redes transformers diferentes (red siamesa); en el caso de DPR, se usa BERT --> obtenemos la similaridad entre pregunta y pasaje calculando el dot product.
- Para entrenar, usamos pasajes positivos ynegativos para cada pregunta o query.
- Optimizar la función de pérdida negative log-likelihood de los pasajes positivos (maximizar la similaridad entre pregunta y ejemplo positivo y decreciendo la similaridad entre pregunta y ejemplos negativos).
- El Dataset de DPR se crea usando SQUAD y otros datasets de QA.
- Ejemplo positivo: el contexto emparejado con la pregunta en el dataset original.
- Ejemplos negativos:
- Pasajes top obtenidos por BM25 para esa pregunta, excluyendo el positivo.
- Pasajes emparejados con otras preguntas en el dataset de entrenamiento.

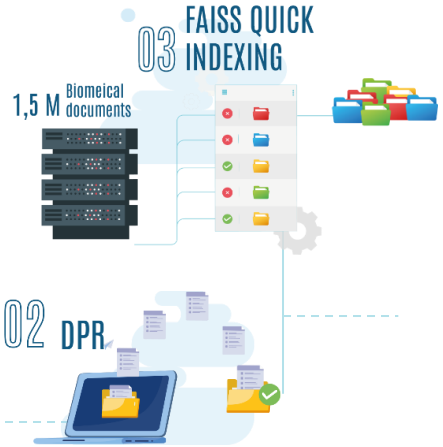




02. DPR

BioMedIA

instituto de ingeniería del conocimiento



- Dataset de 1.5M de documentos del dominio biomédico en español.
- Obj.: Encontrar los textos más relevantes para la pregunta.
- BETO: https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

Metric	dpr-squad	dpr-allqa
F1-Macro	0.880	**0.945***
avgrank	0.274	**0.117***

Table 2
Test results on SQUAD-ES for both DPR models. We measure relevant vs not relevant f1 performance (higher is better), and average rank in the ranking task (lower is better).

03. Búsqueda por DPR

El modelo DPR nos codifica 1.5M de documentos como vectores de 768 números. Ante una pregunta, podemos codificar el texto de la misma también como un vector de 768 número y buscar cuáles de los 1.5M de vectores son más similares.

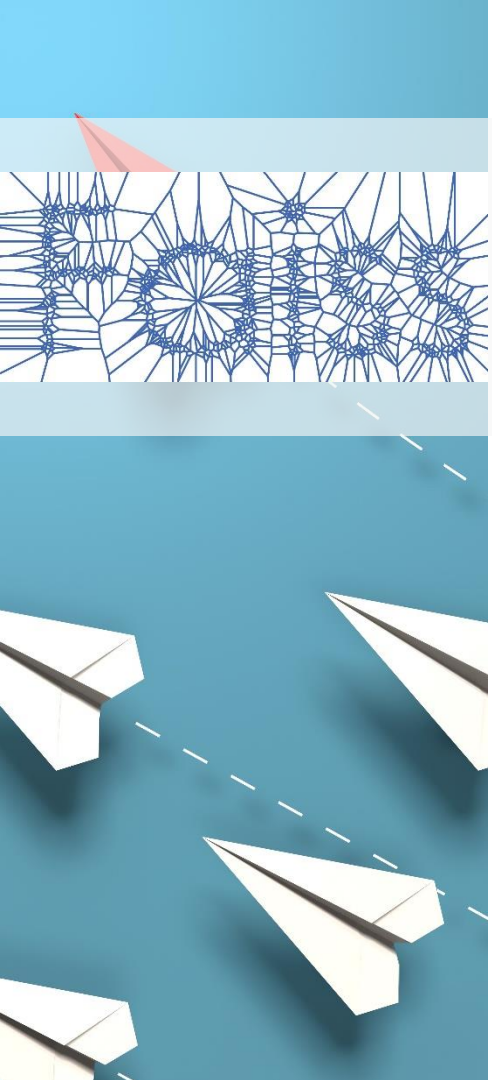
Pero... es muy lento

2.35s
(100 queries, 8 CPUs)

Y es muy pesado en RAM

4.3GB

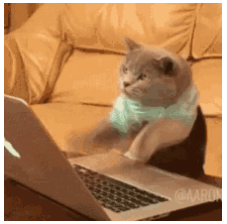
Para crear una solución práctica es necesario un algoritmo de búsqueda más eficiente.



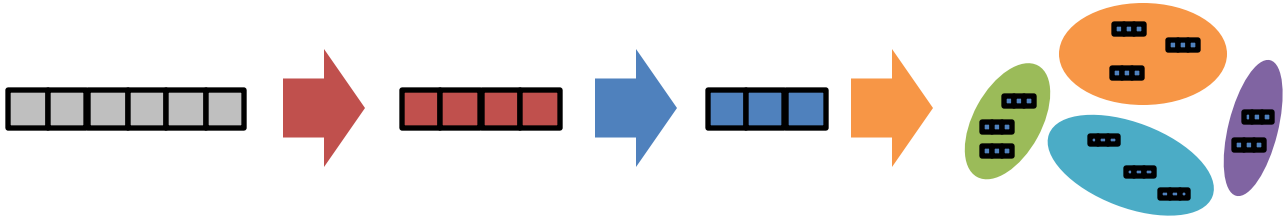
Con la librería FAISS de Meta podemos construir fácilmente un índice que nos permite hacer una búsqueda aproximada de manera muy rápida.

En BioMedIA usamos un índice de tipo

OPQ64_128,IVF4898,PQ64x4fsr



- Reducción de la dimensionalidad a cada vector, para llevarlo de 768 dimensiones a 128, mediante el método OPQ.
- Cuantización de cada vector para expresarlo como 64 códigos de 4 bits cada uno (1 vector = 32 bytes).
- Clusterización con K-means en 4898 clusters (recomendación de FAISS: $4\sqrt{N}$, $N = 1.500.00$)



En el momento de realizar una query, su vector se procesa con el mismo pipeline, y se devuelven los elementos más similares del mismo cluster.

Mejoras del índice FAISS

2.35s
(100 queries, 8 CPUs)

21.5ms
(100 queries, 8 CPUs)

4.3GB

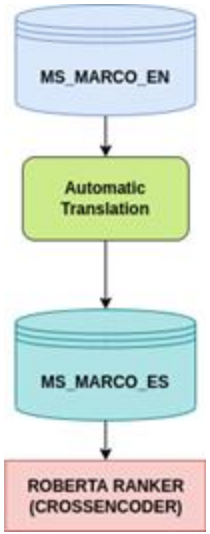
65MB

04. Ranker

BioMedIA

instituto de ingeniería del conocimiento

- Extraer sólo los 5 pasajes top en términos de importancia.
- Multilingual Sentence Transformer.
- Entrenar CrossEncoder con Roberta-base y MS Marco - ES.
- Finalmente: Combinación de ambos.



Model	MRR@10
Multiling-SentenceTrans.	0.5891
Roberta-Ranker (ours)	0.6880
Combination of both	**0.6935***

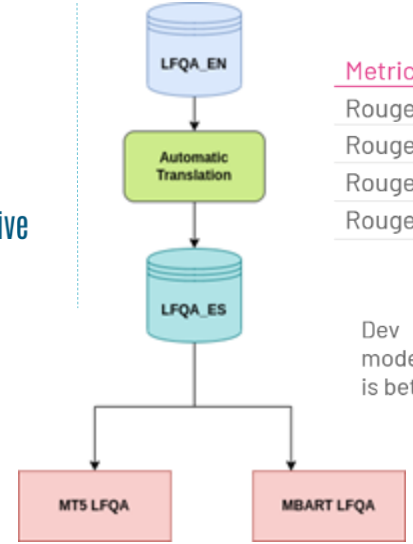
Table 3
Eval results on MSMarco_ES for both Ranker models. Higher is better.

05. Modelo generativo de QA

BioMedIA

instituto de ingeniería del conocimiento

- MT5-base (Google) & MBART-large (Meta)
- LFQA-ES: dataset basado en ELI5, de Reddit.
  - A partir de una serie de artículos de apoyo, generar la respuesta que puede incluir información de todas las fuentes.
- MT5 es el modelo final usado.

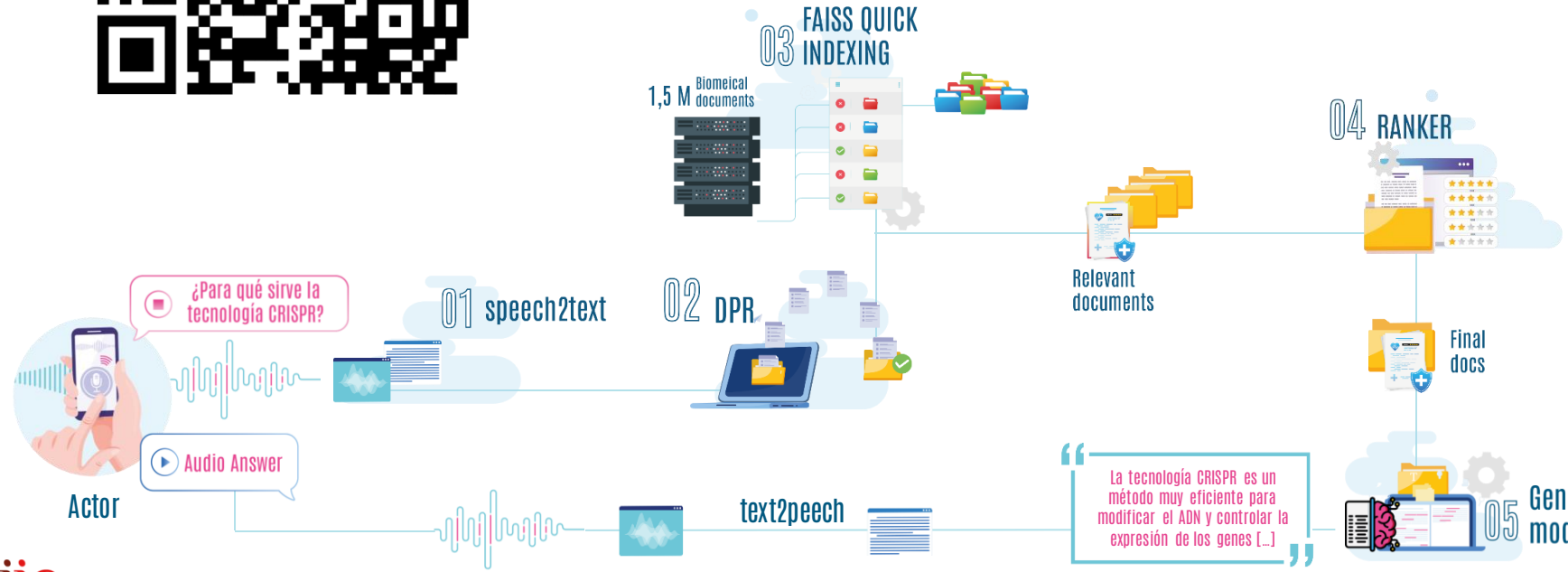


Metric	MT5-base-lfqa	MBART-large-lfqa
Rouge1	10.291*	0.511
Rouge2	1.725*	0.004
RougeL	8.919*	0.511
RougeLSum	7.987*	0.511

Table 4
Dev results on LFQA_ES for both LFQA models in rouge metrics (Lin, 2004). Higher is better.

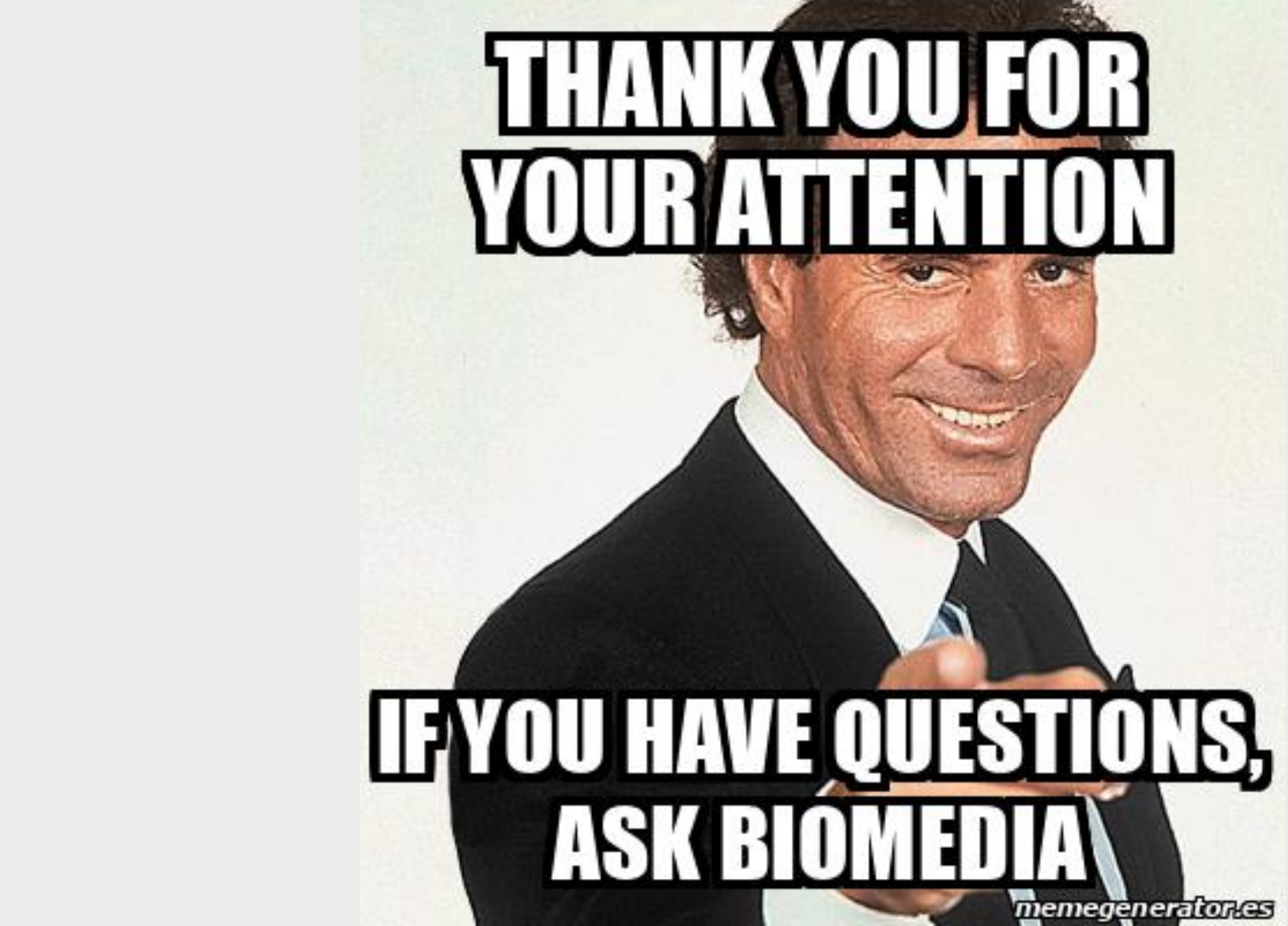
BioMedIA

instituto de ingeniería del conocimiento



¿Qué es el *lupus*?

> *El lupus es una enfermedad del sistema inmunológico. El cuerpo produce anticuerpos o proteínas que atacan a las células sanas del cuerpo, y pueden causar daño tisular o muerte. Cuando el paciente se diagnostica de lupus eritematoso sistémico o cutáneo, la inmunidad del paciente está perturbada.*


THANK YOU FOR
YOUR ATTENTION
IF YOU HAVE QUESTIONS,
ASK BIOMEDIA
memegenerator.es