

¿Cómo procesar datos textuales?

Introducción a los modelos de lenguaje
basados en redes neuronales

Andrea Morales Garzón



Andrea Morales

- Estudiante de Doctorado en Inteligencia Artificial (NLP) en UGR
- Graduada en Ingeniera Informática, mención en Computación y Sistemas Inteligentes
- Máster (profesionalizante) en Ingeniería Informática
- Máster en Ciencia de Datos



@andreamorgar



[andreamorgar.github.io](https://github.com/andreamorgar)



amoralesg@ugr.es



Contenido

1. ¿Por qué NLP?
2. Workflow
3. Representaciones numéricas a partir de texto
4. NLP y redes neuronales
5. Transformers
6. Language Representation Models
7. Desafíos y tendencias

Natural Language Processing (NLP)

Name Entity Recognition

Categorizar palabras del texto

Text classification

¿Es una noticia falsa o no?

Cómo el emoji de la zanahoria se convirtió en un código secreto en internet para camuflar contenido antivacunas

Text generation

Generar el texto de la noticia a partir del titular

Summarization

Resumir el texto de la noticia

Question Answering

¿Qué emoji se utilizó para camuflar contenido antivacunas?

Machine Translation

How the carrot emoji became a secret code on the internet to camouflage anti-vaccine content

¿Por qué NLP?

- Porcentaje mayoritario de la información es textual
- Interacción humana con las máquinas: comprensión y comunicación a través de lenguaje natural
- Análisis de datos masivos: tendencias, estadísticas, categorización de información para posterior visualización, etc.

Aplicaciones en la industria

Mejora de la
experiencia del
usuario en
aplicaciones

Automatización
de tareas
Atención al
cliente

Marketing y
análisis de
opiniones

La revolución del NLP

3 ingredientes estrella

TRANSFORMER

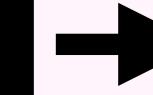
Arquitectura de red neuronal



GitHub Copilot

Modelos pre-entrenados

Pre-training

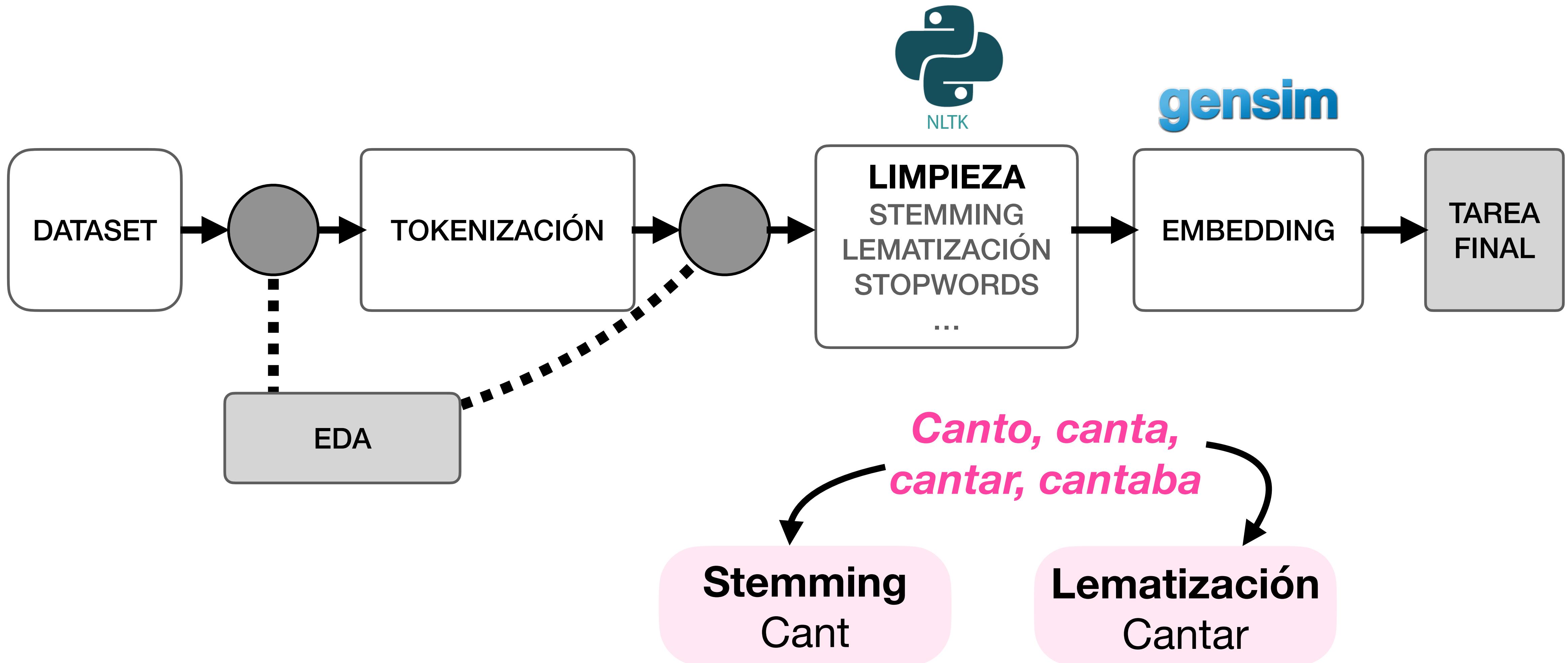


Fine-tuning



Hugging Face

NLP Workflow



Embeddings

¿Cómo procesa un ordenador cadenas de texto?

One hot encodings

| Hoy el día está gris | | | | | |
|----------------------|-----|----|-----|------|------|
| <u>vocab</u> | hoy | el | día | está | gris |
| | 1 | 0 | 0 | 0 | 0 |
| | 0 | 1 | 0 | 0 | 0 |
| | 0 | 0 | 1 | 0 | 0 |
| | | | : | | |
| | | | | | |
| n | | | : | | |

Problemas

- Capacidad de memoria (sparse vectors and matrices)
- Falta de cohesión semántica. No codifica relaciones a nivel semántico
- Problemas de *out-of-vocabulary words*

vocab
1 : hola
2 : hoy
3 : el
4 : la
5 : día

21 : está
22 : gris
n : palabra

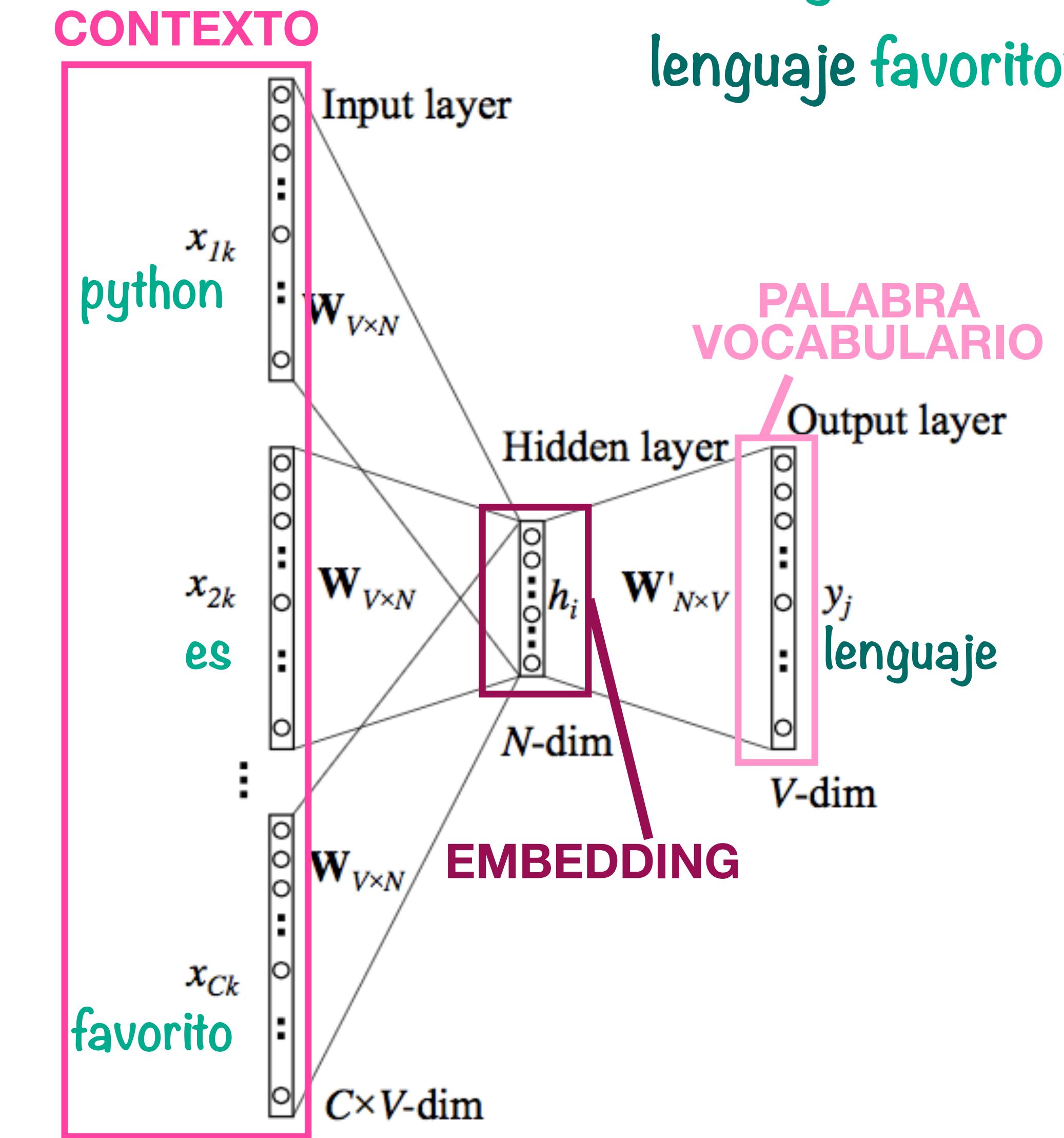
Hoy el día está gris
0 2 3 4 21 22

NLP y redes neuronales

Word2vec



- Espacio multidimensional
- Distancias significativas
- Relaciones significativas
- Palabra del vocabulario → vector



El contexto importa

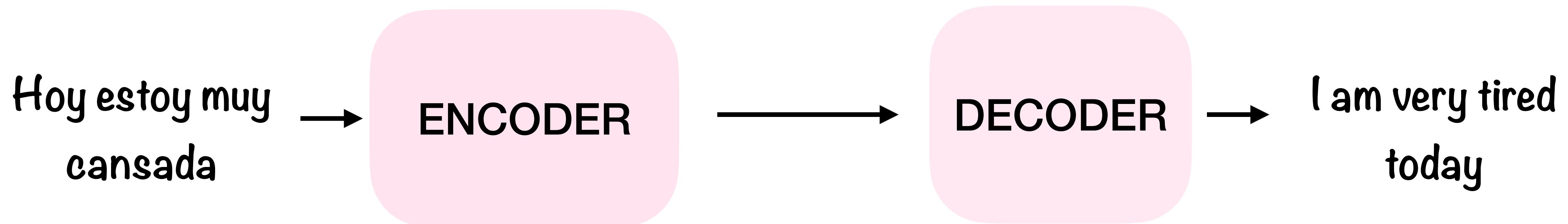


Mi hermano se ha caído en el gimnasio. Hoy **(ella)** no se puede mover.

Trasformers

Arquitectura Encoder-Decoder

- Pensada para tareas seq-to-seq (Traducción Automática)

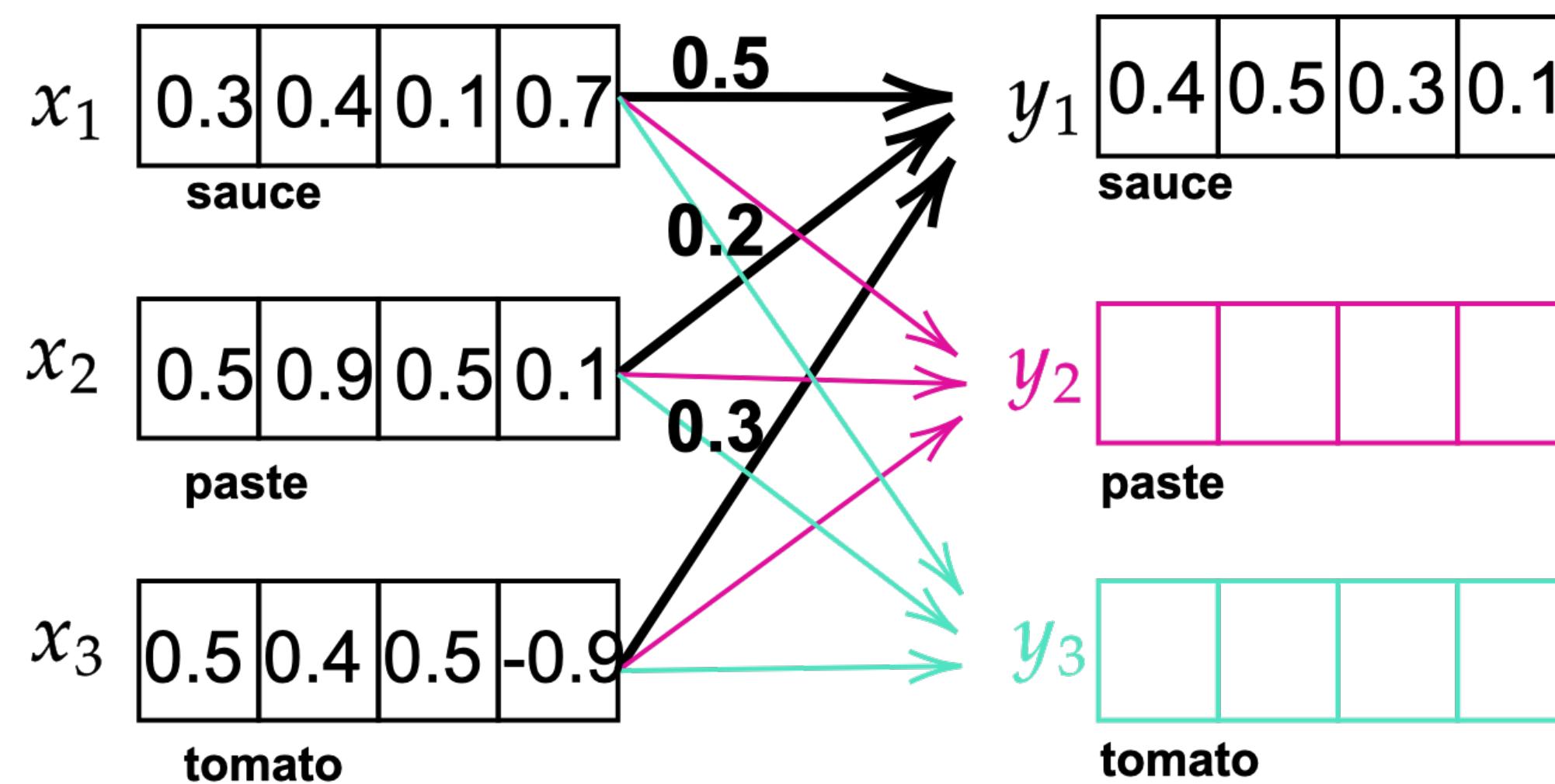


Transformers

Self-attention

- La red asigna un peso/relevancia a los elementos de una secuencia
- “Self” porque se aplica a todos los “hidden state” de un mismo módulo.

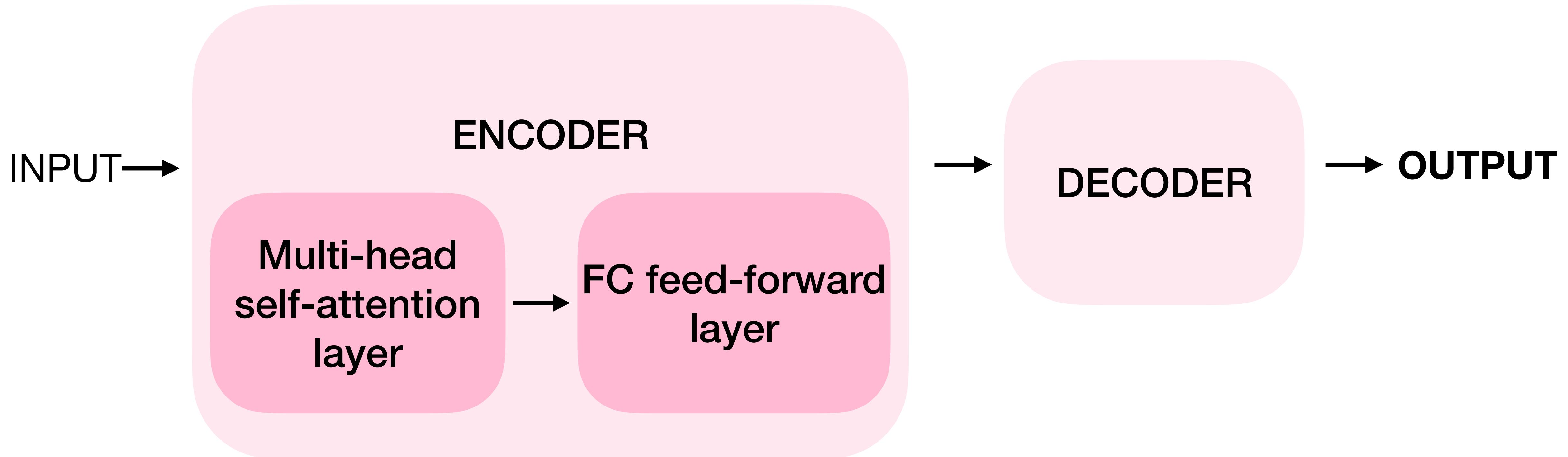
X = Sauce paste tomato



Trasformers

Arquitectura Encoder-Decoder

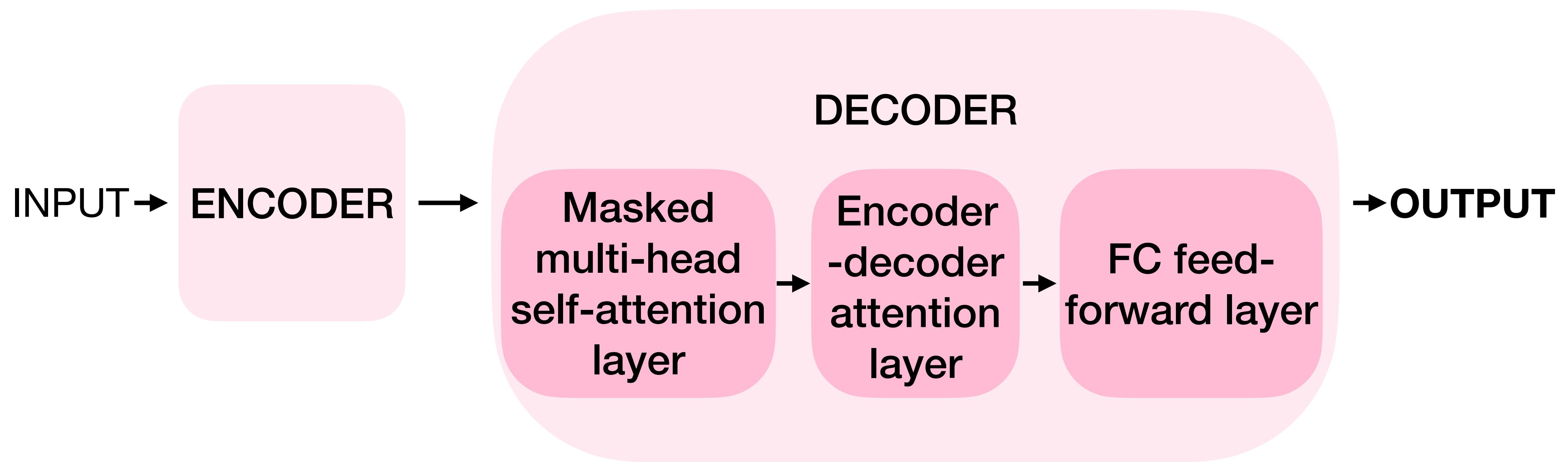
- Pensada para tareas seq-to-seq (Traducción Automática)



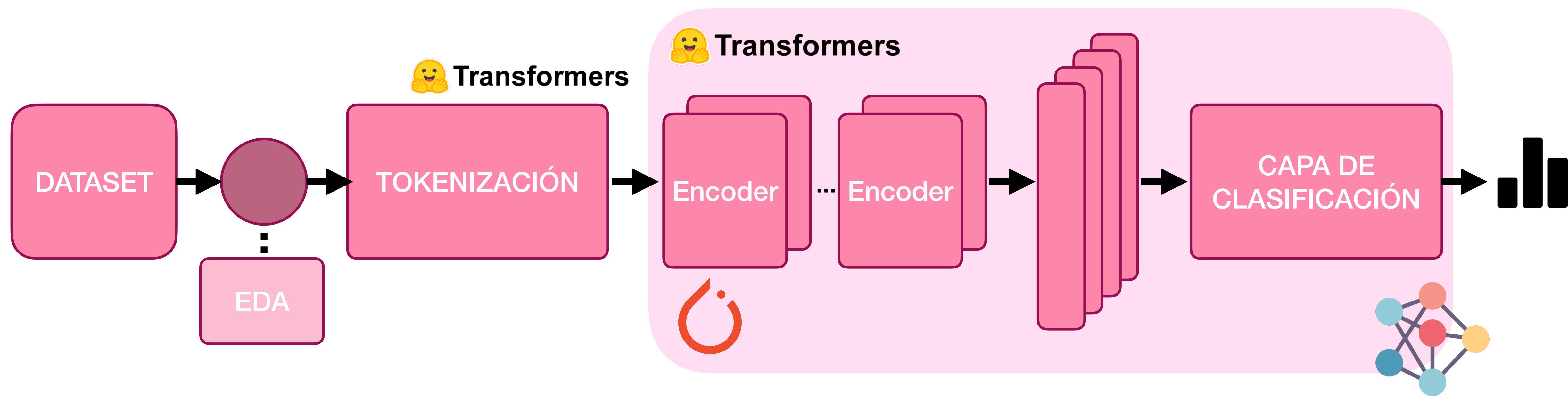
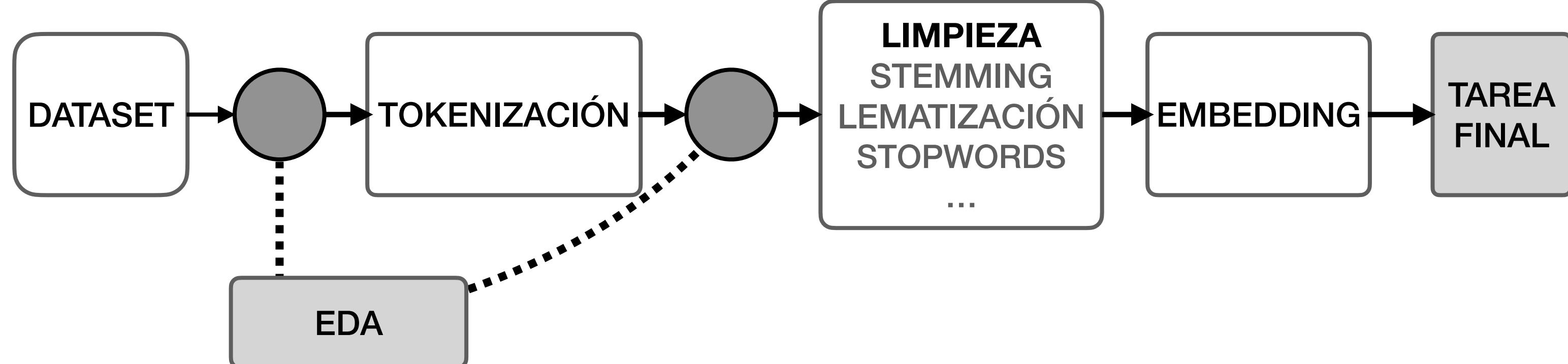
Trasformers

Arquitectura Encoder-Decoder

- Pensada para tareas seq-to-seq (Traducción Automática)



NLP Workflow



Tokenización

De texto a tokens

- A nivel de **carácter**

[‘I’, ‘l’, ‘o’, ‘v’, ‘e’,]

“I love to build embeddings. They are so cool”

- A nivel de **palabra**

[‘I’, ‘love’, ‘to’, ‘build’, ‘embeddings’,]

- A nivel de **subpalabra**

[‘I’, ‘love’, ‘to’, ‘build’, ‘em’ ‘##bed’, ‘##dings’,]

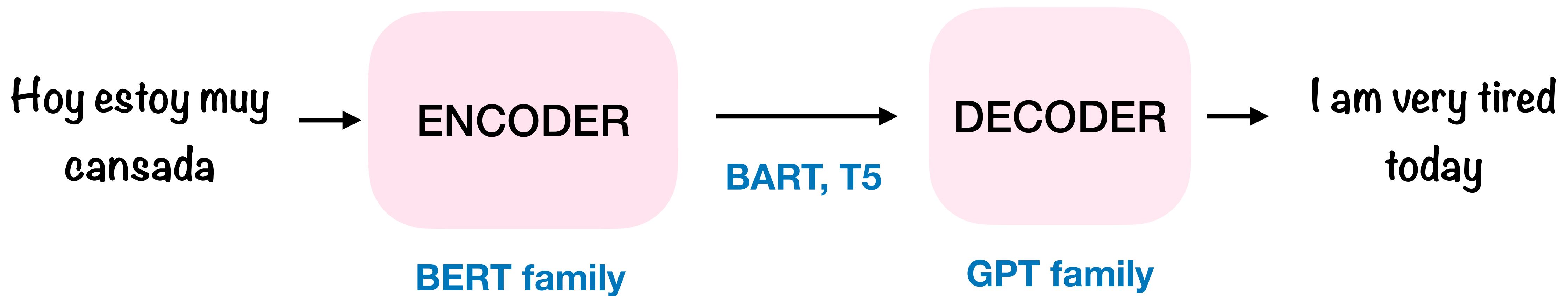
Tokens especiales añadidos por el modelo

[‘<CLS>’, ‘I’, ‘love’, ‘to’, ‘build’, ‘em’ , ‘##bed’, ‘##dings’, ‘<SEP>’, ‘They’, ‘are’, ‘so’, ‘cool’, ‘<SEP>’]

Trasformers

Arquitectura Encoder-Decoder

- Pensada para tareas seq-to-seq (Traducción Automática)



transformers 4.26.0

pip install transformers

Language representation models

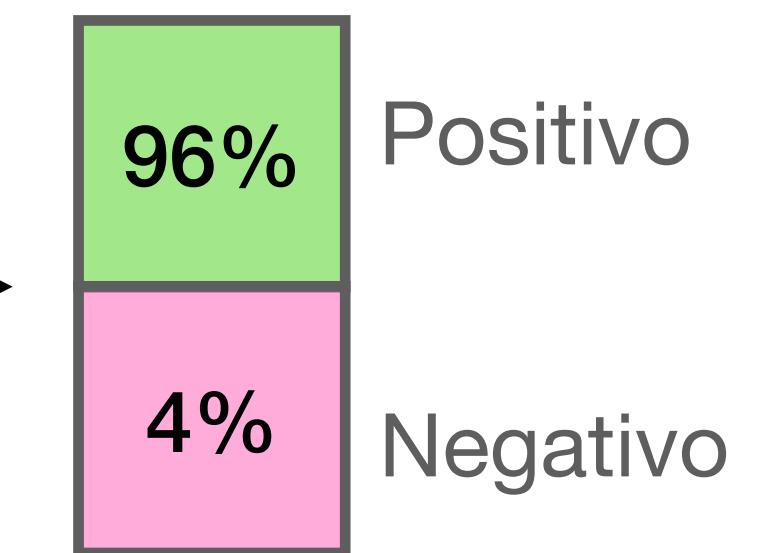
Entrada
(Secuencia de texto)

“I have been waiting for Pydata Granada my whole life”

Pre-trained
BERT

Clasificación
(Feed forward
neuronal network
+ sigmoid/softmax)

Salida
(predicción)



```
from transformers import pipeline
classifier = pipeline("sentiment-analysis")

classifier("I have been waiting for Pydata Granada my whole life.")

[{'label': 'POSITIVE', 'score': 0.9587684869766235}]
```

Desafíos y tendencias

Desafíos

Más allá del texto

Transformers para procesamiento y clasificación de imágenes

Transformers multimodales

Modelos para datos heterogéneos

Curado de datos, lenguas minoritarias y evaluación de los modelos

Tendencias

Generación automática de vídeo a partir de texto

Más ChatGPT y ... ¿GPT-4?

Mejoras en modelos de audio a texto

Para terminar...

AM

tell me a quote about natural language processing



"Language is the soul of the soul." - Noam Chomsky

AM

another one



Load failed



AK @_akhaliq · 26 ene.
pip install

transformers

diffusers

timm

datasets

safetensors

accelerate

optimum

wandb

tokenizers

evaluate

simulate

gradio

is all you need

32

140

1.282

123,9 mil



Referencias y links

- <https://transformersbook.com>
- <https://huggingface.co/docs/transformers/index>
- <https://amatriain.net/blog/transformer-models-an-introduction-and-catalog-2d1e9039f376/#Transformers>
- <https://www.narrativa.com/4-trends-in-nlp-and-nlg-for-2023/>
- <https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0>