Universidad Politécnica de Madrid

Introducción a la inteligencia artificial generativa a través de los grandes modelos de lenguaje

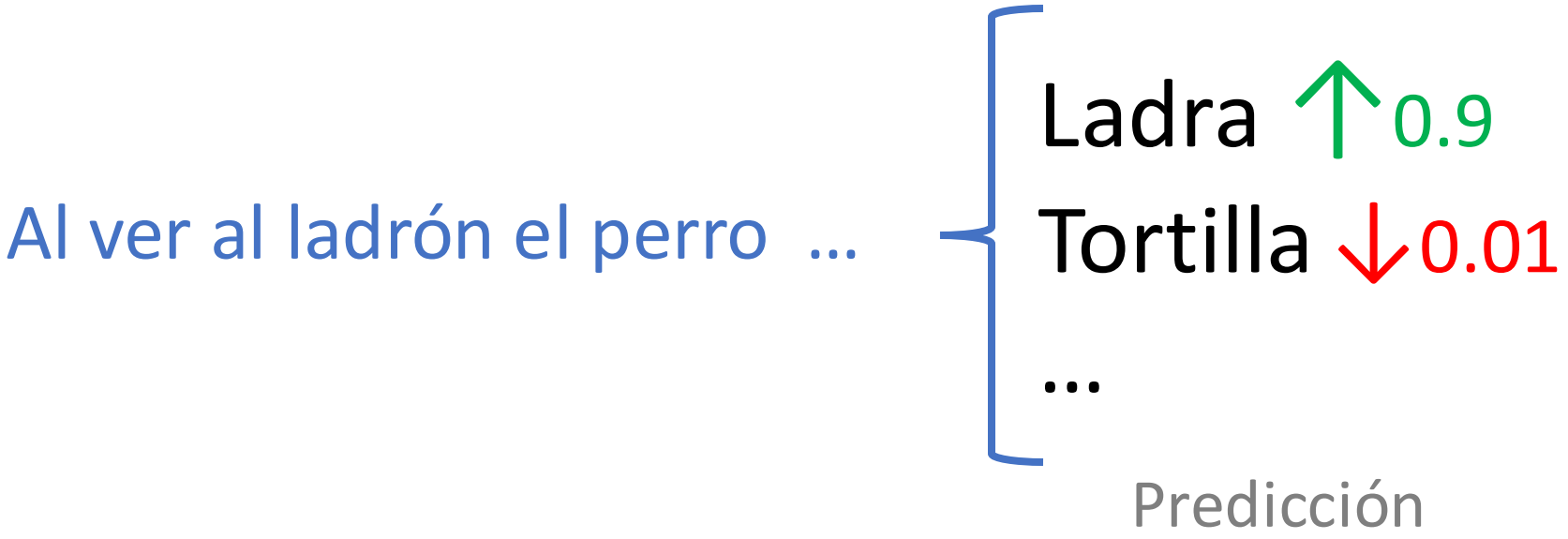
Tokenización y embeddings

Javier Conde
Pedro Reviriego

¿Qué es un LLM? (Recordatorio)

- Los LLM son modelos estadísticos que aprenden a predecir la siguiente palabra*

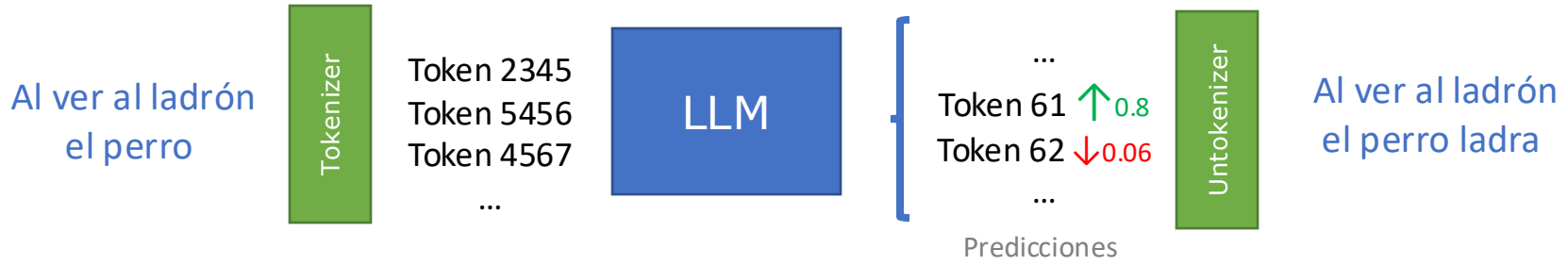
- El modelo estima las probabilidades de que cada palabra sea la siguiente



* En realidad el modelo predice tokens no palabras. Se verá más adelante en el curso

¿Qué es un token?

- Entradas y salidas de un LLM
- Una palabra se representa mediante uno o más tokens
- Un LLM no predice palabras sino tokens



Tokenizadores

- Cada LLM utiliza su propio tokenizador

¿Por qué los tokens?

- Existen muchas palabras, solo la RAE más de 90.000
- Los LLM son multilenguaje
- Es muy costoso tener que predecir la probabilidad para millones de palabras
- Los modelos parten las palabras en tokens y usan los tokens en vez de las palabras
- Con decenas de miles de tokens se pueden generar todas las palabras de todos los idiomas

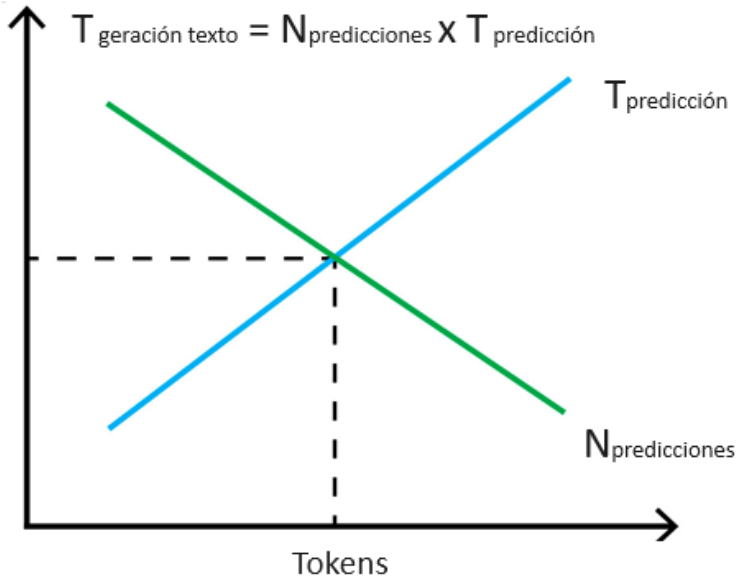
Estrategias de tokenización

- Una de la tareas previas al entrenamiento de un LLM es el diseño del tokenizador

- Diferentes posibilidades:
  - Un token por palabra
  - Un token por sílaba
  - Un token por carácter
  - **Esquemas de tokenización basados en frecuencia de ocurrencia (teoría de la información)**
    - A las palabras con mayor frecuencia de aparición se describirán con menos tokens
    - Ejemplo: si el 80% de las palabras de un idioma empiezan por “RELÁMPAGO” puedo asignarle a esa secuencia un token

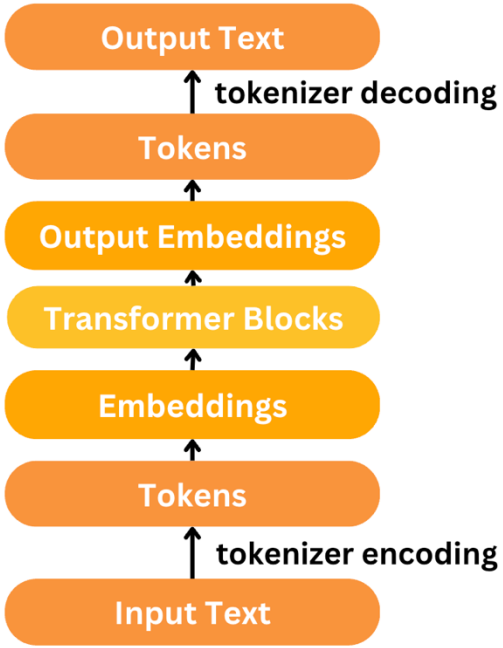
Equilibrio tokens/predicciones

- Si el diccionario de tokens es muy grande en cada predicción se tienen que calcular las probabilidades de todos los tokens ($\uparrow T_{predicción}$, $\downarrow N_{predicciones}$)
- Si el diccionario de tokens es muy pequeño es necesario realizar muchas predicciones para obtener un texto completo ($\downarrow T_{predicción}$, $\uparrow N_{predicciones}$)



Embeddings

- En realidad los tokens no se representan como caracteres sino como símbolos en un espacio continuo vectorial (vectores)
- Representar los tokens como vectores permite realizar operaciones matemáticas con ellos
- Esto permite que los modelos no se limiten exclusivamente a información textual



Fuente: https://docs.mistral.ai

Universidad Politécnica de Madrid

Introducción a la inteligencia artificial generativa a través de los grandes modelos de lenguaje

Tokenización y embeddings

Javier Conde
Pedro Reviriego