1.5 Modelos de Lenguaje, tipos y aplicaciones

En los últimos años, una de las áreas de la Inteligencia Artificial que más ha avanzado es la relacionada con el **lenguaje humano**. Los sistemas actuales son capaces de leer textos, resumir información, traducir entre idiomas o responder preguntas de forma bastante natural. Este conjunto de tecnologías se conoce como **Procesamiento del Lenguaje Natural**, o *Natural Language Processing (NLP)*.

Para entenderlo de forma sencilla, podemos pensar en estos sistemas como **lectores muy rápidos que han leído millones de textos**. A partir de ese entrenamiento, aprenden cómo suelen aparecer las palabras juntas, qué estructuras tienen las frases o cómo se organizan las ideas en un texto.

Un símil útil para explicarlo es el de un estudiante que ha leído muchos libros. Con el tiempo, ese estudiante empieza a reconocer patrones: sabe cómo se construyen las frases, cómo se explican ciertos conceptos o qué palabras suelen aparecer en determinados contextos. Los modelos de lenguaje funcionan de una forma similar, aunque a una escala mucho mayor.

Por ejemplo, cuando utilizamos un sistema como *ChatGPT* para pedir un resumen de un texto o para generar una explicación de un concepto, el modelo no está “pensando” en el sentido humano. Lo que hace es **predecir qué palabras tienen más probabilidad de aparecer a continuación en una frase**, basándose en los patrones que aprendió durante su entrenamiento.

Este tipo de modelos se conocen como **modelos de lenguaje** porque están diseñados precisamente para trabajar con lenguaje. Su tarea básica consiste en **predecir la siguiente palabra dentro de una secuencia de texto**, pero a partir de esta capacidad básica se pueden construir muchas aplicaciones diferentes: traducción automática, asistentes conversacionales, análisis de textos o generación de contenido.

En el ámbito educativo, estas herramientas pueden utilizarse para **explicar conceptos, generar ejemplos, resumir textos o apoyar la elaboración de materiales didácticos**. Sin embargo, también es importante recordar que los modelos de lenguaje no comprenden el mundo como lo hace una persona. Su conocimiento se basa en patrones estadísticos aprendidos a partir de grandes cantidades de datos.

Por ello, una buena forma de trabajar con estas herramientas en el aula es entenderlas como **un asistente que ayuda a explorar el lenguaje y la información**, pero cuyo resultado siempre debe ser revisado críticamente por el usuario. De esta manera, los modelos de lenguaje pueden convertirse en una herramienta interesante para apoyar el aprendizaje y al mismo tiempo reflexionar sobre cómo funcionan los sistemas actuales de Inteligencia Artificial.

Procesamiento del Lenguaje Natural (NLP)

El **Procesamiento del Lenguaje Natural**, conocido habitualmente como **NLP (Natural Language Processing)**, es la rama de la Inteligencia Artificial que se ocupa de que los ordenadores puedan **analizar, comprender y trabajar con textos escritos o hablados en lenguaje humano**.

Podemos imaginarlo como el conjunto de técnicas que permiten a una máquina hacer tareas que normalmente asociamos a la lectura o al análisis del lenguaje. Por ejemplo, identificar las palabras importantes de un texto, clasificar documentos, detectar el tema principal de un artículo o responder preguntas.

Un símil útil para entenderlo en el aula es el de un **profesor que corrige muchos exámenes**. Con el tiempo, el profesor aprende a reconocer rápidamente ciertas palabras clave o estructuras que indican si el alumno ha entendido el tema. Los sistemas de NLP hacen algo parecido: analizan los textos buscando patrones que permitan interpretar su contenido.

Antes de la aparición de los modelos de lenguaje actuales, muchas aplicaciones de **procesamiento del lenguaje natural (NLP)** se basaban en métodos estadísticos relativamente simples pero muy eficaces. Uno de los más conocidos es **TF-IDF (Term Frequency – Inverse Document Frequency)**, una técnica que permite estimar qué palabras son más importantes dentro de un texto comparándolas con el resto de documentos de una colección. La idea es sencilla: una palabra que aparece muchas veces en un documento suele ser relevante para ese texto, pero si esa misma palabra aparece en casi todos los documentos —como ocurre con artículos o preposiciones— su valor informativo es menor. TF-IDF combina estas dos medidas para identificar qué términos caracterizan realmente un documento dentro de un conjunto más amplio de textos.

Durante años, este tipo de técnicas fue fundamental en numerosas aplicaciones de análisis de texto. Se utilizaban en buscadores para ordenar documentos según su relevancia, en sistemas de recomendación, en clasificación automática de textos o para detectar temas dominantes dentro de grandes colecciones de documentos. A partir de estos métodos también surgieron otras tareas habituales del procesamiento del lenguaje natural, como identificar si un mensaje es spam o no, analizar el tono de una opinión para detectar sentimientos positivos o negativos, localizar nombres o fechas dentro de un texto o generar resúmenes automáticos de documentos.

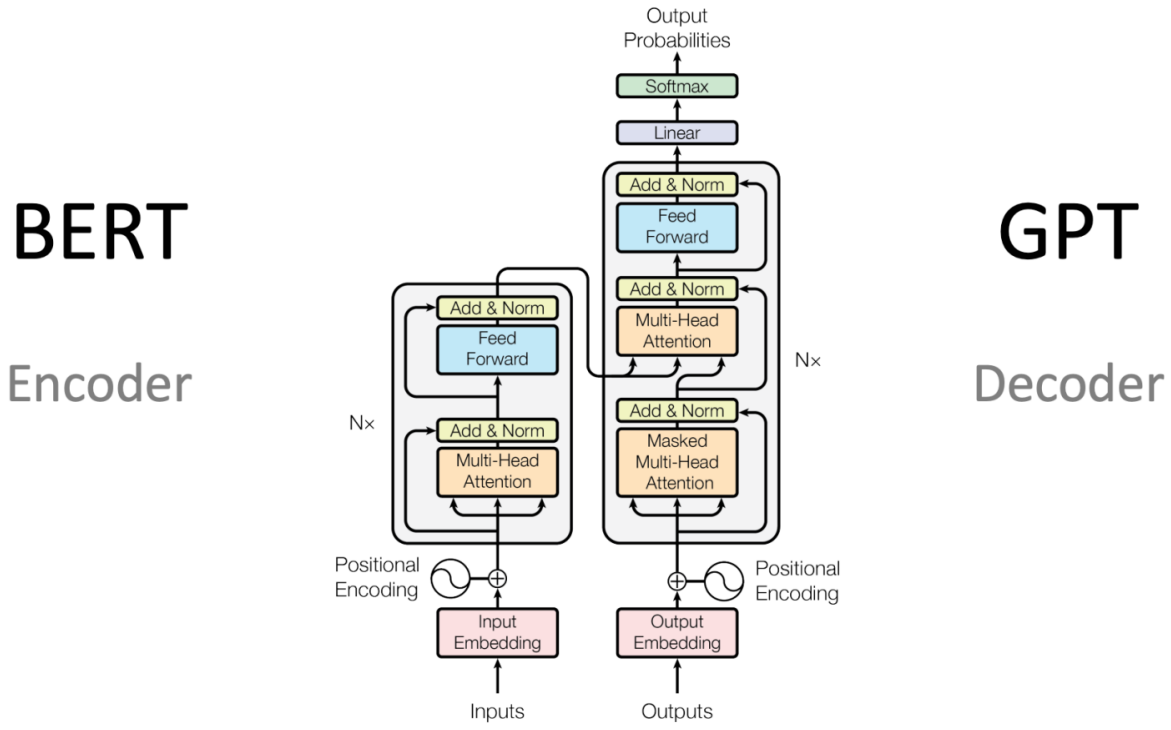
Durante mucho tiempo, todas estas aplicaciones se resolvieron combinando **estadística, lingüística y reglas programadas manualmente**. Aunque estos enfoques eran relativamente simples comparados con los modelos actuales de inteligencia artificial, constituyeron la base de muchas herramientas de análisis de texto y permitieron desarrollar gran parte del procesamiento automático del lenguaje antes de la llegada de las redes neuronales profundas y los modelos generativos modernos.

Sin embargo, el campo del NLP ha experimentado una auténtica revolución desde la aparición de una nueva arquitectura de modelos llamada **Transformers**, presentada en 2017 en el famoso artículo *"Attention is All You Need"*.

Los modelos basados en transformers son capaces de analizar el contexto completo de una frase y comprender mejor las relaciones entre palabras. Gracias a esta arquitectura se han desarrollado los actuales **modelos de lenguaje de gran tamaño (LLM)**, como GPT, Gemini o Claude.

Esto ha permitido que muchas tareas de procesamiento del lenguaje que antes requerían sistemas complejos y específicos ahora puedan resolverse con **un único modelo capaz de realizar múltiples tareas**: traducir, resumir, responder preguntas o generar texto.

En otras palabras, técnicas clásicas como TF-IDF o los modelos estadísticos tradicionales siguen siendo importantes para entender los fundamentos del NLP, pero los modelos actuales basados en transformers han ampliado enormemente las capacidades de los sistemas de lenguaje.



Arquitectura de los modelos Transformers publicada por primera vez por Google en el 2017

Desde el punto de vista educativo, el NLP puede entenderse como **un conjunto de herramientas para analizar grandes cantidades de texto**. En un mundo donde cada día se generan millones de documentos, artículos y mensajes, estas técnicas permiten organizar la información, detectar patrones y extraer conocimiento.

Para el profesorado de asignaturas científicas o tecnológicas, explicar el NLP también puede ser una buena oportunidad para conectar **lingüística, estadística y computación**, mostrando cómo el lenguaje humano puede estudiarse y analizarse mediante modelos matemáticos y algoritmos.

De los Transformers a la IA generativa

Durante muchos años, las técnicas de **Procesamiento del Lenguaje Natural (NLP)** se centraron principalmente en analizar textos: clasificar documentos, detectar palabras clave o traducir frases sencillas. Sin embargo, la aparición de los **modelos basados en transformers** supuso un cambio profundo en este campo y abrió la puerta a lo que hoy conocemos como **IA generativa**.

El punto de inflexión llegó en 2017 con la publicación del artículo científico **“Attention is All You Need”**, donde se presentó la arquitectura de los **transformers**. Este tipo de modelos introdujo un mecanismo llamado **atención**, que permite analizar las relaciones entre todas las palabras de una frase al mismo tiempo. Gracias a esto, los sistemas pueden comprender mejor el contexto

completo de un texto.

Un símil útil para entenderlo es imaginar que, cuando leemos una frase, no analizamos cada palabra de forma aislada. En realidad, nuestro cerebro conecta unas palabras con otras para comprender el significado global. El mecanismo de atención de los transformers intenta hacer algo parecido: **relacionar cada palabra con las demás para interpretar mejor el mensaje**.

Gracias a esta arquitectura, los modelos de lenguaje comenzaron a entrenarse con cantidades enormes de texto procedente de libros, artículos, páginas web o documentos. Durante el entrenamiento, el modelo aprende a **predecir la siguiente palabra dentro de una secuencia**, pero al hacerlo también aprende patrones complejos del lenguaje.

Este proceso dio lugar a los **Large Language Models (LLM)** o modelos de lenguaje de gran tamaño, como GPT, BERT, LLaMA o Gemini. Estos modelos no solo pueden analizar textos, sino también **generarlos**: redactar explicaciones, resumir información, traducir entre idiomas o mantener conversaciones.

Aquí es donde aparece el concepto de **IA generativa**. Mientras que los sistemas de IA tradicionales se centraban en clasificar o analizar información, los modelos actuales pueden **crear contenido nuevo** a partir de lo que han aprendido durante el entrenamiento.

Por ejemplo, un modelo generativo puede:

- redactar un texto explicativo
- generar código de programación
- crear preguntas para un examen
- producir imágenes a partir de descripciones
- sintetizar música o voz

Aunque estas aplicaciones parecen muy distintas, muchas de ellas comparten la misma idea fundamental: **aprender patrones en grandes conjuntos de datos y utilizarlos para generar nuevos resultados**.

En el ámbito educativo, esta evolución ha transformado las posibilidades de uso de la IA. Los modelos de lenguaje ya no solo sirven para analizar textos, sino que pueden actuar como **asistentes para generar materiales didácticos, ejemplos, explicaciones o actividades**.

No obstante, es importante recordar que estos sistemas no “piensan” ni comprenden el mundo como lo hacen las personas. Funcionan identificando patrones estadísticos en los datos con los que fueron entrenados. Por ello, sus resultados siempre deben interpretarse con sentido crítico.

En resumen, la combinación del **NLP tradicional con la arquitectura de los transformers** ha permitido el desarrollo de los actuales sistemas de **IA generativa**, capaces de producir texto, imágenes, audio o vídeo. mediante los llamados LLMs o modelos de lenguaje. Esta evolución representa uno de los avances más significativos de la inteligencia artificial en las últimas décadas y está teniendo un impacto directo en ámbitos como la educación, la ciencia o la comunicación.

Los Modelos de Lenguaje

Los modelos de lenguaje (LLM)

Los **Large Language Models (LLM)** son modelos de inteligencia artificial diseñados para comprender y generar lenguaje natural. Se entrenan con enormes cantidades de texto y utilizan redes neuronales basadas en la arquitectura **transformer** para aprender patrones del lenguaje y producir respuestas coherentes. Estos modelos pueden realizar tareas como responder preguntas, resumir documentos, traducir idiomas o generar código.

Aunque todos los LLM comparten principios tecnológicos similares, pueden clasificarse según distintos criterios: su grado de apertura, la forma en que se ejecutan y algunas características técnicas fundamentales.

Tipos de LLM según su acceso

Modelos propietarios o cerrados

Los modelos cerrados son desarrollados por empresas que no publican completamente el modelo ni los datos con los que se entrenó. Normalmente se utilizan mediante **APIs en la nube** o plataformas online.

Ejemplos conocidos incluyen modelos como GPT de OpenAI, Gemini de Google o Claude de Anthropic. Estos sistemas suelen ofrecer un rendimiento muy alto porque están entrenados con grandes infraestructuras de computación y enormes conjuntos de datos, pero su uso depende de las condiciones y servicios de la empresa que los desarrolla.

Modelos abiertos

Los modelos abiertos (open source u open weight) publican total o parcialmente sus parámetros para que puedan descargarse y ejecutarse localmente o modificarse.

Ejemplos de este tipo de modelos son Llama, Mistral o algunas versiones de Qwen. Estos modelos permiten a investigadores y desarrolladores experimentar con ellos, entrenarlos con nuevos datos o integrarlos en sistemas propios sin depender completamente de servicios externos.

Modelos online y modelos locales

Modelos online

Muchos LLM se utilizan **a través de internet mediante APIs**. El usuario envía una consulta al servidor del modelo y recibe una respuesta generada por el sistema.

Este enfoque tiene varias ventajas:

- no requiere hardware potente
- permite acceder a modelos muy grandes
- el proveedor se encarga de las actualizaciones

Sin embargo, también implica dependencia de internet, posibles costes por uso y menor control sobre los datos enviados.

Modelos ejecutados localmente

Los modelos abiertos pueden ejecutarse **directamente en un ordenador o servidor local**. Esto permite trabajar sin conexión a internet y mantener los datos dentro de la propia infraestructura.

Este enfoque es especialmente interesante para investigación, entornos educativos o aplicaciones que requieren mayor privacidad.

Parámetros técnicos importantes

Para comprender cómo funcionan los LLM conviene conocer algunos conceptos técnicos básicos.

Tokens

Los modelos de lenguaje no trabajan exactamente con palabras completas, sino con **tokens**, que son fragmentos de texto. Un token puede ser una palabra, una parte de palabra o incluso un signo de puntuación. Los modelos generan texto prediciendo el siguiente token más probable en una secuencia.

Contexto

El **context window** o ventana de contexto es la cantidad de tokens que el modelo puede analizar al mismo tiempo. Cuanto mayor es el contexto, más información puede tener en cuenta el modelo al generar una respuesta.

Tokens de entrenamiento

Los LLM se entrenan con cantidades gigantescas de texto, que pueden alcanzar **billones de tokens**. Cuantos más datos de entrenamiento tenga el modelo, mayor será su capacidad para aprender patrones complejos del lenguaje.

Tamaño del modelo

El tamaño de un LLM se mide normalmente por su número de **parámetros**, que son las variables internas de la red neuronal que el modelo ajusta durante el entrenamiento. En general, un mayor número de parámetros permite representar patrones más complejos, aunque también requiere más recursos computacionales.

Tabla de algunos LLM populares

Modelo	Empresa / Organización	Tipo	Características
GPT (ChatGPT)	OpenAI	Cerrado	Muy extendido, multimodal
Gemini	Google DeepMind	Cerrado	Multimodal, gran contexto
Claude	Anthropic	Cerrado	Fuerte enfoque en seguridad
Llama	Meta	Abierto	Muy usado en investigación
Mistral	Mistral AI	Abierto	Modelos eficientes y rápidos
Qwen	Alibaba	Abierto / mixto	Multilingüe y adaptable
Falcon	TII	Abierto	Popular en proyectos open source
DeepSeek	DeepSeek	Abierto	Alto rendimiento en código y razonamiento

Revision #14
Created 2025-12-17 18:24:06 CET by Maria
Updated 2026-03-16 17:44:27 CET by Luis Hueso