Nociones básicas del procesamiento del lenguaje natural moderno

Una introducción.

Por: Erick Merino

Contexto

El procesamiento del lenguaje natural, por sus siglas en inglés NLP, se trata de hacer que el computador sea capaz de interpretar y "hacer algo" con el lenguaje natural como el inglés, español, chino, portugués, etc...

Tokenización

La primera tarea importante es la TOKENIZACIÓN, que se trata de separar las palabras y elementos de un texto. un enfoque directo es separar por espacio, pero también hay que contar con los puntos, comas, comillas, comillas dobles, y otros signos, números, etc. saber diferenciar entre puntos suspensivos y punto final también es parte. Se usan expresiones regulares y diccionarios.

Significado

La segunda parte es obtener el significado de una palabra, para esto hay varios enfoques, los más importantes son:

- STEMING: Cortar la palabra en su raíz usando una serie de reglas preestablecidas, por ejemplo, transformar "estar" en "est".
- LEMMA: Transformar usando IA una palabra en su forma "original" o "fundamental", por ejemplo "eres", "soy", "somos" a "ser" (la forma infinitiva del verbo).
- EMBEDDINGS: Transformar cada palabra a un vector n-dimensional (donde cada dimensión tiene un significado oculto). Por ejemplo, la palabra "perro" tiene varias dimensiones, "es animal", "amigo del ser humano", etc... estos se codifican como números, y las dimensiones se obtienen por un proceso de IA (no se definen previamente, hay que interpretarlas luego de entrenado el modelo). Este es el enfoque que está de moda. Permite también volver desde el vector hasta la palabra que más se acerca a la definición (en base a los números que describen la palabra). Por lo cual tiene ventajas, además se puede "operar" con palabras, ejemplo: rey + mujer = reina.

2018: Transformers

Con esto en mente, el 2017 Google hace su gran avance con los **TRANSFORMERS**, modelos de IA que lo que hacen es que mapean en una oración, la importancia de cada palabra con cada una de las otras, que es "cuánto pesa la primera, segunda, tercera.. palabra en las demás". Con esto se obtiene una matriz 2D cuadrada del tamaño de palabras, con el "mapa de atención" de una oración. Hoy en día los modelos grandes pueden manejar ventanas de contexto de 4000 o hasta 128mil tokens. Una ventana de contexto es el tamaño de la oración o cuántas palabras se toman en cuenta para el mapa de atención.

Ejemplo de mapa de atención

Fuente: https://www.aprendemachinelearning.com/como-funcionan-los-transformers-espanol-nlp-gpt-bert/

2019: BERT

Con esto en mente, el 2019 Google crea el modelo **BERT**, que permite tokenizar (y convertir en embeddings), así como generar texto para oraciones cortas, como traducciones, o "adivinar la palabra que falta" entre otras operaciones. Estos modelos pesan en promedio unos 300 megas, y no requieren una muy grande capacidad de cómputo.

El nombre BERT viene porque usa ventanas de contexto que utilizan las palabras a la izquierda y a la derecha, enfoque distinto a los LLM que para predecir una palabra siguiente sólo usan las palabras previas.

SLM

En base a éstos, se crean los modelos mT5 y SLM (Sequence to Sequence Language model), que son como BERT pero específico para tareas de "secuencia de texto a secuencia de texto", como por ejemplo traducir, resumir, o responder preguntas cortas. Estos modelos pesan un poquito más que BERT pero no mucho más.

Se crearon originalmente para traductores de idiomas.

Modelos grandes del lenguaje

En 2022 entra la GRAN Revolución, y es que usando modelos TRANSFORMERS (DECODERS, es decir sólo usa información de palabras previas) de decenas o cientos de capas de profundidad, crearon CHATGPT, que está entrenado:

- Primero en predecir la siguiente palabra de un texto (dada la ventana de contexto que hablamos antes).
- Una vez bien entrenado en esa tarea, se hace un FINETUNNING (ajuste pequeño de parámetros), para que sepa responder a preguntas, usando algún dataset como por ejemplo “Alpaca” (conjunto de datos de pregunta – respuesta).
- Una vez hecho esto, se puede entrenar de nuevo con aprendizaje reforzado por feedback humano (personas votando las mejores respuestas), para que sepa responder de cierta forma o de mejor forma, o ajustado a parámetros de censura y buenas costumbres, etc.

Open Source (Código abierto)

Luego surgieron modelos Open Source para esto, el primero fue LLAMA (y ahora LLAMA 2) de Meta, en versiones de 7mil millones, 13mil millones y 70mil millones de parámetros. También una empresa europea "Mistral" ha creado los suyos, Mistral 7B y ahora Mixtral 8x7B.

MoE: Mezcla de Expertos

Una técnica "moderna" para crear estos modelos es el Mixture of Experts, el cual, en palabras sencillas se trata de tener un TOKENIZADOR común, y mapas de atención comunes, y luego varias redes neuronales "tipo" LLM que funcionan en paralelo, y al final una red neuronal que "junta" todos los resultados y entrega la(s) palabras resultantes. El más famoso es Mixtral 8x7B, que usa 8 modelos de 7mil millones.

Pequeños modelos del lenguaje

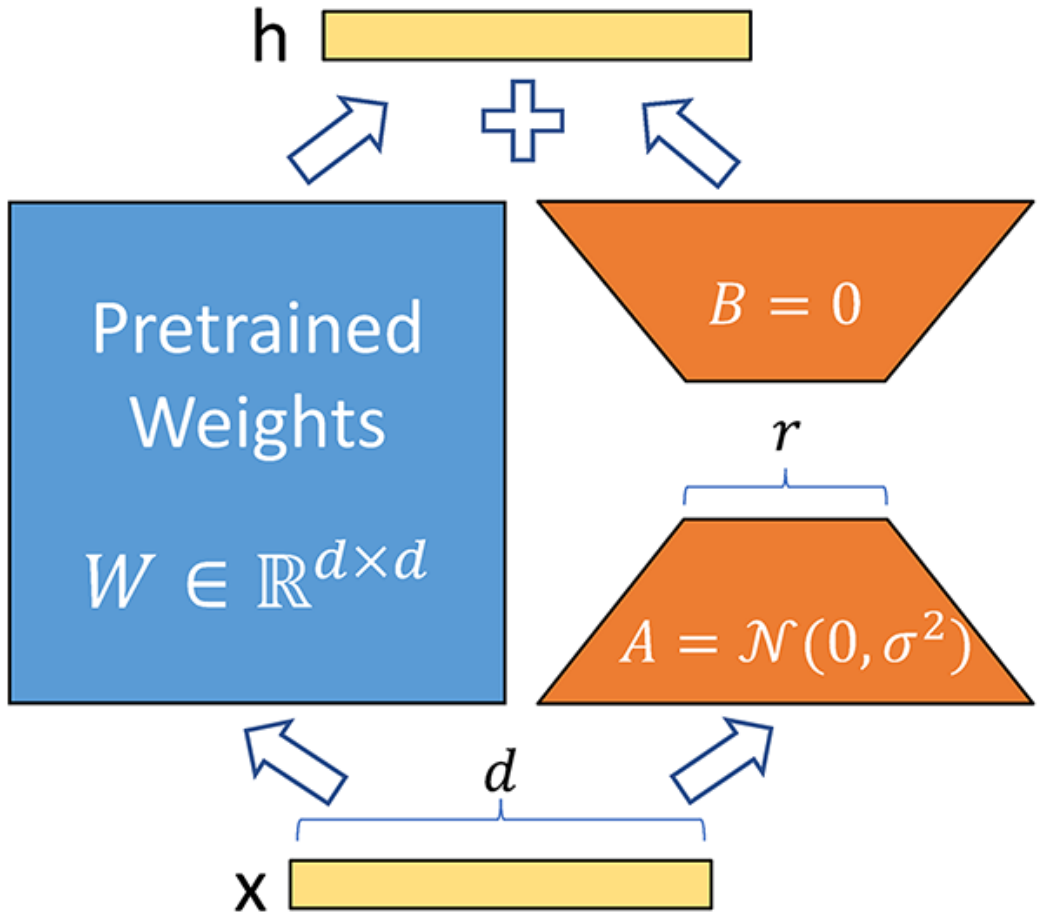
Otra evolución actual es la aparición de pequeños modelos del lenguaje, de 1B (mil millones de parámetros) o similar. Los ejemplos más clásicos son TinyLlama 1.1B (ya hay versiones MoE de 3x1.1B), Phi2 (creado por Microsoft sólo para inglés en principio), y MiniCPM (creo que es de origen chino).

Técnicas relacionadas

Para finalizar, mencionar dos técnicas relacionadas a los LLM:

- LoRa: Una técnica para hacer **finetunning** de modelos LLM con pocos recursos y en poco tiempo. La técnica consiste en pocas palabras de que dado que un LLM es una red neuronal profunda, con una matriz de "pesos" o "parámetros", NxM (tamaño N filas y M columnas), se descompone en una multiplicación de dos matrices NxR * RxM, donde R es el llamado "Rank", y es un número pequeño. Finalmente, como el resultado es más "débil" que la red neuronal total, lo que se hace al momento de ejecutarse, es que se ejecuta el LLM original, y se promedia con el resultado del LoRa, y así se obtiene el resultado final.
- RoPe Scaling: RoPe es una técnica que usan los LLM para fijar el tamaño de su ventana de contexto. La técnica de escalamiento permite modificar el parámetro de la ventana de contexto para lograr ventanas mucho mayores. No es perfecto el resultado, pero funciona bastante bien, y se escalan los mapas de atención al nuevo tamaño de ventana.

LoRa


h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

Bonus: Tareas típicas del NLP

Algunas tareas típicas del procesamiento del lenguaje natural son:

- Summarization: Crear resúmenes
- Traducción
- Análisis de sentimiento/emociones
- Reconocimiento de entidades: Reconocer nombres de cosas, ciudades, personas, etc
- POS tagging: Clasificar las palabras según cumplen función de sustantivo, adjetivo, etc
- Extracción de relaciones / Grafos de conocimiento
- Clasificación de texto en categorías
- Modelamiento de tópicos: Obtener temas/tópicos comunes entre muchos documentos, cada tema es una distribución probabilística de palabras
- Recuperación de la información: Buscadores como Google
- Responder a preguntas / Chatbots
- Obtener la palabra faltante en una oración, completar texto
- Etc...

Gracias

Hasta la próxima