



Procesamiento de Lenguaje Natural

De BERT a LLMs: Explorando los Límites del Procesamiento de Lenguaje Natural

Mauricio Toledo-Acosta
mauricio.toledo@unison.mx

Departamento de Matemáticas
Universidad de Sonora



Section 1

Introducción



- ¿Cómo toman en cuenta el contexto de una palabra modelos como Word2Vec, FastText, GloVe?
- ¿Cómo son los embeddings de *banco*? ¿cómo podemos considerar las diferentes acepciones? ¿concept-net?
- ¿Cómo contribuyen los embeddings de palabras para crear embeddings de documentos? ¿se considera el orden? ¿las relaciones entre palabras?



Machine Translation

Procesamiento
de Lenguaje
Natural

Introducción

Attention is all
you need

One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.

Warren Weaver, 1947

- Rule-based translation (1970-1990).
- Statistical machine translation (1990-2010).
- Deep learning-based translation (2003-).

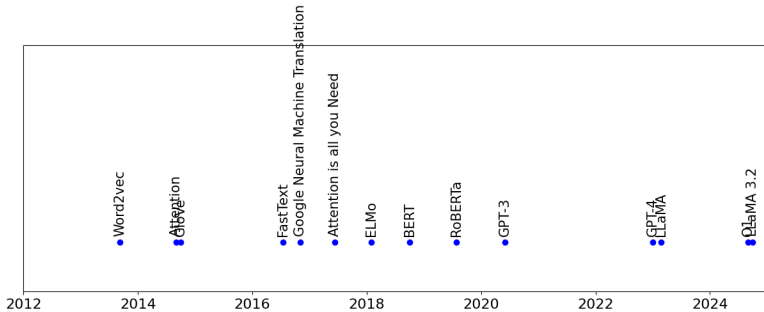


Timeline

Procesamiento de Lenguaje Natural

Introducción

Attention is all
you need





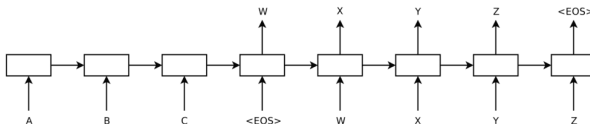
Seq2Seq

Procesamiento de Lenguaje Natural

Introducción

Attention is all
you need

- Seq2Seq es una familia de enfoques de ML en el NLP para: traducción de idiomas, generación de subtítulos para imágenes, modelos conversacionales y resumen de textos.
- Convierten una secuencia de entrada en otra secuencia de salida.



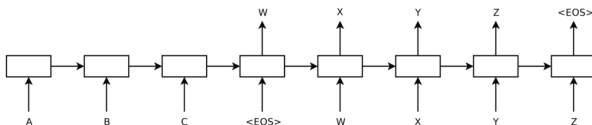


Seq2Seq

Procesamiento
de Lenguaje
Natural

Introducción

Attention is all
you need



- El modelado y la generación de secuencias se realizaron mediante arquitecturas RNN. Esto llevó al problema del desvanecimiento de gradiente.
- La arquitectura LSTM se convirtió en la estrategia estándar para el modelado de secuencias largas hasta la aparición en 2017 de los Transformers.
- Las RNN operan un token a la vez, del primero al último; no pueden operar en paralelo con todos los tokens de una secuencia.

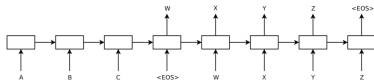


Seq2Seq

Procesamiento de Lenguaje Natural

Introducción

Attention is all
you need



- Los primeros modelos seq2seq carecían de mecanismo de atención.
- El vector de estado sólo es accesible después de procesar la última palabra del texto de origen.
- Este vector conserva la información sobre toda la frase original, en la práctica la información se conserva mal, ya que la entrada es procesada secuencialmente y si la entrada es larga, el vector de salida no podría contener toda la información relevante.



- ELMo (embeddings from language model) es un modelo de embeddings de secuencias de palabras.
- La arquitectura de ELMo logra una comprensión contextual de los tokens.
- ELMo es un LSTM bidireccional multicapa sobre una capa de embeddings de tokens. La salida de todas las LSTM concatenadas consiste en la incrustación de tokens.

Artículo

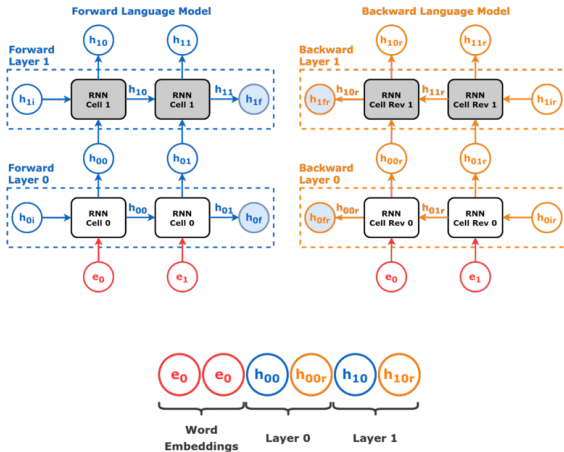


ELMo

Procesamiento
de Lenguaje
Natural

Introducción

Attention is all
you need





- El **mecanismo de atención** es una mejora introducida en 2014 para abordar las limitaciones de la arquitectura básica Seq2Seq.
- Permite al modelo centrarse en diferentes partes de la secuencia de entrada durante el proceso de decodificación. Es un mecanismo que permite que los tokens *hablen* entre sí.
- En 2016, Google Translate se actualizó con **Google Neural Machine Translation**, que reemplazó el modelo basado en *statistical machine translation*. GNMT fue un Seq2Seq donde el codificador y decodificador contenían 8 capas de LSTMs bidireccionales.



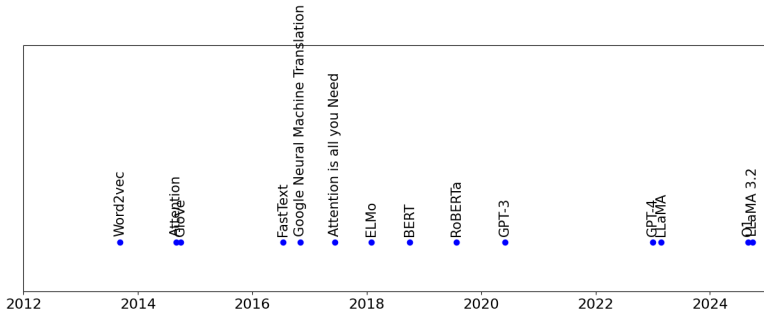


Timeline

Procesamiento de Lenguaje Natural

Introducción

Attention is all
you need





◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻ 15/26



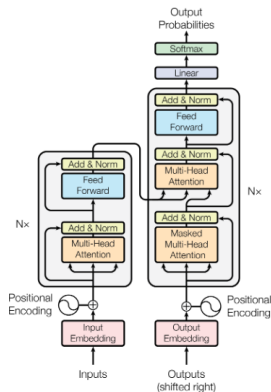
El transformador

Procesamiento
de Lenguaje
Natural

Introducción

Attention is all
you need

- Nueva arquitectura llamada **transformador**, basada en un mecanismo de atención multi-head (atención en paralelo).
- Los transformadores tienen la ventaja de no tener unidades recurrentes, por lo que requieren menos tiempo de entrenamiento que las arquitecturas RNN.
- Fué desarrollado para la traducción, encontró aplicaciones en los LLM, visión computacional, audio, etc.





Scaled Dot-Product Attention

Procesamiento
de Lenguaje
Natural

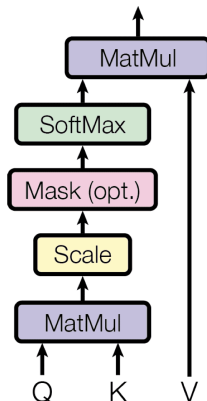
Introducción

Attention is all
you need

Conceptos importantes:

- Query
- Keys
- Values
- Masking
- Self-attention
- Transformers

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$





BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova

Google AI Language

{jacobdevlin, mingweichang, kentonl, kristout}@google.com

Hubo dos implementaciones:

- BERT_{BASE} (110 million parameters)
- BERT_{LARGE} (340 million parameters)

[Artículo, repositorio](#)

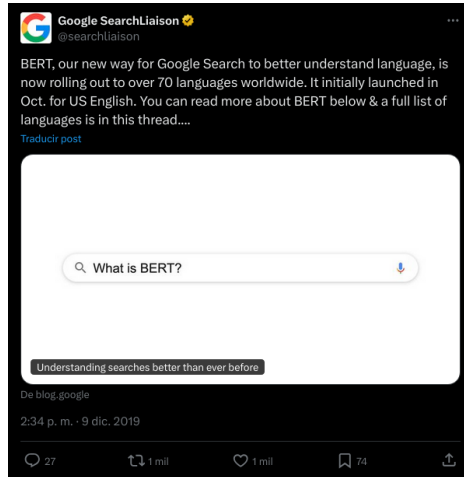


BERT: A blast from the past

Procesamiento
de Lenguaje
Natural

Introducción

Attention is all
you need



El anuncio, noticias 1, noticias 2



Ventajas de BERT

Procesamiento de Lenguaje Natural

Introducción

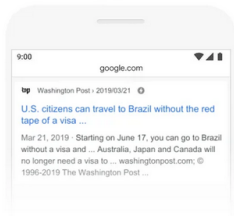
Attention is all
you need

Here's a search for "2019 brazil traveler to usa need a visa." The word "to" and its relationship to the other words in the query are particularly important to understanding the meaning. It's about a Brazilian traveling to the U.S., and not the other way around. Previously, our algorithms wouldn't understand the importance of this connection, and we returned results about U.S. citizens traveling to Brazil. With BERT, Search is able to grasp this nuance and know that the very common word "to" actually matters a lot here, and we can provide a much more relevant result for this query.

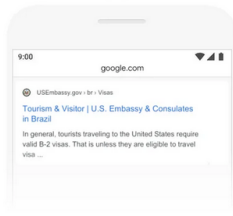


2019 brazil traveler to usa need a visa

BEFORE



AFTER



BERT en las búsquedas



- **RoBERTa** (2019): A Robustly Optimized BERT Pretraining Approach.
- **DistilBERT** (2019): A distilled version of BERT: smaller, faster, cheaper and lighter. [huggingface](#)
- **CamemBERT** (2020): Une variante de RoBERTa entraînée en un corpus français.

A partir de 2018, comenzó la serie de modelos GPT de Transformers *decoder only* de OpenAI.



BERT es una arquitectura de transformador *encoder only*. BERT consta de 4 partes:

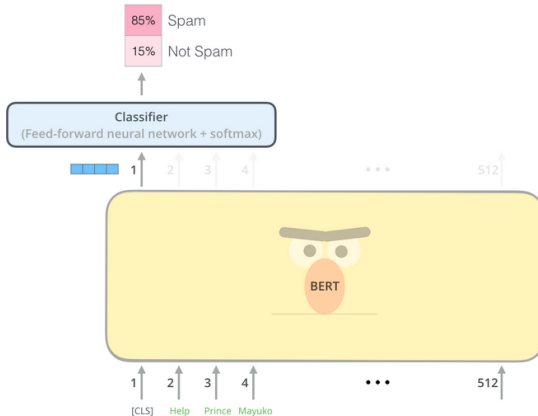


Fine-Tuning

Procesamiento de Lenguaje Natural

Introducción

Attention is all you need





¿Dónde encontrar modelos BERT?

Procesamiento
de Lenguaje
Natural

Introducción

Attention is all
you need



Hugging Face

Hugging Face, Inc. es una empresa conocida por su biblioteca de transformadores creada para aplicaciones de NLP y su plataforma que permite a los usuarios compartir **datasets** y **modelos** de ML. Su campo de acción principal es el NLP, pero también se centra en otras áreas del ML: CV, el aprendizaje por refuerzo y el aprendizaje supervisado.

[repositorio](#)



- Sesgos.



¿Dónde están los modelos?

Procesamiento
de Lenguaje
Natural

Introducción

Attention is all
you need

- Hugging face



Hugging Face

- Ollama



- LangChain



LangChain