

Sistemas de Preguntas y Respuestas con redes LSTM, Mecanismos de Atención y modelos pre-entrenados

Alumno: Navío Torres, José Luis

Asesor: Prof. César Lara Ávila

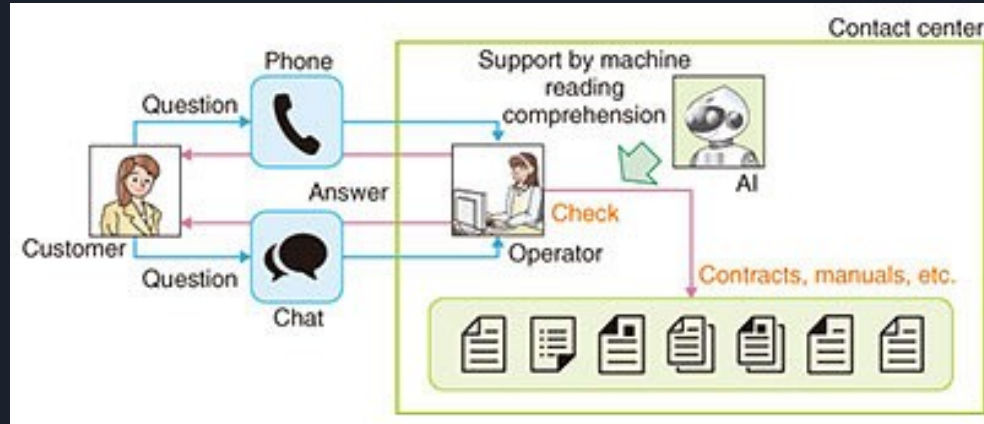


Índice


1. Introducción
2. Marco Teórico
3. Implementación
4. Experimentación y Resultados
5. Conclusiones y Trabajos Futuros

1. Introducción

- Desarrollo de Modelos de Comprensión Lectora como el de Preguntas y Respuestas basadas en texto tienen gran aplicabilidad en la industria.



Aplicación de sistema de Comprensión Lectora en un centro de atención al cliente[1]



1. Introducción

1.1. Problemática

- Pocos Modelos desarrollados para el lenguaje español.
- Dificultad de aplicar modelos grandes en tiempo real.



1.2 Objetivo General

- Implementar arquitecturas de redes neuronales para la tarea de Procesamiento de Lenguaje Natural (NLP) de Preguntas y Respuestas, y realizar un estudio de sus métricas para la implementación en una etapa de despliegue.



1.3 Objetivos Específicos

O1) Implementar y evaluar modelos basados en mecanismos de atención y redes LSTM: DrQA.

O2) Implementar y evaluar modelos basados en la arquitectura de Transformers como son BERT, RoBERTa y ELECTRA.

O3) Realizar un cuadro comparativo e identificar el modelo que proporcione mejores métricas para su uso en aplicaciones prácticas.

O4) Desarrollar una implementación de un prototipo de despliegue con el framework Streamlit.

2. Marco Teórico

2.1 NLP

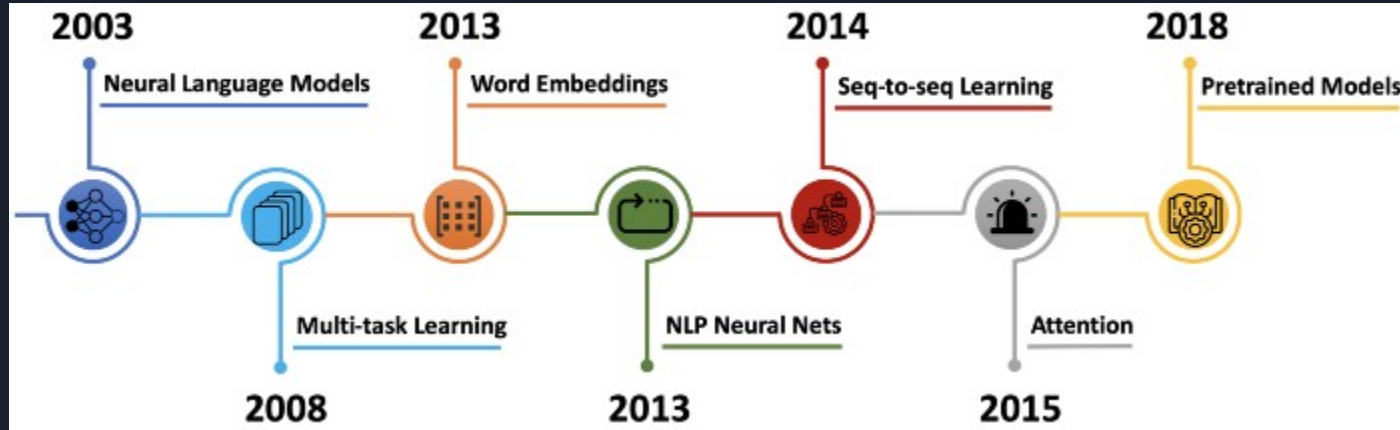


Tareas de NLP



Tarea de Preguntas y Respuestas

2.2 Inteligencia Artificial y Evolución de NLP



Acontecimientos en NLP en la era del deep learning. [2]

[2] A Brief History of Natural Language Processing - <https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-2-f5e575e8e37>

2.3. Conjunto de Datos: SQUAD-es

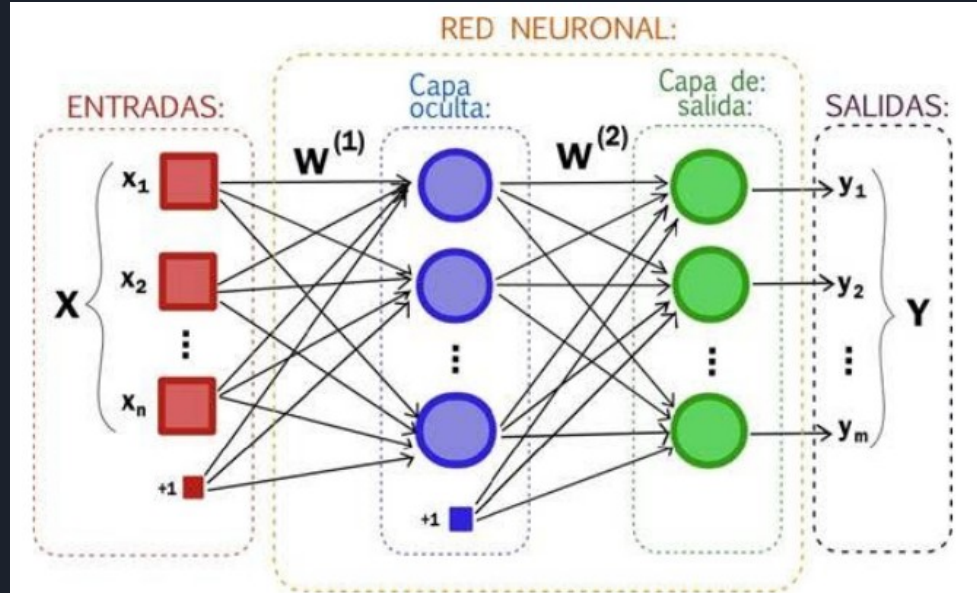
	Español
Contexto	Fryderyk Chopin nació en Zelazowa Wola , 46 kilómetros al oeste de Varsovia, en lo que entonces era el Ducado de Varsovia, un estado polaco establecido por Napoleón. El registro de bautismo de la parroquia da su cumpleaños el 22 de febrero de 1810 , y cita sus nombres en latín Fridericus Franciscus (en polaco, Fryderyk Franciszek). Sin embargo, el compositor y su familia utilizaron la fecha de nacimiento 1 de marzo, [n 2] que ahora se acepta generalmente como la fecha correcta.
Pregunta	1) ¿En qué pueblo nació Frédéric?
Respuesta	1) Zelazowa Wola ,
Pregunta	2) ¿Cuándo se registró su cumpleaños
Respuesta	2) 22 de febrero de 1810 ,
Pregunta	3) ¿Cuál es la forma latina del nombre de Chopin?
Respuesta	3) Fridericus Franciscus



2.3.1 Métricas

- Exact Match: (EM): El porcentaje de predicciones que exactamente concuerda con la respuesta.
- F1: Se calcula la superposición entre la predicción y la respuesta correcta.

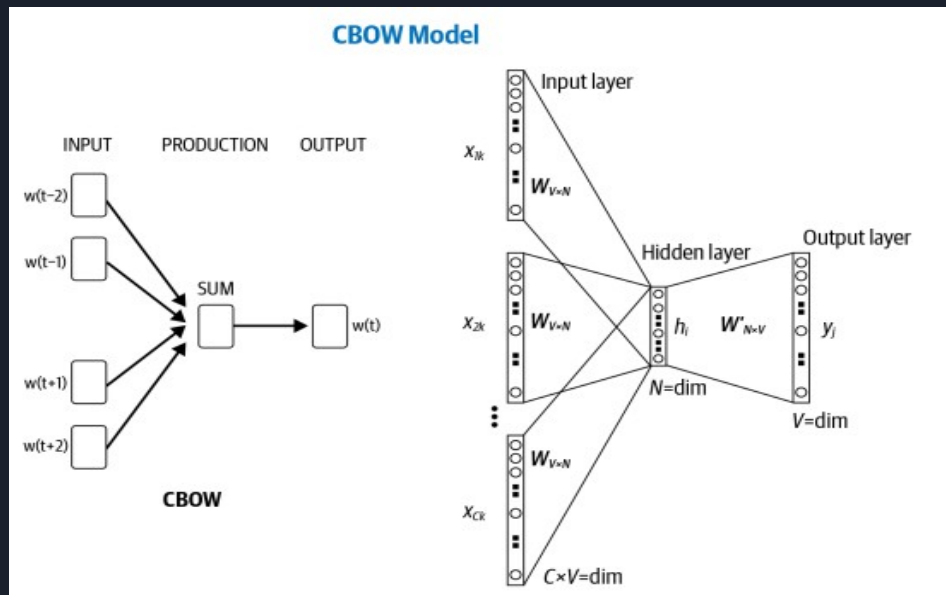
2.4. Redes Neuronales



Red Neuronal de 3 capas: Capa de entrada, Capa Oculta y Capa de Salida [3]

[3] URL : <https://sites.google.com/site/mayinteligenciartificial/unidad-4-redes-neuronales>.

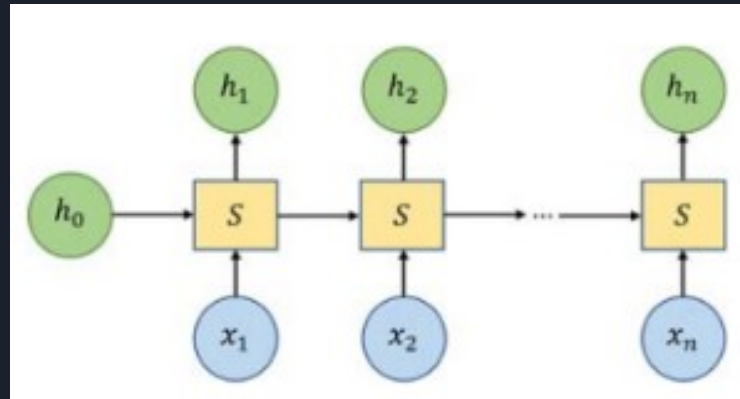
2.4.1 Word Embeddings



Entrenamiento de Word Embeddings [4]

[4] Sowmya Vajjala, Book: Practical Natural Language Processing

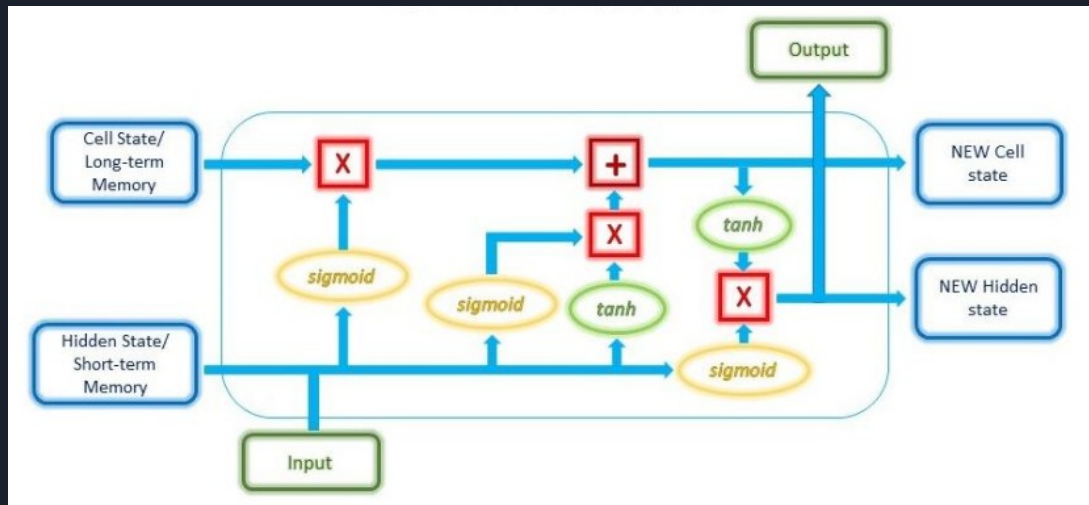
2.5 Redes Neuronales Recurrentes



Red Neuronal Recurrente [5]

[5] Chenguang Zhu. Machine reading comprehension: algorithms and practice. Elsevier, 2021.

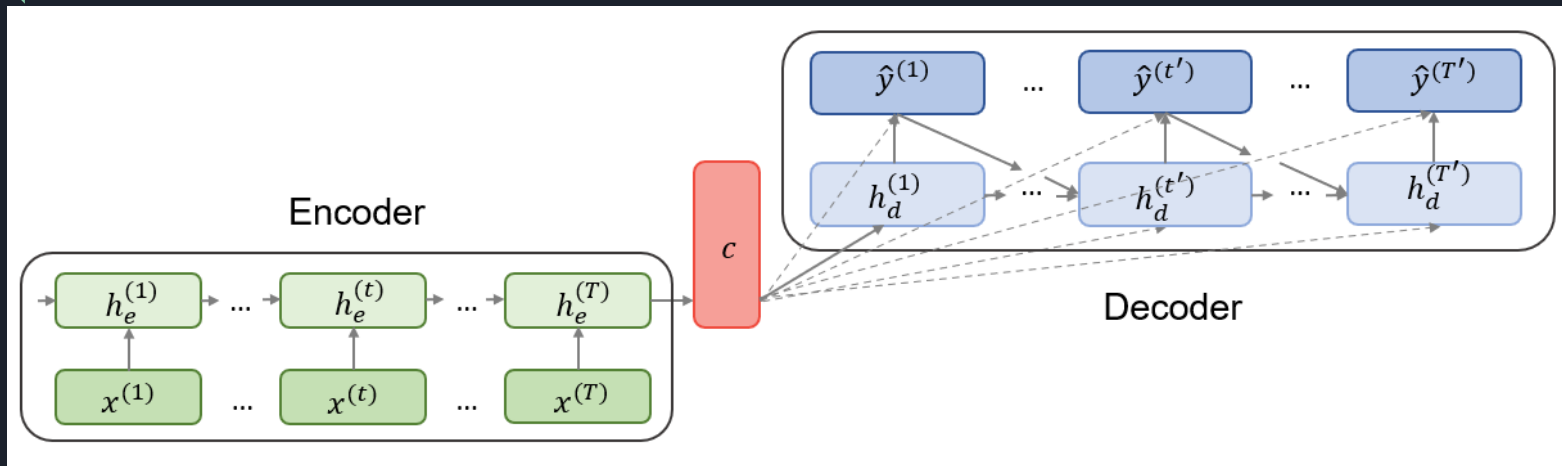
2.5.1 LSTM



Arquitectura interna de LSTM [6]

[6] Gabriel Loya. Long Short-Term Memory: From Zero to Hero with PyTorch. Ago. de 2019.
URL :
<https://blog.floydhub.com/long-short-term-memory-from-zero-to-hero-with-pytorch/>

2.5 Arquitectura Codificador-Decodificador

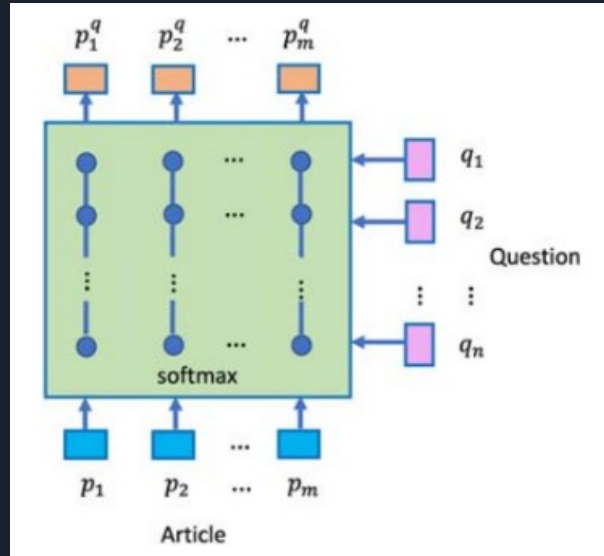


Arquitectura Codificador-Decodificador[7]

[7] Recurrent neural networks and their applications in NLP -

https://compstat-lmu.github.io/seminar_nlp_ss20/recurrent-neural-networks-and-their-applications-in-nlp.html

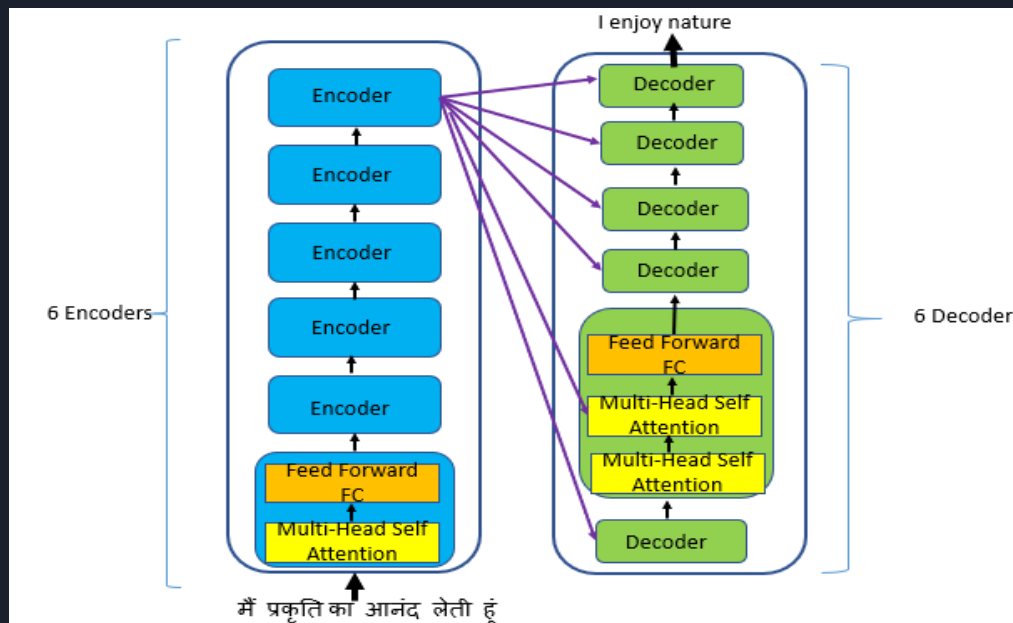
2.6 Mecanismos de Atención



Mecanismo de Atención Cruzada [8]

[8] Chenguang Zhu. Machine reading comprehension: algorithms and practice. Elsevier, 2021.

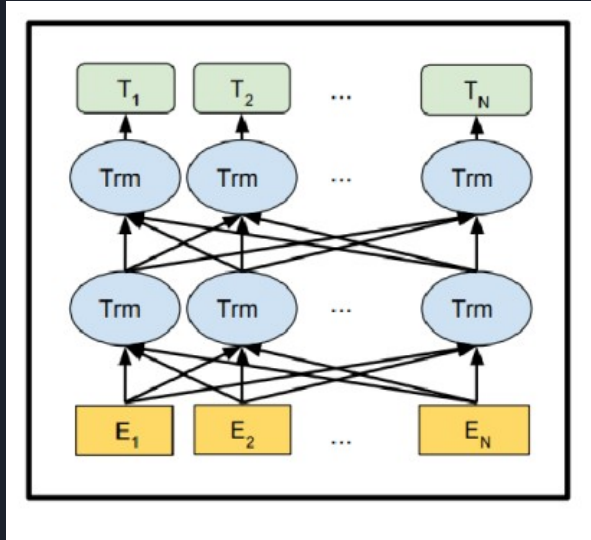
2.7 Modelos Transformer



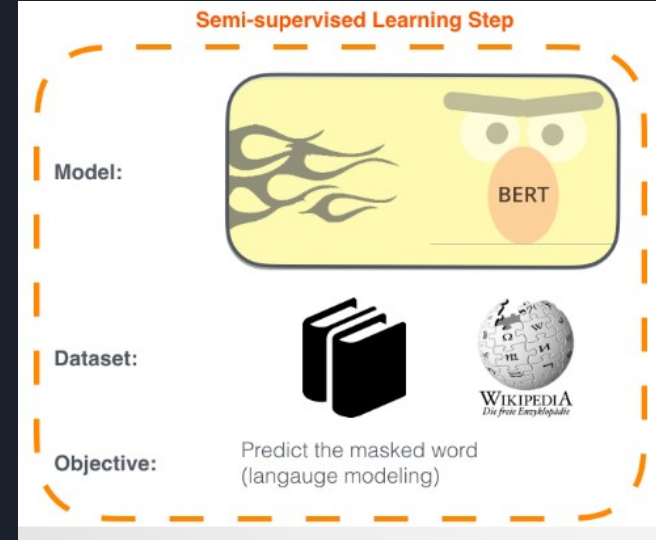
Arquitectura Transformer [9]

2.8. Modelos Pre-Entrenados

2.8.1. BERT(Bidirectional Encoder Representations from Transformers)



Arquitectura BERT[10]



Entrenamiento de BERT[11]

[10]h

<https://yashueth.blog/2019/06/12/bert-explained-faqs-understand-bert-working/#:~:text=What%20is%20BERT%3F,task%2Dspecific%20fine%2Dtuning>

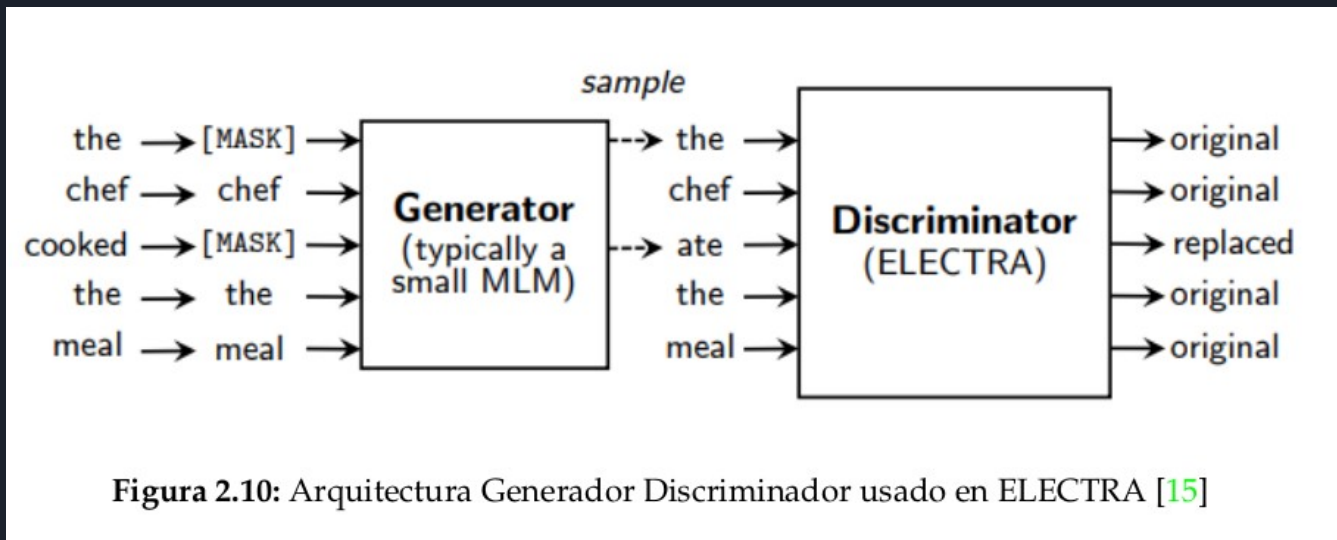
[11] <http://jalammar.github.io/illustrated-bert/>



2.8.2. RoBERTa

- Es una versión mejorada de BERT, por el cambio de hiperparámetros de entrenamiento, un tiempo mayor de entrenamiento sobre una mayor cantidad de corpus de texto (160Gb vs 16 Gb de BERT).
- Adicionalmente se eliminó el entrenamiento con la tarea de NSP.

2.8.3 ELECTRA



Arquitectura Electra [12]

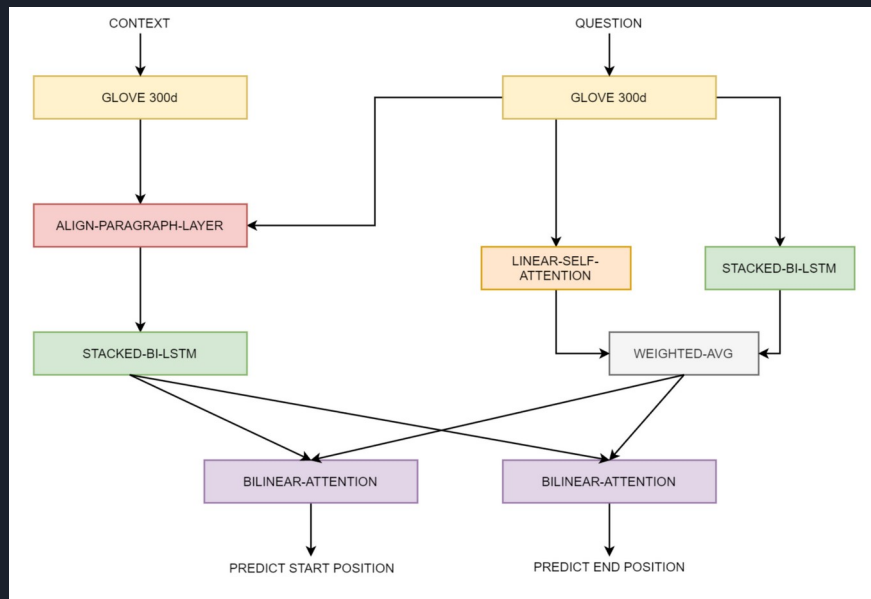
[12] Kevin Clark y col. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. 2020. arXiv: 2003.10555 [cs.CL]



3.Implementación

1. Se utilizaron los conceptos de Word Embeddings, LSTM y Mecanismos de Atención para la implementación del modelo base DrQA.
2. Se utilizaron los modelos basados en Transformers : BERT, RoBERTa y ELECTRA en español y la librería de Huggingface para realizar una etapa de Aprendizaje de transferencia con el dataset SQUAD en español.

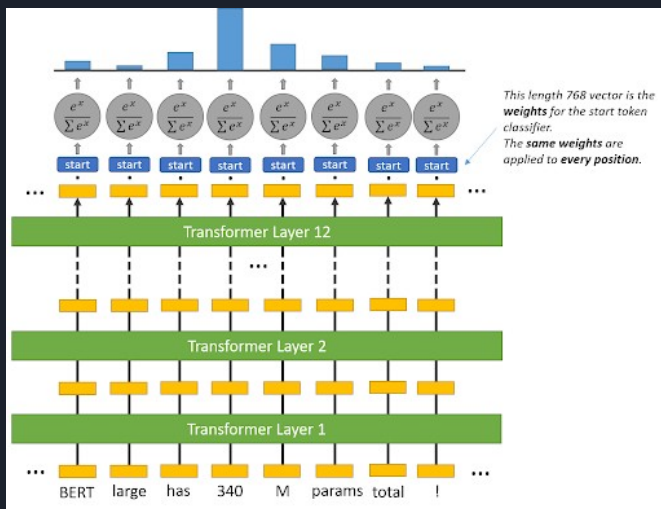
3.1. DrQA



Arquitectura de Dr.QA[13]

[13] <https://github.com/kushalj001/pytorch-question-answering>

3.2 Modelos pre entrenados y Aprendizaje por Transferencia



FineTuning de BERT para Preguntas y Respuestas [14]

[14] <https://mccormickml.com/2020/03/10/question-answering-with-a-fine-tuned-BERT/>

4. Experimentación y Resultados

4.1. Modelo de LSTM con Mecanismos de Atención: Dr QA

Hiperparametros de Entrenamiento:

Hiperparámetro	Valor
Número de épocas	10
Optimizador	Adam
Tamaño de Batch	128
Dimensión de capa oculta	128
Número de direcciones	2
Dropout	3
Dimension de word embeddings	300

Resultados:

Métrica	Resultado
EM	41.7
F1	59.5



4.2. Modelos Transformers

4.2.1. Modelo BERT en español

Hiperparametros de Entrenamiento:

Hiperparámetro	Valor
Número de épocas	2
Optimizador	AdamW
Tamaño de Batch	8
Taza de Aprendizaje	3e-5

Resultados:

Métrica	Resultado
EM	59.5
F1	76.6



4.2.3. Modelo RoBERTa en español

Hiperparametros de Entrenamiento:

Hiperparámetro	Valor
Número de épocas	2
Optimizador	AdamW
Tamaño de Batch	8
Taza de Aprendizaje	3e-5

Resultados:

Métrica	Resultado
EM	49.8
F1	66.4



4.2.4. Modelo ELECTRA en español

Hiperparametros de Entrenamiento:

Hiperparámetro	Valor
Número de épocas	2
Optimizador	AdamW
Tamaño de Batch	8
Taza de Aprendizaje	3e-5

Resultados:

Métrica	Resultado
EM	58.4
F1	75.1



4.2.5. Modelo ELECTRA-small en español

Hiperparametros de Entrenamiento:

Hiperparámetro	Valor
Número de épocas	2
Optimizador	AdamW
Tamaño de Batch	8
Taza de Aprendizaje	3e-5

Resultados:

Métrica	Resultado
EM	46.1
F1	64.2

4.3. Comparativa de Resultados

Curva de entrenamiento: en color fucsia ELECTRA base, en color celeste BERT, en color azul RoBERTa y en color naranja ELECTRA small.



Figura 4.1: Comparativa de los modelos de las funciones de pérdida en la etapa de entrenamiento vs número de pasos

Curva de evaluación: en color fucsia ELECTRA base, en color celeste BERT, en color azul RoBERTa y en color naranja ELECTRA small.



Figura 4.2: Comparativa de los modelos de las funciones de pérdida en la etapa de validación vs número de pasos



4.4. Cuadro Comparativo

Modelo	EM	F1	Peso	T.Entren. (1 época)	T. de Inferencia
DrQA	41.7	59.5	128 Mb	10 min	0.002 s.
BERT en español	59.5	76.6	438 Mb	4h	0.2 s.
RoBERTa en español	49.8	66.4	505 Mb	5h	0.2 s.
ELECTRA-small en español	46.1	64.2	53 Mb	1h	0.04 s.
ELECTRA-base en español	58.4	75.1	437 Mb	4h	0.21 s.



4.5. Discusión de Resultados

- El menor tiempo de entrenamiento y tiempo de inferencia del modelo DrQA se debe a la menor cantidad de parámetros de entrenamiento con respecto a los modelos pre-entrenados.
- El problema del uso de DrQA es el tamaño de los pesos de Word Embeddings que es del orden de 3 Gb.
- Entrenamiento sobre más de 2 épocas genera overfitting en los modelos.
- Los modelos BERTy ELECTRA para español nos brindan las mejores métricas.
- El modelo ELECTRA small supera en métricas al modelo base de redes neuronales LSTM con atención: DrQA
- El modelo Roberta para español no ofrece buenas métricas, esto puede ser ocasionado por una falta de datos en su etapa de entrenamiento del modelado del lenguaje.

4.6. Prototipo de Despliegue

Sistema de Preguntas y Respuestas usando BERT, RoBERTa y ELECTRA para español

Ingrese el texto a analizar

El 28 de julio, tal y como indica la Constitución Política del Perú, el presidente electo Pedro Castillo jurará al cargo ante el Congreso de la República en una sesión solemne en la cual asumirá formalmente el cargo a la cabeza del Poder Ejecutivo.

El evento marcará el inicio de la presidencia de la República para el periodo 2021- 2026 y congregará a otras autoridades de países que han confirmado su presencia. A continuación, un recuento de los jefes de Estado que han asegurado que vendrán a Lima para presenciar la toma de mando de Pedro Castillo.

El presidente de Bolivia, Luis Arce, confirmó que asistirá a la ceremonia en la sede del Poder Legislativo peruano. "Será grato acompañarle en esta fecha histórica, en la que el pueblo hermano también celebra el Bicentenario de su Independencia", señaló en Twitter.

Ingrese pregunta sobre el texto

¿Qué sucederá el 28 de Julio?

4.4.1. Resultados

Modelo:

- ☒ Bert
- ☐ Roberta
- ☐ ElectraBase
- ☐ ElectraSmall

Realizando Predicción:

La respuesta del Sistema es:

presidente electo Pedro Castillo jurará al cargo

El tiempo de inferencia es:

0.19401907920837402

Modelo:

- ☐ Bert
- ☒ Roberta
- ☐ ElectraBase
- ☐ ElectraSmall

Realizando Predicción:

La respuesta del Sistema es:

Congreso de la República

El tiempo de inferencia es:

0.1967320442199707

4.4.1. Resultados

Modelo:

- ☐ Bert
- ☐ Roberta
- ☒ ElectraBase
- ☐ ElectraSmall

Realizando Predicción:

La respuesta del Sistema es:

el presidente electo Pedro Castillo jurará al cargo ante el Congreso de la República

El tiempo de inferencia es:

0.2109997272491455

Modelo:

- ☐ Bert
- ☐ Roberta
- ☐ ElectraBase
- ☒ ElectraSmall

Realizando Predicción:

La respuesta del Sistema es:

Luis

El tiempo de inferencia es:

0.051760196685791016



5. Conclusiones y Trabajos Futuros

5.1. Conclusiones

- Se experimentó con distintas arquitecturas de redes neuronales, entre ellas: redes recurrentes LSTM con mecanismos de atención como DrQA y arquitecturas basadas en Transformers aplicados modelos pre-entrenados como BERT, RoBERTa y ELECTRA en el lenguaje español.
- Se aplicó la técnica de transferencia de aprendizaje para la tarea de Preguntas y Respuestas con el conjunto de datos SQUAD v1, traducido al español.
- Se desarrolló una comparativa de las métricas de las arquitecturas de redes neuronales estudiadas, donde se observó que ELECTRA en español y BERT en español ofrecen los mejores desempeños debido a sus métricas y su tiempo de inferencia.
- Se desarrolló un prototipo de despliegue con el framework Streamlit para un sistema de preguntas y respuestas incluyendo a los modelos que nos brindan las mejores métricas.



5.2. Trabajos futuros

- Usar datasets como SQUAD v2, MLQA, XQUAD para mejora del estudio.
- Sistema de preguntas y Respuestas de Dominio Abierto para aplicaciones en industria como en centros de atención al cliente.
- Desarrollo de Chabot con modelos pre entrenados basados en Transformers.

Gracias

