

Sistema Inteligente de Preguntas y Respuestas utilizando Técnicas de Procesamiento de Lenguaje Natural

Alumno: José Luis Navío Torres
Asesor: Mg. Herminio Paucar Curasma

Índice

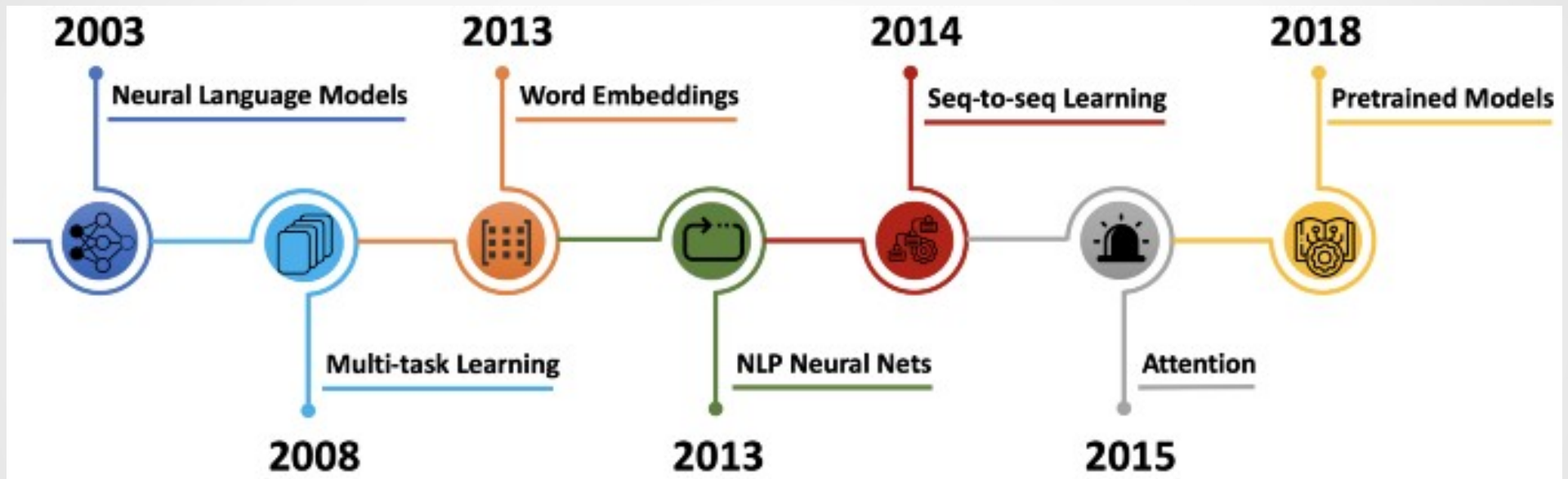
- 1. Justificación
- 2. Objetivos
- 3. Antecedentes
- 4. Modelos
- 5. Conjunto de Datos
- 6. Métricas
- 7. Experimentos
- 8. Resultados y Discusión
- 9. Trabajos Futuros
- 10. Conclusiones

1. Justificación



- Investigación Dominante de NLP → Inglés

Progreso en investigación de Tareas de NLP



- Acontecimientos en NLP en la era del deep learning. [1]

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

2. Objetivos

2.1 Objetivos Generales

- Determinar cuáles son las últimas técnicas utilizadas en Inteligencia Artificial para abordar los problemas de tareas de NLP.
- Determinar el estado del Arte para la tarea de Preguntas y Respuestas.
- Establecer cuáles son los modelos más adecuados para la tarea de Preguntas y Respuestas en español.

2.2 Objetivos Específicos

- Desarrollar un prototipo de despliegue de la tarea estudiada.
- Establecer métricas de la tarea de Preguntas y Respuestas para el lenguaje español.

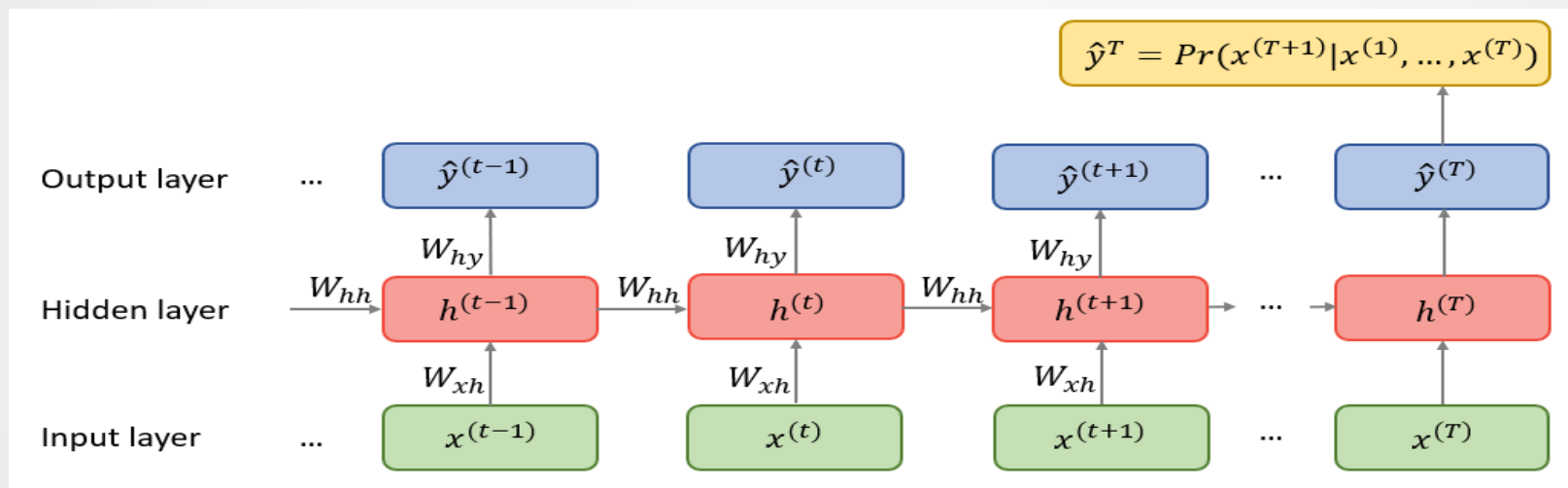
3. Antecedentes

- Sistemas de Preguntas y Respuestas basadas en texto

Data Set	Año	Características
SQuAD 1.1	2016	100k QA
SQuAD 2.0	2018	150k QA
xQuAD	2019	1k QA traducidos del SquAD a 11 idiomas
MLQA	2019	5k QA en 7 idiomas
SquAD-es	2020	SquAD Traducido sintéticamente al español

4. Modelos de Redes Neuronales para NLP

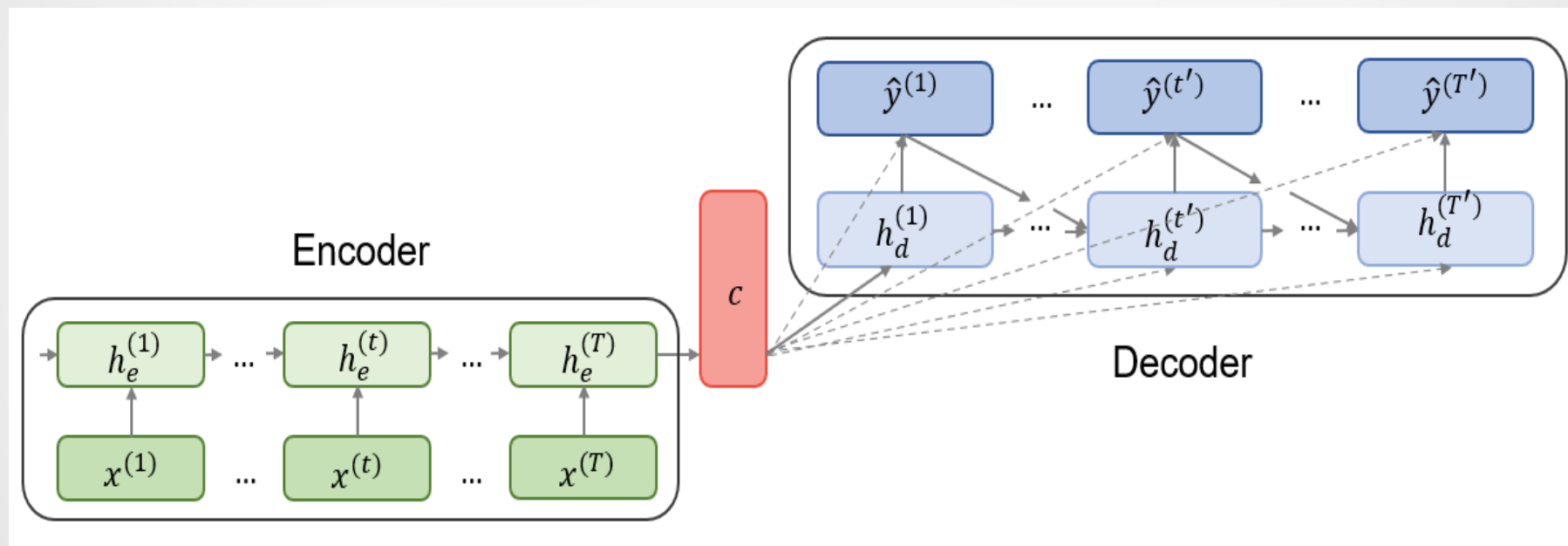
- 3.1. RNN, LSTM, GRU



- RNN desplegado. [2]

- [2] 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

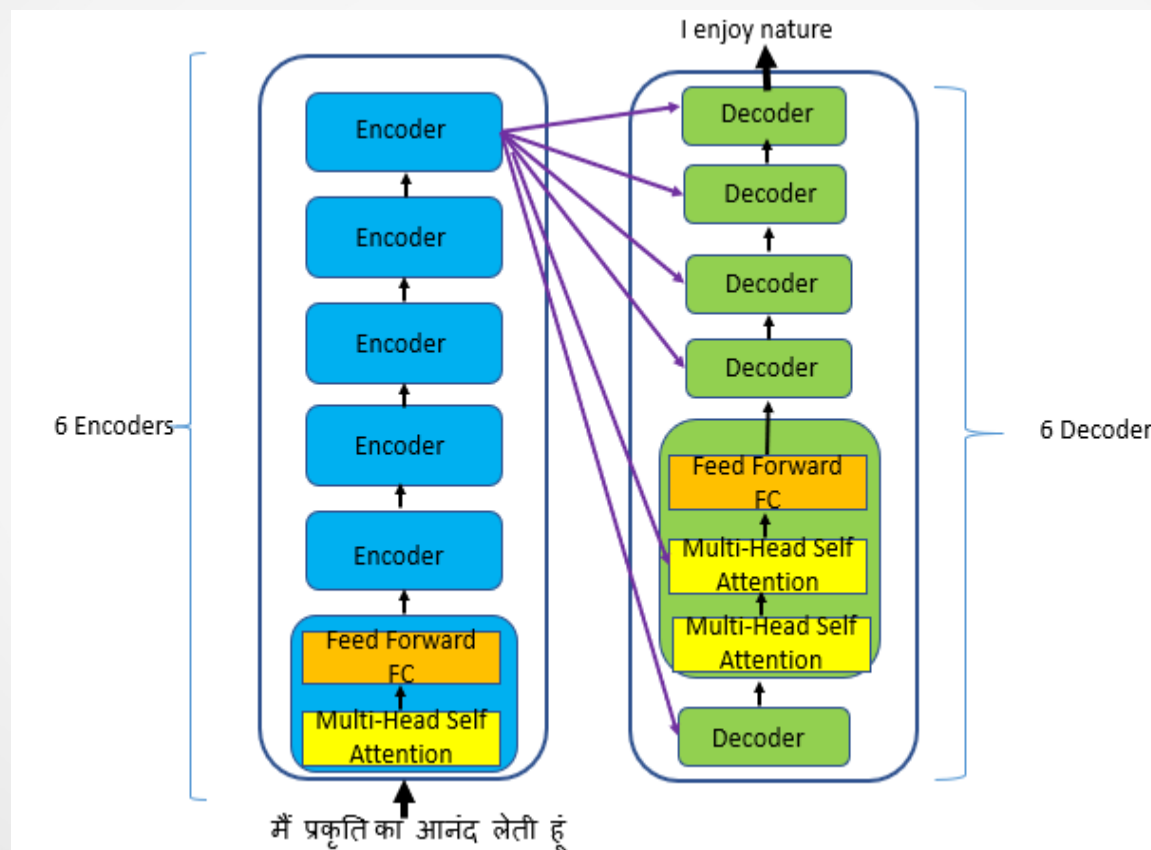
- 3.2. Seq to Seq -2014



- Arquitectura Encoder - Decoder. [3]

- [3] 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

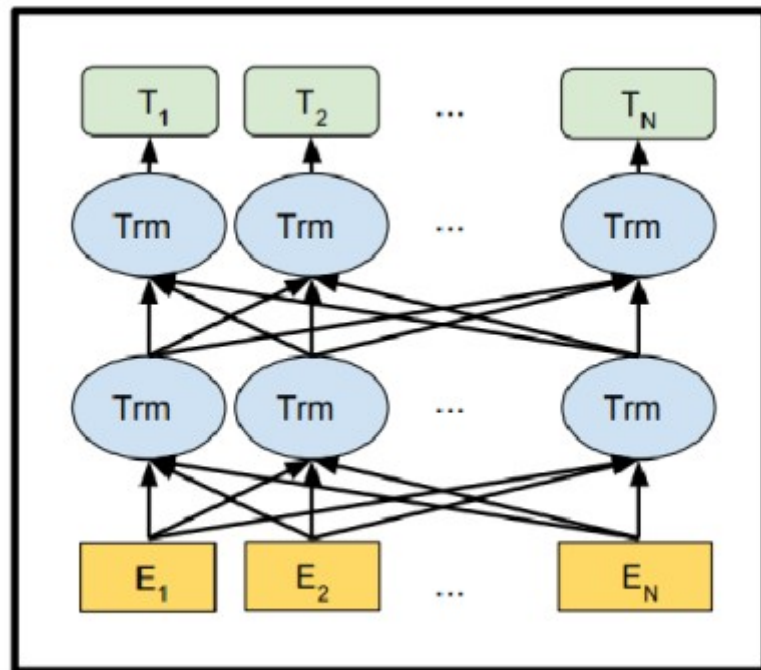
- 3.3. Transformers - 2017



- Arquitectura Transformer [4]

- [4] <https://towardsdatascience.com/simple-explanation-of-transformers-in-nlp-da1adfc5d64f>

- 3.4. BERT (Bidirectional Encoder Representations from Transformers) - 2018



- Arquitectura BERT [5]

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) QIANXIN	90.724	93.011
2 May 05, 2020	SA-Net-V2 (ensemble) QIANXIN	90.679	92.948

- Tabla de Clasificación para SquAD 2.0 [6]

- [5] <https://yashuseth.blog/2019/06/12/bert-explained-faqs-understand-bert-working/#:~:text=What%20is%20BERT%3F,task%2Ds%20specific%20fine%2Dtuning.>
- [6] <https://rajpurkar.github.io/SQuAD-explorer/>

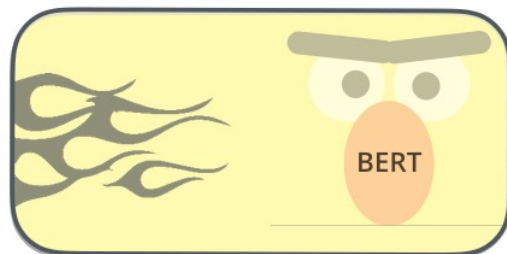
Entrenamiento de BERT y Aprendizaje por Transferencia

1 - **Semi-supervised** training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step

Model:



Dataset:



Objective:

Predict the masked word (language modeling)

2 - **Supervised** training on a specific task with a labeled dataset.

Supervised Learning Step

Classifier

75% Spam
25% Not Spam

Model:
(pre-trained in step #1)



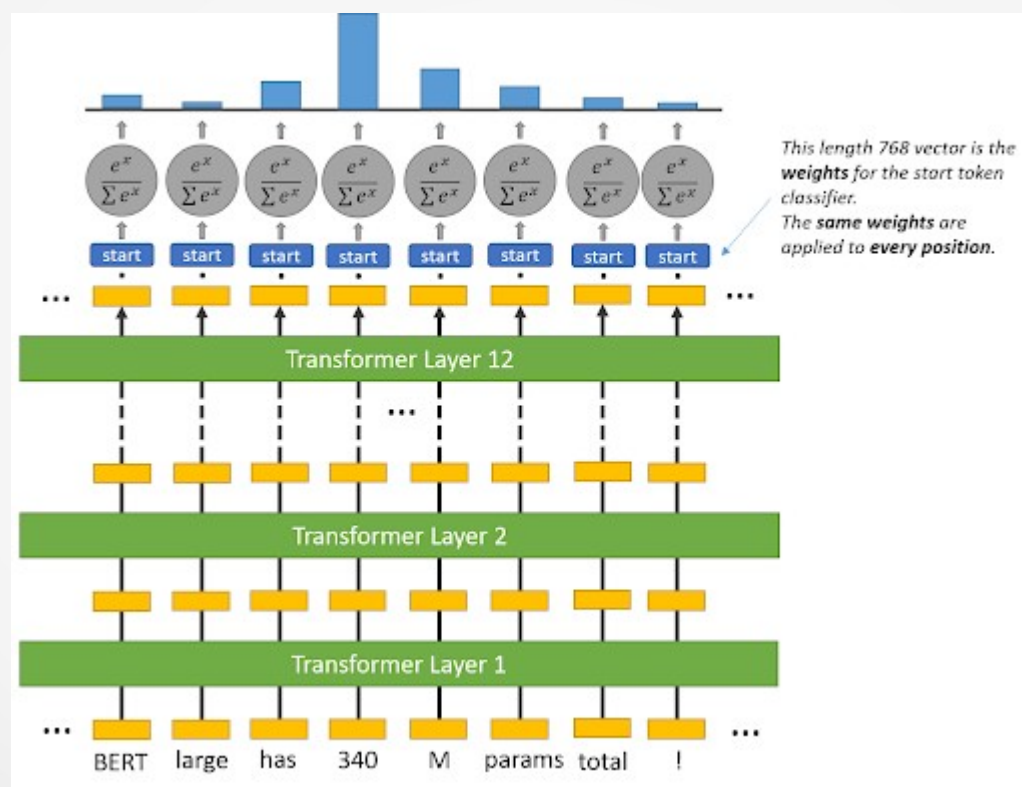
Dataset:

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached...	Not Spam

- Pre-entrenamiento y Finetuning [7]

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

- Fine tuning para Preguntas y Respuestas



- Obtención de respuesta con BERT [8]
- [8] <https://mccormickml.com/2020/03/10/question-answering-with-a-fine-tuned-BERT/>

- 3.5. Otros Modelos basados en BERT

	BERT	RoBERTa	DistilBERT
Size (millions)	Base: 110 Large: 340	Base: 110 Large: 340	Base: 66
Training Time	Base: 8 x V100 x 12 days* Large: 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	Large: 1024 x V100 x 1 day; 4-5 times more than BERT.	Base: 8 x V100 x 3.5 days; 4 times less than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	3% degradation from BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation

- Comparativa con otros modelos basados en BERT.[9]

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

- BETO (Arquitectura de BERT entrenado con Corpus en español) - 2019

Task	BETO-cased	BETO-uncased	Best Multilingual BERT	Other results
POS	98.97	98.44	97.10 [2]	98.91 [6], 96.71 [3]
NER-C	87.24	82.67	87.38 [2]	87.18 [3]
MLDoc	95.60	96.12	95.70 [2]	88.75 [4]
PAWS-X	89.05	89.55	90.70 [8]	
XNLI	(soon)	80.15	78.50 [2]	80.80 [5], 77.80 [1], 73.15 [4]

- Resultados de BETO para Tareas NLP.[10]

- POS: Part of Speech Tagging
- NER: Name Entity Recognition
- MLDoc: MultiLingualDocuments
- PAWS-X: Paraphrase Adversaries from Word Scrambling
- XNLI: The Cross-Lingual NLI Corpus
- [10] <https://github.com/dccuchile/beto>

5. Conjunto de Datos

- SQUAD v1.1 – es :
- Train (23.3Mb) – Test(3.6Mb)

```
▶ print('Hay {:,} ejemplos de entrenamiento.'.format(len(examples)))
```

```
Hay 46,260 ejemplos de entrenamiento.
```

```
▶ print('Hay {:,} ejemplos de test.'.format(len(examples)))
```

```
Hay 6,583 ejemplos de test.
```

```
Title: Universidad _ de _ Notre _ Dame  
ID: 573393184776f41900660da9
```

```
===== Question =====
```

```
¿Cuál era el tamaño de la dotación de Notre Dame cuando Theodore Hesburgh se convirtió en presidente?
```

```
===== Context =====
```

```
El Rev. Theodore Hesburgh, C.S.C., (1917-2015) fue presidente durante 35 años (1952-87) de transformaciones dramáticas. En ese momento el presupuesto anual de operación aumentó en un factor de 18 de $9.7 millones a $176.6 millones, y la dotación en un factor de 40 de $9 millones a $350 millones, y la financiación de investigación. La matrícula casi se duplicó de 4.979 a 9.600, la facultad más que duplicó de 389 a 950, y los grados otorgados anualmente se duplicaron de 1.212 a 2.500.
```

```
===== Answer =====
```

```
9 millones
```


6. 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.

$$F1 = 2 \cdot \frac{Precision \cdot Recuperacion}{Precision + Recuperacion}$$

$$Precision = \frac{Numero\ Tokens\ similares}{longitud\ (tokens\ predecidos)}$$

$$Recuperacion = \frac{Numero\ Tokens\ similares}{longitud\ (tokens\ etiquetados)}$$

7. Experimentos



**The AI community
building the future.**

- tasa de aprendizaje : $5e-5$. La tasa de aprendizaje inicial de Adam.
- max seq length : 384. La longitud total máxima de la secuencia de entrada después de tokenización WordPiece. Las secuencias más largas que esto se truncan y a las secuencias más cortas se les hará un padding.
- docs stride : 128. Para el overlapping entre el contexto y la pregunta.
- train batch size : 12. Tamaño de lote por GPU / CPU para entrenamiento.

Modelo	Exact Match(EM)	F1	Peso del Modelo
BETO(BERT español)	57.9978	73.6394	420MB
BERT multilinguaje	57.4358	73.0959	678MB
DistilBERT multilinguaje	51.8764	67.7269	514MB
RoBERTa base	51.0861	67.4771	476MB

Table 1: Resultado experimental de métricas de distintos modelos sobre la tarea de Preguntas y Respuestas aplicando FineTuning.

- Costo Computacional: Se uso Colab, el cual proporcionó una tarjeta gpu NVIDIA T4 de 16Gb RAM

Modelo	Tiempo de Entrenamiento (x 1 época)
BETO(BERT español)	1.5 horas
BERT multilinguaje	1.5 horas
DistilBERT multilinguaje	1.5 horas
RoBERTa	2 horas

Despliegue

Sistema Inteligente de Preguntas y Respuestas usando el modelo de BETO destilado

Ingrese el texto a analizar

El 6 de marzo, se confirmó oficialmente el primer caso positivo (paciente cero) por COVID-19 en territorio peruano en una conferencia de prensa presidida por el presidente Martín Vizcarra y la entonces ministra de salud Elizabeth Hinostroza. Según el comunicado del gobierno, el paciente de 25 años de edad trabajaba como piloto en LATAM Airlines y contrajo el virus en su viaje de vacaciones por España, Francia y República Checa. Conocido el caso, el gobierno activó inmediatamente el «Plan nacional de preparación y respuesta frente al riesgo de introducción del coronavirus» con el fin de fortalecer los sistemas de vigilancia, contención y respuesta. Los protocolos de seguridad también fueron activados en los aeropuertos, en los lugares con mayor afluencia de turistas y en los terminales marítimos y terrestres.

Ingrese pregunta sobre el texto

¿ Donde estuvo el primer infectado de COVID antes de venir al Perú?

Cargar modelo

La respuesta del Sistema es:

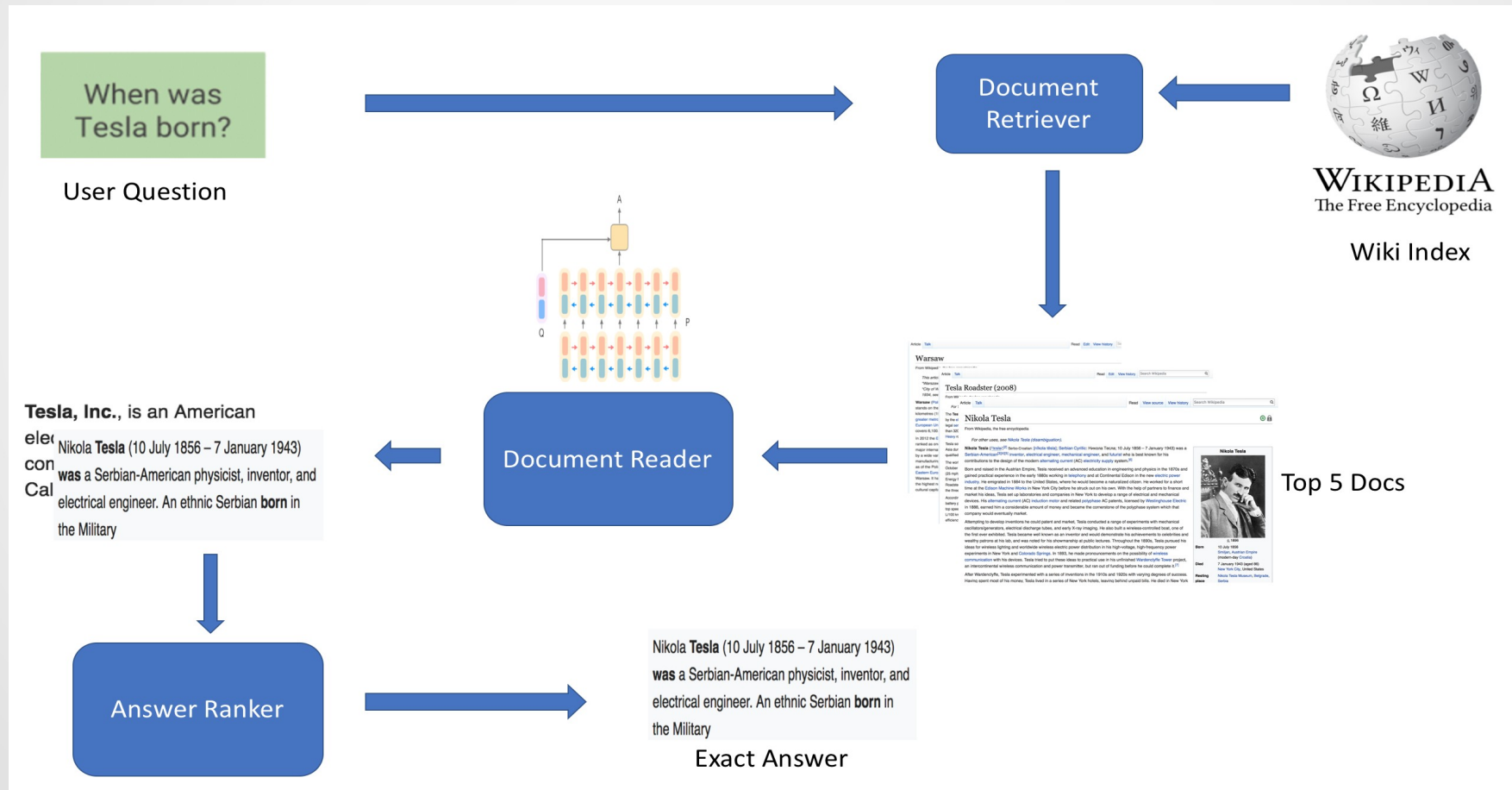
España, Francia y República Checa

8. Resultados y Discusión

- El modelo de BETO obtuvo las mejores métricas.
- El modelo de mBERT mostró una performance muy cercana, sin embargo cuenta con un peso 61% más grande que el de BETO, lo que podría generar problemas para su implementación en la etapa de despliegue.
- Como era de esperar, el modelo destilado proporciona una peor performance que el mBERT, pero reduce el tamaño del modelo en un 25%.
- RoBERTa base muestra unas métricas cercanas a modelo Destilado a pesar de estar entrenado con un corpus en inglés.

9. Trabajos Futuros

• Preguntas y Respuestas de Dominio Abierto



- <https://medium.com/analytics-vidhya/eddie-a-knowledge-backed-question-answering-agent-part-1-b93f68dab5c1>

- QA con múltiples alternativas

Question 12

10 points

The examination of everyday human social interactions on a small scale describes:

- ☐ A functionalism
- ☐ B macrosociology
- ☐ C cultural anthropology
- ☐ D microsociology

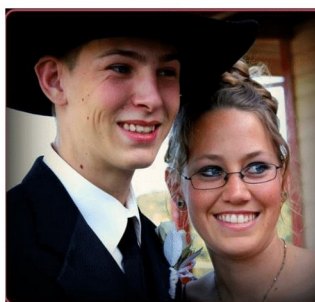
- Visual Question Answering

Who is wearing glasses?

man



woman

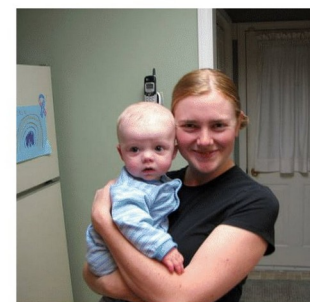


Where is the child sitting?

fridge



arms

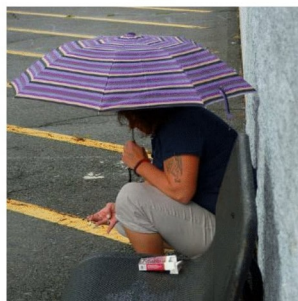


Is the umbrella upside down?

yes



no



How many children are in the bed?



2



1



- ChatBot

Random personality Shuffle  Share 

I work at a museum.
I like to go to the park.
I am stuck in a wheel chair.
I read a lot.
I don t have a lot of friends.

Start chatting
The machine learning model created a consistent persona based on these few lines of bio. You can now chat with this persona below.

the butterfly. do you like to travel?

What is the title ?

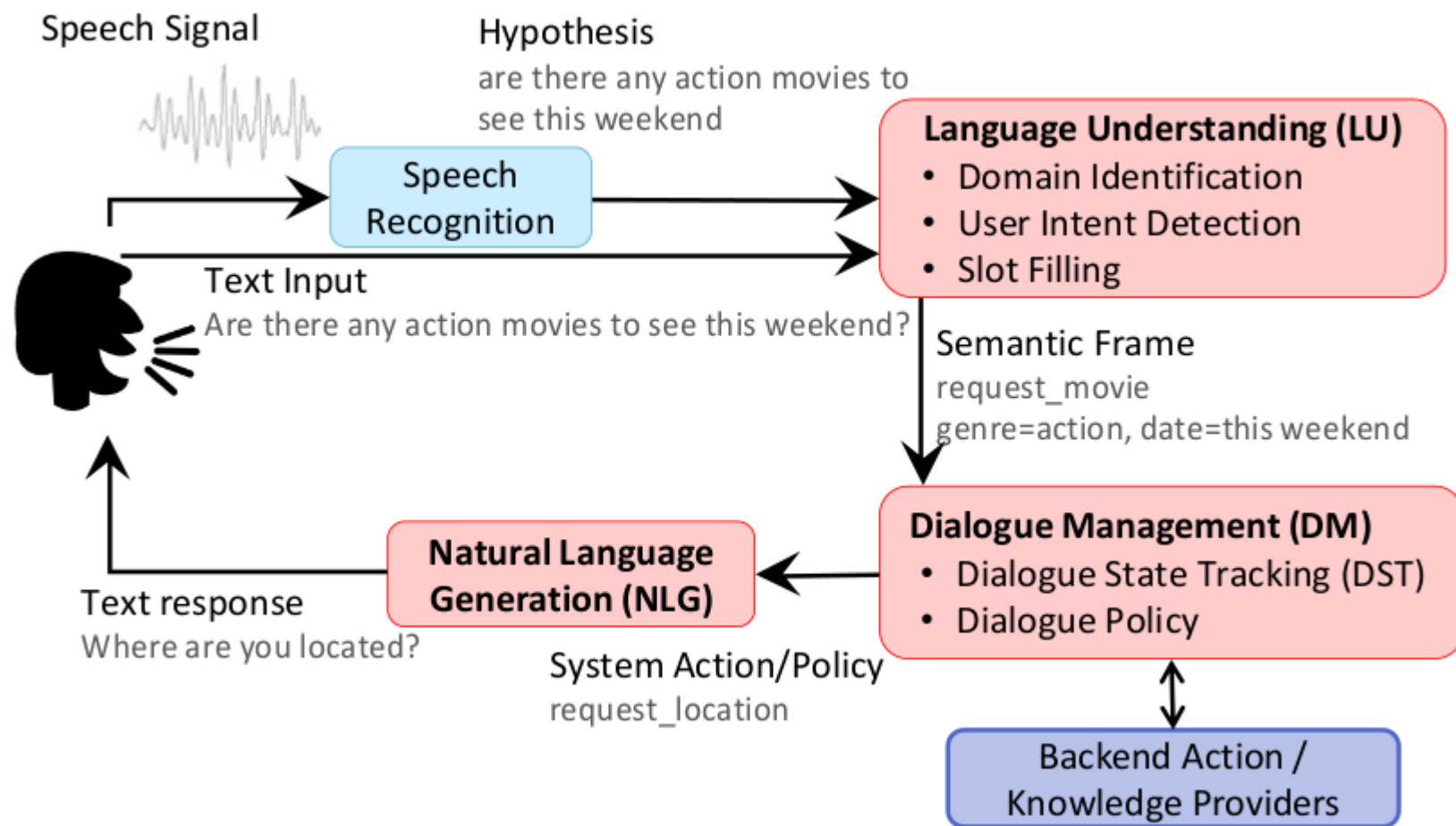
I've been to Mexico once

i like to go to the park with friends.

Send

Suggestion: [i'm stuck in a wheel chair](#)

• Sistemas de Diálogo



10. Conclusiones

- Sistemas de Preguntas y Respuestas pueden ser muy útiles en diversas aplicaciones de la industria, que necesiten un sistema de búsqueda de respuestas automáticas a preguntas de clientes.
- El uso de modelos pre-entrenados como BERT y sus variantes han facilitado la experimentación y el despliegue de diversas tareas de NLP.
- Para la aplicación de preguntas y tareas en el lenguaje español, el uso de conjunto de datos sintéticos, traducidos desde conjuntos de datos de alta calidad en inglés, genera un rendimiento aceptable.
- El modelo que nos ofrece mejores métricas para la tarea en español es BETO, que es una réplica de BERT con un corpus entrenado en español.

