Entrenamiento de un transformer para Q&A

Andrés Alejandro Guzmán González - A01633819

Ernesto Reynoso Lizárraga - A01639915

Joel Isaias Solano Ocampo - A01639289

Tania Sayuri Guizado Hernández - A01640092

In [1]:
```
# Librerías que se usaran en la actividad
import ebooklib
from ebooklib import epub
from transformers import BertForQuestionAnswering, BertTokenizer
import torch
import warnings
```

1.- Obtenga una base de conocimiento o Corpus con información técnica referente a su reto.

In [2]:
```
# Funcion para leer el contenido del archivo
def read_file(file_path):
    with open(file_path, 'r', encoding='utf-8') as file:
        book_corpus = file.read()
        return book_corpus
```

In [3]:
```
# Directorio del corpus a usar
corpus1 = read_file('Medical1.txt')
corpus2 = read_file('Medical2.txt')
corpus3 = read_file('Medical3.txt')
corpus4 = read_file('Landmarks.txt')
corpus5 = read_file('Echocardiography.txt')
corpus6 = read_file('Echocardiography2.txt')
corpus7 = read_file('Medical4.txt')
corpus8 = read_file('Echocardiography3.txt')
corpus9 = read_file('CNNs.txt')
corpus10 = read_file('Segmentation.txt')
```

2-. Utilizando un modelo previamente entrenado de “BertForQuestionAnswering”, administre como corpus el texto usado en las actividades anteriores o algún otro diferente.

In [4]:
```
# Cargar el modelo preentrenado y el tokenizador
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
model = BertForQuestionAnswering.from_pretrained('bert-base-uncased')
```

```
Some weights of BertForQuestionAnswering were not initialized from the model checkpoint at ber
t-base-uncased and are newly initialized: ['qa_outputs.weight', 'qa_outputs.bias']
You should probably TRAIN this model on a down-stream task to be able to use it for prediction
s and inference.
```

In [5]:
```
def answer_question(question, context):
    # Tokenizar la entrada
    inputs = tokenizer(question, context, return_tensors='pt')
```

```
    # Realizar la inferencia
    with torch.no_grad():
        outputs = model(**inputs)

    start_scores = outputs.start_logits
    end_scores = outputs.end_logits

    # Obtener los índices de inicio y fin con las puntuaciones más altas
    start_index = torch.argmax(start_scores)
    end_index = torch.argmax(end_scores) + 1

    # Obtener la respuesta del contexto
    answer_tokens = inputs['input_ids'][0][start_index:end_index]
    answer = tokenizer.decode(answer_tokens, skip_special_tokens=True)
    return answer
```

3-. Plantee 10 preguntas que el transformer debería de responder con respecto al corpus.

In [6]:
```
# Preguntas sobre el corpus
question1 = "What is medical image segmentation?"
question2 = "What are the benefits of medical image segmentation? "
question3 = "How does medical image segmentation work?"
question4 = "What is Landmark Detection?"
question5 = "Why do traditional segmentation methods often fall short?"
question6 = "What is the primary challenge addressed by unsupervised domain adaptation in echo
question7 = "What are the Problems faced in medical image segmentation?"
question8 = "What is an echocardiogram?"
question9 = "How does Echo-ODE improve cardiac ultrasound video segmentation compared to previ
question10 = "What is the rationale behind choosing the 17-segment model for echocardiographic
```

In [7]:
```
print("Pregunta:", question1)
print("Respuesta:", answer_question(question1, corpus1))
```

```
Pregunta: What is medical image segmentation?
Respuesta:
```

In [8]:
```
print("Pregunta:", question2)
print("Respuesta:", answer_question(question2, corpus2))
```

```
Pregunta: What are the benefits of medical image segmentation?
Respuesta:
```

In [9]:
```
print("Pregunta:", question3)
print("Respuesta:", answer_question(question3, corpus3))
```

```
Pregunta: How does medical image segmentation work?
Respuesta: manual options to paint on the data, or semi - automated operations such as thresho
lding and region growing. applications are also available for cardiovascular image segmentatio
n, with particular options for working with different heart cases. for many cases using medica
l data, it may only be necessary to use a few segmentation tools. as previously noted, studyin
g the placement of medical devices can involve a few steps to segment regions of interest in a
bone,
```

In [10]:
```
print("Pregunta:", question4)
print("Respuesta:", answer_question(question4, corpus4))
```

```
Pregunta: What is Landmark Detection?
Respuesta: what is landmark detection? the mechanism of detecting the famous human - made scul
ptures, buildings, and monuments inside an image is defined as landmark detection. you can sim
ply compare it with the famous application of google known as google landmark detection, which
is used by google maps. at the end of this blog, you will be able to create your own landmark
detector
```

In [11]:
```
print("Pregunta:", question5)
```

```
print("Respuesta:", answer_question(question5, corpus5))
```

```
Pregunta: Why do traditional segmentation methods often fall short?
Respuesta: echocardiography is often used clinically for appraising cardiac morphology and fun
ction. for example, the ejection fraction ( ef ) of the left ventricle ( lv ) is one of the fa
mous clinical indices to quantify the lv systolic function. the lvef can be calculated by the
volume of the left ventricle. the width of the left ventricular myocardium also contains patho
logical information. the exact segmentation of the left ventricle endocardium ( lvendo ) and t
he left ventricle epicardium ( lvepi ) provides the clinical quantitative measures mentioned a
bove. however, manual echocardiography labeling requires a doctor with rich clinical expertise
to spend a lot of time,
```

In [12]:
```
print("Pregunta:", question6)
print("Respuesta:", answer_question(question6, corpus6))
```

```
Pregunta: What is the primary challenge addressed by unsupervised domain adaptation in echocar
diography datasets
Respuesta: echocardiography datasets are greatly different from each other in domain styles, s
uch as grayscale distribution. figure 1 shows an example of echocardiography from diverse cent
ers. these images are inconsistent in the gray
```

In [13]:
```
print("Pregunta:", question7)
print("Respuesta:", answer_question(question7, corpus7))
```

```
Pregunta: What are the Problems faced in medical image segmentation?
Respuesta: image segmentation faces several challenges, including : image variability : divers
e image characteristics make creating a universal segmentation algorithm challenging. ambiguit
y and complexity : complex structures, ambiguous boundaries, and
```

In [14]:
```
print("Pregunta:", question8)
print("Respuesta:", answer_question(question8, corpus8))
```

```
Pregunta: What is an echocardiogram?
Respuesta: echocardiogram uses sound waves to create pictures of the heart. this common test c
an show blood flow through the heart and heart valves. your health care provider can use the p
ictures from the test to find
```

In [15]:
```
print("Pregunta:", question9)
print("Respuesta:", answer_question(question9, corpus9))
```

```
Pregunta: How does Echo-ODE improve cardiac ultrasound video segmentation compared to previous
methods?
Respuesta: as a registration estimation problem and present a novel method for diffe
```

In [16]:
```
print("Pregunta:", question10)
print("Respuesta:", answer_question(question10, corpus10))
```

```
Pregunta: What is the rationale behind choosing the 17-segment model for echocardiographic exa
mination of the left ventricle?
Respuesta: employed in practice, it has an anatomical basis, segments can be easily identified
on the basis of obvious echocardiographic landmarks, there is good correspondence with the dis
tribution of coronary arteries,
```

4-. Obtenga las respuestas de esas 10 preguntas en español e inglés (recuerden que sólo se entrena una vez, la idea es ver las diferentes respuestas con entradas de diferentes idiomas):

In [17]:
```
# Mismas preguntas pero en español
question1 = "¿Qué es la segmentación de imágenes médicas?"
question2 = "¿Cuáles son las ventajas de la segmentación de imágenes médicas? "
question3 = "¿Cómo funciona la segmentación de imágenes médicas?"
question4 = "¿Qué es la detección de puntos de referencia?"
question5 = "¿Por qué suelen fallar los métodos tradicionales de segmentación?"
question6 = "¿Cuál es el principal reto que aborda la adaptación de dominios no supervisada en
question7 = "¿Cuáles son los problemas a los que se enfrenta la segmentación de imágenes médic
```

```
question8 = "¿Qué es un ecocardiograma?"
question9 = "¿Cómo mejora Echo-ODE la segmentación de vídeos de ecografía cardíaca en comparac
question10 = "¿Cuál es el fundamento de la elección del modelo de 17 segmentos para el examen
```

In [18]:
```
print("Pregunta:", question1)
print("Respuesta:", answer_question(question1, corpus1))
```

```
Pregunta: ¿Qué es la segmentación de imágenes médicas?
Respuesta:
```

In [19]:
```
print("Pregunta:", question2)
print("Respuesta:", answer_question(question2, corpus2))
```

```
Pregunta: ¿Cuáles son las ventajas de la segmentación de imágenes médicas?
Respuesta:
```

In [20]:
```
print("Pregunta:", question3)
print("Respuesta:", answer_question(question3, corpus3))
```

```
Pregunta: ¿Cómo funciona la segmentación de imágenes médicas?
Respuesta: manual options to paint on the data,
```

In [21]:
```
print("Pregunta:", question4)
print("Respuesta:", answer_question(question4, corpus4))
```

```
Pregunta: ¿Qué es la detección de puntos de referencia?
Respuesta:
```

In [22]:
```
print("Pregunta:", question5)
print("Respuesta:", answer_question(question5, corpus5))
```

```
Pregunta: ¿Por qué suelen fallar los métodos tradicionales de segmentación?
Respuesta:
```

In [23]:
```
print("Pregunta:", question6)
print("Respuesta:", answer_question(question6, corpus6))
```

```
Pregunta: ¿Cuál es el principal reto que aborda la adaptación de dominios no supervisada en co
njuntos de datos de ecocardiografía?
Respuesta: ##grafia? as a result of the diverse imaging devices and imaging protocols, echocar
diography datasets are greatly different from each other in domain styles, such as grayscale d
istribution. figure 1 shows an example of echocardiography from diverse centers. these images
are inconsistent in the gray
```

In [24]:
```
print("Pregunta:", question7)
print("Respuesta:", answer_question(question7, corpus7))
```

```
Pregunta: ¿Cuáles son los problemas a los que se enfrenta la segmentación de imágenes médicas?
Respuesta: ¿ cuales son los problemas a los que se enfrenta la segmentacion de imagenes medica
s? medical image segmentation faces several challenges, including : image variability : divers
e image characteristics make creating a universal segmentation algorithm challenging. ambiguit
y and complexity : complex structures, ambiguous boundaries, and
```

In [25]:
```
print("Pregunta:", question8)
print("Respuesta:", answer_question(question8, corpus8))
```

```
Pregunta: ¿Qué es un ecocardiograma?
Respuesta: echocardiogram uses sound waves to create pictures of the heart. this common test c
an show blood flow through the heart and heart valves. your health care provider can use the p
ictures from the test to find
```

In [26]:
```
print("Pregunta:", question9)
print("Respuesta:", answer_question(question9, corpus9))
```

```
Pregunta: ¿Cómo mejora Echo-ODE la segmentación de vídeos de ecografía cardíaca en comparación
con métodos anteriores?
Respuesta:
```

In [27]:
```
print("Pregunta:", question10)
```

```
print("Respuesta:", answer_question(question10, corpus10))
```

```
Pregunta: ¿Cuál es el fundamento de la elección del modelo de 17 segmentos para el examen ecoc
ardiográfico del ventrículo izquierdo?
Respuesta:
```

- ¿Hubo alguna diferencia?

Después de hacer diversas experimentaciones con el código, logramos ver que las preguntas en español las contestaba un tanto acorde a lo que se preguntaba, pero no tan conciso y realmente no sabemos si debiamos esperar que también pudiese responder en español mientras que las preguntas en inglés algunas dejo la respuesta en blanco otras si tenian coherencia. Cabe recalcar que si nosotros corriamos de nuevo la pregunta ya sea en español o inglés a veces la podía responder. Nuestra deducción es que tras una vez se hizo la pregunta la proxima vez que se le preguntó lo mismo fue capaz de mejorar la interpretación de la misma y asi mismo su capacidad de dar una respuesta.

- ¿Qué lenguaje conviene más y por qué?

Creo que todo depende del idioma del corpus, en el sentido, que si el idioma de las preguntas estan más allegadas al idioma del corpus su entendimiento será más conveniente y rápido entorno a su capacidad de respuesta.

- ¿Cuál era el tamaño del corpus?

Al principio quisimos intentar con un corpus largo, pero el programa nos arrojaba un error que estaba limitado a los 512 tokens entonces relativamente nuestros corpus se tuvieron que adaptar a los tokens limite.

- ¿Cuántas respuestas tienen coherencia?

En inglés el número fue de ocho respuesta coherentes, no obstante, no estaban completadas. Aunque se entendia que quiso dar el acercamiento a la mejor respuesta se corto sin sentido. Mientras que las preguntas en español 3 de las 10 tuvieron coherencia, aqui el único detalle como ya se menciono antes es que sus respuestas estaban en inglés.

- ¿Si cambia el corpus y pregunta lo mismo recibirá una respuesta? Demuestre

In [28]:
```
print("Pregunta:", question1)
print("Respuesta:", answer_question(question1, corpus4))
```

```
Pregunta: ¿Qué es la segmentación de imágenes médicas?
Respuesta: image is defined as landmark detection. you can simply compare it with the famous a
pplication of google known as google landmark detection
```

La respuesta es si, habra una respuesta pero no tendrá nada que ver con la pregunta hecha. A lo que vemos es que toma al menos una parte de la pregunta que entiende y con base a eso trata de dar una respuesta.

- ¿Cuántos lenguajes puede manejar el BERT para resolver preguntas?

Más de 100 diferentes lenguajes.