DESCRIPCIÓN

SELIN COCARCA

1. INTRODUCCIÓN

2. DESCRIPCIÓN DEL DATASET

3. MODELO UTILIZADO

4. RESULTADOS OBTENIDOS

5. RAZÓN DE UTILIZACIÓN EN EL PROYECTO

6. CONCLUSIÓN

1. INTRODUCCIÓN:

Este proyecto implementa un sistema de traducción automática utilizando el modelo preentrenado MarianMT de la biblioteca Transformers de Hugging Face. El objetivo principal es facilitar la traducción fluida entre tres idiomas principales: inglés, español y euskera. A continuación, se detalla el dataset utilizado, las características del modelo MarianMT, los resultados obtenidos y la razón por la cual se eligió este enfoque para el proyecto.

2. DESCRIPCIÓN DEL DATASET

El dataset utilizado en este proyecto proviene del conjunto de datos Flores de Facebook, que contiene textos multilingües alineados en pares entre inglés, español y euskera. Esta estructura de datos es fundamental para entrenar modelos de traducción automática que puedan aprender y mantener correspondencias precisas entre frases en diferentes idiomas. Este dataset fue sacado de Huggin Face, porque se sabe que estos datasets públicos multilingües como Flores son ampliamente utilizados para tareas de procesamiento del lenguaje natural.

3. MODELO UTILIZADO

Opte por el modelo MarianMT de Hugging Face debido a su arquitectura basada en Transformers, optimizada específicamente para traducción automática multilingüe. Este modelo está preentrenado en grandes corpus de texto multilingüe, lo que le permite capturar patrones lingüísticos y realizar traducciones efectivas entre los idiomas seleccionados. Utilizando el tokenizador MarianTokenizer, el modelo convierte texto de entrada en secuencias de tokens compatibles, asegurando una representación adecuada y coherente para su procesamiento.

4. RESULTADOS OBTENIDOS

La evaluación del modelo se centró en métricas cruciales como precisión, recall y F1-score para evaluar la calidad de las traducciones realizadas en los siguientes pares de idiomas: inglés a español y euskera, español a inglés y euskera, y euskera a inglés y español. Estas métricas proporcionaron una evaluación detallada de cómo MarianMT maneja la transferencia de significado y la fluidez lingüística entre los idiomas, destacando su capacidad para mantener la coherencia y precisión en diversos contextos multilingües. Gracias a esto, las traducciones logran afinarse al máximo, asegurando resultados coherentes y significativos para cada idioma destino.

5. RAZÓN DE UTILIZACIÓN EN EL PROYECTO

El modelo MarianMT fue seleccionado por su capacidad demostrada para realizar traducciones precisas y consistentes entre múltiples idiomas, cumpliendo con los requisitos específicos del proyecto de traducción automática entre inglés, español y euskera. Su estructura preentrenada y su tokenización eficiente aseguran un rendimiento óptimo en términos de precisión y velocidad, lo que es crucial para aplicaciones prácticas que requieren procesamiento rápido y preciso del lenguaje en entornos multilingües.

6. CONCLUSIÓN

El enfoque utilizando MarianMT ha demostrado ser efectivo para la traducción automática fluida entre inglés, español y euskera, facilitando la comunicación y el acceso a la información en entornos multiculturales y globales. La evaluación continua y la optimización del modelo son esenciales para mantener y mejorar la calidad de las traducciones en diferentes contextos y pares de idiomas, asegurando que el sistema pueda adaptarse a las necesidades cambiantes de los usuarios y las aplicaciones del mundo real.