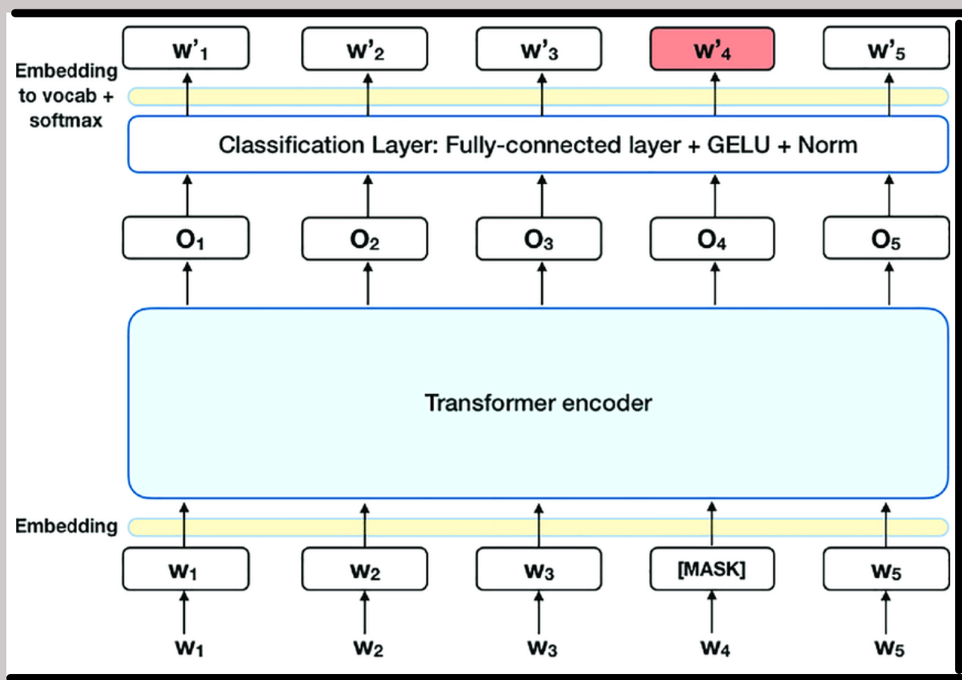
Reconocimiento de entidades nombradas en español usando BERT

Miguel Angel Tovar Onofre - matovaro@unal.edu.co
Andres Felipe Esteban Bautista - afestebanb@unal.edu.co
Juan Sebastian Hernández Reyes - jushernandezre@unal.edu.co

Introducción

Una entidad nombrada, se puede definir como ***una palabra o secuencias de palabras que se identifican como nombre de persona, organización, lugar, fecha, tiempo, porcentaje o cantidad***. Este término fue acuñado en las conferencias MUC (Message Understanding Conferences), competiciones patrocinadas por la DARPA cuyo objetivo era la extracción de información en textos; y actualmente su reconocimiento y clasificación tiene una gran cantidad de usos para el análisis de textos en muchos campos. Aunque los sistemas de reconocimiento de entidades construidos con métodos mas actuales en inglés tienen un desempeño cercano al humano, en español todavia no se ha conseguido tan buenos resultados



Metodología

Primeramente se elaboro un preprocesamiento del dataset (El CoNLL2002) que posee las etiquetas NER, asignando las etiquetas POS a cada una de las palabras en el dataset. Posteriormente el dataset es cargado y analizado para detectar las diferentes etiquetas POS, NER y Tokens presentes, donde a cada una de las palabras detectadas, se le asigno un ID númerico. Con los ID's asignados se crea un diccionario para obtener las oraciones convertidas a valores numéricos. Para realizar el entrenamiento, importamos el modelo preentrenado multilingual de BERT, el cual es entrenado con los diccionarios obtenidos y optimizados a nuestro problema específico por medio Fine Tuning. Los resultados obtenidos son evaluados utizando las metricas puntaje F1, Recall y precisión.

Algunos Resultados:

Mariana Pajón PER , a la final del BMX MISC de los Juegos Olímpicos MISC

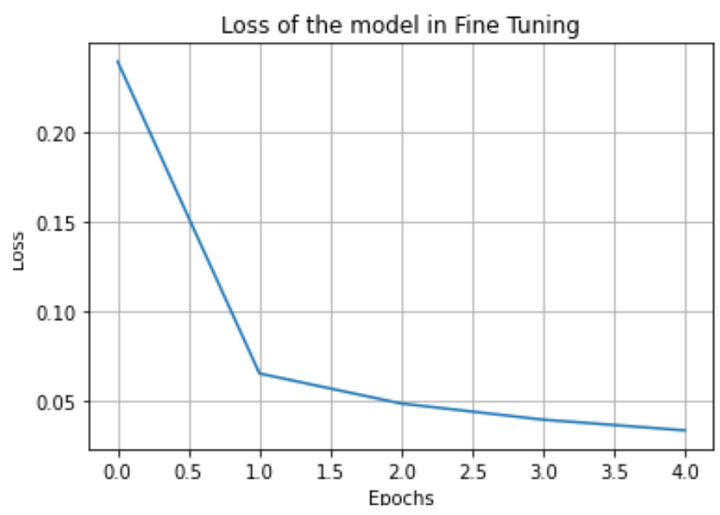
Muere el PES MISC y nace eFootball MISC : así será la apuesta de Konami PER

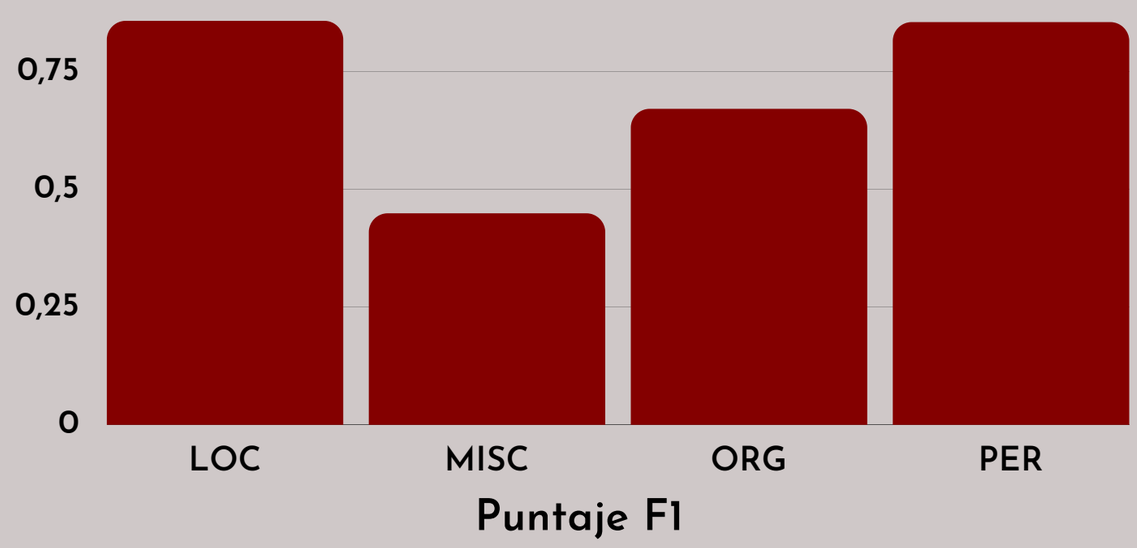
El crimen por el cual se estrenaría la cadena perpetua en Colombia LOC

Discusión

Como se puede apreciar en la figura de la de derecha, la función de pérdida va disminuyendo a medida que las épocas y adicionalmente se puede concluir que después de las 5 épocas las métricas aumentan sin embargo se estaría realizando sobreentrenamiento.

Dentro de los resultados obtenidos, 3 de las cuatro entidades presentan una buena puntuación F1 (Ubicación, organización y personas), sin embargo las miscelanias no tuvieron una buena puntuación como consecuencia de su diversidad, ya que pertenecen a múltiples entidades.





Conclusiones

- Se logra un desempeño aceptable del modelo, pero esta limitado tanto por el dataset, como por el hecho de que es construido bajo un pre-training en otra lengua.
- El modelo logra aprender "tips utiles" para reconocer las entidades, como tendiendo a clasificar una palabra como Nombre cuando esta en mayúscula

Para mas información:

- Revisar el video explicativo del trabajo, disponible en: https://youtu.be/qqn5Y0Hui3k
- Revisar el informe, disponible en: https://drive.google.com/file/d/1l532yD2uktgH0biilTyWmfNJAfWu8KR7/view?usp=sharing

Machine Learning - 2021 - I
Departamento de Ingeniería de Sistemas e Industrial
Facultad de Ingeniería - Sede Bogotá

UNIVERSIDAD NACIONAL DE COLOMBIA