PREDICCIÓN DE PUNTUACIÓN Y CAPITALIZACIÓN EN TEXTO NORMALIZADO

Azul Barracchia, Mario Sigal Aguirre, Milagros Maurer, Tobías Llop

Aprendizaje Automático II

DEPARTAMENTO DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Introducción

El problema que abordamos es la reconstrucción de un texto normalizado, agregando capitalización y puntuación, con modelos de aprendizaje automático.

Utilizamos tres tipos de modelos secuenciales distintos y comparamos sus performances.

Este tipo de problema tiene aplicaciones reales en procesamiento de lenguaje natural, por ejemplo en el post procesamiento de la salida de sistemas de reconocimiento automático del habla.

¿A qué hora vamos a la UBA?

			Punt. inicial	Punt. final	Capitalización
a	[169]		[1]	[0]	[1]
qué	[38188]		[0]	[0]	[0]
hora	[24301]		[0]	[0]	[0]
vamos	[10321], [13386]	MODELO →	[0]	[0,0]	[0,0]
a	[169]		[0]	[0]	[0]
la	[10109]		[0]	[0]	[0]
uba	[189], [10537]		[0]	[0,3]	[3,3]

Texto reconstruido: ¿A qué hora vamos a la UBA?

Preguntas principales

- ¿Cuál es el mejor modelo para cada clase?
- ¿Cuál es el mejor modelo para nuestro problema?

1. Modelos usados

- Random Forest
- RNN Unidireccional
- RNN Bidireccional

2. Fuentes de Datos

Para la extracción de datos usamos PDFs de libros. Cada libro fue separado en oraciones, y nuestros párrafos están hechos de a dos o tres oraciones, que pueden estar incompletas.

3. Random Forest

- Extracción de atributos: Dado un párrafo, para cada token, los atributos son: id del token, posición en la frase, distancia al final del párrafo, id del token anterior, id del token siguiente, ¿es principio de palabra?, ¿es token del medio de la palabra?, ¿es token final?, ¿forma parte de una palabra con más de un token?.

- Elección de hiperparametros: n_estimators = 100, max_depth = None, random_state = 42, min_samples_split = 10

- Entrenamiento: Entrenamos un modelo para cada etiqueta (capitalización, puntuación inicial y puntuación final).

4. RNNs

Arquitectura:

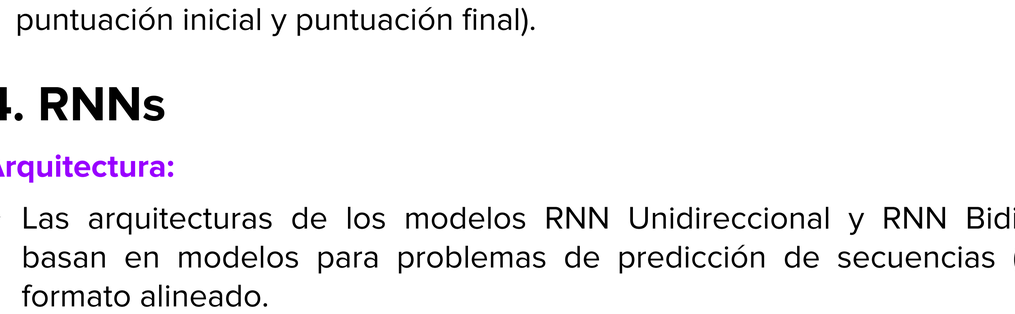
- Las arquitecturas de los modelos RNN Unidireccional y RNN Bidireccional se basan en modelos para problemas de predicción de secuencias (sec2sec) en formato alineado.

- Se utilizan celdas LSTM (Long-Short Term Memory), variando su formato entre Unidireccional y Bidireccional respectivamente en cada modelo.

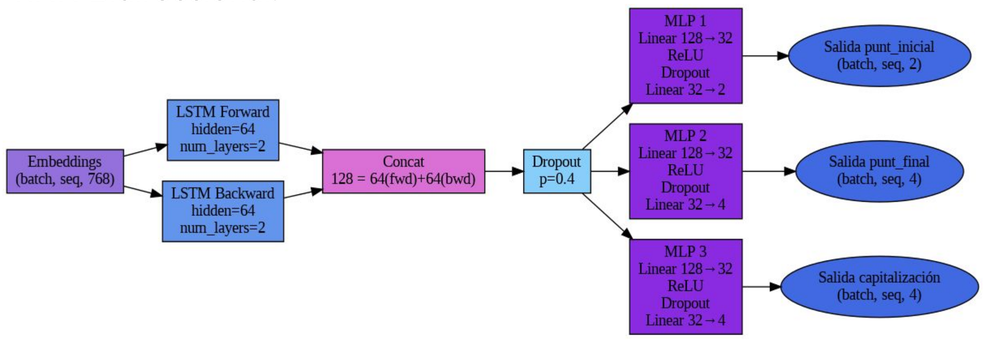
- Posteriormente, cada salida de la LSTM es procesada por un MLP (Multi-Layer Perceptron) específico para cada predicción de etiqueta. Se implementan tres MLPs: uno para la puntuación inicial, otro para la puntuación final y uno para la capitalización.

Visualización de la Arquitectura:

- RNN Unidireccional:



- RNN Bidireccional:



5. Resultados

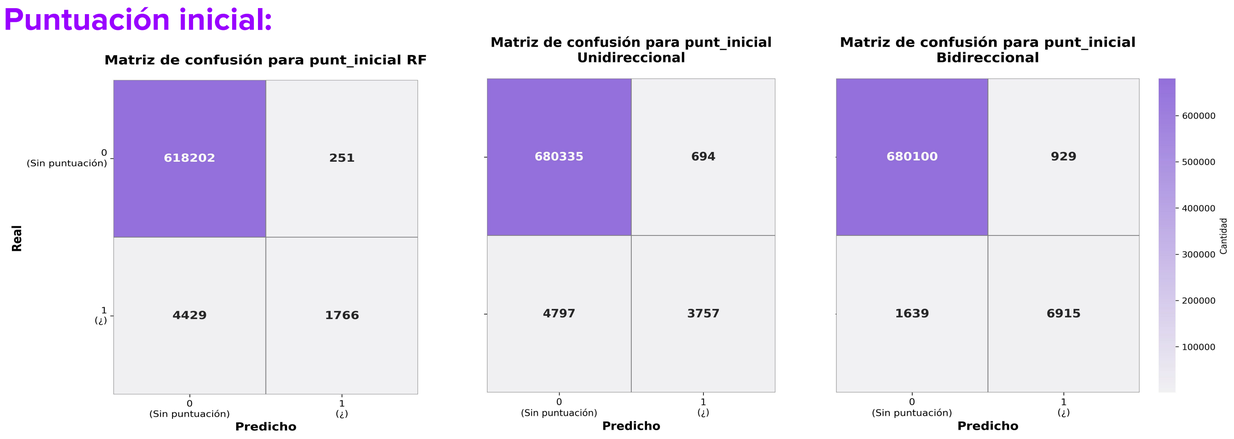
Estos fueron los resultados obtenidos luego de entrenar con el 90% de los datos y testear con el 10%, en el caso de Random Forest, y entrenar con el 80%, validar con el 10% y testear con el 10% restante, en el caso de las RNN.

F1 Score Macro:

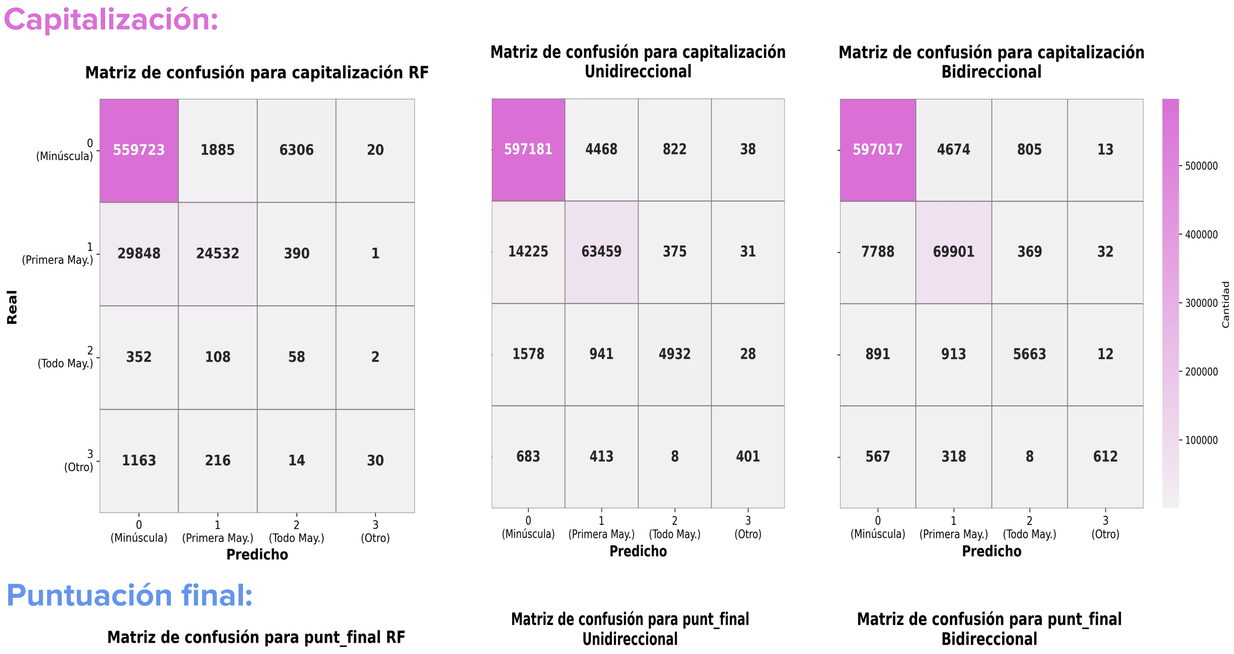
	Puntuación Inicial	Puntuación Final	Capitalización
Random Forest	0.7131	0.4498	0.4060
RNN Uni	0.7869	0.3820	0.7420
RNN Bi	0.9208	0.7996	0.8125

Tabla 1: Tabla con F1 macro como métrica para cada tipo de predicción y modelo.

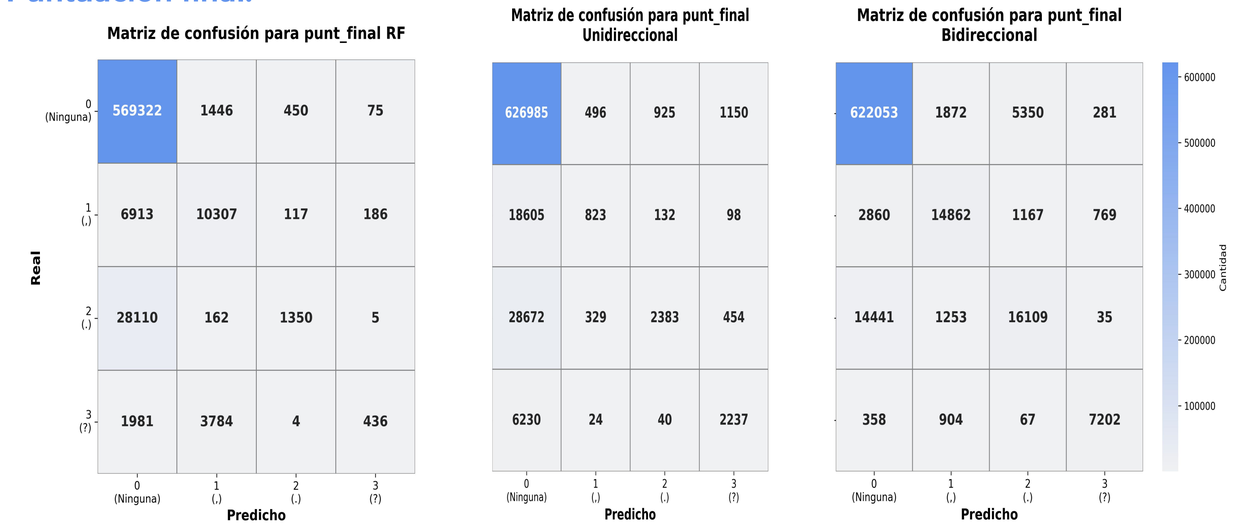
Puntuación inicial:



Capitalización:



Puntuación final:



6. Conclusión

El mejor modelo para el problema de predecir capitalización y puntuación es la red neuronal recurrente bidireccional.

Los resultados muestran que la arquitectura más efectiva es la que aprovecha información pasada y futura en la secuencia.

Para mejorar la precisión, se podría explorar el uso de arquitecturas basadas en Transformers, que han demostrado ser superiores en diversas tareas de procesamiento de lenguaje natural.