Reconocimiento y Clasificación de Entidades en Reportes Médicos

Alumna: Andrea Vanesa Borek
Directora: Ana G. Maguitman
Codirector: Axel Soto

1- Introducción

1.1 - Motivación

En la actualidad, la información constituye un aspecto fundamental en nuestra vida. Para que diversas aplicaciones puedan hacer un uso provechoso de esta información, la misma debe ser previamente procesada. Sin embargo, la extracción de datos manuales puede demorar mucho tiempo. Es por esto que es imprescindible la utilización de sistemas que automaticen el reconocimiento y clasificación de entidades del mundo real contenidas en los textos [3].

Bajo este contexto, el reconocimiento de entidades nombradas (NER de sus siglas en inglés) cobra importancia. Este tiene como objetivo la extracción y categorización de ciertas entidades en un texto. Los sistemas basados en NER son utilizados en campos tales como Inteligencia Artificial y Procesamiento de Lenguaje Natural (NLP de sus siglas en inglés).

Con la era de la digitalización, las instituciones de salud tiene un volumen de información grande, dando lugar a un problema de recolección de información. Dado esto, se puede encontrar diversos corpus sobre los cuales se puede trabajar; sin embargo, se ha encontrado uno en particular, Meddocan, que cuenta con tres dataset con la información necesaria. Además, este corpus fue previamente analizado por un personal competente, por lo que se prevén buenos resultados.

En la actualidad existen distintos modelos NER entrenados; sin embargo, en su mayoría fueron entrenados para el idioma inglés y con etiquetas específicas. A partir de esto, surge la idea de implementar uno específico para el corpus seleccionado, el cual se encuentra en español, con tags específicos del ámbito médico.

1.2 - Objetivo

El objetivo de este proyecto de ingeniería es implementar y analizar algoritmos de reconocimiento de entidades, donde el sistema será entrenado para la clasificación de datos en reportes médicos. En este proyecto se propone integrar técnicas orientadas a la extracción de contenido semántico en reportes médicos, con el objetivo de automatizar dicha tarea.

La implementación de algoritmos de reconocimiento de entidades nombradas puede hacerse en base a distintos modelos de aprendizaje supervisado [10]. Para este proyecto se utilizarán algoritmos que tienen como base la técnica de Campo Aleatorio Condicional (CRF de sus siglas en inglés) y redes neuronales artificiales de memoria a largo y corto término (LSTM de sus siglas en inglés). En particular, se utilizarán los algoritmos CRF y LSTM-CRF, para luego comparar sus resultados y poder decidir qué algoritmo se adapta mejor al contexto presentado.

2 - Conceptos Preliminares

2.1- Minería de Texto

En la actualidad, la información constituye un aspecto fundamental de nuestras vidas, donde las nuevas tecnologías facilitan la generación de la misma; sin embargo, la capacidad para procesarla y utilizarla no ha crecido lo suficiente para llegar a igualarla. Por este motivo, el problema de analizar estos grandes volúmenes de datos es el que se encarga de solucionar la minería de textos, para explorar, analizar, comprender y aplicar el conocimiento obtenido. Por lo que a esta disciplina se la puede entender como el proceso de descubrir patrones en los datos, donde este proceso puede ser automatizado o manual. Luego, van a ser estos patrones los que nos van a permitir realizar nuevas predicciones de datos [1], para tomar decisiones o mejorar la comprensión de los fenómenos que nos rodean.

Como se dijo, la minería de textos es el proceso análisis de datos para descubrir patrones y construir modelos predictivos. Esta disciplina se basa en muchos campos, tales como la matemática, estadística y aprendizaje automático (machine learning de sus siglas en inglés). Sin embargo, en su aplicación sólo se obtienen patrones que serán de poca utilidad mientras no se les encuentre significado y su valor real reside en la información que se puede extraer de ellos: información que ayude a tomar decisiones o mejorar la comprensión de los fenómenos que nos rodean [7]. Estos patrones van a facilitar la transformación de información no estructurada en un formato estructurado. Luego, los modelos obtenidos a partir del proceso de análisis necesita ser clasificado como útil o no; sin embargo, esta tarea necesita una valoración subjetiva por parte del usuario. [8]

En los algoritmos de minería de texto podemos encontrar tres componentes:

1. El modelo, que contiene parámetros que se han obtenido a partir de los datos de entrada.
2. El criterio de preferencia, que sirve para comparar modelos alternativos.
3. El algoritmo de búsqueda, que constituye cualquier otro programa de inteligencia artificial (IA).

El criterio de preferencia suele ser algún tipo de heurística y los algoritmos de búsqueda empleados suelen ser los mismos que en otros programas de inteligencia artificial [8].

La minería de texto es de gran interés para la industria, es por esto que es aplicada en distintas áreas. Algunas de estas son:

- Medicina: caracterización y predicción de enfermedades, probabilidad de respuesta satisfactoria a tratamiento médico.
- Mercadotécnia: identificar clientes susceptibles de responder a ofertas de productos y servicios por correo, fidelidad de clientes, selección de sitios de tiendas, afinidad de productos, etc
- Inversión en casas de bolsa y banca (credit scoring, redes neuronales o regresión logística): análisis de clientes, aprobación de préstamos, determinación de montos de crédito, etc.
- Detección de fraudes y comportamientos inusuales: telefónicos, seguros, en tarjetas de crédito, de evasión fiscal, etc

2.2 - NER

Uno de las tareas a realizar dentro de la Minería de texto es la Extracción de información (IE); donde esta convierte la información no estructurada incrustada en textos en datos estructurados. Un uso de esta tarea puede ser rellenar un base de datos relacional y así permitir un mayor procesamiento.
El primer paso en la extracción de información es detectar las entidades nombradas en el texto. Por lo que, la tarea del reconocimiento de entidad nombrada (NER) es encontrar cada mención de una entidad con nombre en el texto y etiquetar su tipo. Un entidad es cualquier palabra a la que se pueda hacer referencia con un nombre propio: una persona, un lugar, una organización, entre otros. El término se extiende comúnmente para incluir cosas que no son entidades en sí mismo, incluidas fechas, horas y otros tipos de expresiones temporales, e incluso expresiones numéricas como edades. a continuación está el texto que muestra entidades marcadas:

Remitido por: **[NPS Isabel Pavón de Paz]** Servicio de Endocrinología **[HOSP Hospital Universitario de Getafe] [CALLE Ctra. de Toledo, Km 12,5 28905 Getafe] - [TERR Madrid]** (**[PAIS España]**) Correo electrónico: **[CE pavonisa@yahoo.es]**.

En el texto se identificaron 5 menciones de entidades nombradas, incluidas nombre de personal sanitario, hospital, territorio, país y correo electrónico.
Según el contexto, las entidades son útiles para muchas otras tareas de procesamiento del lenguaje. En análisis de sentimientos podríamos querer saber el sentimiento de un consumidor hacia un participante. Las entidades son una primera etapa útil para responder preguntas o para vincular texto a la información en fuentes de conocimiento estructuradas como Wikipedia [3].
Como se dijo, el reconocimiento de entidad nombrada significa encontrar palabras del texto que constituyen entidades y luego clasificar el tipo de la entidad. El reconocimiento es difícil en parte porque puede causar ambigüedad; necesitamos decidir qué es una entidad y qué no lo es, dónde están los límites y qué tipo de entidad es. Por ejemplo, los siguientes casos pueden ser ambiguos si no se tiene un contexto.

Isabel Pavón de Paz: Nombre Sujeto Asistencia, Nombre de personal sanitario
28905: Código Postal, Identificación Sujeto Asistencia, ID de asistencia.

Las etiquetas pueden ser representadas en dos maneras, la primera es la denominadas IO y las segunda es IOB. En la primera, una entidad compuesta por varias palabras se etiqueta con el mismo tag por cada palabra de de la entidad. En la segunda, se introduce la etiqueta para el comienzo (B) y el interior (I) de cada tipo de entidad. Para ambos casos, se utiliza un token distinto (O) de cualquier entidad que represente el concepto de no estar etiquetado. La ventaja de utilizar la segunda representación, es que IOB puede expresar exactamente la misma información. Por un lado, si eliminamos la B, como hace el etiquetado IO, no podemos distinguir entre dos entidades del mismo tipo que está consecutivamente. Por otro lado, una ventaja del etiquetado IO es la cantidad de etiquetas utilizadas, ya que este define N + 1 etiquetas, un valor mucho menor que el de IOB con 2N + 1. Sin embargo, para este proyecto se decide utilizar el etiquetado IOB por sobre IO.

Remitido	O	O

por	O	O
Isavel	B-NPS	NPS
Pavón	I-NPS	NPS
de	I-NPS	NPS
Paz	I-NPS	NPS
Servicio	O	O
de	O	O
Endocrinología	O	O
Hospital	B-HOSP	HOSP
Universitario	I-HOSP	HOSP
de	I-HOSP	HOSP
Getafe	I-HOSP	HOSP

El problema NER puede ser abordado de distintas maneras. Para esto podemos resolverlo en tres grandes algoritmos: basado en funciones (MEMM / CRF), red neural (bi-LSTM) y basado en reglas. Sin embargo, en este proyecto solamente se van abordar los dos primeros.

2.3 - Métricas

Luego de obtener el modelo es importante medir el rendimiento, y para este proyecto, poder compararlos. Esto permite establecer qué enfoque es mejor que otro para ciertos resultados. Sin embargo, es importante destacar que dependiendo de lo que se quiera optimizar es la métrica en que debe enfocarse el análisis.
Una vez que se obtiene el modelo resultante, el sistema va a clasificar las entidades con un tag A, el cual puede dar como resultado a cuatro situaciones:

- True positive: Se predice que es A y efectivamente la entidad es A.
- False Positive: Se predice que es A, pero es la entidad es B.
- True Negative: No se predice que es A y la entidad efectivamente no es A.
- False Negative: No se predice que es A y la entidad es en realidad A.

Por ejemplo, el sistema tiene que predecir la siguiente frase:

Remitido por: Dr. Alonso Cabrera Servicio de Urología Hospital Universitario La Paz ***Avenida Castellana*** *E-28046 Madrid. (España) E-mail: manuelcabreracastillo@gmail.com*

Para este caso tenemos los siguientes tags [B-NPS, I-NPS, B-HOSP, I-HOSP, B-CALLE, I-CALLE, B-PAIS, B-CE, I-CE]. Sin embargo, para simplificar el análisis, se va a concentrar la predicción solamente en la frase "Avenida Castellana". Para esto, el modelo predice lo siguiente:

Palabras	Original	Predicción
Avenida	B-CALLE	B-CALLE
Castellana	I-CALLE	O

Para estas dos predicciones se puede obtener los siguientes valores:

- Para B-CALLE se tiene un True positive, dado que es correcta la etiqueta para esa palabra.
- Para O se tiene un FALSE POSITIVE, dado que se predijo que era O cuando no lo era.
- Para I-CALLE se tiene un False negative, dado que no se reconoció a I-CALLE cuando tendría que haber sido.
- Para el resto de tags que se tenga en el sistema se tiene un dos True negative, dado que no fueron reconocidos y efectivamente no deben serlo.

Palabra	B-CALLE	I-CALLE	O	Restantes Tags
Avenida	True Positive	True Negative	True Negative	True Negative
Castellana	True Negative	False Negative	False Positive	True Negative

En base a estos valores se definen algunas funciones para evaluar un modelo, las cuales son: *exactitud, precision, cobertura (del inglés recall) y* media armónica entre precisión y recal (*f1-score del inglés).*

2.3.1 - Exactitud

Esta métrica es una de las más conocidas y más intuitivas, ya que esta responde a la pregunta ¿Cuántas entidades clasificó bien el modelo? A partir de esto, podemos definirla matemáticamente de la siguiente manera:

$$Exactitud = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

Sin embargo, esta métrica no es útil para conjuntos de datos asimétricos. Siendo esto válido para nuestro contexto, donde van a haber muchos true negative por cada vez que se prediga una etiqueta. Por ejemplo, en el peor de los casos en que se erre la etiqueta, suponiendo que predijo mal B-CALLE con B-NPS, para todas las restantes etiquetas es un true negative. Dado esto, nuestro modelo es sumamente asimétrico. Es por esto, que no se va a considerar esta métrica para el análisis y comparación de los enfoques utilizados para resolver el reconocimiento de entidades nombradas.

2.3.2 - Precisión

La precisión la podemos definir como la relación entre los resultados generados por el sistema que predijeron correctamente las observaciones positivas (True Positive) con respecto a las observaciones positivas pronosticadas totales del sistema, tanto las correctas (True Positive) como las incorrectas (False Positive). Matemáticamente se define de la siguiente manera

$$Precisión = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

La precisión es útil cuando hay un costo muy alto a los Falsos Positivos. Es decir, cuando Decimos que que una palabra es B-NSA, siendo en realidad B-CALLE. En esta última situación tenemos un Falso Positivo asociado a B-NSA.

2.3.3 - Cobertura

La precisión la podemos definir como la proporción de palabras etiquetadas correctamente. Por otra parte el cobertura mide la proporción de observaciones positivas que se predijeron correctamente (True Positive) con respecto a las observaciones que son verdaderas, tanto aquellas pronosticadas de manera correcta (True Positive) como incorrecta (False Negative).

$$Cobertura = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

cobertura es útil cuando hay un costo muy alto a los Falsos Negativos. Es decir, cuando Decimos que que una palabra es B-NSA, siendo en realidad B-CALLE. En esta última situación tenemos un Falso

Negativo asociado a B-CALLE.

2.3.4 - F1-Score

Dependiendo del modelo que se esté representando se va a elegir optimizar la precisión o cobertura. Sin embargo, hay situaciones en las cuales es importante optimizar ambas, siendo este proyecto el caso. Esto es debido a que pueden suceder casos en que una etiqueta tenga valores opuestos para la Precisión y Cobertura. Por ejemplo, este es el resultado parcial de un entrenamiento hecho, donde la columna support corresponde a la cantidad de palabras con su respectiva etiqueta:

	precision	recall	f1-score	support
I-ISA	1.00	0.05	0.09	21
B-PROF	1.00	0.33	0.50	6
I-PROF	1.00	0.29	0.44	7
I-PAIS	1.00	0.40	0.57	5

A partir de esto, F1 es una mejor medida frente a la exactitud si necesitamos buscar un equilibrio entre Precisión y Cobertura , y hay una distribución de clase desigual.
F1-Score se puede definir como el promedio ponderado (o media armónica) de precisión y cobertura. Este puntaje tiene en cuenta tanto los **falsos positivos** como los **falsos negativos** para lograr un equilibrio entre precisión y cobertura. Matemáticamente se define:

$$F1 - Score = \frac{2 * (Precisión * cobertura)}{Precisión + cobertura}$$

2.3.5 - Micro-promedio y macro-promedio

Son dos promedios ligeramente diferentes; por lo tanto, su interpretación es distinta. Un macro-promedio calculará la métrica independientemente para cada clase y luego tomará el promedio. Por lo tanto, tratará a todas las clases por igual, mientras que un micro-promedio agregará las contribuciones de todas las clases para calcular la métrica promedio. En un sistema de multiclases como lo es un algoritmo NER, El macro-promedio trata a todas las clases por igual, mientras que el micro-promedio favorece a las clases más grandes. Por lo que vamos a tener situaciones donde el micro-promedio tenga valores por encima del 0.9, mientras que el macro-promedio sea de aproximadamente 0.6. Por esta razón es que solamente se va a hacer uso del macro-promedio.
Otra métrica observada en este proyecto es el macro-promedio ponderado, este calcula los valores para cada una de las clases, y luego devuelve el promedio considerando la proporción para cada etiqueta en el conjunto de datos. Sin embargo, esta métrica no es útil para el análisis del modelo desarrollado en este proyecto, ya que los valores con menor rendimiento tienen una contribución muy pequeña; por lo tanto, esta métrica va a favorecer las clases con mayores valores. Por ejemplo, tenemos el siguiente caso considerando la métrica F1-Score:

- Clase A: 0.29 - 6 reportes - contribuye 0.01
- Clase B: 0.92 - 478 reportes - contribuye 0.80

- Clase C: 0.17 - 51 reportes - contribuye 0.08
- Clase D:0.20 - 65 reportes - contribuye 0.1

$$P_{macro-promedio-ponderado} = 0.29 * 0.01 + 0.92 * 0.8 + 0.17 * 0.08 + 0.2 * 0.1 = 0.77$$
$$P_{macro-promedio} = \frac{0.29 + 0.92 + 0.17 + 0.2}{4} = 0.39$$

Debido a que el macro promedio va a favorecer los casos con mayor contribución, la métrica va a presentar valores altos usualmente. Por lo tanto, solamente se va a tener en consideración el macro-promedio

2.4 - Aprendizaje Automático

Aprendizaje automático (del inglés machine learning) es una rama de la inteligencia artificial que se ocupa de desarrollar algoritmos que las computadoras pueden utilizar para aprender los patrones en los datos de manera automatizada. Sin embargo ¿Cuándo necesitamos aprendizaje automático? Dos aspectos de un problema dado pueden requerir el uso de programas que aprenden y mejoran en función de su "experiencia": la complejidad del problema y la necesidad de adaptabilidad.
El aprendizaje es, por supuesto, un dominio muy amplio. En consecuencia, el campo de aprendizaje automático se ha ramificado en varios subcampos que se ocupan de diferentes tipos de tareas de aprendizaje. De manera simplificada podemos clasificar el paradigma en dos:

- **Supervisado y no supervisado:** Dado que el aprendizaje implica una interacción entre el alumno y el entorno, uno puede dividir las tareas de aprendizaje de acuerdo con la naturaleza de esa interacción.

De manera más abstracta, al ver el aprendizaje como un proceso de "uso de la experiencia para ganar más experiencia", el aprendizaje supervisado describe un escenario en el que la "experiencia", un ejemplo de entrenamiento, contiene información significativa (por ejemplo, las etiquetas de CALLE/O) que falta en los "ejemplos de prueba", a los que se aplicará la experiencia adquirida. En este contexto, la experiencia adquirida tiene como objetivo predecir la información que falta para los datos de prueba. En tales casos, podemos pensar en el entorno como un maestro que "supervisa" al alumno al proporcionarle información adicional (etiquetas). Sin embargo, en el aprendizaje no supervisado, no hay distinción entre los datos de entrenamiento y prueba. El alumno procesa los datos de entrada con el objetivo de obtener un resumen o una versión comprimida de esos datos. Agrupar un conjunto de datos en subconjuntos de objetos similares es un ejemplo típico de tal tarea [9].
Los algoritmos de aprendizaje automático tienen los siguientes componentes:

- **Conjunto de dominio**: un conjunto arbitrario, X. Este es el conjunto de objetos que podemos etiquetar. Por ejemplo, en el problema de reconocimiento de entidades, el conjunto de dominios será el conjunto de todos los reportes. Por lo general, estos puntos de dominio estarán representados por un vector de strings.
- **Conjunto de etiquetas**: para este proyecto, se restringirá el conjunto de etiquetas para que sea un conjunto de strings. Para nuestro problema de reconocimiento de entidades, tenemos que Y es {B-NSA, I-NSA, B-CALLE, I-CALLE, B-PAIS, I-PAIS, ...}, donde B representa el comienzo de la palabra e I significa el interior.
- **Datos de entrenamiento**: S = ((x1, y1)... (Xm, ym)) es una secuencia finita de pares en X × Y. Siguiendo el ejemplo del alumno, esta es la entrada a la que el alumno tiene acceso (como un conjunto de ejemplos que se han probado). Tales ejemplos etiquetados a menudo se denominan ejemplos de entrenamiento [9].

2.5 - CRF

Una manera de hacer un reconocimiento de entidades nombradas es a través de un modelo oculto de Markov (HMM de sus siglas en inglés); de esta manera se busca identificar la etiqueta más probable para cualquier palabra en una oración dada. HMM es un modelo genérico que define una distribución probabilística p(X,Y), donde X e Y son variables aleatorias, sobre secuencias de observación y su secuencia de etiquetas correspondiente [4]. Por ejemplo X podría ser una secuencia de palabras, mientras que Y una etiqueta dentro de un conjunto.

Un campo aleatorio condicional (CRF de sus siglas en inglés) es un método probabilístico popular para predicciones estructuradas, el cual es utilizado en distintas aplicaciones de procesamiento de lenguaje natural. Este enfoque tiene la capacidad de predecir múltiples variables que dependen unas de otras. Un campo aleatorio condicional puede entenderse como un grafo no dirigido o un modelo aleatorio de Markov, condicionado globalmente por X. En aplicaciones que se utilizan el enfoque CRF, deseamos para predecir un vector y = $\{y_0, y_1, ..., y_n\}$ de variables aleatorias dado un vector de características observado X. . De esta manera definir una probabilidad condicional $p(y \mid x)$ de la siguiente manera:

$$P(y|x, \omega) = \frac{e^{(\omega . f(x,y))}}{\sum_{y' \in y} e^{(\omega . f(x,y'))}}$$

Donde f representará un vector de características global definido por un conjunto de funciones de características $f_1 ... f_n$ donde cada función f_i puede calcularse utilizando todo acerca x_i, el actual y_i, el anterior y_{i-1} y la posición i

$$f(x, y) = \sum_{i=1}^{n} f(y_{i-1}, y_i, x, i)$$

De esta manera, la predicción consiste en maximizar la siguiente ecuación:

$$\widehat{y} = arg_y max(y \mid x, \omega)$$

Sin embargo, es necesario calcular el valor del parámetro ω, el cual va a ser determinado por los datos de entrenamiento $\{(x_1, y_1), ..., (x_n, y_n)\}$.

2.5.1 - Entrenamiento

El propósito de entrenamiento consiste en identificar todos los valores de ω. Por lo general, se puede establecer según el conocimiento del dominio. Sin embargo, en nuestro caso, aprendemos w de los datos de entrenamiento. Los datos de revisión completamente etiquetados son $\{(x_1, y_1), ..., (x_n, y_n)\}$ Dado que en los CRF, definimos la probabilidad condicional $p(y|x, \omega)$, el objetivo del aprendizaje de parámetros es maximizar la probabilidad condicional basada en los datos de entrenamiento:

$$L(\omega) = \sum_{i=1}^{m} logp(x_i, y_i, \omega)$$

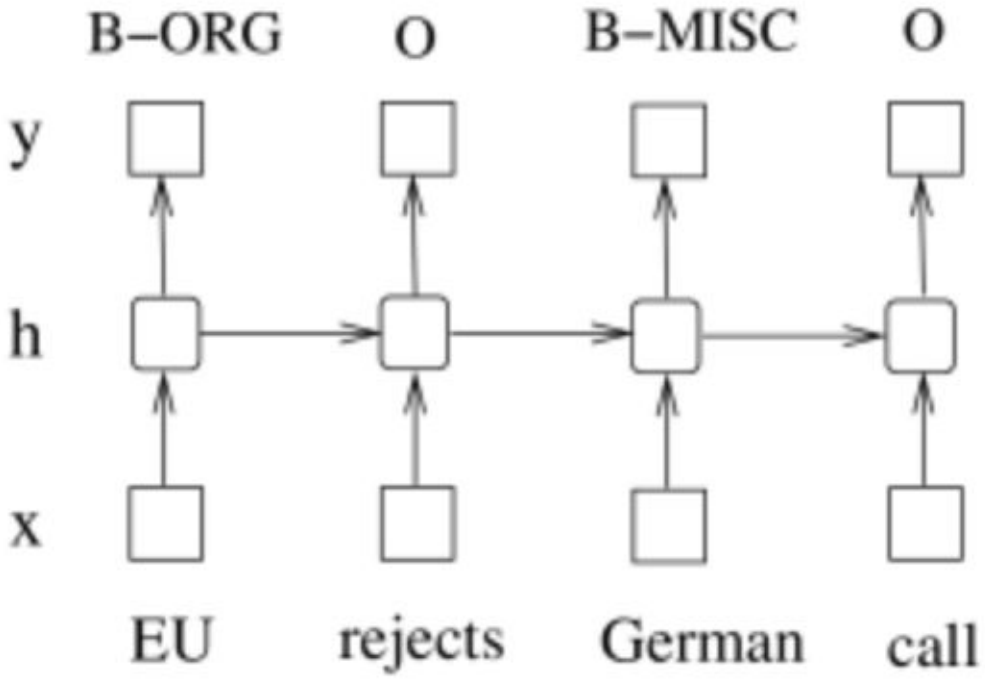
Por lo que, la estimaciones de este parámetro se se calcula como:

$$\omega^* = arg_{\omega \in R} max \sum_{i=1}^{m} logp(x_i, y_i, \omega) - \frac{\lambda}{2}|w|^2$$

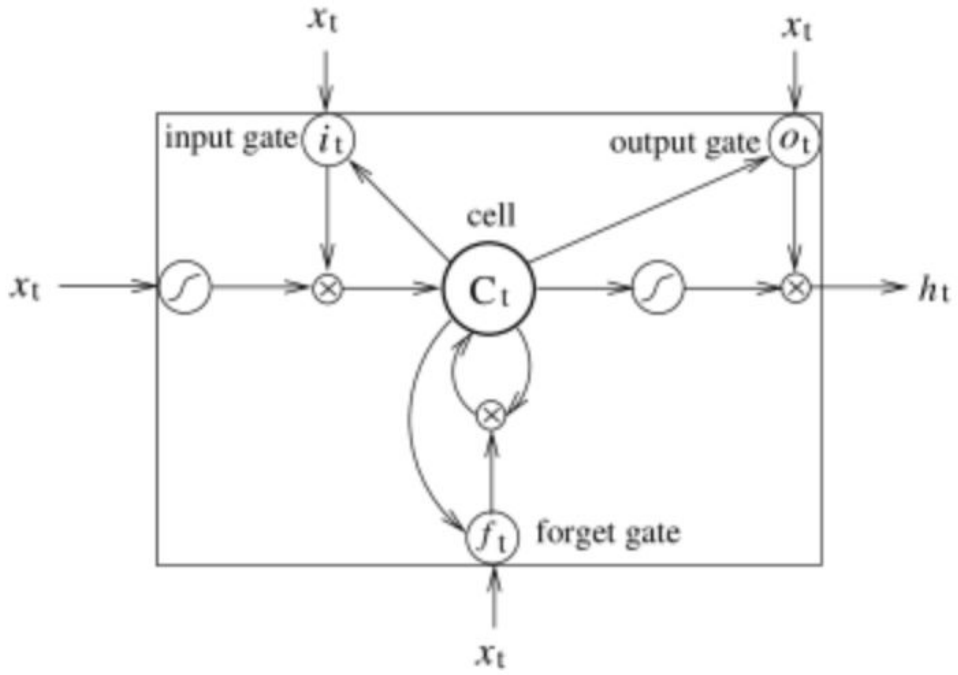
donde $\frac{\lambda}{2}|w|^2$ es un término de regularización L2. La ecuación es cóncava, por lo que w tiene un conjunto único de valores óptimos globales, por lo que el enfoque estándar para encontrar ω^* es calcular su correspondiente gradiente.

2.6 - LSTM

Las redes neuronales recurrentes (RNN) se han empleado para producir resultados prometedores en una variedad de tareas que incluyen, entre otras, el modelo del lenguaje. Un RNN mantiene una memoria basada en la información del historial, lo que permite al modelo predecir la salida actual condicionada por las características de larga distancia.
La imagen a continuación muestra la estructura RNN que tiene una capa de entrada x , una capa oculta h y una capa de salida y. En el contexto de etiquetado de entidad con nombre, x representa las palabras de entrada e y representa etiquetas.
La figura ilustra un sistema de reconocimiento de entidad con nombre en el que cada palabra está etiquetada con otro (O) o uno de los cuatro tipos de entidad: Persona (PER) , Ubicación (LOC) , Organización (ORG) y Misceláneo (MISC) . La sentencia *"EU rejects German call to boycott British lamb"* está etiquetada como B-ORG O B-MISC O O B-MISC O , donde las etiquetas B- , I- indican las posiciones inicial e interior de las entidades. [6]



Una capa de entrada representa entidades en el tiempo t . Podrían ser de codificación única para la función de palabra. Una capa de entrada tiene la misma dimensionalidad que el tamaño de la entidad. Una capa de salida representa una distribución de probabilidad sobre las etiquetas en el tiempo t. Tiene la misma dimensionalidad que el tamaño de las etiquetas. Una RNN introduce la conexión entre el estado oculto anterior y el estado oculto actual (y, por lo tanto, los parámetros de peso de capa recurrentes). Esta capa recurrente está diseñada para almacenar información del historial. Las redes de memoria a corto plazo son las mismas que las RNN, excepto que las actualizaciones de la capa oculta se reemplazan por celdas de memoria especialmente diseñadas. La imagen a continuación ilustra una sola celda de memoria LSTM.

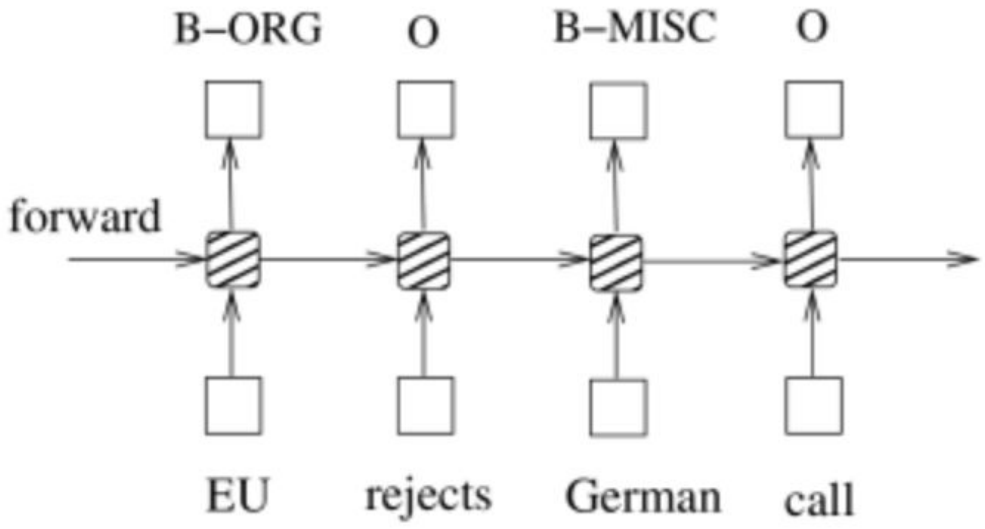


Teniendo en cuenta lo ilustrado anteriormente, una celda de memoria LSTM se implementa de la siguiente manera:

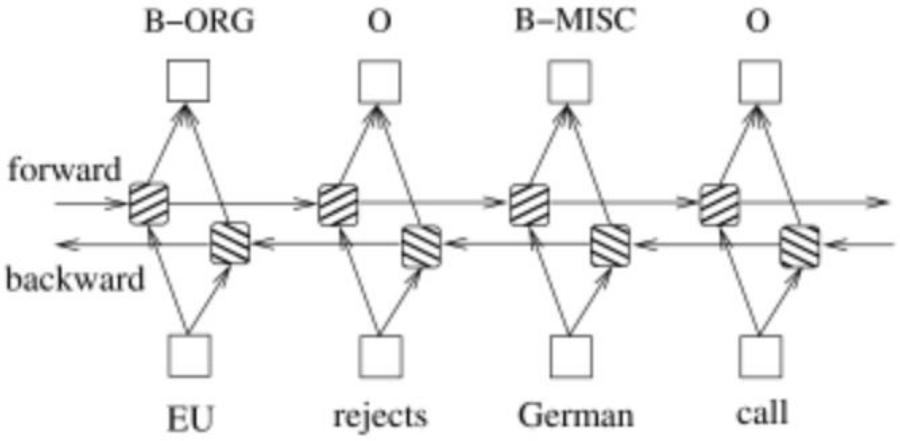
$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1}\, W_{ci}c_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1}\, W_{cf}c_{t-1} + b_{f)}$$
$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_{c)}$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1}\, W_{ci}c_{t-1} + b_o)$$
$$h_t = o_t tanh(c_t)$$

donde σ es la función sigmoidea por elementos e i, f o, c son la puerta de entrada, la puerta de olvido, la puerta de salida y los vectores de celda, todos los cuales tienen el mismo tamaño que el vector oculto h. x_t es el vector de entrada en el tiempo t y h_t representa el vector de estado oculto. Los pesos w_i, w_f, w_o, w_c y sesgo b_i, b_f, b_o, b_c son los parámetros a aprender. Dada una secuencia de vectores de entrada (x_1, x_2, …, x_n), LSTM calcula un vector de representación de contexto h_t para cada entrada x_t.

La imágen a continuación muestra un modelo de etiquetado de secuencia LSTM que emplea las celdas de memoria LSTM mencionadas anteriormente (cajas rayadas con esquinas redondeadas).

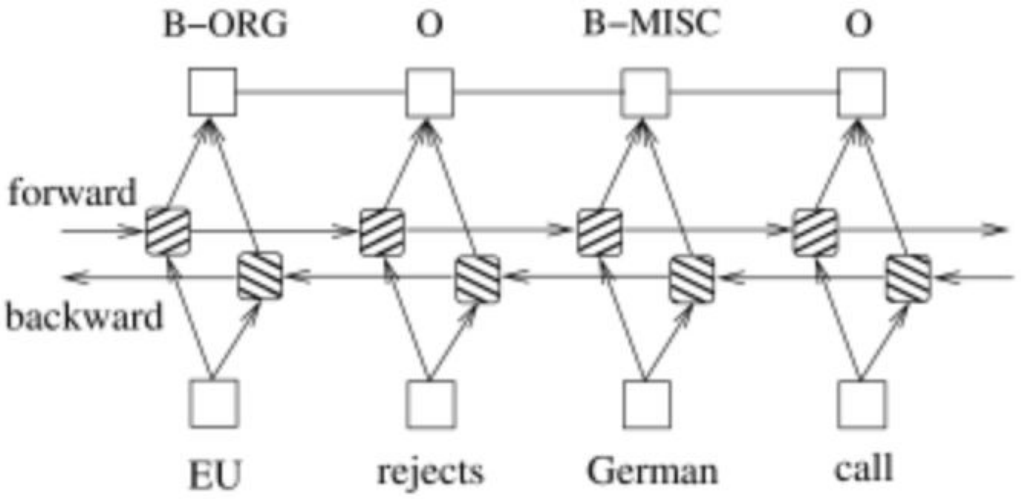


En la tarea de etiquetado de secuencias, tenemos acceso a las características de entrada pasadas y futuras durante un tiempo determinado, por lo tanto, podemos utilizar una red LSTM bidireccional como se propone. Al hacerlo, podemos hacer uso eficiente de características pasadas (a través de estados hacia adelante) y características futuras (a través de estados hacia atrás) para un marco de tiempo específico. Entrenamos redes LSTM bidireccionales utilizando la retropropagación a través del tiempo (BPTT). Los pases hacia adelante y hacia atrás sobre la red desplegada a lo largo del tiempo se llevan a cabo de manera similar a los pases hacia adelante y hacia atrás de la red normal, excepto que necesitamos desplegar los estados ocultos para todos los pasos de tiempo. También necesitamos un tratamiento especial al principio y al final de los puntos de datos. En nuestra implementación, avanzamos y retrocedemos para oraciones completas y solo necesitamos restablecer los estados ocultos a 0 al comienzo de cada oración. Tenemos una implementación por lotes que permite procesar múltiples oraciones al mismo tiempo.



2.6.1 - REDES BI-LSTM-CRF

En una red LSTM-CRF se combina una red bidireccional LSTM y una red CRF para formar una red BI-LSTM-CRF. Además de las características de entrada pasadas y la información de etiqueta de nivel de oración utilizada en un modelo LSTM-CRF, un modelo BI-LSTM-CRF puede usar las características de entrada futuras. Las características adicionales pueden aumentar la precisión del etiquetado.



3 - Meddocan Corpus

Para este proyecto se dispone de un corpus llamado Meddocan compuesto por reportes médicos en español; el cual fue anonimizado para proteger la información sensible, por lo que cualquier dato que permita identificar al paciente fue eliminada o remplazada [2]. Meddocan es patrocinado por el gobierno de España; por lo cual, todos los reportes médicos encontrados en el dataset fueron registrado en dicho país y en Español.
Para este corpus un médico practicante ha seleccionado manualmente los reportes que fueron incluidos, el cual trabajó junto a un documentalista para anonimizarlos [3].

3.1 - Descripción del corpus

Meddocan es un corpus compuesto por 1000 reportes médicos divididos aleatoriamente en los sets: development, train y test; donde el dataset development y test está compuesto por 250 casos y el dataset train por 500. En cada dataset podemos encontrar 2 archivos por cada reporte, un .txt que contiene el reporte en texto plano y un .ann que contiene las anotaciones con sus respectivos tags, donde cada reporte está compuesto por alrededor de 500 palabras. Por un lado, hay que destacar que para este proyecto el tamaño del corpus es un limitante, ya que debido a los algoritmos seleccionados para el tratamiento de dicho proyecto, a mayor cantidad de reportes para el entrenamiento, mejores resultados pueden ser percibidos.

3.2 -Anotación y Entidades

Las entidades sobre las cuales se trabajaron son los que se propone en el corpus Meddocan. Se decidió trabajar con una anotación IOB debido a que esta es simple y permite distinguir el comienzo y final de una entidad, donde la letra B representa el comienzo del tag y la letra I el interior del mismo. Se ha agregado un tag denominado O para representar contenido no relevante.
Debido a que algunos tags, dentro del corpus, son muy extensas, se decidió usar una abreviacions de cada uno, teniendo como referencia sus iniciales. Por lo que, el mapeo entre los tags identificado en el corpus y el utilizado en el algoritmo está representado en la siguiente tabla

MEDDOCAN	COMIENZO-ALGORITMO	INTERIOR-ALGORITMO
NOMBRE_SUJETO_ASISTENCIA	B-NSA	I-NSA
EDAD_SUJETO_ASISTENCIA	B-ESA	I-ESA
FAMILIARES_SUJETO_ASISTENCIA	B-FSA	I-FSA
NOMBRE_PERSONAL_SAN	B-NPS	I-NPS

ITARIO		
FECHAS	B-F	I-F
PROFESION	B-P	I-P
CENTRO_SALUD	B-CS	I-CS
HOSPITAL	B-HOSP	I-HOSP
INSTITUCIÓN	B-INSTITUCIÓN	I-INSTITUCIÓN
ID_TITULAR_PERSONAL_SANITARIO	B-ITPS	I-ITPS
NOMBRE_PERSONAL_SANITARIO	B-NPS	I-NPS
CALLE	B-CALLE	I-CALLE
TERRITORIO	B-TERR	I-TERR
PAIS	B-PAIS	I-PAIS
NUMERO_TELEFONO	B-NT	I-NT
CORREO_ELECTRONICO	B-CE	I-CE
ID_CONTACTO_ASISTENCIAL	B-ICA	I-ICA
ID_ASEGURAMIENTO	B-IA	I-IA
OTRO_SUJETO_ASISTENCIA	B-OSA	I-OSA

3.3 - Observaciones

Por un lado, para el correcto desarrollo de los algoritmos se tuvo especial cuidado en que los dataset de desarrollo y evaluación tengan las mismas entidades. Por otro lado, en la página oficial del corpus se puede observar tags que en este proyecto no son usados, estos son los siguientes:

- ID_EMPLEO_SANITARIO
- IDENTIF_VEHICULOS_NRSERIE_PLACAS
- IDENTIFICACION_DISPOSITIVOS_NSERIE
- NRO_BENEFICIENCIA_PLAN_SALUD
- URL_WEB
- DIREC_PROT_INTERNET
- IDENTIF_BIOMETRICOS

- OTRO_NUMERO_IDENTIFICACION
- NUMERO_FAX

Para el desarrollo de este proyecto se decidió utilizar los tres sets ofrecidos: development, train y test. Para el caso del conjunto de reportes development se utilizó más de 400 reportes médicos; los restantes reportes no incluidos fueron excluidos debido a que presentaron inconvenientes en el entrenamiento.

4 - Procesamiento de datos

Para hacer el entrenamiento de los algoritmos es necesario que los reportes estén previamente taggeadas cada una de las palabras. Sin embargo, el dataset no dispone una estructura acorde para el entrenamiento, por lo que se requirió un procesamiento de los reportes presentados.
El procesamiento de datos, junto con el desarrollo del código, fue la primer actividad desarrollada para el proyecto. Esto fue debido al no tener casos sobre los cuales entrenar los algoritmos, no se podía detectar anomalías o posibles mejoras. En primera instancia se trabajó exclusivamente sobre el dataset de train, el cual se dividió a este en dos parte, el 80% para entrenar y el 20% para testear el modelo resultante. Luego de completar el desarrollo de los algoritmos, se siguió procesando el dataset de test.
Durante el procesamiento de reportes se pudo predecir que ciertos tags van a ser poco precisos, esto es debido a los pocos casos de entrenamiento que había.

4.1 - Estructura de los datos

Los dos algoritmos sobre los cuales se trabajó requieren la misma estructura de datos para la entrada. Debido a que el corpus presenta el reporte en formato de texto, separado de las etiquetas. A partir de esto se trabajó para que los reportes estén en formato de tabla, donde la primer columna estaban cada una de las palabras y símbolos y la segunda el tag correspondiente. A modo de ejemplo tenemos el siguiente extracto de reporte:

Datos del paciente.

Nombre: María Carla.

Apellidos: Perez….

Sentence #	Word	TAG
1	Datos	O
1	del	O
1	paciente	O
1	.	O
1	Nombre	O
1	:	O
1	María	B-NSA
1	Carla	I-NSA
1	.	O

4.2 - Tags

4.2.1 - Obstáculos

Uno de los obstáculos presentados para el desarrollo del proyecto fue la estructura de datos que los algoritmos requerían. ya que este era muy distinto al que el dataset ofrecía. Para esto, en primera instancia, se buscó alguna herramienta que ayude al proceso; sin embargo, debido a que no se encontró una herramienta que se adapte a los requerimientos, se optó por realizar un script personalizado para este problema.

4.2.2 - Script

Como se ha mencionado previamente, cada dataset del corpus tiene los reportes en dos archivos. En el primero archivo .txt se encuentra el mismo en texto plano, en el segundo archivo .ann se encuentran los tags identificados. Sin embargo, esta anotación no es válida para que sea utilizado en nuestro. Por esto es que se decidió utilizar un script que ayude en la tarea de transformar la estructura del reporte en una válida como entrada de datos para los algoritmos. El script tiene como objetivo transformar el texto plano en un texto que un texto que tenga una palabra o símbolo por fila, de esta manera poder taggear cada una de las palabras y que el algoritmo pueda detectarlas. para esto se utilizó un algoritmo basado en una lista blanca de expresiones regulares.

5 - Implementación

5.1 - Algoritmo CRF

El algoritmo CRF implementado se puede dividir en tres partes.

- Primer Parte: En la primera se procesan los datos para que estos puedan ser utilizados por el algoritmo hasta tener listadas las sentencias y los tags en dos variables distintas. En esta etapa es en la que se define el vector de características. Este es el proceso más importante para los enfoques de aprendizaje automático porque el diseño de una característica afecta en gran medida los resultados del etiquetado. Para este proyecto extraemos los siguientes atributos: word.lower() word[-3:] word[-2:] word.isupper(), word.istitle() word.isdigit()
- Segunda Parte: Una vez que se tiene los datos procesados, se está listo para entrenar y obtener un modelo. CRFsuite leerá los datos de entrenamiento, generará el estado necesario (etiqueta de atributo) y las características de transición (etiqueta bigram) en función de los datos, maximizará la probabilidad de registro de la distribución de probabilidad condicional y almacenará el modelo.
- Tercera Parte: En la última parte se evalúa el modelo, para esto se utilizan las métricas precisión, recall y f1-score, dado que el resultado accuracy obtenido no es buen indicador. Para esto se muestra una tabla con todos los valores relevantes.

5.1.1 - Librerías

Para este proyecto se decidió utilizar la librería sklearn para la implementación de CRF, la cual es extensa y muy flexible con respecto a los parámetros para su utilización. Además, ofrece distintas métricas, facilitando el análisis del mismo. De esta librería se utilizó la función flat_classification_report, que permite tener una visualización de todos los tags, junto con el micro y macro avg, variables útiles para el análisis.
La librería sklearn es una implementación de campos aleatorios condicionales para etiquetar datos secuenciales, la cual fue elegida debido a las siguientes características:

- una implementación de Linear-chain (first-order Markov) CRF, donde esta hace uso del algoritmo Limited-memory BFGS (L-BFGS) para el entrenamiento.
- Formato de datos simple para entrenamiento y etiquetado. El formato de datos es similar a los utilizados en otras herramientas de aprendizaje automático;
- Evaluación de desempeño en capacitación. CRFsuite puede generar precisión, recuperación, puntajes F1 del modelo evaluado en los datos de la prueba.

5.1.2 - Parámetros

La librería seleccionada ofrece múltiples parámetros para modificar el comportamiento por defecto de la implementación. Sin embargo, se decidió trabajar solamente sobre los siguientes: [5]

- algorithm: algoritmo de entrenamiento
- c1: El coeficiente para la regularización de L1. Si se especifica un valor distinto de cero, CRFsuite cambia al método de Quasi-Newton Orthant-Wise Limited-memory (OWL-QN). El valor predeterminado es cero (sin regularización L1).
- c2: El coeficiente para la regularización de L2.

- max_iterations: El número máximo de iteraciones para algoritmos de optimización.
- all_possible_transitions: especifique si se genera características de transición que ni siquiera ocurren en los datos de entrenamiento.

Esta librería ofrece distintos algoritmos para la implementación, los cuales son los siguientes:

- 'lbfgs' - Gradient descent using the L-BFGS method
- 'l2sgd' - Stochastic Gradient Descent with L2 regularization term
- 'ap' - Averaged Perceptron
- 'pa' - Passive Aggressive (PA)
- 'arow' - Adaptive Regularization Of Weight Vector (AROW)

Este proyecto se va a implementar CRF utilizando el método L-BFGS como es mostrado en los conceptos preliminares; esto es debido a la conclusión llegada luego de leer documentación pertinente. Sin embargo, se va a mostrar un ejemplo con la implementación 'l2sgd' con el objetivo de mostrar otras opciones.

Luego, se va a mostrar un caso utilizando hiperparametrización. Esto permite calcular los mejores valores de C1 y C2, así llegar lo más posible a unos valores óptimos.

5.2 - Algoritmo CRF-LSTM

El algoritmo CRF-LSTM, al igual que el anterior enfoque, la implementación se puede dividir en tres partes:

- Primer parte: La primer tarea que se debe hacer es procesar los datos, esto incluye: formateo de datos, crear una lista de duplas para organizar los datos de entrada y diferenciar las oraciones entre sí, convertir todos los identificadores a tipo numérico y, por último, estandarizar la longitud de las sentencias.
- Segunda parte: Luego de procesar los datos, se procede a crear el modelo. Para esto, se utilizarán 6 capashasta conseguir el modelo final.
    - Capa Input: La capa input (capa de entrada) toma un parámetro de forma que es una dupla que indica la dimensionalidad de los datos de entrada.
    - Capa Embedding: básicamente es una búsqueda de diccionario que toma enteros como entrada y devuelve los vectores asociados. Se necesitan tres **parámetros** :
        - **input_dim** : Tamaño del vocabulario en los datos de texto, es decir; n_words + 1
        - **output_dim** : Dimensionalidad
        - **input_length** : longitud de input_sequence, es decir; longitud de la oración más larga
    - Capa BI-LSTM: Toma cinco parámetros :
        - unidades : dimensionalidad del espacio de salida
        - return_sequences : si return_sequence = True, devuelve la secuencia completa de salida, de lo contrario, devuelve la última salida en la secuencia de salida.
        - dropout : Fracción de las unidades a soltar para la transformación lineal de las entradas. Se encuentra entre 0 y 1.
        - recurrent_dropout : Fracción de las unidades a soltar para la transformación lineal del estado recurrente. Se encuentra entre 0 y 1.
        - kernel_initializer : Inicializador para la matriz de ponderaciones del kernel, utilizada para la transformación lineal de las entradas.
    - Capa TimeDistributed: Es un contenedor que nos permite aplicar una capa a cada elemento de nuestra secuencia de forma independiente. Se utiliza en la clasificación de secuencia para mantener relaciones uno a uno en entrada y salida.
    - Capa CRF: No hemos aplicado ninguna personalización a la capa CRF. Hemos pasado el número de clases de salida a la capa CRF.
- Tercer Parte: Una vez obtenido el modelo, se procede a evaluarlo. Para esto se utiliza un ejemplo y se ejecuta sobre él. Luego, se muestra los gráficos junto con las métricas [11].

5.2.1 - librerías

Para este proyecto se decidió utilizar la librería Keras, API de redes neuronales de alto nivel, escrita en Python y capaz de ejecutarse sobre TensorFlow. algunas de las ventajas que ofrece esta y por las cuales fue elegida son:

- Permite la creación de prototipos fácil y rápida (a través de la facilidad de uso, la modularidad y la extensibilidad).
- Se ejecuta en CPU y GPU.

Además, se seguirá utilizando la librería sklearn para el cálculo de las métricas, donde se le adicionará los gráficos proporcionados por la librería matplotlib.

5.2.2 - Parámetros

Para la implementación del algoritmo de LSTM se utiliza la función fit, que permite entrar el modelo. Esta tiene distintos parámetros de entrada, sobre los cuales solo se ha modificadon modificado dos. Estos parámetros son:

- epochs: Número de épocas para entrenar al modelo. Una época es una iteración sobre todos los datos x e y proporcionados.
- max_len: Tamaño máximo de los reportes.

Con el objetivo de analizar distintos resultados, se ha realizado distintos escenarios cambiando la cantidad de épocas y la cantidad de caracteres de los textos.

5.2.3 - Obstáculos

Esta librería permite ejecutarse en GPU; sin embargo, se tuvo problemas con esto, aumentando el tiempo de procesamiento.

6 - Entrenamiento

La documentación ofrecida por la librería es completa, además de encontrar muchos ejemplos en la web para tener de referencia.
el entrenamiento o generación de un modelo comienza con brindar los casos de entrenamientos. Dichos casos deben estar formateados según se explicó previamente. Como se dijo con anterioridad, cuanto mayor sea la cantidad de casos de entrenamientos, mejores resultados en el modelo se va a obtener.
A partir de las sentencias y los tags listados, ambos algoritmos proceden a generar el modelo que podrá ser utilizado para el reconocimiento de entidades en nuevos reportes médicos.
Luego, una vez que se genera el modelo, se debe evaluarlo a través de casos de test de un nuevo dataset con las mismas características que el de entrenamiento. Es importante que los casos utilizados en la evaluación no estén presentes en los de entrenamiento. También es importante que haya coherencia entre el entrenamiento y la evaluación; Por ejemplo, que el dataset tenga más casos para poder hacer observaciones y aprender.
Si bien, muchas referencias en la documentación, proponían dividir aleatoriamente el dataset en 2 partes; donde la primera es el conjunto de entrenamiento y el segundo el de set. Siendo los porcentajes aproximadamente 85% y 25% respectivamente. Este enfoque solamente se utilizó para la etapa de desarrollo, para poder evaluar el modelo sin necesidad de un conjunto de datos para test.

6.1 - Obstáculos

Para el entrenamiento de ambos algoritmos se presentaron una serie de obstáculos; sin embargo, hay que destacar que el enfoque de LSTM-CRF tuvo más dificultades, especialmente en cuanto al hardware necesario, haciendo que el proceso de desarrollo demande mucho más tiempo.
Otro de los inconvenientes presentados fue el haber trabajado con librerías distintas para CRF y CRF-LSTM; esto retrasó el desarrollo de las evaluaciones e hizo que la comparativa de ambos modelos más difícil.

6.1.1 - Hardware

Para la primera instancia de este proyecto, el cual consistía en el el desarrollo de los algoritmos, se utilizó una computadora con placa de video integrada. Para esta primer parte, el hardware no fue un obstáculo significativo. Luego, cuando se procedió a entrenar, el hardware comenzó a ser un problema. En primer lugar, el dataset en la etapa de desarrollo era de 30 reportes, mientras que en la etapa de entrenamiento se utilizó aproximadamente 400 reportes, debido a esto el tiempo de entrenamiento aumentó significativamente. Particularmente, en el caso de LSTM-CRF, pasó de requerir unos pocos minutos a requerir más de 20 minutos para la misma cantidad de épocas.
En segundo lugar, la cantidad de épocas también aumentó significativamente, aumentó de 100 a 400, lo que significó un aumento a más de 2 horas para entrenar el mismo algoritmos.
Dada esta situación, el hardware ya no fue suficiente para obtener resultados rápidos, por lo que se utilizó en ciertas ocasiones, especialmente en la etapa de desarrollo, otro hardware con mejores características. Esto ayudó a obtener resultados más rápido y facilitar el proceso de análisis.

6.1.2 - Tamaño del Corpus

Debido a las características de ambos algoritmos, el tamaño del corpus fue un inconveniente. Por lo que se hubiese tenido un mejores resultados con un corpus de mayor tamaño, de esta manera el entrenamiento podría tener mejores resultados.

7 - Evaluación

Luego de haber entrenado el modelo a partir del conjunto de datos, se obtiene el modelo resultante. A partir de este se realiza una serie análisis con el objetivo de comparar los enfoques elegidos.
Para la evaluación se decide utilizar la F1-score, la cual fue definida previamente y establece una función armónica entre la precisión y cobertura. Para cada modelo se presentaran los valores obtenidos a partir del mismo conjunto de datos de entrenamiento. A partir de estas se determinará qué enfoque tiene mejores resultados.
Es importante mencionar que los modelos intermedios, los cuales fueron surgieron debido al desarrollo de ambos enfoques, no son mostrados. Esto es debido a que estos tenían errores; sin embargo, se presentará a modo de réplica y comparativa el proceso llevado para analizar los modelos.

7.1 - CRF

Para los primeros casos el modelo fue entrenado considerando otros proyectos con características similares, a partir de la recolección de información y pruebas con distintas características, se decide establecer los valores de C1, C2 y la cantidad de iteraciones. Sin embargo, luego se utilizó la técnica de optimización de hiperparámetros, la cual permite identificar los mejores valores para C1 y C2.

7.1.1 - Caso 1

Parámetros de entrenamiento para el algoritmo CRF:

```
Algorithm = 'lbfgs'
C1 = 0.4
C2 = 0.4
max_iterations = 100
all_possible_transitions = false
```

	precision	recall	f1-score	support
B-CALLE	0.96	0.89	0.92	407
B-CE	0.94	0.92	0.93	237
B-CS	0.00	0.00	0.00	5
B-ESA	0.93	0.94	0.94	494
B-F	0.96	0.91	0.94	572
B-FSA	0.67	0.37	0.48	59
B-HOSP	0.97	0.84	0.90	121
B-IA	0.83	0.88	0.85	183
B-ICA	0.00	0.00	0.00	36
B-INSTITUCION	0.33	0.10	0.15	51
B-ISA	0.70	0.77	0.74	274

B-ITPS	0.89	0.89	0.89	228
B-NPS	0.95	0.91	0.93	480
B-NSA	0.90	0.95	0.92	478
B-NT	0.95	0.59	0.73	32
B-OSA	0.00	0.00	0.00	2
B-PAIS	0.95	0.90	0.93	336
B-PROF	1.00	0.17	0.29	6
B-SSA	0.95	0.96	0.96	428
B-TERR	0.91	0.76	0.83	910
I-CALLE	0.92	0.93	0.93	2404
I-CE	0.95	0.99	0.97	1110
I-CS	0.00	0.00	0.00	26
I-ESA	0.87	0.93	0.90	448
I-F	0.97	0.97	0.97	2083
I-FSA	0.50	0.03	0.05	37
I-HOSP	0.91	0.84	0.87	411
I-IA	0.83	0.88	0.85	370
I-INSTITUCION	0.28	0.09	0.14	121
I-ISA	0.00	0.00	0.00	21
I-ITPS	0.88	0.89	0.88	442
I-NPS	0.94	0.92	0.93	1118
I-NSA	0.93	0.95	0.94	260
I-NT	0.92	0.59	0.72	59
I-OSA	0.00	0.00	0.00	4
I-PAIS	1.00	0.40	0.57	5
I-PROF	1.00	0.29	0.44	7
I-TERR	0.81	0.35	0.49	144
O	0.99	1.00	0.99	115298
accuracy			0.98	129707
macro avg	0.73	0.61	0.64	129707
weighted avg	0.98	0.98	0.98	129707

A pesar de que el tiempo requerido para entrenar no es un factor esencialmente importante, se destaca el poco tiempo de entrenamiento que lleva entrenarlo, con una media de 3 minutos. Esto facilitó mucho el desarrollo del algoritmo y la variación en los parámetros de entrada.

7.1.2 - Optimización de hiperparámetros

Para mejorar la calidad, se seleccionar los parámetros de regularización mediante búsqueda aleatoria y validación cruzada triple; de esta manera se podrá tener valores más óptimos. En particular, se buscará tener los mejore C1 y C2, que maximice la métrica F1-Score. Este proceso demanda bastante tiempo de CPU y RAM, requiriendo para el hardware utilizado aproximadamente 2 horas.

```
best params: {'c1': 0.07067075666654962, 'c2': 0.05765391389345155}
model size: 0.81M
```

	precision	recall	f1-score	support
B-CALLE	0.95	0.89	0.92	407
B-CE	0.93	0.94	0.93	237
B-CS	0.00	0.00	0.00	5
B-ESA	0.93	0.94	0.93	494
B-F	0.96	0.93	0.94	572
B-FSA	0.71	0.59	0.65	59
B-HOSP	0.98	0.84	0.91	121
B-IA	0.86	0.91	0.88	183
B-ICA	0.00	0.00	0.00	36
B-INSTITUCION	0.38	0.16	0.22	51
B-ISA	0.72	0.78	0.75	274
B-ITPS	0.92	0.92	0.92	228
B-NPS	0.95	0.91	0.93	480
B-NSA	0.91	0.95	0.93	478
B-NT	0.95	0.59	0.73	32
B-OSA	0.00	0.00	0.00	2
B-PAIS	0.95	0.92	0.93	336
B-PROF	0.75	0.50	0.60	6
B-SSA	0.95	0.97	0.96	428
B-TERR	0.89	0.80	0.84	910
I-CALLE	0.90	0.94	0.92	2404
I-CE	0.95	0.99	0.97	1110
I-CS	0.00	0.00	0.00	26
I-ESA	0.86	0.94	0.89	448
I-F	0.98	0.97	0.97	2083
I-FSA	0.50	0.08	0.14	37
I-HOSP	0.94	0.83	0.88	411
I-IA	0.86	0.91	0.88	370
I-INSTITUCION	0.34	0.13	0.19	121
I-ISA	1.00	0.05	0.09	21
I-ITPS	0.91	0.92	0.91	442
I-NPS	0.94	0.91	0.93	1118
I-NSA	0.91	0.95	0.93	260
I-NT	0.95	0.63	0.76	59
I-OSA	0.00	0.00	0.00	4
I-PAIS	1.00	0.40	0.57	5
I-PROF	0.67	0.57	0.62	7
I-TERR	0.81	0.43	0.56	144
O	0.99	1.00	0.99	115298

```
    accuracy                           0.99    129707
   macro avg       0.75      0.65      0.67    129707
weighted avg       0.98      0.99      0.98    129707
```

7.1.3 - Matriz de Confusión

La matriz de confusión (confusion matrix del inglés) es una herramiento que nos permite visualizar el desempeño del modelo. En particular, esta librería muestra una matriz por cada una de las etiquetas; donde el cuadrante [0, 0] es la cantidad de True negative, el cuadrante [1, 0] es la cantidad de false negative, el cuadrante [1, 1] es la cantidad de True positive, el cuadrante [0, 1] es la cantidad de false positive

True Negative	False Positive
False Negative	True Positive

A continuación se mostrará los valores observados en un entrenamiento del modelo CRF.

B-NSA
Total de apariciones 478

129178	51
25	453

I-NSA
Total de apariciones 260

129429	18
17	243

Con este ejemplo podemos observar y confirmar que claramente el modelo es asimétrico, por lo que no es recomendado utilizar la métrica de la exactitud como parámetro de rendimiento.
Uno de los obstáculos presentados para esta función, es que los tags entre el conjunto de reportes original y la predicción deben ser exactamente iguales.

7.1.4 Implementación Alternativa

Si bien, en este proyecto se enfoca la utilización de CRF con una implementación del algoritmo lbfgs, se decide hacer una prueba con una implementación alternativa mostrar otra opción que presenta la

librería. En este caso se puede observar que presenta valores análogos al primer caso. Este modelo fue entrenado con los parámetros sugeridos por la documentación oficial de la librería:

```
Algorithm = 'l2sgd'
C2 = 1
max_iterations = 1000
all_possible_transitions = false
```

	precision	recall	f1-score	support
B-CALLE	0.96	0.89	0.92	407
B-CE	0.93	0.91	0.92	237
B-CS	0.00	0.00	0.00	5
B-ESA	0.92	0.94	0.93	494
B-F	0.96	0.92	0.94	572
B-FSA	0.69	0.31	0.42	59
B-HOSP	0.97	0.84	0.90	121
B-IA	0.82	0.87	0.85	183
B-ICA	0.00	0.00	0.00	36
B-INSTITUCION	0.36	0.10	0.15	51
B-ISA	0.71	0.78	0.74	274
B-ITPS	0.89	0.89	0.89	228
B-NPS	0.94	0.91	0.92	480
B-NSA	0.87	0.94	0.91	478
B-NT	1.00	0.53	0.69	32
B-OSA	0.00	0.00	0.00	2
B-PAIS	0.97	0.89	0.93	336
B-PROF	1.00	0.17	0.29	6
B-SSA	0.95	0.96	0.95	428
B-TERR	0.91	0.74	0.81	910
I-CALLE	0.90	0.94	0.92	2404
I-CE	0.95	0.99	0.97	1110
I-CS	0.00	0.00	0.00	26
I-ESA	0.86	0.93	0.89	448
I-F	0.97	0.97	0.97	2083
I-FSA	0.00	0.00	0.00	37
I-HOSP	0.91	0.84	0.87	411
I-IA	0.83	0.87	0.85	370
I-INSTITUCION	0.35	0.13	0.19	121
I-ISA	0.00	0.00	0.00	21
I-ITPS	0.88	0.89	0.88	442
I-NPS	0.93	0.92	0.92	1118
I-NSA	0.92	0.95	0.93	260
I-NT	0.89	0.58	0.70	59
I-OSA	0.00	0.00	0.00	4

I-PAIS	1.00	0.40	0.57	5
I-PROF	1.00	0.29	0.44	7
I-TERR	0.85	0.35	0.49	144
O	0.99	1.00	0.99	115298
accuracy			0.98	129707
macro avg	0.72	0.61	0.64	129707
weighted avg	0.98	0.98	0.98	129707

7.2 - LSTM-CRF

7.2.1 - Caso 1

Para este enfoque se realizaron dos entrenamientos; el primero consistía en solamente utilizar el conjunto de entrenamiento, dividiéndolo en dos partes. La primer mitad consistía en el 85% del conjunto de datos, el cual era utilizado para entrenar al modelo, mientras que el restante 15% es utilizado para evaluar el modelo.
En esta evaluación se puede observar que el conjunto de datos a entrenar es considerablemente más pequeño que el conjunto correspondiente al de Test otorgado por Medoccan. El hecho de que se tenga pocos reportes para evaluar el modelo afecta al valor de la métrica F1-Score.
Para este entrenamiento se debe considerar que la evaluación tiene una cantidad muy pequeña de reportes; por lo que los valores altos pueden no ser completamente válidos.
Para el resto de entrenamientos que se hicieron se utilizó el segundo conjunto de datos para la evaluación, siendo el mismo utilizado en CRF. Esto permitió tener una mejor evaluación del modelo.
En este caso en vez de tener asignada la etiqueta O para las palabras que no estaban asignadas a ningún tag, fue asignado con el tag PAD.

Datos de entrada

```
batch_size = 64
epoches = 300
max_len = 800
embedding = 50
```

	precision	recall	f1-score	support
B-CALLE	0.91	0.82	0.86	82
B-CE	0.94	0.85	0.89	59
B-ESA	0.97	0.95	0.96	96
B-F	0.99	0.86	0.92	115

B-FSA	0.15	0.22	0.18	9
B-HOSP	0.87	0.91	0.89	22
B-IA	0.96	0.90	0.93	30
B-ICA	0.00	0.00	0.00	3
B-INSTITUCION	0.75	0.50	0.60	12
B-ISA	0.92	0.55	0.69	40
B-ITPS	0.98	0.98	0.98	43
B-NPS	0.93	0.85	0.89	109
B-NSA	0.99	0.94	0.96	72
B-NT	1.00	0.33	0.50	3
B-PAIS	0.89	0.92	0.91	64
B-PROF	0.00	0.00	0.00	1
B-SSA	0.97	0.98	0.97	88
B-TERR	0.93	0.72	0.81	195
I-CALLE	0.98	0.88	0.93	503
I-CE	0.96	0.95	0.95	279
I-ESA	0.98	0.95	0.96	97
I-F	0.99	0.93	0.96	400
I-FSA	0.00	0.00	0.00	2
I-HOSP	0.80	0.89	0.84	74
I-IA	0.96	0.91	0.94	58
I-INSTITUCION	0.67	0.40	0.50	30
I-ISA	1.00	0.33	0.50	6
I-ITPS	0.95	0.87	0.91	84
I-NPS	0.92	0.82	0.87	248
I-NSA	0.95	0.89	0.92	44
I-NT	1.00	0.50	0.67	4
I-PROF	0.00	0.00	0.00	1
I-TERR	0.46	0.25	0.32	24
PAD	0.99	1.00	0.99	29103
accuracy			0.99	32000
macro avg	0.79	0.67	0.71	32000
weighted avg	0.99	0.99	0.99	32000

7.1.2 - Caso 2

En este segundo caso se puede observar que la el valor de F1-Score bajó considerablemente, esto se puede observar en la columna Support. Por lo que, se puede decir que el caso 1 no tenía un buen conjunto de datos para la evaluación.

Datos de entrada:

```
batch_size = 512
epoches = 300
max_len = 800
embedding = 50
```

	precision	recall	f1-score	support
O	0.99	1.00	0.99	179227
B-NSA	0.62	0.60	0.61	429
I-NSA	0.78	0.66	0.71	234
B-ISA	0.10	0.00	0.01	246
B-IA	0.09	0.05	0.07	165
I-IA	0.90	0.33	0.48	335
B-CALLE	0.89	0.84	0.86	386
I-CALLE	0.89	0.76	0.82	2274
B-TERR	0.68	0.56	0.61	870
B-F	0.87	0.89	0.88	533
I-F	0.93	0.94	0.93	1936
B-PAIS	0.82	0.69	0.75	319
B-ESA	0.89	0.89	0.89	458
I-ESA	0.76	0.86	0.81	415
B-SSA	0.86	0.94	0.90	399
B-NPS	0.93	0.81	0.86	464
I-NPS	0.92	0.78	0.84	1081
B-ITPS	0.95	0.89	0.92	216
I-ITPS	0.94	0.94	0.94	418
B-CE	0.21	0.20	0.21	237
I-CE	0.74	0.96	0.84	1110
B-HOSP	0.78	0.42	0.55	120
I-HOSP	0.57	0.57	0.57	405
I-TERR	0.31	0.04	0.06	141
B-OSA	0.00	0.00	0.00	0
B-FSA	0.21	0.15	0.18	59
B-INSTITUCION	0.50	0.02	0.04	50
I-INSTITUCION	0.07	0.01	0.02	119
B-NT	0.62	0.25	0.36	32
I-NT	0.60	0.20	0.30	59
I-FSA	0.00	0.00	0.00	0
B-ICA	0.00	0.00	0.00	31
B-PROF	0.00	0.00	0.00	5
I-PROF	0.00	0.00	0.00	5

I-ISA	0.00	0.00	0.00	18
I-PAIS	0.00	0.00	0.00	4
I-OSA	0.00	0.00	0.00	0
B-CS	0.00	0.00	0.00	0
I-CS	0.00	0.00	0.00	0
micro avg	0.98	0.98	0.98	192800
macro avg	0.50	0.42	0.44	192800
weighted avg	0.97	0.98	0.98	192800

7.2.3 - Caso 3

En este tercer caso solamente se aumentó la cantidad de épocas, con el objetivo de observar si esto hacía mejorar el resultado del modelo resultante.

Datos de entrada

```
batch_size = 64
epoches = 500
max_len = 800
embedding = 50
```

	precision	recall	f1-score	support
O	0.99	1.00	0.99	179153
B-NSA	0.86	0.81	0.83	429
I-NSA	0.83	0.81	0.82	234
B-ISA	0.50	0.01	0.02	246
B-IA	0.93	0.70	0.80	165
I-IA	0.93	0.77	0.85	335
B-CALLE	0.88	0.83	0.86	386
I-CALLE	0.95	0.79	0.86	2274
B-TERR	0.87	0.66	0.75	870
B-F	0.94	0.91	0.93	533
I-F	0.98	0.96	0.97	1936
B-PAIS	0.94	0.86	0.90	319
B-ESA	0.83	0.91	0.87	458
I-ESA	0.65	0.88	0.75	415
B-SSA	0.96	0.94	0.95	399
B-NPS	0.93	0.80	0.86	464
I-NPS	0.94	0.65	0.77	1081
B-ITPS	0.97	0.76	0.85	216

```
       I-ITPS       0.94      0.53      0.67       418
         B-CE       0.68      0.75      0.71       237
         I-CE       0.89      0.95      0.92      1110
       B-HOSP       0.85      0.73      0.79       120
       I-HOSP       0.84      0.66      0.74       405
       I-TERR       0.61      0.20      0.30       141
        B-OSA       0.00      0.00      0.00         2
        B-FSA       0.43      0.44      0.43        59
B-INSTITUCION       0.28      0.10      0.15        50
I-INSTITUCION       0.21      0.08      0.11       119
         B-NT       0.69      0.28      0.40        32
         I-NT       0.83      0.34      0.48        59
        I-FSA       0.00      0.00      0.00        37
        B-ICA       0.00      0.00      0.00        31
       B-PROF       0.17      0.20      0.18         5
       I-PROF       0.40      0.40      0.40         5
        I-ISA       0.00      0.00      0.00        18
       I-PAIS       0.00      0.00      0.00         4
        I-OSA       0.00      0.00      0.00         4
         B-CS       0.00      0.00      0.00         5
         I-CS       0.00      0.00      0.00        26

    accuracy                           0.98    192800
   macro avg       0.61      0.51      0.54    192800
weighted avg       0.98      0.98      0.98    192800
```

7.2.4 - Caso 4

Para el cuarto entrenamiento se aumentó la cantidad de épocas y la máxima cantidad de caracteres por reportes. Sin embargo, se observó que esto no afectó significativamente el resultado, a pesar de haber reportes excluidos con un máximo de 800 caracteres

Datos de entrada

```
batch_size = 512
epoches = 1000
max_len = 800
embedding = 50
```

```
              precision    recall  f1-score   support

           O       0.98      1.00      0.99    111784
       B-NSA       0.97      0.92      0.95       429
       I-NSA       0.95      0.89      0.92       234
```

B-ISA	0.10	0.02	0.04	246
B-IA	0.90	0.68	0.78	165
I-IA	0.92	0.52	0.67	335
B-CALLE	0.90	0.80	0.85	386
I-CALLE	0.95	0.81	0.87	2274
B-TERR	0.93	0.67	0.78	870
B-F	0.89	0.89	0.89	533
I-F	0.97	0.95	0.96	1936
B-PAIS	0.94	0.87	0.90	319
B-ESA	0.95	0.91	0.93	458
I-ESA	0.89	0.87	0.88	415
B-SSA	0.95	0.95	0.95	399
B-NPS	0.94	0.86	0.90	464
I-NPS	0.95	0.76	0.85	1081
B-ITPS	0.97	0.94	0.96	216
I-ITPS	0.96	0.79	0.87	418
B-CE	0.28	0.18	0.22	237
I-CE	0.79	0.96	0.87	1110
B-HOSP	0.91	0.68	0.78	120
I-HOSP	0.80	0.63	0.71	405
I-TERR	0.82	0.26	0.40	141
B-OSA	0.00	0.00	0.00	0
B-FSA	0.53	0.32	0.40	59
B-INSTITUCION	0.20	0.16	0.18	50
I-INSTITUCION	0.15	0.11	0.12	119
B-NT	0.70	0.44	0.54	32
I-NT	0.78	0.54	0.64	59
I-FSA	0.00	0.00	0.00	0
B-ICA	0.00	0.00	0.00	31
B-PROF	0.33	0.20	0.25	5
I-PROF	0.50	0.40	0.44	5
I-ISA	0.04	0.39	0.08	18
I-PAIS	0.00	0.00	0.00	4
I-OSA	0.00	0.00	0.00	0
B-CS	0.00	0.00	0.00	0
I-CS	0.00	0.00	0.00	0
micro avg	0.97	0.97	0.97	125357
macro avg	0.61	0.52	0.55	125357
weighted avg	0.97	0.97	0.97	125357

7.2.5 - Obstáculos

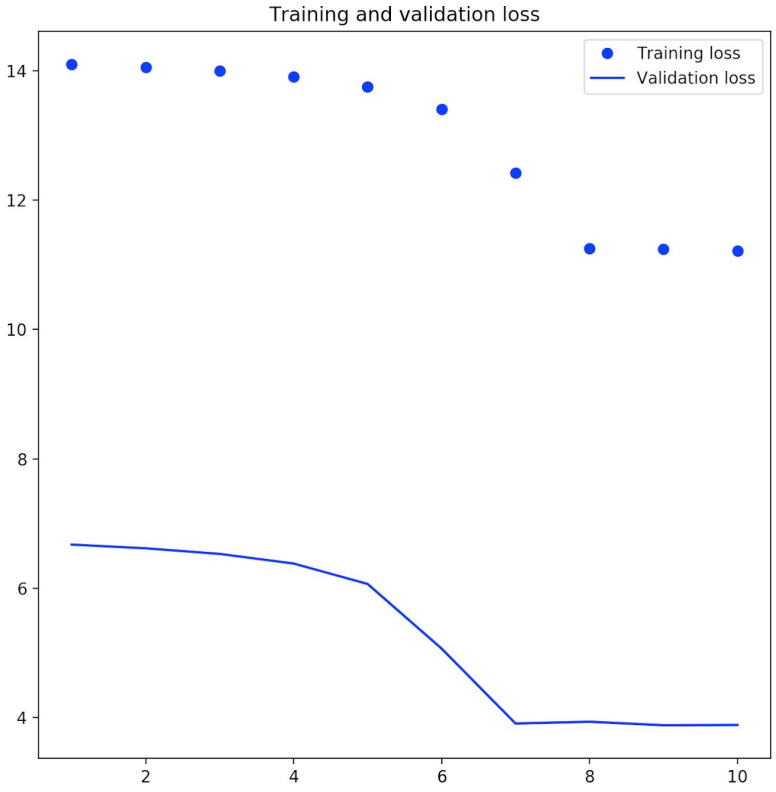
Para este enfoque se tuvo dos obstáculos; el primero, es el hecho de tener un conjunto de datos muy reducido. Este algoritmo tiene la característica de requerir muchos datos de entrenamiento para obtener buenos resultados.
El segundo obstáculo es que requiere mucho tiempo de entrenamiento, tardando entre 15 a 20 segundo cada época. Este inconveniente, junto con el hecho de que se requiere una gran cantidad de épocas para tener buenos resultados, hizo que el desarrollo de este algoritmo fuese lento.

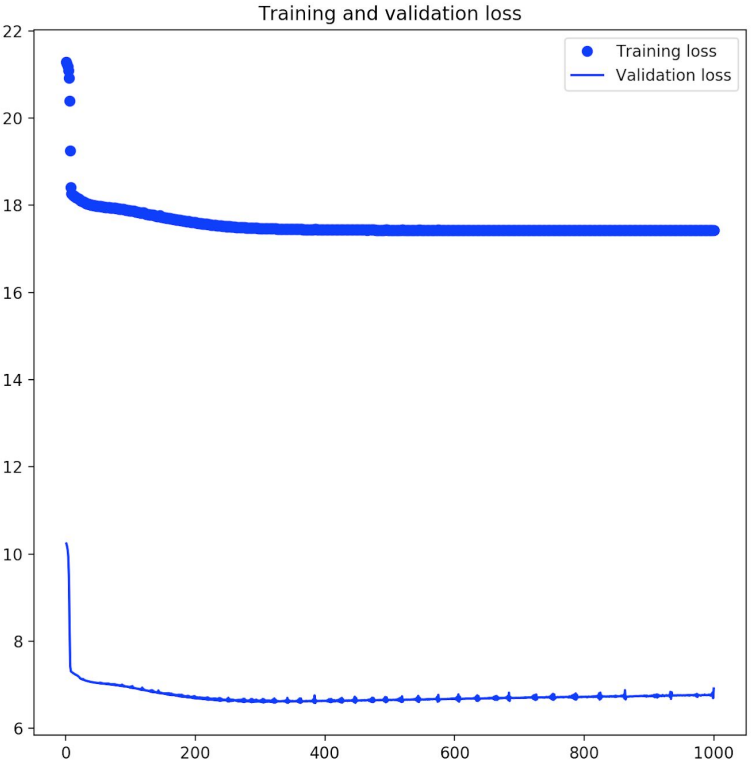
7.2.6 - Gráfico

Para completar el análisis se agregó un gráfico denominado pérdida de entrenamiento y validación (del inglés training and validation loss). Este gráfico nos va a permitir observar cómo resulta el entrenamiento. A partir de este podemos tener los siguientes resultados:

- Si Training loss >> Validation loss, el sistema está underfitting
- Si Training loss << Validation loss, el sistema está overfitting
- Si Training loss ~ Validation loss, el sistema probablemente esté bien

En la primer imágen se muestra como el sistema va evolucionando en las primeras épocas. En la segunda imágen se observa un sistema entrenado con 1000 épocas. A partir de esto podemos concluir que el sistema está underfitting.
Dado este gráfico y el hecho de que el sistema no mejora a partir de la época 400, se puede deducir que el modelo no va a mejorar si no se le agrega más datos reportes al dataset.


Training and validation loss
Training loss
Validation loss
14
12
10
8
6
4
2
4
6
8
10


Training and validation loss
Training loss
Validation loss
22
20
18
16
14
12
10
8
6
0
200
400
600
800
1000

7.3 - Ejemplo de Aplicación

CRF

A continuación se muestra la aplicación de un modelo entrenado con CRF según el caso 1 mostrado previamente.

```
Word            ||True ||Pred
==============================
Nombre          : O     O
:               : O     O
Esteban         : B-NSA B-NSA
.               : O     O
Apellidos       : O     O
:               : O     O
Franco          : B-NSA B-NSA
Romero          : I-NSA I-NSA
.               : O     O
NHC             : O     O
:               : O     O
94646874        : B-ISA B-ISA
.               : O     O
NASS            : O     O
:               : O     O
25              : B-IA  B-IA
38594235        : I-IA  I-IA
54              : I-IA  I-IA
.               : O     O
Domicilio       : O     O
:               : O     O
Calle           : B-CALLE B-CALLE
Gustavo         : I-CALLE I-CALLE
Pittaluga       : I-CALLE I-CALLE
,               : I-CALLE I-CALLE
6               : I-CALLE I-CALLE
.               : O     O
Localidad       : O     O
/               : O     O
Provincia       : O     O
:               : O     O
MÃ¡laga         : B-TERR B-TERR
.               : O     O
CP              : O     O
:               : O     O
29130           : B-TERR B-TERR
```

```
.               : O     O
Datos           : O     O
asistenciales   : O     O
.               : O     O
Fecha           : O     O
de              : O     O
nacimiento      : O     O
:               : O     O
21              : B-F   B-F
/               : I-F   I-F
9               : I-F   I-F
/               : I-F   I-F
1975            : I-F   I-F
.               : O     O
PaÃs           : O     O
:               : O     O
EspaÃ±a         : B-PAIS B-PAIS
.               : O     O
Edad            : O     O
:               : O     O
42              : B-ESA B-ESA
aÃ±os           : I-ESA I-ESA
Sexo            : O     O
:               : O     O
H               : B-ESA B-SSA
.               : O     O
Fecha           : O     O
de              : O     O
Ingreso         : O     O
:               : O     O
13              : B-F   B-F
/               : I-F   I-F
1               : I-F   I-F
/               : I-F   I-F
2018            : I-F   I-F
.               : O     O
Episodio        : O     O
:               : O     O
1520368541      : B-ICA B-ISA
.               : O     O
MÃ©dico         : O     O
:               : O     O
Jose            : B-NPS B-NPS
Miguel          : I-NPS I-NPS
Mora            : I-NPS I-NPS
OrdÃ³Ã±ez       : I-NPS I-NPS
```

```
NÂ°Col          : O     O
:              : O     O
29             : B-ITPS B-ITPS
29             : I-ITPS I-ITPS
78451          : I-ITPS I-ITPS
.              : O     O
Antecedentes   : O     O
:              : O     O
bebedor        : O     O
de             : O     O
mÃ¡s           : O     O
de             : O     O
100            : O     O
g              : O     O
de             : O     O
etanol         : O     O
al             : O     O
dÃa           : O     O
,              : O     O
sin            : O     O
otros          : O     O
antecedentes   : O     O
de             : O     O
interÃ©s       : O     O
.              : O     O
Historia       : O     O
Actual         : O     O
:              : O     O
Hombre         : B-SSA B-SSA
de             : O     O
42             : B-ESA B-ESA
aÃ±os          : I-ESA I-ESA
que            : O     O
refiere        : O     O
un             : O     O
cuadro         : O     O
de             : O     O
3              : O     O
dÃas           : O     O
de             : O     O
evoluciÃ³n     : O     O
(              : O     O
coincidiendo   : O     O
con            : O     O
un             : O     O
aumento        : O     O
```

```
en              : O     O
la              : O     O
ingesta         : O     O
habitual        : O     O
de              : O     O
alcohol         : O     O
)               : O     O
de              : O     O
febrÃcula     : O     O
,               : O     O
ictericia       : O     O
,               : O     O
dolor           : O     O
abdominal       : O     O
en              : O     O
hipocondrio     : O     O
derecho         : O     O
y               : O     O
distensiÃ³n     : O     O
abdominal       : O     O
.               : O     O
ExploraciÃ³n    : O     O
fÃsica         : O     O
:               : O     O
En              : O     O
la              : O     O
exploraciÃ³n    : O     O
fÃsica         : O     O
destaca         : O     O
:               : O     O
ictericia       : O     O
muco            : O     O
-               : O     O
cutÃ¡nea        : O     O
franca          : O     O
,               : O     O
encefalopatÃa : O     O
grado           : O     O
I               : O     O
y               : O     O
ascitis         : O     O
.               : O     O
No              : O     O
se              : O     O
aprecia         : O     O
aumento         : O     O
```

```
del             : O     O
hÃgado        : O     O
,               : O     O
ni              : O     O
del             : O     O
bazo            : O     O
.               : O     O
La              : O     O
auscultaciÃ³n  : O     O
cardiorrespiratoria: O     O
era             : O     O
normal          : O     O
.               : O     O
Resumen         : O     O
de              : O     O
pruebas         : O     O
complementarias: O     O
:               : O     O
En              : O     O
la              : O     O
analÃtica     : O     O
de              : O     O
ingreso         : O     O
se              : O     O
obtuvieron      : O     O
los             : O     O
siguiente       : O     O
resultados      : O     O
:               : O     O
leucocitos      : O     O
26              : O     O
.               : O     O
820             : O     O
por             : O     O
Âµl             : O     O
(               : O     O
neutrÃ³filos   : O     O
,               : O     O
77              : O     O
%               : O     O
)               : O     O
,               : O     O
actividad       : O     O
de              : O     O
protrombina     : O     O
40              : O     O
```

```
%               : O     O
,               : O     O
creatinina      : O     O
plasmÃ¡tica     : O     O
1               : O     O
,               : O     O
9               : O     O
mg              : O     O
/               : O     O
dl              : O     O
;               : O     O
urea            : O     O
91mg            : O     O
/               : O     O
dl              : O     O
;               : O     O
bilirrubina     : O     O
total           : O     O
21mg            : O     O
/               : O     O
dl              : O     O
;               : O     O
AST             : O     O
164             : O     O
U               : O     O
/               : O     O
l               : O     O
;               : O     O
ALT             : O     O
103             : O     O
U               : O     O
/               : O     O
l               : O     O
;               : O     O
GGT             : O     O
152             : O     O
UI              : O     O
/               : O     O
dl              : O     O
.               : O     O
Marcadores      : O     O
virales         : O     O
hepÃ¡ticos      : O     O
negativos       : O     O
.               : O     O
La              : O     O
```

```
radiografÃa   : O     O
de             : O     O
tÃ³rax         : O     O
mostrÃ³        : O     O
un             : O     O
infiltrado     : O     O
alveolar       : O     O
basal          : O     O
derecho        : O     O
y              : O     O
la             : O     O
ecografÃa     : O     O
abdominal      : O     O
un             : O     O
hÃgado        : O     O
pequeÃ±o       : O     O
,              : O     O
con            : O     O
ecogenicidad   : O     O
aumentada      : O     O
y              : O     O
estructura     : O     O
homogÃ©nea     : O     O
;              : O     O
vÃa           : O     O
biliar         : O     O
normal         : O     O
;              : O     O
vena           : O     O
porta          : O     O
permeable      : O     O
,              : O     O
esplenomegalia : O     O
,              : O     O
abundante      : O     O
ascitis        : O     O
.              : O     O
En             : O     O
la             : O     O
tomografÃa    : O     O
axial          : O     O
computarizada  : O     O
(              : O     O
TAC            : O     O
)              : O     O
de             : O     O
```

```
abdomen         : O     O
se              : O     O
pudo            : O     O
apreciar        : O     O
un              : O     O
hÃgado         : O     O
pequeÃ±o        : O     O
de              : O     O
contorno        : O     O
lobulado        : O     O
.               : O     O
EvoluciÃ³n      : O     O
:               : O     O
Con             : O     O
el              : O     O
juicio          : O     O
clÃnico        : O     O
de              : O     O
hepatitis       : O     O
alcohÃ³lica     : O     O
aguda           : O     O
sobre           : O     O
hepatopatÃa    : O     O
crÃ³nica        : O     O
con             : O     O
criterios       : O     O
de              : O     O
gravedad        : O     O
y               : O     O
Maddrey         : O     O
score           : O     O
de              : O     O
71              : O     O
ingresa         : O     O
en              : O     O
el              : O     O
servicio        : O     O
de              : O     O
digestivo       : O     O
.               : O     O
Se              : O     O
instaura        : O     O
un              : O     O
tratamiento     : O     O
con             : O     O
corticoides     : O     O
```

```
,              : O     O
nutriciÃ³n     : O     O
enteral        : O     O
y              : O     O
antibioterapia : O     O
empÃrica      : O     O
.              : O     O
Al             : O     O
quinto         : O     O
dÃa           : O     O
de             : O     O
estancia       : O     O
se             : O     O
observa        : O     O
una            : O     O
progresiÃ³n    : O     O
de             : O     O
la             : O     O
encefalopatÃa : O     O
hepÃ¡tica      : O     O
hasta          : O     O
grado          : O     O
III            : O     O
y              : O     O
fracaso        : O     O
renal          : O     O
agudo          : O     O
en             : O     O
el             : O     O
contexto       : O     O
de             : O     O
un             : O     O
sÃndrome      : O     O
hepatorrenal   : O     O
tipo           : O     O
I              : O     O
.              : O     O
En             : O     O
la             : O     O
analÃtica     : O     O
se             : O     O
aprecia        : O     O
un             : O     O
empeoramiento  : O     O
de             : O     O
la             : O     O
```

```
funciÃ³n       : O     O
renal         : O     O
con           : O     O
valores       : O     O
de            : O     O
creatinina    : O     O
plasmÃ¡tica   : O     O
de            : O     O
4             : O     O
,             : O     O
19            : O     O
mg            : O     O
/             : O     O
dl            : O     O
;             : O     O
urea          : O     O
140           : O     O
mg            : O     O
/             : O     O
dl            : O     O
y             : O     O
bilirrubina   : O     O
total         : O     O
de            : O     O
35            : O     O
,             : O     O
3             : O     O
mg            : O     O
/             : O     O
dl            : O     O
,             : O     O
persistiendo  : O     O
los           : O     O
valores       : O     O
elevados      : O     O
de            : O     O
transaminasas : O     O
y             : O     O
la            : O     O
alteraciÃ³n   : O     O
en            : O     O
el            : O     O
tiempo        : O     O
de            : O     O
protromina    : O     O
.             : O     O
```

```
Se              : O     O
aÃ±ade          : O     O
al              : O     O
tratamiento     : O     O
pentoxifilina   : O     O
,               : O     O
(               : O     O
como            : O     O
terapia         : O     O
anti            : O     O
-               : O     O
TNF             : O     O
)               : O     O
,               : O     O
terlipresina    : O     O
y               : O     O
albÃ°mina       : O     O
,               : O     O
(               : O     O
como            : O     O
tratamiento     : O     O
para            : O     O
el              : O     O
SHR             : O     O
tipo            : O     O
I               : O     O
)               : O     O
y               : O     O
MARS            : O     O
,               : O     O
previa          : O     O
solicitud       : O     O
de              : O     O
consentimiento  : O     O
informado       : O     O
.               : O     O
Realizamos      : O     O
3               : O     O
sesiones        : O     O
,               : O     O
a               : O     O
dÃas           : O     O
alternos        : O     O
,               : O     O
con             : O     O
ingreso         : O     O
```

```
del             : O     O
paciente        : O     O
en              : O     O
la              : O     O
Unidad          : O     O
de              : O     O
Cuidados        : O     O
Intensivos      : O     O
(               : O     O
UCI             : O     O
)               : O     O
para            : O     O
cada            : O     O
sesiÃ³n         : O     O
.               : O     O
La              : O     O
duraciÃ³n       : O     O
aproximada      : O     O
de              : O     O
cada            : O     O
sesiÃ³n         : O     O
fue             : O     O
12              : O     O
horas           : O     O
o               : O     O
hasta           : O     O
la              : O     O
coagulaciÃ³n    : O     O
del             : O     O
filtro          : O     O
.               : O     O
Se              : O     O
utilizÃ³        : O     O
acceso          : O     O
venoso          : O     O
femoral         : O     O
,               : O     O
usado           : O     O
exclusivamente  : O     O
para            : O     O
este            : O     O
fin             : O     O
.               : O     O
La              : O     O
tÃ©cnica        : O     O
de              : O     O
```

```
reemplazo       : O     O
renal           : O     O
asociada        : O     O
fue             : O     O
HDF             : O     O
veno            : O     O
-               : O     O
venosa          : O     O
continua        : O     O
.               : O     O
La              : O     O
anticoagulaciÃ³n: O     O
fue             : O     O
variable        : O     O
con             : O     O
epoprostenol    : O     O
sÃ³dico         : O     O
5               : O     O
ng              : O     O
/               : O     O
kg              : O     O
/               : O     O
minuto          : O     O
,               : O     O
con             : O     O
heparina        : O     O
Na              : O     O
5U              : O     O
/               : O     O
kg              : O     O
/               : O     O
hora            : O     O
o               : O     O
con             : O     O
ambos           : O     O
.               : O     O
El              : O     O
enfermo         : O     O
estuvo          : O     O
hemodinÃ¡micamente: O     O
estable         : O     O
durante         : O     O
las             : O     O
sesiones        : O     O
,               : O     O
sin             : O     O
```

```
incidencias    : O     O
destacables    : O     O
.              : O     O
Remitido       : O     O
por            : O     O
:              : O     O
Jose           : B-NPS B-NPS
Miguel         : I-NPS I-NPS
Mora           : I-NPS I-NPS
OrdÃ³Ã±ez      : I-NPS I-NPS
.              : O     O
Avda           : B-CALLE B-CALLE
.              : I-CALLE I-CALLE
AndalucÃa     : I-CALLE I-CALLE
,              : I-CALLE I-CALLE
146            : I-CALLE I-CALLE
.              : I-CALLE I-CALLE
UrbanizaciÃ³n  : I-CALLE I-CALLE
Pinos          : I-CALLE I-CALLE
de             : I-CALLE I-CALLE
AlhaurÃn      : I-CALLE I-CALLE
.              : O     I-CALLE
AlhaurÃn      : B-TERR I-CALLE
de             : I-TERR I-CALLE
la             : I-TERR I-CALLE
Torre          : I-TERR I-CALLE
.              : O     I-CALLE
29130          : B-TERR I-CALLE
MÃ¡laga        : B-TERR B-TERR
.              : O     O
EspaÃ±a        : B-PAIS B-PAIS
.              : O     O
Correo         : O     O
electrÃ³nico   : O     O
:              : O     O
jumo111975     : B-CE  B-CE
@              : I-CE  I-CE
hotmail        : I-CE  I-CE
.              : I-CE  I-CE
com            : I-CE  I-CE
.              : I-CE  O
```

LSTM-CRF

A continuación se muestra la aplicación de un modelo utilizando LSTM-CRF entrenado según el caso 3 mostrado previamente.

```
Word            ||True ||Pred
==============================
Datos           : O     O
del             : O     O
paciente        : O     O
.               : O     O
Nombre          : O     O
:               : O     O
Esperanza       : B-NSA B-NSA
.               : O     O
Apellidos       : O     O
:               : O     O
Tuy             : B-NSA B-NSA
RIvera          : I-NSA O
.               : O     O
NHC             : O     O
:               : O     O
8349570         : B-ISA O
.               : O     O
NASS            : O     O
:               : O     O
84              : B-IA  B-NSA
56872485        : I-IA  O
67              : I-IA  I-IA
.               : O     O
Domicilio       : O     O
:               : O     O
Av              : B-CALLE B-CALLE
/               : I-CALLE I-CALLE
de              : I-CALLE I-CALLE
CÃ³rdoba        : I-CALLE I-CALLE
,               : I-CALLE I-CALLE
23              : I-CALLE I-CALLE
,               : I-CALLE I-CALLE
3               : I-CALLE I-CALLE
,               : I-CALLE I-CALLE
A               : I-CALLE I-CALLE
.               : O     O
Localidad       : O     O
/               : O     O
```

```
Provincia       : O     O
:               : O     O
Madrid          : B-TERR B-TERR
.               : O     O
CP              : O     O
:               : O     O
28041           : B-TERR B-TERR
.               : O     O
Datos           : O     O
asistenciales   : O     O
.               : O     O
Fecha           : O     O
de              : O     O
nacimiento      : O     O
:               : O     O
25              : B-F   B-F
/               : I-F   I-F
8               : I-F   I-F
/               : I-F   I-F
1932            : I-F   O
.               : O     O
PaÃs           : O     O
de              : O     O
nacimiento      : O     O
:               : O     O
EspaÃ±a         : B-PAIS B-PAIS
.               : O     O
Edad            : O     O
:               : O     O
81              : B-ESA B-ESA
aÃ±os           : I-ESA I-ESA
Sexo            : O     O
:               : O     O
M               : B-SSA B-SSA
.               : O     O
Fecha           : O     O
de              : O     O
Ingreso         : O     O
:               : O     O
30              : B-F   B-F
/               : I-F   I-F
11              : I-F   I-F
/               : I-F   I-F
2013            : I-F   I-F
.               : O     O
MÃ©dico         : O     O
```

```
:               : O     O
Mario           : B-NPS B-NPS
MartÃn         : I-NPS I-NPS
HernÃ¡ndez      : I-NPS I-NPS
NÂºCol          : O     O
:               : O     O
28              : B-ITPS B-ITPS
28              : I-ITPS I-ITPS
62253           : I-ITPS I-ESA
.               : O     O
Informe         : O     O
clÃnico        : O     O
del             : O     O
paciente        : O     O
:               : O     O
Mujer           : B-SSA B-SSA
de              : O     O
82              : B-ESA B-ESA
aÃ±os           : I-ESA I-ESA
con             : O     O
antecedentes    : O     O
de              : O     O
hipertensiÃ³n   : O     O
arterial        : O     O
,               : O     O
fibrilaciÃ³n    : O     O
auricular       : O     O
en              : O     O
tratamiento     : O     O
anticoagulante  : O     O
,               : O     O
diabetes        : O     O
mellitus        : O     O
e               : O     O
insuficiencia   : O     O
renal           : O     O
crÃ³nica        : O     O
.               : O     O
Ingresa         : O     O
por             : O     O
cuadro          : O     O
de              : O     O
ictericia       : O     O
obstructiva     : O     O
secundaria      : O     O
a               : O     O
```

coledocolitiasis: O O
. : O O
Se : O O
realiza : O O
CPRE : O O
terapÃ©utica : O O
con : O O
sedacciÃ³n : O O
consciente : O O
(: O O
midazolam : O O
- : O O
propofol : O O
y : O O
remifentanilo : O O
) : O O
, : O O
tÃ©cnicamente : O O
laboriosa : O O
, : O O
con : O O
una : O O
duraciÃ³n : O O
de : O O
150 : O O
minutos : O O
. : O O
De : O O
forma : O O
brusca : O O
al : O O
finalizar : O O
la : O O
intervenciÃ³n : O O
presenta : O O
un : O O
cuadro : O O
de : O O
inestabilidad : O O
hemodinÃ¡mica : O O
y : O O
respiratoria : O O
. : O O
Se : O O
realizÃ³ : O O
con : O O

```
carÃ¡cter      : O     O
de            : O     O
urgencia      : O     O
una           : O     O
tomografÃa    : O     O
axial         : O     O
computerizada : O     O
(             : O     O
TAC           : O     O
)             : O     O
abdominal     : O     O
con           : O     O
contraste     : O     O
intravenoso   : O     O
directo       : O     O
,             : O     O
en            : O     O
el            : O     O
que           : O     O
se            : O     O
observa       : O     O
una           : O     O
imagen        : O     O
hipervascular : O     O
en            : O     O
el            : O     O
polo          : O     O
anterior      : O     O
del           : O     O
bazo          : O     O
compatible    : O     O
con           : O     O
una           : O     O
zona          : O     O
de            : O     O
contusiÃ³n     : O     O
esplÃ©nica     : O     O
,             : O     O
con           : O     O
lÃquido       : O     O
libre         : O     O
intraabdominal : O     O
,             : O     O
fundamentalmente: O     O
en            : O     O
regiÃ³n        : O     O
```

```
periesplÃ©nica : O     O
,             : O     O
gotiera       : O     O
paracÃ³lica    : O     O
izquierda     : O     O
,             : O     O
perihepÃ¡tico  : O     O
y             : O     O
tambiÃ©n       : O     O
en            : O     O
pelvis        : O     O
con           : O     O
densidad      : O     O
intermedia    : O     O
compatible    : O     O
con           : O     O
sangre        : O     O
.             : O     O
La            : O     O
paciente      : O     O
se            : O     O
trasladÃ³      : O     O
a             : O     O
quirÃ³fano     : O     O
con           : O     O
deterioro     : O     O
del           : O     O
estado        : O     O
general       : O     O
.             : O     O
Se            : O     O
realizÃ³       : O     O
laparotomÃa   : O     O
urgente       : O     O
,             : O     O
evacuÃ¡ndose   : O     O
hemoperitoneo : O     O
de            : O     O
2000          : O     O
ml            : O     O
por           : O     O
desgarro      : O     O
en            : O     O
borde         : O     O
anterior      : O     O
esplÃ©nico     : O     O
```

```
en              : O     O
la              : O     O
zona            : O     O
visualizada     : O     O
por             : O     O
TAC             : O     O
.               : O     O
Se              : O     O
practica        : O     O
esplenectomÃa : O     O
,               : O     O
colecistectomÃa: O     O
,               : O     O
duodenostomÃa : O     O
y               : O     O
extracciÃ³n     : O     O
del             : O     O
molde           : O     O
biliar          : O     O
con             : O     O
posterior       : O     O
colocaciÃ³n     : O     O
de              : O     O
tubo            : O     O
de              : O     O
Kher            : O     O
.               : O     O
No              : O     O
evidencia       : O     O
de              : O     O
perforaciÃ³n    : O     O
duodenal        : O     O
.               : O     O
En              : O     O
postoperatorio : O     O
presentÃ³       : O     O
un              : O     O
cuadro          : O     O
de              : O     O
shock           : O     O
sÃ©ptico        : O     O
por             : O     O
peritonitis     : O     O
biliar          : O     O
,               : O     O
tratado         : O     O
```

```
empÃricamente : O     O
con            : O     O
piperacilina   : O     O
-              : O     O
tazobactam     : O     O
,              : O     O
y              : O     O
necesidad      : O     O
de             : O     O
relaparotomÃa : O     O
,              : O     O
hallÃ¡ndose    : O     O
fÃstula       : O     O
biliar         : O     O
por            : O     O
salida         : O     O
del            : O     O
tubo           : O     O
de             : O     O
Kher           : O     O
.              : O     O
Fue            : O     O
dada           : O     O
de             : O     O
alta           : O     O
de             : O     O
la             : O     O
Unidad         : O     O
de             : O     O
ReanimaciÃ³n   : O     O
a              : O     O
los            : O     O
12             : O     O
dÃas          : O     O
del            : O     O
ingreso        : O     O
.              : O     O
El             : O     O
informe        : O     O
de             : O     O
anatomÃa      : O     O
patolÃ³gica    : O     O
de             : O     O
la             : O     O
pieza          : O     O
de             : O     O
```

```
esplenectomÃa : O     O
objetivÃ³      : O     O
la            : O     O
existencia    : O     O
en            : O     O
su            : O     O
cara          : O     O
interna       : O     O
de            : O     O
una           : O     O
soluciÃ³n      : O     O
de            : O     O
continuidad   : O     O
de            : O     O
la            : O     O
cÃ¡psula      : O     O
.             : O     O
El            : O     O
resto         : O     O
del           : O     O
parÃ©nquima   : O     O
no            : O     O
presentaba    : O     O
alteraciones  : O     O
.             : O     O
Responsable   : O     O
clÃnico       : O     O
:             : O     O
Dr            : O     O
.             : O     O
Mario         : B-NPS B-NPS
MartÃn        : I-NPS I-NPS
HernÃ¡ndez    : I-NPS I-NPS
Correo        : O     O
electrÃ³nico  : O     O
:             : O     O
mariomhdez    : B-CE  B-CE
@             : I-CE  I-CE
yahoo         : I-CE  I-CE
.             : I-CE  I-CE
es            : I-CE  I-CE
```

7.4 - Resultados

Lo primero que es importante destacar entre la comparativa de ambos enfoques es que CRF es más sencillo de desarrollar y, debido al poco tiempo que lleva entrenar y de evaluar, el proceso de análisis es más rápido. En el caso de LFTM-CRF, es un algoritmo con más restricciones y propenso a tener errores. Por ejemplo, en esta implementación el conjunto de datos y test deben tener los mismos tags o se un subconjunto del conjunto de datos de Entrenamiento. Además, al guardar el modelo, también se debe guardar el indexado de las palabras y de los tags.

Teniendo como referencia la métrica F1-Score, se puede decir que el enfoque CRF tiene un mejor modelo resultante. Esto puede ser debido a la características del enfoque LSTM-CRF, ya que este requiere más reportes que CRF. Teniendo un conjunto de datos más grande se podría tener resultados distintos.

Se puede comprobar empíricamente que las métricas accuracy y weighted avg no son buen parámetro para comparar los modelos; ya que en todos los casos dio un resultado mayor a 90%. En el caso de accuracy, se alcanzaba el 90% en solo la época 8, cuando F1-Score era de solo 2%.

Para algunos tags ningún modelo es bueno prediciéndolos; esto es debido a los pocos casos de entrenamientos. Sin embargo, esto se podía pronosticar observando los reportes de entrenamiento.

A pesar de presentar una métrica de alrededor del 65% en f1-score para el modelo CRF, los ejemplos de utilización fueron buenos.

8 - Conclusión

- El sistema entrenado con CRF presentó mejores resultados que el entrenado con LFTM-CRF. Esto es debido a que se necesita mejorar la calidad del conjunto de entrenamiento; sin embargo, con mejores casos de entrenamiento el resultado puede ser distinto. Además, el desarrollo de CRF es sustancialmente más sencillo.
- Los modelos obtenidos pueden ser mejorados, esto se concluye a partir de los datos presentados. Por un lado, para aquellas entidades cuya performance fue baja, los resultados podrían mejorar si se incorporan nuevos ejemplos de entrenamiento.. Por otro lado, el vector de características puede ser extendido.
- Los tags NSA, CALLE, CE, ITPS, F, SSA, ESA, NPS presentaron resultados bastante satisfactorios en la métrica f1-score. Esto se podía predecir observando el conjunto de entrenamientos, debido a la gran cantidad de casos de entrenamiento. Por el contrario, los tags OSA, ICA, CS, INSTITUCION y PROF presentaron resultados muy bajos debido a los pocos casos de entrenamiento que habían. Para el resto de tags, los resultados fueron buenos.
- Para el algoritmo de LSTM, un entrenamiento con más de 500 épocas no presenta mejoras significativas. Es decir, que entrenarlo con 500 épocas es suficiente para obtener sus mejores resultados.
- Por un lado, teniendo en cuenta que el modelo CRF presentaba mejores resultados, la predicción hecha puede categorizarse como buena. Por otro lado, la predicción hecha por el modelo LSTM-CRF presenta una notable baja en el rendimiento, haciendo su uso poco útil.
- Las lecciones aprendidas en este proyecto serán útiles para desarrollar métodos que servirán para entrenar modelos basados en otro tipo de reportes médicos

9 - Anexo

9.1 - CRF

9.1.2 - Matriz de Confusión

El siguiente modelo fue entrenado con CRF bajo las siguientes características:

```
Algorithm = 'lbfgs'
c1 = 0.4
c2 = 0.4
max_iterations = 100
all_possible_transitions = false
```

```
[[[ 13579    830]
  [   448 114850]]

 [[129178     51]
  [    26    452]]

 [[129427     20]
  [    15    245]]

 [[129343     90]
  [    48    226]]

 [[129491     33]
  [    19    164]]

 [[129271     66]
  [    38    332]]

 [[129284     16]
  [    45    362]]

 [[127091    212]
  [   154   2250]]

 [[128731     66]
  [   209    701]]

 [[129113     22]
  [    38    534]]
```

```
[[127568     56]
 [    52   2031]]

[[129356     15]
 [    32    304]]

[[129179     34]
 [    29    465]]

[[129195     64]
 [    31    417]]

[[129256     23]
 [    14    414]]

[[129197     30]
 [    44    436]]

[[128515     74]
 [    96   1022]]

[[129460     19]
 [    23    205]]

[[129227     38]
 [    44    398]]

[[129458     12]
 [    16    221]]

[[128529     68]
 [     9   1101]]

[[129583      3]
 [    18    103]]

[[129267     29]
 [    69    342]]

[[129553     10]
 [    91     53]]

[[129705      0]
 [     2      0]]

[[129636     12]
```

```
 [    31     28]]

[[129642     14]
 [    43      8]]

[[129543     43]
 [   105     16]]

[[129673      2]
 [    13     19]]

[[129642      6]
 [    24     35]]

[[129666      4]
 [    34      3]]

[[129671      0]
 [    36      0]]

[[129701      0]
 [     5      1]]

[[129700      0]
 [     5      2]]

[[129684      2]
 [    21      0]]

[[129702      0]
 [     2      3]]

[[129703      0]
 [     4      0]]

[[129702      0]
 [     5      0]]

[[129681      0]
 [    26      0]]

[[129707      0]
 [     0      0]]

[[129707      0]
 [     0      0]]
```

```
 [[129707      0]
  [     0      0]]

 [[129707      0]
  [     0      0]]

 [[129707      0]
  [     0      0]]]
```

	precision	recall	f1-score	support
B-CALLE	0.96	0.89	0.92	407
B-CE	0.95	0.93	0.94	237
B-CS	0.00	0.00	0.00	5
B-ESA	0.93	0.94	0.94	494
B-F	0.96	0.93	0.95	572
B-FSA	0.70	0.47	0.57	59
B-HOSP	0.97	0.85	0.91	121
B-IA	0.83	0.90	0.86	183
B-ICA	0.00	0.00	0.00	36
B-INSTITUCION	0.36	0.16	0.22	51
B-ISA	0.72	0.82	0.77	274
B-ITPS	0.92	0.90	0.91	228
B-NPS	0.94	0.91	0.92	480
B-NSA	0.90	0.95	0.92	478
B-NT	0.90	0.59	0.72	32
B-OSA	0.00	0.00	0.00	2
B-PAIS	0.95	0.90	0.93	336
B-PROF	1.00	0.17	0.29	6
B-SSA	0.95	0.97	0.96	428
B-TERR	0.91	0.77	0.84	910
I-CALLE	0.91	0.94	0.92	2404
I-CE	0.94	0.99	0.97	1110
I-CS	0.00	0.00	0.00	26
I-ESA	0.87	0.93	0.90	448
I-F	0.97	0.98	0.97	2083
I-FSA	0.43	0.08	0.14	37
I-HOSP	0.92	0.83	0.87	411
I-IA	0.83	0.90	0.86	370
I-INSTITUCION	0.27	0.13	0.18	121
I-ISA	0.00	0.00	0.00	21
I-ITPS	0.91	0.90	0.91	442
I-NPS	0.93	0.91	0.92	1118
I-NSA	0.92	0.94	0.93	260
I-NT	0.85	0.59	0.70	59

I-OSA	0.00	0.00	0.00	4
I-PAIS	1.00	0.60	0.75	5
I-PROF	1.00	0.29	0.44	7
I-TERR	0.84	0.37	0.51	144
O	0.99	1.00	0.99	115298
accuracy			0.98	129707
macro avg	0.73	0.63	0.65	129707
weighted avg	0.98	0.98	0.98	129707

El siguiente modelo fue entrenado con LSTM-CRF bajo el caso 3 presentado en la sección de entrenamiento de dicho algoritmo:

	precision	recall	f1-score	support
B-CALLE	0.88	0.83	0.86	386
B-CE	0.68	0.75	0.71	237
B-CS	0.00	0.00	0.00	5
B-ESA	0.83	0.91	0.87	458
B-F	0.94	0.91	0.93	533
B-FSA	0.43	0.44	0.43	59
B-HOSP	0.85	0.73	0.79	120
B-IA	0.93	0.70	0.80	165
B-ICA	0.00	0.00	0.00	31
B-INSTITUCION	0.28	0.10	0.15	50
B-ISA	0.50	0.01	0.02	246
B-ITPS	0.97	0.76	0.85	216
B-NPS	0.93	0.80	0.86	464
B-NSA	0.86	0.81	0.83	429
B-NT	0.69	0.28	0.40	32
B-OSA	0.00	0.00	0.00	2
B-PAIS	0.94	0.86	0.90	319
B-PROF	0.17	0.20	0.18	5
B-SSA	0.96	0.94	0.95	399
B-TERR	0.87	0.66	0.75	870
I-CALLE	0.95	0.79	0.86	2274
I-CE	0.89	0.95	0.92	1110
I-CS	0.00	0.00	0.00	26
I-ESA	0.65	0.88	0.75	415
I-F	0.98	0.96	0.97	1936
I-FSA	0.00	0.00	0.00	37
I-HOSP	0.84	0.66	0.74	405
I-IA	0.93	0.77	0.85	335
I-INSTITUCION	0.21	0.08	0.11	119
I-ISA	0.00	0.00	0.00	18

```
       I-ITPS       0.94      0.53      0.67       418
        I-NPS       0.94      0.65      0.77      1081
        I-NSA       0.83      0.81      0.82       234
         I-NT       0.83      0.34      0.48        59
        I-OSA       0.00      0.00      0.00         4
       I-PAIS       0.00      0.00      0.00         4
       I-PROF       0.40      0.40      0.40         5
       I-TERR       0.61      0.20      0.30       141
            O       0.99      1.00      0.99    179153

    accuracy                           0.98    192800
   macro avg       0.61      0.51      0.54    192800
weighted avg       0.98      0.98      0.98    192800

[[[ 11451   2196]
  [   322 178831]]

 [[ 11451   2196]
  [   322 178831]]

 [[ 11451   2196]
  [   322 178831]]

 ...

 [[191558    132]
  [    51   1059]]

 [[191558    132]
  [    51   1059]]

 [[191558    132]
  [    51   1059]]]
```

10 - Bibliografía

[1] Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal Data Mining Practical Machine Learning Tools and Techniques, Fourth Edition.
[2] Meddocan, www.temu.bsc.es/meddocan/, 17 de febrero 2020.
[3] Daniel Jurafsky, James H. Martin, Speech and Language Processing, tercera edición.
[4] Hanna M. Wallach, Conditional Random Fields: An Introduction, 2004.
[5] Mikhail Korobov, www.sklearn-crfsuite.readthedocs.io, 17 de febrero 2020
[6] Zhiheng Huang, Wei Xu, Kai Yu Bidirectional LSTM-CRF Models for Sequence Tagging
[7] Satnam Alag, Collective Inteligence in Action, edición 2009.
[8] Violeta Valcárcel Asencios, DATA MININGYEL DESCUBRIMIENTO DEL CONOCIMIENTO
[9] Shai Shalev-Shwartz, Shai Ben-David, Understanding Machine Learning: From Theory to Algorithms.
[10] Nadeau, D., Sekine, S (2007). A survey of named entity recognition and classification. Linguisticae Investigationes 30(1):3–26
[11] Luole Qi and Li Chen, A Linear-Chain CRF-Based Learning Approach for Web Opinion Mining.