Tarea 1: Detección de Entidades Nombradas

Sergio Soler Rocha
Universidad Nacional de Educación a Distancia

Resumen

El objetivo de esta práctica es utilizar un etiquetador de entidades nombradas, viendo la utilidad de las herramientas para procesamiento del lenguaje aplicadas a un problema planteado en una competición científica, y analizar los resultados, estudiando las causas de los casos de error y cómo pueden mejorarse

1. CoNLL-2002

1.1. Objetivo de la competición

En el año 2002, se propuso la tarea del *reconocimiento de entidades nombradas*, es decir, se planteaba la idea de identificar palabras o frases en un texto dado que contuvieran nombres propios de personas, organizaciones, localizaciones, fechas y cantidades. Como ejemplo tenemos:

> [PER Clement Lenglet] está viviendo una extraña situación en el [ORG Aston Villa]. El club de [LOC Reino Unido]...

La frase contiene tres entidades, *Clement Lenglet* como persona, *Aston Villa* como organización, y *Reino Unido* como localización. Nos concentraremos en cuatro tipos de entidades nombradas: personas, ubicaciones, organizaciones y nombres de entidades diversas que no pertenece a los tres grupos anteriores. El formato IOB es el siguiente:

- PER: Persona
- ORG: Organización
- LOC: Localización
- MISC: Miscelánea

Se proporcionaron conjuntos de datos de entrenamiento y prueba en los idiomas español y holandés, con el objetivo de crear modelos de aprendizaje automático supervisado y comparar los resultados obtenidos.

1.2. Los Datos

Los datos constan de tres diferentes archivos para cada idioma, un archivo de entrenamiento, un archivo de desarrollo y otro archivo de prueba. El archivo de desarrollo se usa para ajustar los diferentes parámetros del método elegido. Una vez encontrado los mejores parámetros, éstos son usados para entrenar los datos de entrenamiento y testearlos con los datos de prueba. Los datos de prueba son separados con anterioridad de los demás datos para evitar que los parámetros sean ajustados usando éstos.

Los datos contienen una palabra por línea, con líneas vacías que representan los límites de las oraciones. Cada línea contiene una etiqueta que indica si la palabra está dentro de una entidad con nombre o no. La etiqueta también codifica el tipo de entidad. Por ejemplo:

- Clement B-PER
- Lenglet I-PER

- está O
- viviendo O
- una O
- extraña O
- situación O
- en O
- el O
- Aston B-ORG
- Villa I-ORG
- . O
- El O
- club O
- de O
- Reino B-LOC
- Unido I-LOC

La letra O significa que no se trata de ninguna entidad (outside of named entities), la letra B significa que inicia la entidad nombrada (beginning of named entities), mientras que la letra I (inside of named entities) simboliza todas las palabras que pertenecen a la misma entidad que siguen a la B.

Los datos en español, en los que nos vamos a centrar de ahora en adelante, son una colección de artículos informativos puestos a disposición por la Agencia de Noticias española EFE. Los artículos son de mayo de 2000. La anotación fue realizada por el Centro de Investigación TALP2 de la Universitat Politècnica de Catalunya (UPC) y el Centro de Lenguaje y Computación (CLiC3) de la Universidad de Barcelona (UB). Los datos contienen únicamente palabras y etiquetas de entidad. Los archivos de datos de entrenamiento, desarrollo y prueba contienen 273,037, 54,837 y 53,049 líneas respectivamente.

El rendimiento de la tarea se mide con $F_{\beta=1}$, que es igual a:

$$F_\beta = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall} \tag{1}$$

Precisión es el porcentaje de entidades nombradas encontradas por el sistema de aprendizaje que son correctas. *Recall* es el porcentaje de entidades nombradas presentes en el corpus que encuentra el sistema. Una entidad nombrada es considerada correcta solo si coincide exactamente con la entidad correspondiente en el archivo de datos.

1.3. Los Resultados de la competición

A la hora de abordar el problema de clasificación propuesto, se emplearon diferentes tipos de algoritmos, entre ellos, support vector machines, árboles de decisión, clasificadores AdaBoost, modelos de máxima entropía y métodos de Markov de primer orden [1]. El modelo que obtuvo los mejores resultados fue desarrollado por Carreras, Márquez y Padrón. Abordaron la tarea compartida utilizando AdaBoost aplicado a árboles de decisión de profundidad fija. Su sistema utilizó diversas características de entrada, información contextual, pistas internas de palabras, clases de entidades anteriores, etiquetas de partes del discurso (solo en holandés) y listas de palabras externas (solo en español). Procesaron los datos en dos etapas: primero el reconocimiento de entidades y luego la clasificación. Su sistema logró los mejores resultados en esta tarea compartida para los conjuntos de datos de prueba en español y holandés (conjunto de pruebas en español: $F_1 = 81{,}39$; conjunto de pruebas en holandés: $F_1 = 77{,}05$).

2. Detección de Entidades Nombradas con SpaCy

El objetivo de la tarea es replicar los objetivos de la competición mencionada en la sección anterior, pero en nuestro caso, lo haremos únicamente con los datos en español del conjunto de prueba. Esto se debe a que ya contamos con un modelo preentrenado de SpaCy para la detección de entidades nombradas.

SpaCy es una biblioteca de procesamiento de lenguaje natural (NLP) de código abierto y alta eficiencia diseñada para realizar tareas relacionadas con el procesamiento del texto de manera rápida y precisa. Algunas de las características clave de SpaCy incluyen:

1. Análisis lingüístico: SpaCy ofrece análisis lingüístico detallado, que incluye tokenización, lematización, análisis morfosintáctico y reconocimiento de entidades nombradas. Esto permite descomponer el texto en sus componentes lingüísticos básicos y extraer información sobre las palabras y frases en un texto.

2. Modelos preentrenados: SpaCy proporciona modelos preentrenados para varios idiomas, lo que facilita el procesamiento de texto en diferentes lenguajes y la realización de tareas NLP sin necesidad de entrenar modelos desde cero.

3. Edad del BronceEficiencia: SpaCy se destaca por su velocidad y eficiencia, lo que lo hace adecuado para el procesamiento de grandes cantidades de texto. Es ampliamente utilizado en aplicaciones que requieren procesamiento de texto a gran escala.

4. Extensibilidad: SpaCy es altamente extensible y permite a los usuarios personalizar sus modelos y añadir reglas específicas para adaptarse a necesidades particulares.

5. Soporte para tareas NLP: SpaCy es útil para realizar una variedad de tareas de procesamiento de lenguaje natural, como análisis de sentimiento, extracción de información, resumen de texto, traducción automática y más.

2.1. Descripción del código desarrollado

Lo primero que hacemos es importar la librería spacy y crear un objeto *nlp* en español, tal y como se ve en la Figura 1:

```
import spacy

nlp = spacy.load("es_core_news_sm")
```

Figura 1: Importación de y creación del objeto nlp.

El segundo paso consiste en cargar los datos de prueba. En este caso, están en formato de texto (*esp.testb.txt*), donde cada línea contiene una palabra seguida de su etiqueta, tal y como se ha visto en el apartado anterior. Ahora creamos una lista llamada *tokens*, que contiene en primer lugar la palabra y en segundo lugar su etiqueta, tal y como se muestra en la Figura 2

```
# Cargamos los datos de prueba
with open ("data/esp.testb.txt", "r", encoding='utf-8') as f:
    lines = f.readlines()

# Separamos de las palabras de las palabras de sus etiquetas
tokens = []

for linea in lines:
    partes = linea.strip().split()
    if len(partes) == 2:
        palabra, etiqueta = partes
        tokens.append((palabra, etiqueta))
```

Figura 2: Cargar los datos y dividir las palabras y sus respectivas etiquetas

Una vez separadas las palabras y etiquetas necesitamos sacar el texto plano para ser pasado al objeto *nlp*, para eso usamos función Doc de *spacy*, tal y como se ve en la Figura 3

```
# creamos un doc con spacy a partir de las palabras ya separadas
text = spacy.tokens.Doc(nlp.vocab, words=[token[0] for token in tokens])

# pasamos la funcion nlp al texto base
doc = nlp(text)

print(text[0:20])
```

```
La Coruña , 23 may ( EFECOM ) . - Las reservas " on line " de billetes aéreos a
```

Figura 3: Transformar a texto plano y pasarlo por nlp

Al crear *doc*, el texto ya ha sido tokenizado y etiquetado automáticamente por *spaCy*. Antes de comparar los datos de prueba con los datos de predicción, nos aseguramos de que los textos sean idénticos. Para ello, creamos una función llamada *are_identical*, como se muestra en la Figura 4.

```
def are_identical(doc, tokens):
    """
    Devuelve True si ambos textos son idénticos
    """
    i = 0
    add = 0
    for word in doc:
        if word.text == tokens[i][0]:
            add += 1

        i += 1


    if add == len(doc):
        return True
    else:
        return False
```

```
are_identical(doc, tokens)
```

```
True
```

Figura 4: El texto de los datos de prueba y de los datos a predecir son idénticos

Una vez que estamos seguros de que ambos textos son idénticos, procedemos a visualizar el etiquetado de entidades nombradas realizado por *spaCy*. Para esto, utilizamos los métodos *ent_iob*, que nos devuelve la entidad en formato IOB, y también el método *ent_type*, que nos devuelve el tipo de entidad, como se aprecia en la Figura 5.

```
for token in doc:
    print(token.text ,token.ent_iob_+"-"+token.ent_type_)
```

```
La O-
Coruña B-MISC
, O-
23 O-
may O-
( O-
EFECOM B-ORG
) O-
```

Figura 5: Etiquetado de las entidades nombradas hecho por SpaCy

Ahora procedemos a crear una lista con las etiquetas de los datos de prueba a la que llamaremos *true_seqs*. De la misma forma, creamos otra lista con las etiquetas predichas por el modelo de *spaCy* a la que llamaremos *pred_seqs*, como se muestra en la Figura 6.

```
true_seqs =[]
for token in tokens:
    true_seqs.append(token[1])

pred_seqs = []
for token in doc:
    pred_seqs.append(token.ent_iob_+"-"+token.ent_type_)
```

Figura 6: Lista de los valores verdaderos y predichos respectivamente

2.2. Resultados de evaluación del etiquetado.

Una vez que hemos creado las respectivas listas de los datos de prueba y los datos predichos, procederemos a compararlos calculando la F_1 (como en la competición). Para ello, utilizaremos el script *conlleval.py* ya creado con ese propósito, Figura 7

```
conlleval.evaluate(true_seqs, pred_seqs)
```

```
processed 51533 tokens with 3558 phrases; found: 4172 phrases; correct: 2017.
accuracy:  55.60%; (non-O)
accuracy:   6.67%; precision:  48.35%; recall:  56.69%; FB1:  52.19
              LOC: precision:  51.63%; recall:  70.11%; FB1:  59.47  1472
             MISC: precision:   8.45%; recall:  23.30%; FB1:  12.40  935
              ORG: precision:  76.07%; recall:  43.36%; FB1:  55.23  798
              PER: precision:  59.05%; recall:  77.69%; FB1:  67.10  967

(48.34611697027804, 56.68915120854413, 52.1862871927555)
```

Figura 7: Evaluación de los resultados

Observamos que obtenemos un promedio de $F_1 = 52{,}19$, lo que sugiere que hay margen para mejorar en la tarea de reconocimiento de entidades, especialmente en la categoría MISC, donde se obtiene una puntuación de $F_1 = 12{,}4$. Una de las posibles causas podría ser el desequilibrio en los datos de entrenamiento utilizados para el modelo de SpaCy; es probable que haya menos ejemplos de la clase MISC en comparación con las otras clases. Otra posible razón del bajo rendimiento al identificar palabras con la etiqueta MISC es la naturaleza misma de dicha categoría. MISC se utiliza para agrupar entidades que no encajan claramente en las categorías principales. Puede abarcar una variedad de elementos como fechas, eventos, artefactos, términos especializados, entre otros. Esta categoría suele ser empleada cuando el modelo no puede asignar la entidad a una categoría específica o cuando la entidad es diversa y no se ajusta a las etiquetas predefinidas. La amplitud y diversidad de su definición dificultan su categorización. La amplia gama de elementos que abarca MISC aumenta la complejidad de su identificación. Dado que la precisión más baja se observa en 'MISC', analizamos los falsos positivos (cuando el modelo predice 'MISC' pero corresponde a otra entidad). En este sentido, los errores más recurrentes incluyen la combinación de artículo determinado con un sustantivo común, como por ejemplo: 'El plan de prevención', 'el director general', 'La periodista Pilar Blanco', 'los parlamentarios', entre otros. Estos ejemplos evidencian la dificultad que enfrenta el modelo al categorizar la entidad 'MISC'.Al revisar los falsos negativos, notamos que con facilidad algunas entidades catalogadas como MISC pueden confundirse con otro tipo de entidad. Por ejemplo, cuando aparecen palabras como 'España' u otras localizaciones, tiende a clasificarlas como 'LOC' sin prestar suficiente atención al contexto. En este caso, la entidad completa 'Abierto de España' debería ser etiquetada como 'MISC'. Similarmente, con 'PER', se presenta confusión, como por ejemplo, la palabra 'Goya' es erróneamente interpretada como el pintor, cuando en realidad se refiere al premio de cine, que corresponde a la categoría 'MISC'. En el caso de 'ORG', se observa una situación similar, como en el caso de 'Junta de Acreedores', clasificado erróneamente como una organización. Es importante mencionar que en particular en este aspecto, a veces es difícil determinar si una entidad es una organización o no sin un contexto adecuado o la información necesaria.

La primera idea para mejorar su rendimiento sería con un entrenamiento adicional, ampliando el conjunto de datos. A partir de ahí, se podría también realizar un ajuste de hiperparámetros como la tasa de aprendizaje, el número de iteraciones y otros parámetros del entrenamiento, para ver cómo afectan al rendimiento del modelo.

La segunda idea es probar con un modelo más grande. El modelo 'es_core_news_sm' es un modelo pequeño y rápido, pero puede que no sea lo suficientemente robusto para todas las tareas. Intenta utilizar un modelo más grande, como 'es_core_news_md' o 'es_core_news_lg'. Al usar ambos modelos obtenemos las siguientes evaluaciones en las

```
processed 51533 tokens with 3558 phrases; found: 3932 phrases; correct: 2218.
accuracy:  63.40%; (non-O)
accuracy:   7.60%; precision:  56.41%; recall:  62.34%; FB1:  59.23
              LOC: precision:  55.68%; recall:  76.01%; FB1:  64.27  1480
             MISC: precision:  13.01%; recall:  30.09%; FB1:  18.17  784
              ORG: precision:  75.68%; recall:  49.36%; FB1:  59.75  913
              PER: precision:  79.60%; recall:  81.77%; FB1:  80.67  755

(56.408952187182095, 62.338392355255756, 59.22563417890521)
```

Figura 8: Evaluación de los resultados con el modelo 'es_core_news_md'

```
processed 51533 tokens with 3558 phrases; found: 3845 phrases; correct: 2405.
accuracy:  68.02%; (non-O)
accuracy:   8.15%; precision:  62.55%; recall:  67.59%; FB1:  64.97
              LOC: precision:  61.83%; recall:  78.60%; FB1:  69.21  1378
             MISC: precision:  14.95%; recall:  29.50%; FB1:  19.84  669
              ORG: precision:  78.23%; recall:  60.07%; FB1:  67.96  1075
              PER: precision:  84.65%; recall:  83.27%; FB1:  83.95  723

(62.54876462938882, 67.59415401911187, 64.97365932729976)
```

Figura 9: Evaluación de los resultados con el modelo 'es_core_news_lg'

Observamos una clara mejora en ambos modelos con respecto al modelo 'es_core_news_sm', con un aumento notable en todas las métricas de precisión y recall (y por ende, en F_1). Alcanzamos valores de $F_1 = 59{,}23$ y $F_1 = 64{,}97$ en los modelos 'es_core_news_md' y 'es_core_news_lg', respectivamente. Sin embargo, a pesar de esta mejora, aún se observa un rendimiento deficiente al evaluar la clase MISC, obteniendo valores de $F_1 = 18{,}17$ y $F_1 = 19{,}84$ en los modelos 'es_core_news_md' y 'es_core_news_lg', respectivamente.

3. Parte Opcional

Para realizar la anotación manual, hemos seleccionado fragmentos del artículo de Wikipedia sobre la banda Iron Maiden. Utilizamos el formato IOB para etiquetar las entidades nombradas, anotando cada palabra en el archivo de texto adjunto con la tarea denominada 'iron_maiden.text'.

El texto evaluado es el siguiente:

Iron Maiden es una banda británica de heavy metal fundada en 1975 por el bajista Steve Harris. Considerada una de las agrupaciones más importantes y representativas del género, han vendido más de 100 millones de discos en todo el mundo, a pesar de haber contado con poco apoyo de los medios masivos durante la mayor parte de su carrera.

La banda ha basado su éxito en llegar directamente a los aficionados, grabando discos de alta calidad y realizando destacadas actuaciones en vivo.

La agrupación ha obtenido diversos reconocimientos a lo largo de su carrera, como el Premio Ivor Novello en la categoría de «Logro Internacional» en 2002. En 2005 fueron incluidos en el Hollywood's RockWalk en Sunset Boulevard, Los Ángeles. En 2009 fue ganadora del premio «Mejor Performance en Vivo» en los BRIT Awards, el premio musical más importante del Reino Unido. En el año 2011 también obtuvieron un Grammy, en la categoría de «Mejor interpretación de Metal», por la canción «El Dorado». Además, ha ganado el premio de mejor banda metal británica del año en varias ocasiones, en los Metal Hammer Golden Gods Awards, entre otros reconocimientos.

En 2023, fueron incluidos por la Royal Mail junto a un selecto grupo de bandas británicas, catalogadas como las más influyentes de todos los tiempos, y que finalmente quedó conformado por The Beatles, Pink Floyd, Queen, The Rolling Stones y Iron Maiden.

Durante sus más de 45 años de trayectoria, Iron Maiden ha sido identificada gráficamente por su famosa mascota «Eddie the Head», un personaje antropomórfico que ha aparecido en la gran mayoría de las portadas de sus álbumes y sencillos, así como en sus presentaciones en vivo.

Tras varias audiciones y cambios en su formación, ésta finalmente se consolidó con el vocalista Paul Di'Anno, los guitarristas Dave Murray y Dennis Stratton y el batería Clive Burr, siempre bajo el liderazgo del bajista y principal compositor Steve Harris. Luego de muchas giras por todo el Reino Unido, en 1979 la banda publicó un EP llamado The Soundhouse Tapes, y en 1980, su álbum debut homónimo, el cual llegó al cuarto puesto de las listas británicas, sin mediar promoción masiva alguna. Ese mismo año, Stratton fue reemplazado por el guitarrista Adrian Smith, con quien publicaron el álbum Killers (1981). Luego, y tras la salida de Di Anno, ese mismo año, el cantante Bruce Dickinson entró para ocupar el puesto de vocalista para el álbum The Number of the Beast de 1982, el cual llegó al número uno de las listas británicas, marcando el inicio de una serie de lanzamientos de impacto. Para el año 1983 la banda lanzó el álbum Piece of Mind, que contaba como novedad con la salida del batería Clive Burr y la entrada de Nicko McBrain en su reemplazo. A partir de allí, se consolidó la alineación más exitosa que ha tenido la agrupación, la cual ha realizado numerosas giras y álbumes. Iron Maiden ha grabado 17 álbumes de estudio y es considerada una de las bandas más influyentes no solo para el metal y sus respectivos subgéneros, sino también para diversas agrupaciones de rock, e incluso artistas de otros estilos.

La historia de Iron Maiden parte en el año 1971, cuando Steve Harris, inspirado en bandas como The Who, Thin Lizzy, UFO, Black Sabbath, Jethro Tull, Genesis, King Crimson, Led Zeppelin, y Deep Purple, entre otras, adquirió la imitación de un bajo Fender Precision Bass por unas 40 libras esterlinas, y tras dejar atrás la opción de la batería, para la cual no contaba con el espacio suficiente. Inicialmente Harris también tuvo la ilusión de ser jugador de fútbol del West Ham, sin embargo, comenzó a dedicar todos sus esfuerzos a su otra gran pasión, la música. Esto condujo a la formación de una agrupación musical que llamó Gypsy's Kiss en 1972, cuyo primer concierto fue en el mítico local Cart & Horses en Maryland Point, Stratford.

La consagración definitiva de Iron Maiden alrededor del mundo llegó con su tercer álbum de estudio, The Number of the Beast, publicado el 29 de marzo de 1982. La gira promocional del disco fue titulada The Beast On The Road, comenzando en Inglaterra para culminar diez meses más tarde en Japón, siendo su segunda visita a este país. Con el sencillo «Run to the Hills», Iron Maiden llegó hasta el número 7 en el Top 40 británico. Pero fue en plena gira, y mientras su autobús se quedaba varado en la carretera, que la banda se enteró de que el álbum había pasado a encabezar las listas británicas.

Bruce propuso también el regreso de Adrian Smith, quien en ese momento era uno de los guitarristas de su banda solista y con el que grabó los exitosos álbumes Accident of Birth (1997) y Chemical Wedding (1998). La banda decidió además conservar a Janick Gers en su formación, creando así un trío en las guitarras. Ese mismo año la banda lanzó su videojuego Ed Hunter, y como promoción, se realizó una gira de impacto celebrando la vuelta de Bruce y Adrian titulada The Ed Hunter Tour, con alrededor de 30 fechas en Norteamérica y Europa. Una vez culminada la gira, el grupo se volcó de lleno en la preparación del que sería su duodécimo disco, titulado Brave New World. El trabajo, publicado el 29 de mayo del año 2000, alcanzó la séptima posición en las listas británicas y la primera en varios países del mundo, retomando además el sonido característico de la agrupación.

En febrero de 2016, la banda se embarcó en la gira The Book of Souls World Tour, tocando en 35 países de América del Norte y del Sur, Asia, Australasia, África y Europa, incluyendo sus primeras presentaciones en China, El Salvador y Lituania y su esperado regreso a Costa Rica en el Estadio Ricardo Saprissa. Al igual que en las giras Somewhere Back in Time de 2008-09 y The Final Frontier de 2010-11, el grupo viajó en un avión a medida, pilotado por Dickinson y apodado Ed Force One, aunque utilizaron un avión jumbo Boeing 747-400. La banda completó la gira en 2017 con otros espectáculos europeos y norteamericanos.

Repitiendo el proceso del apartado 2 obtenemos los siguientes resultados:

```
processed 1151 tokens with 95 phrases; found: 120 phrases; correct: 73.
accuracy:  80.39%; (non-O)
accuracy:  14.25%; precision:  60.83%; recall:  76.84%; FB1:  67.91
              LOC: precision:  90.00%; recall:  90.00%; FB1:  90.00  20
             MISC: precision:  39.53%; recall:  60.71%; FB1:  47.89  43
              ORG: precision:  70.37%; recall:  70.37%; FB1:  70.37  27
              PER: precision:  63.33%; recall:  95.00%; FB1:  76.00  30

(60.83333333333333, 76.84210526315789, 67.90697674418604)
```

Figura 10: Evaluación del texto con el modelo 'es_core_news_sm'

Observamos que los resultados de la evaluación utilizando el modelo preentrenado 'es_core _news_sm' muestran mejoras respecto a la evaluación anterior, con un $F_1 = 67{,}91\,\%$. Aunque el rendimiento más bajo sigue siendo el de la precisión en 'MISC', ahora mejora significativamente al 39,53 %. Las posibles causas de este mejor rendimiento podrían ser, por un lado, que el texto evaluado es aproximadamente 45 veces más pequeño que el del apartado anterior y, por lo tanto, podría no ser aún lo suficientemente representativo. Por otro lado, es posible que este texto seleccionado sea más fácil de interpretar que el anterior, ya que observamos que muchas entidades pueden ser fácilmente confundidas.

Finalmente estudiamos como es el rendimiento en el modelo que mejores resultados tuvo al analizar el texto anterior, 'es_core _news_lg'.

```
processed 1151 tokens with 96 phrases; found: 106 phrases; correct: 83.
accuracy:  89.76%; (non-O)
accuracy:  15.99%; precision:  78.30%; recall:  86.46%; FB1:  82.18
              LOC: precision:  90.91%; recall:  95.24%; FB1:  93.02  22
             MISC: precision:  57.89%; recall:  78.57%; FB1:  66.67  38
              ORG: precision:  87.50%; recall:  77.78%; FB1:  82.35  24
              PER: precision:  90.91%; recall: 100.00%; FB1:  95.24  22

(78.30188679245283, 86.45833333333334, 82.17821782178217)
```

Figura 11: Evaluación del texto con el modelo 'es_core_news_lg'

El modelo muestra en general muy buenos resultados, siendo $F_1 = 82{,}18$. A continuación mostramos los errores cometidos por este ultimo modelo.

Los falsos positivos de 'LOC' son: *Hollywood's, RockWalk, el*, el artículo iría con *Estadio Ricardo Saprissa*, que son errores comprensibles.

Los falsos positivos de 'MISC' son: *La, banda, La, agrupación, Ese, mismo, año, Para, el, año, 1983, La, historia, de, Iron, Maiden, La, consagración, definitiva, de, La, gira, promocional, del, disco, La, banda, Ese, mismo, año, la, banda, Una, vez, culminada, la, gira, El, trabajo, jumbo, La y banda.* La mayoría de ellos tienen la estructura ya comentada articulo + sustantivo común.

Los falsos positivos de 'ORG' son: *Logro, Internacional y Grammy* .

Los falsos positivos de 'PER' son: *El, Dorado, Jethro y Tull.*

Los falsos negativos de 'MISC' son: *Logro (B-ORG), Internacional (I-ORG), Grammy (B-ORG), interpretación (O), de (O), El (B-PER), Dorado (I-PER), de (O),*

Los falsos negativos de 'ORG' son: *Hollywood's (B-LOC), RockWalk (I-LOC), Iron (I-MISC), Maiden (I-MISC), Jethro (B-PER) y Tull (I-PER).*

No hay falsos negativos ni en 'PER', ni en 'LOC'.

4. Conclusiones

Los resultados promedio muestran un F1 de 52.19, indicando un margen de mejora en la identificación de entidades, especialmente en la categoría MISC, donde se obtiene un puntaje de F1 = 12.4. Una posible causa podría ser el desequilibrio en los datos de entrenamiento, con menos ejemplos de la clase MISC en comparación con otras categorías. La diversidad y amplitud de MISC dificultan su identificación, ya que abarca una variedad de elementos como fechas, eventos y términos especializados.

La baja precisión en MISC se evidencia en los falsos positivos, que incluyen combinaciones de artículo determinado con sustantivo común. Al revisar los falsos negativos, notamos que algunas entidades catalogadas como MISC pueden confundirse con otras categorías.

Una mejora sugerida sería realizar un entrenamiento adicional con un conjunto de datos más amplio y ajustar los hiperparámetros del modelo para mejorar el rendimiento. Aunque se observa una clara mejora con respecto al modelo 'es_core_news_sm' en 'es_core_news_md' y 'es_core_news_lg', aún persiste un rendimiento deficiente al evaluar la clase MISC.

En cuanto a la anotación manual, se extrajeron fragmentos del artículo de Wikipedia sobre Iron Maiden utilizando el formato IOB. Sin embargo, la evaluación con el modelo preentrenado 'es_core_news_sm' muestra un F1 de 67.91 %, con una mejora en la precisión de 'MISC' al 39.53 %. Esto puede atribuirse a la naturaleza y tamaño del texto evaluado, posiblemente más fácil de interpretar que el anterior.

La evaluación con el modelo 'es_core_news_lg' muestra un rendimiento general de F1 = 82.18. Los errores principales incluyen falsos positivos y negativos en diversas categorías ('LOC', 'MISC', 'ORG', 'PER'), con estructuras de palabras que generan confusiones. Sin embargo, no se detectaron falsos negativos en 'PER' ni 'LOC'.

Referencias

[1] Erik F. Tjong Kim Sang *Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition*, segunda edición, University of Antwerp, 2002.

[2] Dr. W.J.B. Mattingly «Introduction to SpaCy», Recuperado de https://spacy.pythonhumanities.com/intro.html