Arquitecturas profundas para el reconocimiento de entidades nombradas

Santiago Moreno
Facultad de ingeniería
Universidad de Antioquia
Medellín, Colombia
santiago.moreno3@udea.edu.co

Nestor Calvo
Facultad de ingeniería
Universidad de Antioquia
Medellín, Colombia
nestor.calvo@udea.edu.co

***Index Terms*—NER, NLP, Deep Learning**

I. Contexto

En el análisis de texto uno de los primeros problemas a resolver es la identificación de las entidades a las que se refieren las palabras en cada frase. Este reconocimiento da paso a un análisis con más detalle del texto, con técnicas como la identificación de relaciones entre entidades, o el reconocimiento de entidades más específicas agrupadas dentro de una super-entidad previamente reconocida.

El reconocimiento de entidades nombradas (NER por sus siglas en inglés) se enfoca en reconocer menciones de entidades en un texto dado, como, ubicaciones, personas, organizaciones, entre otras [1]. Por ejemplo, en la figura 1, se identifica que Jorge Robledo es una persona y la segunda aparición de la palabra Robledo se refiere a una ubicación.

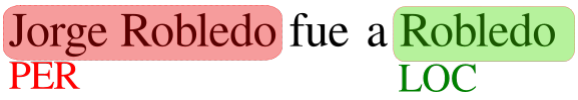

Figura 1. Ejemplo NER

Para desarrollar un modelo que realice el NER se debe tener en cuenta que exisitirán tantas clases como entidades haya, y aquellas palabras que no tengan una entidad asociada se etiquetarán con la entidad OTHERS (O).

El formato mas usado para este tipo de problema es el formato CONLL, este consiste en un archivo de texto plano en donde hay un documento de 2 columnas, la primera columna corresponde a la palabra y la segunda corresponde a su entidad (clase). Cada frase está compuesta por una serie de filas (palabras con etiqueta) cuyo fin de frase e inicio de la siguiente está representado por una que contiene solo un salto de línea.

Dado que una entidad puede estar compuesta por varias palabras las etiquetas pueden tener diversos formatos para reconocer esto. De los mas conocidos es el IOB2, el cual agrega una letra al inicio de cada entidad, en donde B representa el inicio de una entidad (begin) e I representa las palabras siguientes al inicio de la entidad (inside). Las palabras etiquetadas como OTHERS mantienen si etiqueta O y no tiene esta característica.

II. Objetivo Machine Learning

Ya que para reconocer las entidades en un texto se debe tener en cuenta el contexto de toda una frase, para ello es necesario implementar técnicas que modelen datos secuenciales.

Teniendo un dataset en el formato CONLL se busca implementar un modelo que etiquete cada palabra dentro de una frase, teniendo en cuenta el contexto que brinda la frase misma.

III. Dataset

El dataset utilizado es tomado de [4] y consiste en una serie de documentos legales en alemán previamente etiquetados y tomados de decisiones juridícas. La base de datos contiene en total 20 entidades, 19 específicas mas la entidad ”OTROS”.

En total el dataset tiene 66,723 frases con 2,157,048 palabras, cada palabra está etiquetada en el formato IOB2, esta se depura para eliminar este formato y facilitar la obtención de métricas de desempeño. La organización de la base de datos es conforme al formato CONLL.

IV. Métricas de desempeño

Las métricas de desempeño que se harán uso para evaluar la eficiencia del modelo se pueden dividir en metricas para el modelo de machine learning y metricas creadas dependiendo del problema o metricas de negocio.

IV-A. Machine Learning

Al tratarse de un problema de multiples clases se creará una matriz de confusion la cual contiene

- Precision: Esta metrica busca responder que porcentaje de positivos fueron identificados correctamente, permite conocer un numero de las entidades o clases que fueron correctamente clasificadas.

$$\frac{TP}{TP+FP} \tag{1}$$

- Recall: Esta metrica calcula el porcentaje de positivos reales que fueron identificados, la diferencia entre recall y precision es que recall tiene en cuenta los falsos negativos(aquellos positivos que fuero clasificados como otra clase).

$$\frac{TP}{TP+FN} \tag{2}$$

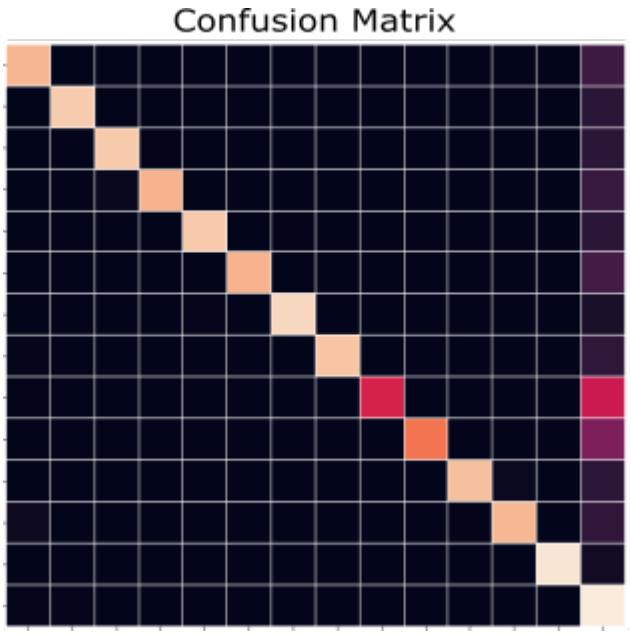

Figura 2. Matriz de confusion para multiples variables

- F1 score: Esta metrica es la media harmonica entre la precision y el recall.

$$\frac{2 * TP}{2 * TP + FP + FN} \tag{3}$$

IV-B. Negocio

Para evaluar el modelo en la vida real se puede seccionar por cada documento, internamente en un documento existen muchas entidades las cuales el modelo intentará clasificar y se medirá si este documento quedó bien reconocido o no basado en si el porcentaje de entidades clasificadas correctamente es mayor a un umbral. Al final se puede sacar el porcentaje de documentos bien clasificados los cuales serán una metrica mas entendible para personas que no tienen tanto conocimiento de machine learning.

V. Referencias y resultados previos

Referente al problema del NER se han tratado diferentes enfoques. Las primeras propuestas de solución se realizaron con un enfoque basado en reglas [2] [4], luego se trataron algoritmos supervisados con extracción de características [3] [4] y después con el aumento de la capacidad computacional y el número de datos se comenzó a utilizar un enfoque profundo [4].

En [4] se utilizó una base de datos etiquetada con el formato BIO. Allí se implementó un modelo basado en campos aleatorios condicionales (CRF por sus siglas en inglés), con 19 entidades obtuvieron un F1-score de 93.05 %. En este mismo trabajo se implementó un enfoque profundo, generando word embeddings con Word2Vec [5] e implementando una BiLSTM con una capa de CRF, con esta arquitectura se obtuvo un F1-score de 93.75 %.

Referencias

[1] Nadeau, D., Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26.

[2] Kim, J. H., Woodland, P. C. (2000). A rule-based named entity recognition system for speech input. In Sixth International Conference on Spoken Language Processing.

[3] Krishnan, V., Manning, C. D. (2006, July). An effective two-stage model for exploiting non-local dependencies in named entity recognition. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (pp. 1121-1128).

[4] Leitner, E., Rehm, G., Moreno-Schneider, J. (2019, September). Fine-grained named entity recognition in legal documents. In International Conference on Semantic Systems (pp. 272-287). Springer, Cham.

[5] Reimers, N., Eckle-Kohler, J., Schnober, C., Kim, J., Gurevych, I. (2014). Germeval-2014: Nested named entity recognition with neural networks.