Detección de entidades nombradas en textos judiciales

Claudio Sebastián Castillo

2022-11-03

Resumen

En este documento presentamos nuestro Plan de Tesis para la Maestría en Minería de Datos de la UTN -Regional Paraná- en el marco del área de Procesamiento del Lenguaje Natural (NLP por sus siglas en ingles), en el tópico vinculado a reconocimiento de entidades nombradas (NER). El objetivo del trabajo es extraer información que hoy permanece inexplorada en los textos judiciales con el fin de facilitar su tratamiento y reutilización. Para ello vamos a aplicar metodologías basadas en *modelos de lenguaje* del tipo *Transformer*(Vaswani et al. 2017), explorando distintas configuraciones para determinar los mejores resultados.

Introducción

El Poder Judicial tiene la función constitucional de brindar justicia y al hacerlo velar por el Estado de Derecho y la resolución pacífica de conflictos. Esa tarea fundamental para la vida en comunidad tiene como producto central a la *sentencia judicial*, documento donde un juez reconstruye una situación problemática y fija la solución jurídica que corresponde. Tan importante es este documento que tiene la fuerza de una ley particular para las personas involucradas en el conflicto y el poder de un mensaje acerca de "lo justo" para toda la sociedad. Por eso Rosatti, juez de la Corte Suprema de Justicia de la Nación, dice que *"las sentencias deben ser profundas y claras"* porque *"todos deben saber qué está prohibido, de ahí la importancia del lenguaje"*. (Rosatti 2022)

En efecto, la *sentencia judicial* es un documento público que concentra todos los datos relevantes de una *proceso judicial*, desde referencias a las partes, lugares y fechas de una causa, hasta complejas descripciones de hechos y derechos. Estos elementos se articulan mediante un discurso eminentemente técnico, que procede -a priori- a partir de una argumentación racional de la forma premisas-conclusión. Tales atributos tornan a las *sentencias judiciales* objetos de un valor epistemológico significativo, y a su agregación en bases de datos en potenciales *reservorios* de conocimiento.

Pero esta importancia contrasta con la escasez de desarrollos orientados a la explotación de las sentencias judiciales a partir de sus atributos lingüísticos. Así, procesos firmemente afianzados en los Poderes Judiciales de Argentina - por ejemplo aquellos vinculados a la extracción y recuperación de información- todavía enfrentan desafíos importantes de digitalización que desplazan temas menos difundidos como aquellos relativos a las *tecnologías aplicadas el lenguaje*. No es extraño entonces que Samy (2021) enumere como *retos* en esta materia los siguientes: *1) El número limitado de recursos y herramientas de PLN adaptados al dominio en general; 2) La predominancia del inglés, ya que la mayoría de los recursos y las herramientas disponibles se desarrollan para el tratamiento de textos en inglés; 3) Una adopción ralentizada de las tecnologías inteligentes en el sector legal y administrativo en comparación con otros sectores como el sector biomédico o financiero. Estos retos han influido en que la consolidación de la tarea NERC en el dominio legal ha tardado unos años en comparación con otros dominios.*

! Extraer de: Samy (2021) pto.2

Todo ello sin mencionar las exigentes condiciones legales y reglamentarias que regulan la disposición y tratamiento de bases de datos judiciales, entre la que están: Protección de datos personales, datos sensibles, ...

En este sentido Samy (2021) menciona entre los *retos* para el avance 1) El número limitado de recursos y herramientas de PLN adaptados al dominio en general; 2) La predominancia del inglés, ya que la mayoría de los recursos y las herramientas disponibles se desarrollan para el tratamiento de textos en inglés; 3) Una adopción ralentizada de las tecnologías inteligentes en el sector legal y administrativo en comparación con otros sectores como el sector biomédico o financiero.

En muchos aspectos las bases de datos judiciales permaneces como reservorios inexplorados

[104]. Esto es particularmente evidente en Argentina pero podría extenderse a otros países de habla hispana, que en los últimos años han comenzado a revertir esta dificultad a partir de diversas iniciativas (. Plan de Impulso a las Tecnologías del Lenguaje (Gobierno España, n.d.)

Objetivo

Factibilidad y relevancia

En cuanto a las metodologías y técnicas, los métodos empleados para abordar la tarea de NERC han ido desarrollando desde modelos basados en reglas con patrones de expresiones regulares, listas o gazetteers hacia modelos de aprendizaje automático supervisado y semisupervisado como Hidden Markov Models (HMM), Support Vector Machine (SVM) y Conditional Random Field (CRF) siendo este último de los más eficientes en NERC (Roy, 2021). En los últimos años, el uso de las redes neuronales con el aprendizaje profundo y la integración de modelos del lenguaje con los WordEmbeddings ha supuesto un cambio en el paradigma del PLN en general y en las tareas específicas como NERC (Roy, 2021). Samy (2021)

El papel de NERC es imprescindible en el desarrollo de sistemas legales inteligentes. Dado el gran volumen de textos que se suele manejar en este dominio, ha surgido un interés, cada vez mayor, por el procesamiento de textos legales, en general y por la tarea NERC, en particular. Samy (2021)

Este interés se fundamenta en el gran potencial de las técnicas de PLN y su capacidad de ofrecer soluciones inteligentes que beneficien a usuarios claves del sector como los abogados, los jueces, los juristas, los documentalistas jurídicos, además del sector de la administración pública que, aunque no trate textos estrictamente jurídicos, sí maneja textos administrativos con un alto contenido legal como es el caso de la contratación pública o los convenios. Por tanto, los avances en el procesamiento de textos legales constituyen un gran potencial para agilizar procesos internos de la administración pública, simplificar los procedimientos y mejorar el acceso de la ciudadanía a la información legal y administrativa. Para impulsar la apertura de Samy (2021).

Estado del arte

Desafíos

[texto DavidPerezFernandez]

No obstante, los trabajos en esta área se enfrentan con retos como: 1) El número limitado de recursos y herramientas de PLN adaptados al dominio en general; 2) La predominancia del inglés, ya que la mayoría de los recursos y las herramientas disponibles se desarrollan para el tratamiento de textos en inglés; 3) Una adopción ralentizada de las tecnologías inteligentes en el sector legal y administrativo en comparación con otros sectores como el sector biomédico o financiero. Estos retos han influido en que la consolidación de la tarea NERC en el dominio legal ha tardado unos años en comparación con otros dominios. De ahí, el presente estudio pretende afrontar la tarea en los textos legales españoles teniendo como objetivo principal

el reconocimiento y la clasificación de cinco tipos básicos de entidades nombradas en textos legislativos españoles.

Metodología

baseline (ver @ Francesca L)

Para tener una referencia con la que comparar el rendimiento de sistemas de aprendizaje autom atico complejos, como las arquitecturas neuronales descriptas arriba, se decidio utilizar como modelo base la misma arquitectura mostrada en la Figura 3, pero con la diferencia que los modelos internos ahora est an basados en M aquinas de vectores soporte (SVM por sus siglas en ingles). Las maquinas de vectores soporte son una familia de algoritmos de aprendizaje supervisado, donde la idea principal del algoritmo es que a partir de los datos de entrenamiento se intenta encontrar un hiperplano optimo que maximice el margen (Maximal Margin Classifier). El margen se define como la distancia entre el hiperplano de separaci on (lımite de decision) y las muestras de entrenamiento de cada una de las clases que se quieren separar m as cercanas a este hiperplano, que son los llamados vectores de soporte.

---------Adem as se desarrollar a un aprendedor automatico simple *baseline*. Este aprendedor, a comparacion de las tecnicas de aprendizaje profundo, no requiere mucha experiencia ni tiempo para su construcci on y posee menos par ametros que las redes neuronales, que generalmente en arquitecturas complejas tienen una gran cantidad de par ametros. Este servira como referencia para los demas aprendedores.

Área de Estudio

Samy (2021)

El reconocimiento de entidades nombradas (NER por sus siglas en ingles), tambien conocido como extracci on de entidades, es una tarea de extracci on de informaci on que busca localizar y clasificar en categorıas predefinidas como personas, organizaciones, lugares, expresiones de tiempo y cantidades, las entidades nombradas encontradas en un texto. El reconocimiento de entidades nombradas a menudo se divide conceptualmente en dos problemas distintos: detecci on de nombres, y clasificaci on de los nombres seg un el tipo de entidad al que hacen referencia. Es por eso que muchas veces en la literatura se lo conoce como reconocimiento y clasificaci on de entidades nombradas (NERC por sus siglas en ingles). Una tercera fase que se desprende del reconocimiento y clasificaci on de entidades nombradas se conoce como anotaci on semantica (entity linking en ingles) donde se anota una entidad con una referencia a alg un link de una base de conocimiento que contenga una definici on sem antica de la entidad (Carreras et al., 2003). La primera fase generalmente se reduce a un problema de segmentaci on: los nombres son una secuencia contigua de tokens, sin solapamiento ni anidamiento, de modo que Banco de la Nacion Argentina es un nombre unico, a pesar del hecho de que dentro de este nombre aparezca la subcadena Argentina que es a su vez el nombre de un paıs. La segunda fase se trata de asignar una categorıa, de entre un conjunto predeterminado, a cada una de las entidades previamente reconocidas en la fase uno. El reconocimiento y clasificaci on de entidades nombradas se puede aprovechar de varias maneras, incluyendo el suministro de enlaces de hipertexto a la informaci on almacenada sobre por ejemplo un artıculo en particular. Por ejemplo, una menci on del “Banco de la Naci on Argentina” podrıa resolverse en un link a la pagina de Wikipedia que contenga un artıculo sobre esta entidad.

Obtención y preparación de datos

Procesamiento de datos

Entrenamiento de modelos

Gutiérrez-Fandiño et al. (2021)

Evaluación

Software

El trabajo se desarrollará utilizando el lenguaje de programación Python 3 y frameworks o librerías asociadas. .... Para el entrenamiento y evaluación de los modelos de NLP propuestos se utilizarán las plataformas Google Colab y Kaggle las cuales brindan acceso gratuito a GPU's de alto rendimiento.

Cronograma

Gobierno España. n.d. “Plan de Tecnologías Del Lenguaje - Página Principal Del Plan de Impulso de Las Tecnologías Del Lenguaje.” https://plantl.mineco.gob.es/Paginas/index.aspx.

Gutiérrez-Fandiño, Asier, Jordi Armengol-Estapé, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. “Spanish Legalese Language Model and Corpora,” October. http://arxiv.org/abs/2110.12201.

Rosatti, Horacio. 2022. “Las sentencias judiciales deben ser profundas y claras.” https://www.jusentrerios.gov.ar/2022/10/27/horacio-rosatti-las-sentencias-judiciales-deben-ser-profundas-y-claras/.

Samy, Doaa. 2021. “Reconocimiento y clasificación de entidades nombradas en textos legales en español.” *Procesamiento del Lenguaje Natural*, 103–14. https://doi.org/10.26342/2021-67-9.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need,” December. https://doi.org/10.48550/arXiv.1706.03762.