

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 inglés), en el tópico vinculado a reconocimiento de entidades nombradas (NER). El objetivo del trabajo es extraer información que hoy permanece inexplorada en textos legales con el fin de facilitar su tratamiento y reutilización. Para ello vamos a: 1) construir un corpus de textos legales que nos permita probar soluciones para la tarea de anotación de entidades, y 2) desarrollar los algoritmos más apropiados para el tratamiento de dicho corpus buscando las configuraciones con mejor performance.

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.

Desafortunadamente existe una gran asimetría entre la importancia que tienen las sentencias judiciales como fuente de información y la escasez de desarrollos orientados a explotar sus atributos lingüísticos. Esa asimetría se explica, en gran medida, por la escasez de recursos en general para el NLP en español, lo que agrava las dificultades en el ámbito legal como uno de sus sub-dominios . En línea con esto Samy (2021) menciona entre los *retos* que enfrentan los proyectos de NLP en el ámbito legal en español a: 1) *El número limitado de recursos y herramientas 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*; y 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*. Por su parte Cardellino et al. (2017) refuerza esta enumeración destacando que *existen muy pocos corpus legales anotados con anotaciones para entidades* [lo que] *constituye una importante barrera para la Extracción de Información del texto legal* Cardellino et al. (2017). Dificultades similares destacan Leitner, Rehm, and Moreno-Schneider (2019) , Serrano et al. (2022), entre otros.

A pesar de estas dificultades, durante los últimos años han surgido importantes avances en el NLP dando al campo nuevo impulso. Nuevos modelos, con nuevas arquitecturas para optimizar los procesos de aprendizaje Serrano et al. (2022), no solo han logrado mejorar las métricas de evaluación para distintas tareas de NLP, sino también han logrado una performance igual o superior al de una persona. Junto con el nuevo conocimiento, nuevos proyectos públicos¹ y privados² con eje en las *tecnologías del lenguaje* han ganado presencia no solo en el campo científico sino también en el discurso público, colocando al NLP como tópico de relevancia en el amplio abanico de la *inteligencia artificial*.

Por todo esto, el presente Plan de Tesis está dirigido a realizar un aporte al área del NLP en español aplicado al dominio legal, empleando para ello las herramientas que hayan demostrado mejor performance en estudios comparados y considerando su potencial de aplicación en el contexto institucional de nuestro país. A continuación desarrollaremos el alcance de este proyecto.

Objetivo

1. Construir un corpus de textos legales que nos permita probar soluciones para la tarea de anotación de entidades, y
2. Desarrollar los algoritmos más apropiados para el tratamiento de dicho corpus buscando las configuraciones con mejor performance.

Factibilidad y relevancia

El proyecto que hemos propuesto es factible porque disponemos de los medios científicos y tecnológicos requeridos para su ejecución.

Respecto del conocimiento requerido cabe destacar que el área de investigación en NER, a pesar de sus jóvenes 30 años, ha visto un acelerado crecimiento en materia de recursos y enfoques disponibles (Roy 2021). Desde las primeras soluciones basadas en modelos estadísticos lineales y reglas de asignación ad-hoc, pasando por modelos de aprendizaje supervisado y semi-supervisado en grandes corpus, hasta los actuales modelos de redes neuronales profundas con arquitectura *Transformer* Serrano et al. (2022), existe un amplio espectro de enfoques para nutrir nuestros abordajes del problema.

Estos avances han impulsado en los últimos años un florecimiento del NLP en español, revirtiendo parcialmente las carencias apuntadas más arriba como veremos en detalle en el punto sobre *estado del arte*.

Respecto de la factibilidad para construir un corpus legal útil para este trabajo tenemos evidencia de las múltiples opciones disponibles. En nuestro experimento sobre [CONSTRUCCION CORPUS LEGAL] hemos producido un corpus legal no anotado de sentencias judiciales y sus respectivos metadatos empleando el método de *scraping*. Esta información consta de Información similar se publica por todos los STJs del país. Entendemos que este experimento evidencia la disponibilidad material y tecnológica para construir el corpus.

Similar a lo anterior, contamos con un ejemplo análogo de construcción de corpus en el trabajo de tesis de [Haag, tesis CORDOBA].

Luego de exponer los argumentos acerca de la factibilidad de proyecto es preciso mencionar las razones que hacen a su relevancia.

- 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)

¹En el ámbito del NLP en español se detacan proyectos interinstitucionales de alcance nacional (Gobierno España, n.d.) e internacional (<https://www.mirelproject.eu/>) dirigidos a promover el avance de las *tecnologías del lenguaje*.

²Enumerar....

- 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).
- The goal is to make knowledge workers, who process and make use of these documents, more efficient and more effective in their day to day work, this also includes the analysis of domain-specific NEs, see [5,31] for related approaches in the area of content curation technologies. Leitner, Rehm, and Moreno-Schneider (2019)

Estado del arte

! Extraer de: Samy (2021) pto.2

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.

Legal language is unique and differs greatly from newspaper language. This also relates to the use of person, location and organization NEs in legal text, which are relatively rare. It does contain such specific entities as designations of legal norms and references to other legal documents (laws, ordinances, regulations, decisions, etc.) that play an essential role. Leitner, Rehm, and Moreno-Schneider (2019)

Despite the development of NER for other languages and domains, the legal domain has not been exhaustively addressed yet. This research also had to face the following two challenges. (1) There is no uniform typology of semantic concepts related to NEs in documents from the legal domain; correspondingly, uniform annotation guidelines for NEs in the legal domain do not exist either. (2) There are no freely available datasets consisting of documents from the legal domain, in which NEs have been annotated. Thus, the research goal is to examine NER with a specific focus on German legal documents. This includes the elaboration of the corresponding concepts, the construction of a dataset, developing, evaluating and comparing state of the art models for NER. Leitner, Rehm, and Moreno-Schneider (2019)

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,

Metodología

baseline (ver @ Francesca L)

Para tener una referencia con la que comparar el rendimiento de sistemas de aprendizaje automático complejos, como las arquitecturas neuronales descritas arriba, se decidió utilizar como modelo base la misma arquitectura mostrada en la Figura 3, pero con la diferencia que los modelos internos ahora están basados en Máquinas de Vectores Soporte (SVM por sus siglas en inglés). Las máquinas 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 óptimo que maximice el margen (Maximal Margin Classifier). El margen se define como la distancia entre el hiperplano de separación (límite de decisión) y las muestras de entrenamiento de cada una de las clases que se quieren separar más cercanas a este hiperplano, que son los llamados vectores de soporte.

-----Además se desarrollará un aprendizador automático simple *baseline*. Este aprendizador, a comparación de las técnicas de aprendizaje profundo, no requiere mucha experiencia ni tiempo para su construcción y posee menos parámetros que las redes neuronales, que generalmente en arquitecturas complejas tienen una gran cantidad de parámetros. Este servirá como referencia para los demás aprendizadores.

Área de Estudio

Samy (2021)

El reconocimiento de entidades nombradas (NER por sus siglas en inglés), también conocido como extracción de entidades, es una tarea de extracción de información que busca localizar y clasificar en categorías 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: detección de nombres, y clasificación de los nombres según el tipo de entidad al que hacen referencia. Es por eso que muchas veces en la literatura se lo conoce como reconocimiento y clasificación de entidades nombradas (NERC por sus siglas en inglés). Una tercera fase que se desprende del reconocimiento y clasificación de entidades nombradas se conoce como anotación semántica (entity linking en inglés) donde se anota una entidad con una referencia a algún link de una base de conocimiento que contenga una definición semántica de la entidad (Carreras et al., 2003). La primera fase generalmente se reduce a un problema de segmentación: los nombres son una secuencia contigua de tokens, sin solapamiento ni anidamiento, de modo que Banco de la Nación Argentina es un nombre único, a pesar del hecho de que dentro de este nombre aparezca la subcadena Argentina que es a su vez el nombre de un país. La segunda fase se trata de asignar una categoría, de entre un conjunto predeterminado, a cada una de las entidades previamente reconocidas en la fase uno. El reconocimiento y clasificación de entidades nombradas se puede aprovechar de varias maneras, incluyendo el suministro de enlaces de hipertexto a la información almacenada sobre por ejemplo un artículo en particular. Por ejemplo, una mención del “Banco de la Nación Argentina” podría resolverse en un link a la página de Wikipedia que contenga un artículo sobre esta entidad.

Obtención y preparación de datos

Procesamiento de datos

To adapt categories for the legal domain, the set of NE classes was redefined in the approaches described above. Thus, Dozier et al. [13] focused on legal NEs (e.g., judge, lawyer, court). Cardellino et al. [8] extended NEs on NERC level to document, abstraction, and act. It is unclear what belongs to these classes and how they were separated from each other. Glaser et al. [18] added reference [23]. However, this was understood as a reference to legal norms, so that further references (to decisions, regulations, legal literature, etc.) were not covered. Leitner, Rehm, and Moreno-Schneider (2019)

3.1 Semantic Categories Legal documents differ from texts in other domains, and from each other in terms of text-internal, and text-external criteria [7,12,15,21], which has a huge impact on linguistic and thematic design, citation, structure, etc. This also applies to NEs used in legal documents. In law texts and administrative regulations, the occurrence of typical NEs such as person, location and organization is very

low. Court decisions, on the other hand, include these NEs, and references to national or supranational laws, other decisions, and regulations. Two requirements for a typology of legal NEs emerge from these peculiarities. First, the categories used must reflect those entities that are typical for decisions. Second, a typology must concern the entities whose differentiation in decisions is highly relevant. Leitner, Rehm, and Moreno-Schneider (2019, 275)

Taxonomía de las categorías: 19! ver Leitner, Rehm, and Moreno-Schneider (2019)

Ojo con la anonimización de sentencias debido a protección de datos personals.

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

- Cardellino, Cristian, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. “A Low-Cost, High-Coverage Legal Named Entity Recognizer, Classifier and Linker.” In *ICAAIL-2017 - 16th International Conference on Artificial Intelligence and Law*, 22. Londres, United Kingdom. <https://hal.archives-ouvertes.fr/hal-01541446>.
- 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>.
- Leitner, Elena, Georg Rehm, and Julian Moreno-Schneider. 2019. “Fine-Grained Named Entity Recognition in Legal Documents.” In *Semantic Systems. The Power of AI and Knowledge Graphs*, edited by Maribel Acosta, Philippe Cudré-Mauroux, Maria Maleshkova, Tassilo Pellegrini, Harald Sack, and York Sure-Vetter, 272–87. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-33220-4_20.
- 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/>.
- Roy, Arya. 2021. “Recent Trends in Named Entity Recognition (NER),” January. <http://arxiv.org/abs/2101.11420>.
- 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>.
- Serrano, Alejandro Vaca, Guillem Garcia Subies, Helena Montoro Zamorano, Nuria Aldama Garcia, Doaa Samy, David Betancur Sanchez, Antonio Moreno Sandoval, Marta Guerrero Nieto, and Alvaro Barbero Jimenez. 2022. “RigoBERTa: A State-of-the-Art Language Model for Spanish,” June. <http://arxiv.org/abs/2205.10233>.