Detección de entidades nombradas en textos judiciales

Claudio Sebastián Castillo

2022-11-03

Tema elegido

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), en el tópico vinculado a reconocimiento de entidades nombradas (NER). El objetivo del trabajo es implementar algoritmos de aprendizaje automático orientados a reconocer y extraer información de textos legales. El dominio legal es un campo de producción de grandes volúmenes de información textual, cuyo contenido y alcance afecta directamente a las relaciones sociales. Esta información es eminentemente no estructurada por lo que su exploración y explotación enfrenta grandes desafíos. Hasta donde conocemos, en nuestro país no existen experiencias de explotación de los atributos lingüísticos del texto en el dominio legal. Por todo ello, en este trabajo nos proponemos implementar en el ámbito legal los algoritmos que hayan acreditado los mejores resultados en la tarea de reconocimiento de entidades nombradas. Para ello vamos a: 1) construir un corpus legal que nos permita probar distintas soluciones para la tarea de reconocimiento y anotación de entidades, y 2) desarrollar *nuevas implementaciones* de algoritmos de aprendizaje para el tratamiento de dicho corpus buscando las configuraciones con mejor performance.

Fundamentación y justificación del tema

El reconocimiento y la clasificación de entidades nombradas (en adelante NER/NERC por las siglas en inglés *Named Entity Recoginition and Classification*) es una tarea principal en el área del Procesamiento del Lenguaje Natural (en adelante NLP por *Natural Language Processing*). Normalmente consiste en la identificación automática -sin intervención humana- de entidades nombradas en textos y su asignación a determinadas categorías semánticas. Estas unidades refieren a cualquier entidad que puede aludirse mediante nombres propios -e.g. personas, organizaciones y localizaciones- aunque en la práctica se han extendido para incluir expresiones numéricas como fechas y cantidades, y diversos tipos de entidades según el dominio de interés. Esta referencia a entidades puede consistir en expresiones lingüísticas simples o complejas, confiriendo a la tarea de reconocimiento automático un nivel de importante dificultad (Jurafsky and Martin 2021). A pesar de ello y gracias a los buenos resultados obtenidos con el uso de redes neuronales profundas (*Deep Neural Networks* o DNN) y la integración de modelos del lenguaje con *WordEmbeddings*, los últimos años han presenciado un cambio de paradigma en el NLP en general y en las tareas NERC en particular (Roy 2021).

Este escenario propicio y estimulante en materias de *nuevas tecnologías del lenguaje* impulsa la mirada sobre ámbitos donde el texto extiende su soberanía. Entre ellos el Estado y particularmente el Poder Judicial. En efecto, éste último 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*, un tipo de documento textual 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, resaltando la importancia del lenguaje, dice que: *"las sentencias deben ser profundas y claras"* porque *"todos deben saber qué está prohibido"* (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, y en particular para el español legal. 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) refuerzan 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.* Dificultades similares destacan Leitner, Rehm, and Moreno-Schneider (2019) , Serrano et al. (2022), entre otros.

A pesar de estas dificultades, y como mencionamos antes, los últimos años han sido testigos de importantes avances en el NLP. Nuevos modelos, con arquitecturas inéditas tendientes a optimizar los procesos de aprendizaje Serrano et al. (2022), han logrado mejorar las métricas de evaluación para distintas tareas de análisis semántico, obteniendo resultados que igualan o superan a los obtenidos por un hablante nativo promedio. Junto con ese nuevo conocimiento, nuevos proyectos públicos[1] y privados[2] 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 interés general.

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 los algoritmos que hayan demostrado mejor performance en estudios comparados. 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 reconocimiento y anotación de entidades nombradas, 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.

[1]En el ámbito del NLP en español se detacan proyectos interinstitucionales de alcance nacional como el Plan de Impulso de las Tecnologías del Lenguaje del Gobierno de España dirigido a fomentar el desarrollo del procesamiento del lenguaje natural y la traducción automática en lengua española y lenguas cooficiales (https://plantl.mineco.gob.es/tecnologias-lenguaje/PTL/Paginas/plan-impulso-tecnologias-lenguaje.aspx). A nivel internacional el Proyecto de la Comunidad Europea MIREL (https://www.mirelproject.eu/) dirigidos a promover el avance de las *tecnologías del lenguaje* .

[2]OpenAI, DeepMind, Ought, Hugging Face, Cohere, entre una larga lista de proyectos disruptivos. Un artículo que repasa empresas y desarrollos puede consultarse aquí: https://hbr.org/2022/04/the-power-of-natural-language-processing.

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.

Aunque estos avances tienen como lenguaje objeto al idioma inglés, no han dejado de impulsar en los últimos años un florecimiento del NLP en español. Gracias a ello, las carencias antes apuntadas han comenzado a revertirse parcialmente, ofreciéndonos material de valor para nuestro proyecto cuyo detalle presentaremos en el *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 de construcción de corpus legal[3] hemos trabajo con el portal de jurisprudencia del Poder Judicial de la provincia de Buenos Aires que brinda acceso público a fallos judiciales[4]. En el experimento construimos un dataset no anotado de 4299 sentencias judiciales y sus respectivos metadatos empleando el método de *scraping*. Esta información consta de fallos judiciales completos y metadatos de la causa (materia, tipo de fallo, número de la causa, caratula, magistrado y tribunal actuantes, entre otros). Información similar a la obtenida de la justicia de Buenos Aires se publica por distintos Poderes Judiciales del país, por lo que entendemos que existe disponibilidad material y tecnológica para construir el corpus.

Abonando lo anterior, contamos con un ejemplo análogo de construcción de corpus en el trabajo de tesis de grado de Karen Haag bajo la dirección de Cristian Cardellino en la Facultad de Matemática, Astronomía, Física y Computación Universidad Nacional de Córdoba en el año 2019[5]. El trabajo aborda el mismo tópico que estamos presentando en este documento (NLP-NER), y el corpus empleado fue construido ad-hoc para la tesis mediante *scraping* del portal Infoleg[6] según se detalla en el punto 3.

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

En primer lugar, desde una perspectiva estrictamente funcional, el desarrollo de estrategias de NLP-NER aplicadas al dominio legal realizaría un aporte inestimable al procesamiento de información textual. El servicio de justicia se asienta sobre el intercambio de información entre los distintos actores de un proceso: jueces, abogados y auxiliares de justicia. Éstos generan documentos que se agregan a un registro de base de datos y componen la historia de un proceso. Estos registros, que conforman un *expediente digital*, actualmente alimentan grandes bases documentales cuya explotación lingüística es incipiente. Contar con herramientas de NLP-NER aplicados al ámbito legal permitiría desarrollar nuevos productos y servicios para las distintas partes interesadas, al tiempo que optimizaría los (escasos) recursos disponibles en los Poderes Judiciales (Leitner, Rehm, and Moreno-Schneider 2019). Por ejemplo, proyectos vinculados a gestión electrónica, celeridad de procesos, regulación de flujos de información y trabajo se podrían beneficiar drásticamente del desarrollo de sistemas inteligentes que permitiesen extraer y procesar atributos lingüísticos de los documentos legales (Samy 2021).

En segundo lugar, desde una perspectiva institucional, el desarrollo de procesos internos sensibles a los atributos lingüísticos del texto liberaría posibilidades de automatización sin precedentes, con inocultable impacto en la calidad del servicio brindado al ciudadano. En efecto, los grandes procesos de reforma procesal de los últimos años -tendientes a oralizar los procesos judiciales- han establecido deberes de actuación exigentes a los órganos de justicia. Principios como la *oficiocidad, celeridad, concentración, economía procesal* y *plazo razonable* [7] demandan una capacidad de procesamiento de la información cada vez más robusta y eficiente,

[3]Accesible en github aquí: https://github.com/castillosebastian/nlp_research/blob/master/Crear_corpus_judicial_experimento0.ipynb

[4]https://juba.scba.gov.ar/Busquedas.aspx

[5]Accesible aquí https://rdu.unc.edu.ar/bitstream/handle/11086/15323/Haag%2C%20K.%20Y.%20Reconocimiento%20de%20entidades%20nombradas%20en%20texto%20de%20dominio%20legal.pdf?sequence=1&isAllowed=y

[6]http://www.infoleg.gob.ar/

[7]El Tribunal Superior de Justicia de la Provincia de Córdoba, mediante Ac. Reg. N° 1550 Serie "A" de fecha 19/02/2019 aprobó el "Protocolo de gestión del Proceso Civil Oral", que enuncia y explica aquellos principios. Ideas rectoras que se pueden

capaz de reconocer, clasificar, ordenar y distribuir información textual no estructurada. [8] Este escenario es un espacio fértil para el desarrollo de las *tecnologías del lenguaje* que estudiaremos en este proyecto, tecnologías que ya están generando profundos cambios en otras esferas de la vida social.

Por último, aunque el foco de nuestro trabajo está puesto en documentos judiciales, no es difícil advertir que el grueso de la administración pública (y buena parte de la privada) funciona en torno a documentos administrativos no estucturados o semi-estructurados. Por esto, los avances en el procesamiento de documentos judiciales presentan un gran potencial de transferencia hacia otros dominios de la administración en general, con los correspondientes beneficios vinculados al tratamiento y explotación de la información (Samy 2021).

Estado del arte

Samy (2021) pto.2

aspectos vacantes?

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, ...

constatar en las demás jurisdicciones provinciales.

[8] Nueva gestión judicial : oralidad en los procesos civiles, Héctor M. Chayer ... [et al.] ; coordinación general de Héctor M. Chayer ; Juan Pablo Marcet. - 2a ed ampliada. Ciudad Autónoma de Buenos Aires : Ediciones SAIJ, 2017.

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

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,andact. 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 *ICAIL-2017 - 16th International Conference on Artificial Intelligence and Law*, 22. Londres, United Kingdom. https://hal.archives-ouvertes.fr/hal-01541446.

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.

Jurafsky, Daniel, and James H. Martin. 2021. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2021st ed.

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.