Detección automática de metadatos identificatorios en sentencias 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 *reconocimiento de entidades nombradas* (*Named Entity Recoginition* o NER). El objetivo del trabajo es implementar algoritmos de aprendizaje automático orientados a reconocer y extraer *datos identificatorios* de sentencias judiciales. El dominio legal es un campo de producción de un gran volumen de información textual, cuyo contenido y alcance impacta de manera definitiva en la vida de muchas personas. Esta información es eminentemente no estructurada por lo que su exploración y explotación enfrenta grandes desafíos. Entre esos desafíos, la falta de recursos para el abordaje automatizado de textos en español, y particularmente la falta de un *corpus legal anotado* en este idioma han sido barreras difíciles de sortear para las tareas NER.[1] Por ello, en este trabajo propondremos cambiar el alcance de la tarea de anotación y construir un *corpus legal semi-anotado* en español que nos permita desarrollar algoritmos de aprendizaje supervisado. Reduciremos el set de *entidades nombradas* al conjunto de *entidades que integran los metadatos identificatorios de una sentencia judicial*. Datos que, dada su importancia en el ámbito legal, están generalmente disponibles y convenientemente individualizados. Así, disminuyendo la dimensión del problema que enfrenta la tarea NER convencional donde los recursos son escasos, nos centraremos en resolver un problema NER no-convencional donde existen recursos suficientes. Hemos denominado a este enfoque como *reconocimiento de metadatos* o MER (*Metadata Entity Recognition*). Consideramos que lejos de ser una operación trivial plantea un avance significativo en materia de implementación pues resuelve un problema real: la generación manual de metadados identificatorios de sentencias. Al mismo tiempo, constituye un avance en materia de conocimiento por la generación de un *corpus legal semi-anotado* y el desarrollo de algoritmos de aprendizaje supervisado para una tarea NER no-convencional.

Fundamentación

El *reconocimiento y la clasificación de entidades nombradas* (NER) es una de las tareas más comunes en del Procesamiento del Lenguaje Natural (Vajjala and Balasubramaniam 2022). Normalmente consiste en la identificación automática -sin intervención humana- de entidades nombradas en textos[2] y su asignación a determinadas categorías semánticas. Dada una secuencia de tokens $s = (w1, w2, \ldots, wN)$ la tarea NER implica producir una lista de la forma $(I_s, I_e, type)$ para cada unidad, donde I_s e I_e representan la posición inicial y final del token dentro de la secuencia y *type* es la categoría asignada al token de un conjunto predefinido de categorías.

Las unidades o tokens abordadas en las tareas NER refieren a entidades que generalmente puede aludirse mediante *nombres propios*: personas, organizaciones y localizaciones. Sin perjuicio de ello, en la práctica se

[1] En nuestro país no existen experiencias publicadas sobre aplicación de estrategias NER en el dominio legal.

[2] A partir de las expresiones lingüísticas que sirven, en una determinada comunidad lingüística, para referenciarlas.

han extendido su alcance para incluir expresiones numéricas referidas a fechas y cantidades, o distintos tipos de entidades que varían según el dominio de interés (e.g. enfermedades en el ámbito de la Medicina). Esta referencia a entidades puede consistir en expresiones lingüísticas simples (un token) o complejas (más de un token), confiriendo a la tarea de reconocimiento automático un nivel de dificultad importante (Jurafsky and Martin 2021).

Según J. Li et al. (2020a) podemos agrupar las técnicas aplicadas en NER en cuatro tipos, a saber: 1) enfoques basados en reglas lingüísticas que no necesitan datos anotados porque descansan en las regularidades del lenguaje (e.g. *expresiones regulares* o *regex*), 2) enfoques basados en aprendizaje no supervisado que emplean algoritmos del mismo tipo (e.g. *clustering*) y tampoco emplean datos anotados, 3) enfoques basados en ingeniería de atributos y aprendizaje supervisado[3] que emplean datos anotados, y 4) enfoques de aprendizaje profundo[4] y representación distribuida de datos (*embeddings*[5]) capaces de generar automáticamente la representación óptima de la información disponible.

Estos últimos enfoques han generado un cambio de paradigma en el NLP en general y en las tareas NER en particular (Roy 2021). Las nuevas arquitecturas, basadas en redes neuronales profundas (*deep neural network* -DNN- o *deep learning* -DL-), han ganado en poder de representación a partir de distintas capas de procesamiento que aplican a los datos y las transformaciones que generan entre capas. Gracias a ello, se han producido mejoras sistemáticas en la eficacia de estos modelos en distintas tareas de análisis semántico (Kiela et al. 2021; Serrano et al. 2022a). Según J. Li et al. (2020b) la fortaleza que presenta el *aprendizaje profundo* en las tareas NER se vincula a tres razones: 1) NER se beneficia de transformaciones no-lineales de los datos (mapeo no-lineal entre imput-output) rasgo que los modelos de DL explotan a partir de la complejidad del procesamiento en capas y las funciones no-lineales de activación, 2) DL ahorra un esfuerzo significativo en la ingeniería de atributos (esfuerzo costoso en otros tipos de modelos) debido a su capacidad de aprender automáticamente respresetaciones significativas de la información disponible, y finalmente 3) los modelos DL pueden optimizarse completamente, de principio a fin[6], habilitando así arquitecturas complejas de procesamiento.

Junto con estas transformaciones en el plano del conocimiento y las metodologías disponibles, nuevos proyectos públicos y privados con eje en las tecnologías NLP han ganado presencia en el campo científico y en el discurso público. En el dominio del lenguaje español se destacan proyectos interinstitucionales de alcance nacional como el *Plan de Impulso de las Tecnologías del Lenguaje del Gobierno de España* o *IBERLEGAL* dirigidos a fomentar el desarrollo del procesamiento del lenguaje natural y la traducción automática en lengua española. A nivel internacional el Proyecto de la Comunidad Europea *MIREL* dirigidos a promover el avance de las *tecnologías del lenguaje,* similar a *MARCELL*, entre otros. En el ámbito privado sobresalen organizaciones consolidadas, dedicadas parcial o exclusivamente a brindar servicios de NLP, a saber: OpenAI, DeepMind, Ought, Hugging Face, Cohere, entre una larga lista de proyectos disruptivos.[7]

Este escenario estimulante para el NLP impulsa la mirada sobre ámbitos con uso intensivo del soporte textual como potenciales espacios de aplicación. Entre ellos, el Estado y particularmente el Poder Judicial presentan una larga tradición de procedimientos escritos con hondo impacto en la vida de las personas. Por eso, en el presente trabajo propondremos implementar estrategias NER no-convencionales en el ámbito judicial con el fin de generar herramientas de exploración-explotación que contribuyan a mejorar los servicios al ciudadano.

[3]Dentro de este grupo, que emplea una estrategia de clasificación multiclases, se han empleado distintos algoritmos, entre los que se encuentran: *Hidden Markov Models (HMM), Decision Trees, Maximum Entropy Models, Support Vector Machines (SVM)*, y *Conditional Random Fields (CRF)*.

[4]Normalmente basadas en Redes Neuronales Profundas.

[5]Ésta constituye una técnica de representación de las palabras a partir de vectores densos dentro de un espacio vectorial predefinido.

[6]Empleando para ello distintos algoritmos, por ejemplo *descenso de gradiente*.

[7]Un artículo que repasa empresas y desarrollos puede consultarse aquí: https://hbr.org/2022/04/the-power-of-natural-language-processing.

Situación Problemática

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 o fallo judicial*: 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"*.[8]

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.[9] Tales atributos tornan a las *sentencias judiciales* en insumos esenciales para la actividad judicial y piezas de información de amplia publicidad. Por ello, su individualización precisa y práctica es muy importante no solo para permitir su fácil reconocimiento y comunicación, sino también para asegurar su referencia clara y unívoca.

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; 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. (2022b), entre otros.

Es importante remarcar de lo anterior que la falta de corpus legal anotado es particularmente crítico para avanzar en tareas NER. Las soluciones más novedosas y con mejores resultados para este problema emplean generalmente algoritmos de aprendizaje supervisado, que tienen a disposición datos de entrenamiento etiquetados. Es decir, datos donde las entidades nombradas ya se encuentran identificadas y asignadas a la categoría semántica de interés, y el algoritmo aprende a reconocer los patrones que asocian datos con categorías. En el caso del español legal dicho corpus no existe, y su creación implica un trabajo manual de anotación y validación de alto costo económico y de recursos. En este contexto se inscriben distintos proyectos tendiente a superar la dificultad. Por ejemplo, Samy, Arenas-García, and Pérez-Fernández (2020) destaca la creación de la primera tarea compartida en el marco de *IberLegal*[10] con el propósito de crear un corpus de textos en español y evaluar la tarea NER enfocándose en cinco categorías: legislación, organización/entidades legales, Personas, Lugares y Expresiones Temporales. A pesar de estos esfuerzos a la fecha no se dispone de un corpus legal anotado con entidades.

Ante este problema, en el presente trabajo propondremos un enfoque distinto de la tarea NER que nos permitirá generar un c*orpus legal semi-anotado* para desarrollar estrategias de aprendizaje supervisado. Para ello, la tarea de anotación de entidades en este trabajo tendrá un alcance específico, en el que no se requiere procesamiento manual ,y descansa exclusivamente en el tratamiento computacional del lenguaje. Este direccionamiento de nuestro esfuerzo implicará la *anotación* de *aquellas entidades nombradas en el texto que coincidan con los metadatos regularmente empleados para identificar sentencias.* Estos *datos identificatorios* constituyen entidades *per se*, empleadas intensivamente en el dominio legal. Regularmente aparecen

[8] Horacio Rosatti, presidente de la Corte Suprema de Justicia de la Nación (CSJN), apertura del "XV Congreso Nacional de secretarios letrados y relatores de Cortes y Superiores Tribunales de Justicia Provinciales y CABA", STJER, 27/10/2022, accesible aquí

[9] Estos elementos se articulan mediante un discurso eminentemente técnico, que procede -a priori- a partir de una argumentación racional de la forma premisas-conclusión.

[10] https://temu.bsc.es/iberlegal/

pre-anotados junto a cada sentencia o fallo judicial como metadatos para su identificación. Constituyen datos protocolares que sirven como referencia para designar un documento legal particular. Como muestra el gráfico que agregamos abajo, los mismos normalmente incluyen: fecha de la sentencia, partes del proceso (actora y demandada), tipo de proceso, órgano que dictó la sentencia y jueces que firmaron la sentencia.

Figure 1: Suprema Corte de Buenos Aires: metadatos y sentecia judicial.

Hemos denominado a este subtipo de procesamiento NER como "*reconocimiento de metadatos*" o MER (*Metadata Entity Recognition*). Consideramos que lejos de ser una operación trivial plantea un avance significativo en materia de implementación pues resuelve un problema real: la generación manual de metadados identificatorios de sentencias. Al mismo tiempo, constituye un avance en materia de conocimiento por la generación de un *corpus legal semi-anotado* y el desarrollo de algoritmos de aprendizaje supervisado para una tarea NER no-convencional.

Finalmente, es preciso destacar que la tarea propuesta resulta fundamental en el ámbito institucional pues implica -nada más ni nada menos- la posibilidad de automatizar la extracción de metadatos identificactorios. Además, dicha extracción permitirá realizar tareas de más alto nivel como clasificación, comparación y búsquedas.

Objetivos

1. Construir un *corpus legal semi-anotado* con *datos identificatorios* de sentencias judiciales en base a la estrategia MER para probar soluciones de aprendizaje supervisado, y
2. Desarrollar distintos algoritmos de aprendizaje supervisado 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 requeridos para su ejecución y la convicción de su utilidad. Respecto de los medios requeridos cabe destacar que el área de investigación en NLP-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 lineales y reglas lingüísticas, hasta los actuales modelos no-lineales basados en redes neuronales profundas y respresentación distribuida de datos (*word/character embeddings*), existe un amplio espectro de enfoques para nutrir nuestros abordajes del problema ((Serrano et al. 2022b), (J. Li et al. 2020a)).

Aunque estos avances tienen como lenguaje objeto al 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 semi-anotado* para este proyecto tenemos evidencia de las múltiples opciones disponibles. En nuestro experimento de construcción de corpus legal[11] hemos trabajo con el portal de jurisprudencia del Poder Judicial de la provincia de Buenos Aires que brinda acceso público a fallos judiciales[12]. En esa experiencia construimos un dataset no anotado de 4299 sentencias judiciales y sus respectivos metadatos empleando el método de *scraping*. La información colectada está disponible en un repositorio de libre acceso.[13] Esta información consta de fallos judiciales completos y metadatos de la causa (materia, tipo de fallo, número de la causa, caratula, fecha de sentencia, magistrado y tribunal actuantes, entre otros), que brindarían -con el tratamiento apropiado- los insumos para una anotación parcial de entidades conforme al enfoque y objetivo propuesto en el presente trabajo.

Información similar a la obtenida de la justicia de Buenos Aires se publica por distintos Poderes Judiciales del país. Pese a las diferencias de formatos en general se repiten protocolos de publicación que permitirá disponer de un grupo de datos relativamente estable. Por esto 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[14]. El trabajo aborda el tópico NLP-NER aplicado al dominio legislativo, y el corpus empleado fue construido ad-hoc para la tesis mediante *scraping* del portal Infoleg[15] según se detalla en el punto 3.

Luego de exponer los argumentos acerca de la factibilidad del proyecto es preciso mencionar las razones que hacen a su relevancia. En tal sentido mencionamos que nuestro trabajo realizará un doble aporte al campo NLP en español. Por un lado crearemos y haremos público un *corpus semi-anotado* con *entidades identificatorias* de fallos judiciales. Este dataset no resolverá la necesidad de un corpus legal anotado que hoy se plantea en el dominio, pero sí dejará disponible un recurso que permitirá probar y desarrollar soluciones de

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

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

[13] https://github.com/castillosebastian/legal_corpus

[14] 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

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

menor escala ligadas a MER. Posibilidad que dada su vinculación con una necesidad institucional concreta y actual que viven las instituciones judiciales sin dudas será un aporte valioso para ellas. Por otro lado, buscaremos también realizar un aporte al campo NLP-MER en español aplicando algoritmos de aprendizaje profundo siguiendo los últimos avances en esta materia y publicando el código para su análisis y reutilización.

Respecto de la relevancia para las instituciones de justicia cabe señalar, en primer lugar, que el desarrollo de estrategias de NLP-MER brindará una herramienta de gran valor para el procesamiento de datos identificatorios de sentencias judiciales. El servicio de justicia se asienta sobre el intercambio de dicha información entre los distintos actores de un proceso: jueces, abogados y auxiliares de justicia. Ello sin mencionar a los intercambios que se generan en la red de instituciones públicas que participan de este servicio (i.e. Registros Públicos, Colegios Profesionales, Instituciones de Salud Pública y Asistenciales, entre otras). Contar con herramientas de NLP-MER aplicadas al ámbito legal permitiría automatizar el proceso de extracción de *datos identificatorios*, al tiempo que optimizará recursos en los Poderes Judiciales (Leitner, Rehm, and Moreno-Schneider 2019). Por ejemplo, proyectos vinculados a gestión electrónica, celeridad y regulación de flujos de información[16] se beneficiarán con el desarrollo de sistemas inteligentes que permitan la identificación automática de sentencias judiciales (Samy 2021).

En segundo lugar, desde una perspectiva institucional de más amplio alcance, el desarrollo de estrategias NLP-MER podría extenderse a otros documentos con metadatos disponibles. En este caso, la posibilidad de desarrollar algoritmos sensibles a los atributos lingüísticos del texto liberaría posibilidades de automatización sin precedentes, elevando la calidad del servicio de justicia. En efecto, los grandes procesos de reforma judicial en los últimos años han establecido deberes de actuación exigentes a los órganos de justicia que demandan una capacidad de procesamiento de información sin precedentes.[17] Este escenario es un espacio fértil para el desarrollo de las *tecnologías del lenguaje* que estudiaremos en este proyecto.

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 con el soporte de documentos administrativos no estucturados o semi-estructurados. Por eso, los avances en el procesamiento de documentos judiciales presentan gran potencial de transferencia hacia otros dominios de la administración pública, multiplicando los potenciales beneficios sociales que tiene nuestro desarrollo (Samy 2021).

Estado del arte

Teniendo en cuenta las detalladas reseñas elaborada por J. Li et al. (2020a) y Roy (2021) sobre la tarea NER en general, vamos a centrarnos en los avances generados respecto de aplicaciones al ámbito legal.

En lenguas distintas al español, podemos mencionar a Leitner, Rehm, and Moreno-Schneider (2019) que proponen aplicar dos modelos basados en *campos aleatorios condicionales* (*Conditional Random Fields* o CRFs) y *redes neuronales recurrentes de memoria de corto y largo plazo* (BLSTM). El corpus legal empleado en este estudio consistió en 750 sentencias en idioma alemán[18], anotadas manualmente. La tipología de clases empleada incluyó 19[19] categorías que pueden consultarse en su idioma original.[20] Para esas dos arquitecturas se probaron tres modelos, obteniendo los mejores resultados con BLSTM (F1 95.4/95.9). Por su parte Chalkidis et al. (2021) proponen un proyecto de extracción de datos específicos de contratos en inglés, formulando así un antecedente valioso para nuestro trabajo. Informan detallados experimentos -con optimización de hiperparámetros- para distintos modelos de redes neuronales recurrentes: LSTMs, DILATED-CNNs, TRANSFORMER y BERT. Reportan los mejores resultados con el primer modelo, emple-

[16] Desarrollos que en las instituciones de justicia adquieren cada vez mayor visibilidad e importancia - ver Soto, Andres, *Nuevas Tecnologías y Gerenciamiento de la Ofician Judicial,* publicado en *Nueva gestión judicial : oralidad en los procesos civiles,* Héctor M. Chayer [et al.] , CABA, Ediciones SAIJ, 2017.

[17] 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.

[18] Con 66.7 mil horaciones y 2.1 millones de tokens.

[19] Que podrían traducirse como: persona, juez/a, abogado/a, país, ciudad, calle, paisaje, organización, empresa, institución, Corte, marca, ley, ordenanza, Norma Legal de la Comunidad Europea, regulación, contrato, decisión judicial y bibliografía legal.

[20] https://github.com/elenanereiss/Legal-Entity-Recognition

ando representaciones distribuidas de palabras específicas del dominio (*word embeddings* mediante algoritmo *wor2vec*(Mikolov et al. 2013)).

Otras implementaciones relevantes encontramos en Pais et al. (2021) que plantean una arquitectura basada en *redes neuronales recurrentes* BLSTM y una capa final CRF, con representación distribuida de datos (*word/character embedings*), diccionarios y afijos en idioma Rumano. Como dato interesante, los investigadores evalúan ensambles de modelos y logran los mejores resultados de los experimentos (F1 90.36). Por su parte Cardellino et al. (2017) proponen un trabajo para documentos en inglés dirigido no solo a la tarea de reconocimiento y clasificación, sino también vinculación de entidades a una ontología (LKIF-Wikipedia)[21], reportando resultados en torno a F1 80% para distintos niveles de granularidad. La arquitectura experimentada en este trabajo incluyó *máquinas de soporte vectorial* (SVM), *redes neuronales* (NN) y redes entrenadas sobre imputs basados en representación distribuida de palabras *(word embeddings)*. Finalmente, otros proyectos interesantes atacan problemas específicos en la tarea NER: Barriere and Fouret (2019) proponen un modelo MICA (*May I Check Again*) para resolver errores de escritura y tipeo en entidades mediante la generación contextualizada de candidatos, y Skylaki et al. (2020) aborda la tarea NER en documentos PDF con pérdida de información (*noisy text*) empleando una estrategia que no implica anotación de entidades sino su generación *en* el texto.

Dentro de los pocos trabajos actuales aplicados al español vamos a encontrar a Samy (2021) que se enfoca en textos legislativos. Este trabajo parte de un corpus legislativo en español no anotado y emplea una metodología híbrida para su anotación según cada tipo de entidad de la siguiente forma: 1) las *normas* se anotan con expresiones regulares y listas de nombres oficiales, 2) las *fechas* mediante expresiones regulares, 3) los *organismos* mediante listas oficiales, 4) los *lugares* mediante listas y 5) las *personas* mediante librería de procesamiento spaCy y listas. Además de lo anterior se validó manualmente y de forma parcial la anotación (no se ofrecen detalles del alcance). El resultado de todo este esfuerzo es un texto enriquecido con 10+ millones de entidades marcadas en los 144.5 mil textos de Leyes-Decretos. El procesamiento fue realizado a partir de *redes neuronales convolucionales profundas* (*Deep CNN* implementadas por la librería spaCy v3) con representación distribuida de datos (Serrà and Karatzoglou 2017). En este trabajo se reporta un F1 general de 93%, aunque los resultados no son reproducibles pues no se publicaron datos ni algoritmos.

Otra experiencia aplicada al ámbito legislativo -esta vez en nuestro país- encontramos en la tesis de grado de Karen Haag[22], bajo la dirección de Cristian Cardellino. Aunque los resultados no son concluyentes por las dificultades afrontadas en la anotación del corpus, el trabajo se enfocó exclusivamente en la mención de normas (ley, decreto, resolución, etc.).

Desafíos

Como mencionamos antes, la falta de disponibilidad de un corpus legal anotado en español constituye una importante barrera para el trabajo. Dicha carencia no solo afecta la posibilidad material de implementar estrategias de aprendizaje supervisado, sino también impide evaluar las soluciones desarrolladas a falta de métricas de referencia para comparar resultados. Esta dificultad ha sido remarcada por investigadores en relación al español(Samy 2021; Samy, Arenas-García, and Pérez-Fernández 2020), pero también en otros idiomas (Leitner, Rehm, and Moreno-Schneider 2019).

Entendemos que nuestro enfoque de la tarea MER nos permitirá sortear este problema, y disponer de un corpus semi-anotado que permita entrenar nuestros algoritmos y evaluar su eficacia. En la sección metodológica expondremos las tareas que cumpliremos para esto.

Respecto de la posibilidad de emplear librerías especializadas para la tarea de anotación (e.g. spaCy, Stanza, entre otras) entendemos que los resultado no satisfactorios informados por otros estudios son una alerta

[21] Tarea que en el ámbito legal es particularmente necesaria dadas las relaciones fuertemente estructuradas y jerarquías que predominan en este dominio, aunque su realización implica un alto costo en tiempo y recursos para desarrollarse y mantenerse.

[22] Trabajo realizado en la Universidad Nacional de Córdoba, FAMAF. En este proyecto se elaboró un corpus ad-hoc obtenido de InfoLEG y se enfocó la tarea NER en la detección de tipos de normas. Para ello se emplearon expresiones regulares, y un anotador semántico basado en CRF (disponible en la librería en Java, Stanford NER-CRF).

importante a tener en cuenta (Samy 2021). En cualquier caso, buscaremos evaluar su implementación en caso de ser empleadas.

Finalmente, existen restricciones legales que deben tenerse en cuenta para el tratamiento de datos personales (Ley 25326). En este punto cabe tener presente que la información que emplearemos en nuestro estudio es información pública, suministrada por órganos oficiales y sus respectivos portales institucionales. No obstante lo cual, el desarrollo de algoritmos y sus resultados tendrá especialmente en cuenta que los mismos no impliquen sesgos potencialmente perjudiciales para personas o grupos sociales.

Metodología

Dividiremos nuestro trabajo en dos partes asociadas a los dos objetivos propuestos: la primera parte resolveremos la construcción de un *corpus legal semi-anotado* y en la segunda desarrollaremos distintos algoritmos de aprendizaje supervisado para el tratamiento de dicho corpus.

Primera Parte: *construcción de corpus semi-anotado*

Para armar nuestro corpus buscaremos ampliar la base de datos disponible de 4299 fallos judiciales con un nuevo proceso de extracción de datos web o *scraping.* Emplearemos para ello Python 3.6, con las librerías *beautifulsoup4* y *Selenium.*

Organizaremos el proceso de extracción de información pública de tal manera de respetar los parámetros de diseño del servicio que ofrece la sección de Jurisprudencia de la Corte Suprema de Justicia de Buenos Aires. Aunque el mismo no solicita identificación agregaremos datos de identificatorios al momento de efectuar la conexión y fijaremos horario de extracción entre las 22 y las 23.30 (banda horaria de tráfico reducido en los servicios públicos web).

Para anotar nuestro corpus emplearemos los *metadatos identificatorios de las sentencias judiciales* obtenidas en el paso anterior. Los cinco tipos de *entidades* que buscaremos anotar son:

- "fecha de sentencia": referencia cronológica del acto judicial,
- "partes del proceso": generalmente actora(s) y demandada(s). Personas (físicas o jurídicas) que intervienen en una causa judicial y dirección del vínculo judicial.
- "tipo de proceso": tipología del proceso según normativa aplicable.
- "órgano que dictó la sentencia": referencia a la estructura orgánica judicial y al mismo tiempo ubicación geográfica de la ciudad asiento del órgano.
- "jueces de sentencia": magistrados que elaborar la sentencia.

La selección de estos metadatos se basa en su amplia disponibilidad e importancia como información identificatoria. Los protocolos empleados en el dominio judicial para la cita de precedentes contemplan normalmente a este grupo de datos para individualizar un caso, por lo que su empleo es extendido. Además de ello, como se puede advertir de la breve descripción ofrecida, cada metadato aporta información valiosa de distintas dimensiones del caso, agregando valor semántico a la anotación.

Para la tarea de anotación normalizaremos[23] el texto y los metadatos, teniendo en mira especialmente que la recuperación de referencias textuales y su ulterior asignación a una entidad sea precisa y exhaustiva. Toda sentencia judicial incluye un universo de múltiples referencias, expresiones polisémicas, ambigüedad e indeterminación (Samy 2021). Por ello, una premisa para la tarea de asignación *token = entidad*, será

[23] Usualmente refiere a la tarea de dividir un documento en oraciones, separar las palabras en tokens (unidades lingüísticas), corregir y/o estandarizar formatos.

minimizar los errores y maximizar la recuperación de información. En ese sentido utilizaremos dos estrategias de asignación basadas en *coincidencia exacta* y *coincidencia parcial* respectivamente que nos permitirá enriquecer la descripción de los metadatos.[24] Seleccionaremos aquella con el puntaje F1 más alto (ver definición de métricas más abajo).

Segunda Parte: *desarrollo de modelos*

El primer paso en esta etapa será elaborar un modelo base y una medida de desempeño inicial (*baseline*) para la tarea propuesta. En este caso emplearemos el corpus semi-anotado sin creación de nuevos atributos, y con arquitectura de redes neuronales recurrentes (RNN). Esta arquitectura está específicamente diseñada para problemas de predicciones de secuencias (como son los datos textuales) (Lample et al. 2016).

Seguidamente iniciaremos dos rondas secuenciales de experimentos en las que buscaremos enriquecer los datos de entrada para nuestros modelos y la arquitectura de procesamiento según las propuestas más prometedoras (Ma and Hovy 2016; Lample et al. 2016; Leitner, Rehm, and Moreno-Schneider 2019). En general, J. Li et al. (2020b) señalan que las arquitecturas de aprendizaje profundo empleadas para la tarea NER suponen estos elementos:

- imputs o datos enriquecidos con distintas estrategias (*word/character embeddings,* etiquetado de partes del discurso (POS-tag), disccionarios, entre otros),
- codificador de contexto (*context encoder)* con distintos modelos de redes neuronales: CNN, RNN, *Language Model*, *Transformer*.
- decodificador de categorías o etiquetas (*tag decoder)* con distintas alternativas softmax, CFR, RNN, entre otras.

Por nuestra parte, siguiendo a Leitner, Rehm, and Moreno-Schneider (2019) y X. Li et al. (2020), iniciaremos nuestra primer ronda de experimentos con los siguientes modelos:

- BLSTM-CRF, y
- *Transformer*

Ambos modelos tienen arquitecturas diferentes. Mientras que los modelos BLSTM se basan en *redes neuronales recurrentes* (RNN) con complejas cajas de procesamiento (que codifican y decodifican sus imputs), los modelos basados en *Transformers* (Vaswani et al. 2017) emplean capas de procesamiento que utilizan un mecanismo de auto-atención (*self-attentions*). Los primeros constituyen la opción más común dentro de los modelos de aprendizaje profundo para tareas NER (J. Li et al. 2020a). Los últimos han mostrado mejores resultados que los primeros en ciertos contextos donde se procesan muchos datos como grandes corpus (X. Li et al. 2020). En cualquier caso, ambos modelos están fuertemente determinados por la representación de los datos o imputs, por lo que buscaremos incorporar formas de representación distribuida de datos (*word/character embeddings)* y modelos de lenguaje pre-entrenados. Recursos valioso para esta tarea serán los modelos de lenguaje en español legal presentados por Samy, Arenas-García, and Pérez-Fernández (2020) y Gutiérrez-Fandiño et al. (2021).

En base a los resultados obtenidos en esta primer ronda realizaremos una nueva ronda de experimentos buscando mejorar los resultados obtenido con las primeras configuraciones.

Para todo este trabajo utilizaremos Python 3, y dependencias asociadas a la tarea (e.g. spaCy (Neumann et al. 2019), T-NER (Ushio and Camacho-Collados 2021), entre otras). Para el entrenamiento y evaluación de los modelos propuestos emplearemos, siempre que sea posible, las plataformas Google Colab y Kaggle que brindan acceso gratuito a GPU's de alto rendimiento.

[24] Dado que las entidades identificatorias son metadatos que reproducen datos de una sentencia, pueden presentarse bajo una forma de expresión de menor valor expresivo (eg. abreviaturas) que el dato original. En este caso, generaremos una versión enriquecida del metadato.

Evaluación

Las métricas son importante en los problemas de aprendizaje automático permitiéndonos cuantificar el rendimiento de nuestros modelos. Por eso, detallaremos a continuación las formulas que empleamos en la evaluación.

Considerando que:

- Positivos verdaderos (TP) son aquellas predicciones que realiza el modelo como positivas y y realmente lo son.
- Negativos verdaderos (TN) son predicciones que realiza el modelo como negativas y realmente lo son.
- Positivos falsos (FP) son aquellas predicciones que realiza el modelo como positivas, pero en realidad son negativas.
- Negativos falsos (FN) son aquellas predicciones que realiza el modelo como negativas, pero en realidad son positivas.

Nuestras métricas son:

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precisión = \frac{TP}{TP + FP}$$

$$Exahustividad = \frac{TP}{TP + FN}$$

$$F1Score = 2\frac{Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

Además de estas métricas que usan coincidencias exactas para evaluar desempeño (*exact match*), utilizaremos métricas que usan coincidencias indulgentes (*lenient match*). Estas métricas se calculan con el índice de *Jaccard*, empleado para medir el grado de similitud entre las entidades del texto de referencia y las entidades predichas. La formula del índice es la siguiente:

$$J(ref, pred) = \frac{overlap(ref, pred)}{length(ref) + length(pred) - overlap(ref, pred)}$$

Donde J mide la relación entre la intersección y la unión de las entidades de la referencia y de la predicción. Lograr el valor 1 indicaría que se encuentra el grado máximo de similitud entre la referencia y la predicción, es decir, los límites de las entidades coinciden exactamente.

Cronograma

mes	actividad	descripción
1	Extracción datos web	Extracción de información web para ampliar corpus disponible
2-3	Creación corpus	Creación corpus semi-anotado

4	Habilitación infraestructura	Se creará la infraestructra para los experimentos y registro para reproducibilidad
5	Modelo base	Creación del modelo base
6-7	Experimentos Ronda 1	Primer Ronda de Experimentos, análisis y evaluación
8-9	Experimentos Ronda 2	Segunda Ronda de Experimentos, análisis y evaluación
10-11-12	Redacción de Tesis	Elaboración del documento de Tesis

Barriere, Valentin, and Amaury Fouret. 2019. "May i Check Again? – a Simple but Efficient Way to Generate and Use Contextual Dictionaries for Named Entity Recognition. Application to French Legal Texts," September. http://arxiv.org/abs/1909.03453.

Cardellino, Cristian, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. "A Low-Cost, High-Coverage Legal Named Entity Recognizer, Classifier and Linker." In, 22. Londres, United Kingdom. https://hal.archives-ouvertes.fr/hal-01541446.

Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2021. "Neural Contract Element Extraction Revisited: Letters from Sesame Street," February. http://arxiv.org/abs/2101.04355.

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. https://web.stanford.edu/~jurafsky/slp3/.

Kiela, Douwe, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, et al. 2021. "NAACL-HLT 2021." In, 41104124. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.324.

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. "Neural Architectures for Named Entity Recognition," April. http://arxiv.org/abs/1603.01360.

Leitner, Elena, Georg Rehm, and Julian Moreno-Schneider. 2019. "Fine-Grained Named Entity Recognition in Legal Documents." In, 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.

Li, Jing, Aixin Sun, Jianglei Han, and Chenliang Li. 2020a. "A Survey on Deep Learning for Named Entity Recognition," March. http://arxiv.org/abs/1812.09449.

———. 2020b. "A Survey on Deep Learning for Named Entity Recognition," March. http://arxiv.org/abs/1812.09449.

Li, Xiaoya, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. "A Unified MRC Framework for Named Entity Recognition," May. http://arxiv.org/abs/1910.11476.

Ma, Xuezhe, and Eduard Hovy. 2016. "End-to-End Sequence Labeling via Bi-Directional LSTM-CNNs-CRF," May. https://doi.org/10.48550/arXiv.1603.01354.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space," September. http://arxiv.org/abs/1301.3781.

Neumann, Mark, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing." In, 319–27. https://doi.org/10.18653/v1/W19-5034.

Pais, Vasile, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. "Named Entity Recognition in the Romanian Legal Domain." In, 918. Punta Cana, Dominican Republic: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.nllp-1.2.

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*, January, 103–14. https://doi.org/10.26342/2021-67-9.

Samy, Doaa, Jerónimo Arenas-García, and David Pérez-Fernández. 2020. "Legal-ES: A Set of Large Scale Resources for Spanish Legal Text Processing." In, 3236. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lt4gov-1.6.

Serrà, Joan, and Alexandros Karatzoglou. 2017. "Getting Deep Recommenders Fit: Bloom Embeddings for Sparse Binary Input/Output Networks," June. http://arxiv.org/abs/1706.03993.

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. 2022b. "RigoBERTa: A State-of-the-Art Language Model for Spanish," June. http://arxiv.org/abs/2205.10233.

———. 2022a. "RigoBERTa: A State-of-the-Art Language Model for Spanish," June. http://arxiv.org/abs/2205.10233.

Skylaki, Stavroula, Ali Oskooei, Omar Bari, Nadja Herger, and Zac Kriegman. 2020. "Named Entity

Recognition in the Legal Domain Using a Pointer Generator Network," December. http://arxiv.org/abs/2012.09936.

Ushio, Asahi, and Jose Camacho-Collados. 2021. "T-NER: An All-Round Python Library for Transformer-Based Named Entity Recognition." In, 5362. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-demos.7.

Vajjala, Sowmya, and Ramya Balasubramaniam. 2022. "What Do We Really Know about State of the Art NER?" May. http://arxiv.org/abs/2205.00034.

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.