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 ha sido una barrera infranqueable para el avance en 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.

Introducción

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, n.d.). Normalmente consiste en la identificación automática -sin intervención humana- de entidades nombradas en textos[2] y su asignación a determinadas categorías semánticas. Estas unidades refieren generalmente a entidades que puede aludirse mediante *nombres propios* -e.g. personas, organizaciones y localizaciones- aunque en la práctica se han extendido para incluir expresiones numéricas referidas a fechas y cantidades, o distintos tipos de entidades que varían 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 dificultad importante (Jurafsky and Martin 2021).

[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.

Según Li et al. (2020) podemos agrupar las técnicas aplicadas en NER en cuatro tipos, a saber: 1) enfoques basados en reglas que no necesitan datos anotados porque descansan en la formulación particular de reglas lingüísticas, 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], y 4) enfoques de aprendizaje profundo[4] y representación distribuida de datos (*Embeddings*) capaces de generar automáticamente la representación óptima de la información disponible para el ulterior tratamiento mediante aprendizaje supervisado.

Estos últimos enfoques han generado un cambio de paradigma en el NLP en general y en las tareas NER en particular gracias a los buenos resultados obtenidos (Roy 2021). Estas nuevas arquitecturas optimizan los procesos de aprendizaje (Serrano et al. 2022), mejorando las métricas de evaluación para distintas tareas de análisis semántico. Incluso han logrado resultados que igualan a los obtenidos por un ser humano (Kiela et al. 2021). Junto con ese nuevo conocimiento, nuevos proyectos públicos[5] y privados[6] con eje en las *tecnologías del lenguaje* han ganado presencia en el campo científico y en el discurso público, colocando al NLP como tópico de interés general.

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.

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"*.[7]

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.[8] 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

[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]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 o IBERLEGAL dirigidos a fomentar el desarrollo del procesamiento del lenguaje natural y la traducción automática en lengua española y lenguas cooficiales. A nivel internacional el Proyecto de la Comunidad Europea MIREL dirigidos a promover el avance de las *tecnologías del lenguaje,* similiar a MARCELL, entre otros.

[6]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.

[7]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

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

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__low-cost__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__fine-grained__2019?**) , Serrano et al. (2022), 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*[9] 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* contituyen entidades *per se*, empleadas intensivamente en el dominio legal. Regularmente aparecen 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.

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

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

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

2. Desarrollar 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 científicos y tecnológicos requeridos para su ejecución.

Respecto del conocimiento requerido 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 embedings*), existe un amplio espectro de enfoques para nutrir nuestros abordajes del problema [Serrano et al. (2022)](Li et al. 2020).

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 [Gutiérrez-Fandiño et al. (2021)](Gutiérrez-Fandiño et al., n.d.)(Serrano et al. 2022), 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[10] hemos trabajo con el portal de jurisprudencia del Poder Judicial de la provincia de Buenos Aires que brinda acceso público a fallos judiciales[11]. 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 el repositorio https://github.com/castillosebastian/legal_corpus. Esta información consta de fallos judiciales completos y metadatos de la causa (materia, tipo de fallo, número de la causa, caratula, fecha

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

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

de sentencia, magistrado y tribunal actuantes, entre otros), que brindarían 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[12]. El trabajo aborda el mismo tópico que estamos presentando en este documento NLP-NER aplicado al dominio legal, y el corpus empleado fue construido ad-hoc para la tesis mediante *scraping* del portal Infoleg[13] 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.

Respecto de la relevancia del trabajo para el campo NLP-NER esperamos hacer un doble aporte. Por un lado creando y haciendo público el 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 menor escala ligada a MER. En cualquier caso, la reducción del alcance que estamos proponiendo lejos de ser caprichosa, se relaciona con una necesidad institucional concreta y actual que viven las instituciones judiciales. Además de esto, en segundo lugar, buscaremos realizar un aporte al campo NLP-MER en español aplicando algoritmos de aprendizaje profundo siguiendo los últimos avances en esta materia.

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 la red de instituciones públicas que interactuan permanentemente en la materialización de los actos de 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 desarrollar nuevos productos y servicios para las distintas partes interesadas, al tiempo que optimizaría los recursos disponibles en los Poderes Judiciales (**leitner_fine-grained_2019?**). Por ejemplo, proyectos vinculados a gestión electrónica, celeridad y regulación de flujos de información[14] se podrían beneficiar drásticamente 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 que se brinda al ciudadano. En efecto, los grandes procesos de reforma judicial los últimos años han establecido deberes de actuación exigentes a los órganos de justicia. Principios como la *oficiocidad, celeridad, concentración* y *plazo razonable* [15] demandan una capacidad de procesamiento de información sin precedentes. [16] 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 docu-

[12] 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

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

[14] 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.

[15] 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 constatar en las demás jurisdicciones provinciales.

[16] 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.

mentos 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 Li et al. (2020) 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, anotadas manualmente. La tipología de clases empleada incluyó 19[17] categorías que pueden consultarse en su idioma original.[18] Para esas dos arquitecturas se probaron tres modelos, obteniendo los mejores resultados con BiLSTM (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, empleando representaciones distribuidas de palabras específica del dominio (*word embedings* mediante algoritmo *wor2vec*).

Otras implementaciones vinculadas al dominio 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 (palabras y caracteres), diccionarios y afijos en idioma Rumano. Como dato interesante, estos 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 dirigido no solo a la tarea de renocimiento y clasificación de entidades sino también su vinculación a una ontología (LKIF-Wikipedia)[19], reportando resultados en torno a F1 80% para distintos niveles de granularidad. La arquitectura experimentada en este trabajo incluyó a Máquinas de Soporte Vectorial (SVM), redes neuronales (NN) y redes entrenadas sobre imputs basados en representación distribuida de palabras (*Embedings).* 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: 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 esfuero 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 libreria spaCy v3) con uso de representación distribuida (Serrà and Karatzoglou 2017). En este trabajo se reporta un F1 general de 0.93, alcanzando 0.97/.098 para entidades como leyes y fechas respectivamente.

Otra experiencia aplicada al ámbito legislativo -esta vez en nuestro país- encontramos en la tesis de grado de Karen Haag[20], bajo la dirección de Cristian Cordellino. Aunque los resultados no son concluyentes por

[17]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.

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

[19]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.

[20]Trabajo realizado en la Universidad Nacional de Córdoba, FAMAF. En este proyecto se elaboró un corpus ad-hoc obtenido

las dificultades afrontadas en la anotación del corpus, el trabajo se enfocó exclusivament 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 [(**samy2020:p4?**)](Samy 2021), pero también en otros idiomas (**leitner_fine-grained_2019?**).

Entendemos que nuestro enfoque de la tarea MER nos permitirá sortear esta dificultad, y disponer de un corpus semi-anotado que permita entrenar nuestros algoritmos y al mismo tiempo evaluar su eficacia.

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 importante a tener en cuenta (Samy 2021). En cualquier caso, buscaremos evaluar su implementación en caso se 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, evaluaremos que el desarrollo de algoritmos y sus resultados no impliquen sesgos potencialmente perjudiciales para personas o grupos sociales.

Metodología

baseline (ver @ Francesca L)

Ver Samy (2021)

PreAnotación

- Tipología de categorías semanticas acorde al dominio.
- expresiones regulares: ej. fechas
- listas: leyes, organismos, lugares (países, comunidades autónomas, provincias y localidades, tipos de vía, etc. Para utilizar estas listas, fue necesario un proceso de depuración porque presentaban los mismos problemas señalados anteriormente. Además, se ha optado por excluir algunos nombres de localidades por su ambigüedad y por el posible ruido que puede causar en forma de falsos positivos. Por ejemplo, se han eliminado de la lista localidades como “María”, “Javier” o “Caso”.)
- nombres propios (SGP), cargos y puestos

Entrenamiento de Modelo

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

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).

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_fine-grained_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_fine-grained_2019?**)

Taxonomía de las categorías: 19! ver (**leitner_fine-grained_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

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.

Gutiérrez-Fandiño, Asier, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. n.d. "MarIA: Spanish Language Models," 22.

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.

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

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. 2022. "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.

Vajjala, Sowmya, and Ramya Balasubramaniam. n.d. "What Do We Really Know about State of the Art NER?" 11.