Procesamiento del Lenguaje Natural, Revista nº 67, septiembre de 2021, pp. 103-114 recibido 10-05-2021 revisado 07-06-2021 aceptado 08-06-2021

Reconocimiento y clasificación de entidades nombradas en textos legales en español

Named Entities Recognition and Classification in Spanish Legal Texts

Doaa Samy
Cairo University, Giza, Egypt
Instituto de Ingeniería del Conocimiento (IIC), Madrid, Spain
doaasamy@cu.edu.eg

Resumen: El reconocimiento y la clasificación de las entidades nombradas (NER/NERC) es una tarea principal en las áreas del Procesamiento del Lenguaje Natural (PLN) y la Extracción de la Información. El papel de NERC en el dominio legal es imprescindible en el desarrollo de sistemas legales inteligentes. El presente trabajo pretende dar un primer paso hacia establecer un "baseline" para la tarea NERC en el español jurídico. El objetivo principal consiste en proporcionar un recurso lingüístico anotando cinco tipos básicos de entidades nombradas en los textos legislativos en español peninsular. Los cinco tipos de entidades nombradas son: Personas, Organizaciones, Lugares, Fechas absolutas y Referencias a leyes, decretos, órdenes, normativas y artículos. Se adopta una metodología híbrida que reúne tres técnicas principales: Patrones de expresiones regulares, listas de fuentes externas y el entrenamiento de tres modelos NERC utilizando la librería abierta spaCy v3. De los tres modelos entrenados, el mejor ha obtenido un f-score de 0.93 alcanzando en algunos tipos como las menciones a leyes o fechas valores de 0.98 y 0.97 respectivamente. El peor de los modelos ha alcanzado una media de f-score de 0.85 que sigue siendo un resultado satisfactorio comparado con el estado de la cuestión.
Palabras clave: Entidades Nombradas, Procesamiento de textos legales, Procesamiento del español jurídico, Extracción de la información en textos legales.

Abstract: Named Entity Recognition and Classification (NER/NERC) is a major task in Natural Language Processing (NLP) and Information Extraction (IE). In the legal domain, NERC is indispensable in developing legal intelligent systems. This study pretends to take a first step towards a baseline for Spanish NERC in the legal domain. The main objective is to provide a linguistic resource by annotating five basic categories of Named Entities in Spanish legislative texts. These five categories are Person, Organization, Location, Dates (absolute expressions) and, finally References to aws, decrees, regulations, etc. To achieve this goal, we adopt a hybrid approach by combining three techniques: hand-crafted patterns through regular expressions, look-up lists and training of three NERC models using the architecture of spaCy. The best model achieved a general f-score of 0.93 with some types of entities such as Legal entities and Dates reaching up to 0.98 and 0.97 respectively. The worst model achieved a general f-score of 0.85, which is still satisfactory given the state of the art.
Keywords: Named Entities, Legal Text Processing, Information Extraction in Legal Texts, Spanish Legal Text Processing.

1 Introducción

El reconocimiento y la clasificación de las entidades nombradas (en adelante NER/NERC por las siglas en inglés: *Named Entity Recoginition and Classification*) es una tarea principal en las áreas del Procesamiento del Lenguaje Natural (PLN) y la Extracción de la Información. El término de entidades nombradas fue acuñado por primera vez en la serie de los congresos MUC (*Message Understanding Conference*) en el año 1995 para referirse al proceso de extraer unidades relevantes de información a partir de textos no-

ISSN 1135-5948. DOI 10.26342/2021-67-9

estructurados (Sekine, 2004) (Nadeau y Sekine, 2007). Estas unidades incluyen nombres propios de personas, organizaciones, lugares o expresiones numéricas como fechas o cantidades, etc.

Dada su relevancia en el análisis semántico, la tarea NERC se ha convertido en una piedra angular para aplicaciones inteligentes como los sistemas de Pregunta-Respuesta (QA), la generación de resúmenes automáticos, la mejora de los sistemas de recuperación de la información, la traducción automática, la anonimización de textos, la generación de grafos de conocimientos, etc.

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

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

Este interés se fundamenta en el gran potencial de las técnicas de PLN y su capacidad de ofrecer soluciones inteligentes que beneficien a usuarios claves del sector como los abogados, los jueces, los juristas, los documentalistas jurídicos, además del sector de la administración pública que, aunque no trate textos estrictamente jurídicos, sí maneja textos administrativos con un alto contenido legal como es el caso de la contratación pública o los convenios.

Por tanto, los avances en el procesamiento de textos legales constituyen un gran potencial para agilizar procesos internos de la administración pública, simplificar los procedimientos y mejorar el acceso de la ciudadanía a la información legal y administrativa. Para impulsar la apertura de datos públicos, la transformación digital inteligente y la agilización de procesos administrativos, legales y judiciales, existen iniciativas y programas a nivel europeo como el portal de e-Justice. Además, el programa de Europa Digital "Digital Europe Programme" pone énfasis en el papel de la inteligencia artificial en la administración pública para mejorar la interacción digital entre ciudadanos y administración pública.

Son numerosas las soluciones inteligentes que puede ofrecer el PLN, en general, y la tarea NERC, en particular, al ámbito legal y administrativo. Identificar sentencias parecidas para fundamentar un caso, enlazar documentos a través de las entidades, construir grafos de documentos, anonimizar datos personales o generar una línea temporal de las leyes y los hechos en un documento legal son solo algunos ejemplos de cómo el PLN y NERC pueden asistir tanto a abogados, juristas como a jueces en realizar sus tareas.

Pese a las oportunidades que supone el dominio legal, son pocos los estudios, recursos y herramientas de PLN en este dominio, sobre todo en español. Sin embargo, en los últimos años, han aparecido algunas iniciativas. En diciembre de 2019 y con el fin de impulsar el desarrollo de recursos y herramientas de PLN en el dominio legal en español, catalán, vasco y gallego, se organizó la jornada "IberLegal" dentro del marco de las actividades del Plan español de Tecnologías del Lenguaje. Las actas de la jornada ofrecen un abanico de temas de interés como la extracción de terminología legal, búsquedas inteligentes en documentos y recuperación de información legal, herramientas para asistir a la ciudadanía en la redacción de textos para la administración pública y, por último, expresiones temporales en textos legales (PlanTL, 2019).

Otra iniciativa es el corpus Legal-ES (Samy et al. 2020), considerado como un meta corpus que reúne varias fuentes del dominio en lengua española con más de dos mil millones de palabras recopiladas a partir de fuentes de datos abiertos españolas, europeas, hispanoamericanas e internacionales. Estas fuentes representan una variedad de textos jurídicos que incluyen textos legislativos, jurisprudenciales (sentencias) y textos administrativos. Además, el estudio presenta resultados preliminares sobre cálculos de *Embeddings* del español jurídico y un modelo

de tópico entrenado sobre el conjunto de legislación.

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.

El trabajo se estructura en ocho secciones además de la introducción y las conclusiones. Las primeras secciones presentan un enfoque teórico con análisis del estado de la cuestión de la tarea de NERC legal, en general y la NERC legal en español. El resto de las secciones se centran en aspectos prácticos donde se describe el trabajo y los experimentos realizados detallando el alcance, la metodología, los datos utilizados y las fases del estudio desde la extracción de los datos, pasando por la pre-anotación, la validación, el entrenamiento hasta la evaluación y la visualización final de los resultados obtenidos por los modelos entrenados.

2 Estado del arte: NERC en el dominio legal

Para ofrecer una visión panorámica acerca del desarrollo de los estudios de NERC en el dominio legal, subrayamos algunas iniciativas y estudios en esta área teniendo en cuenta tres criterios: a) La **lengua** objeto de análisis, b) el **tipo de texto legal** (textos legislativos, textos jurisprudenciales (sentencias), resoluciones, contratos, convenios, registros legales, etc.) y c) la **evolución de las técnicas**.

En primer lugar, destacamos los estudios que han tratado el tema en otras lenguas y, en segundo lugar, nos centramos en los estudios que han abordado la tarea en español.

2.1 NERC de textos legales en otras lenguas

El año 2006 marca la celebración de la primera tarea de evaluación dedicada a la recuperación de textos en el dominio legal "TREC Legal", organizada por el Instituto Estadounidense de Estándares y Tecnología (NIST) (Cormack et al., 2010). Desde esa fecha, el interés por las entidades nombradas en los textos legales ha seguido cobrando mayor importancia por su relevancia en la extracción y la recuperación de la información.

En 2010 se publicó el volumen titulado "Semantic Processing of Legal Texts" (Francesconi et al., 2010) incluyendo uno de los estudios pioneros sobre la identificación y resolución de las entidades nombradas en textos legales en **inglés** (Dozier et al., 2010).

En cuanto a técnicas, en esos años dominaban los modelos de aprendizaje automático clásico y se combinaban con reglas. Dozier et al. (2010) empleaban reglas, listas y modelos estadísticos (SVM) para reconocer jueces, abogados, empresas, tribunales y áreas de jurisdicción alcanzando valores de *f-score* por encima del 0,90. El estudio se centra en textos de jurisprudencia basada en casos (*case law*), deposiciones/declaraciones, juicios y alegatos.

Siguiendo la misma aproximación, Quaresma y Gonçalves (2010) proponen utilizar rasgos sintácticos combinando el uso de un analizador sintáctico (parser) con un modelo SVM para reconocer nombres, lugares, fechas y referencias a documentos en textos de **convenios internacionales** del corpus europeo Eur-lex en **cuatro lenguas**: inglés, alemán, portugués e italiano.

Por otro lado, en esos años, se observa el auge de las ontologías y los datos enlazadas. Landthaleret al., (2016) generan redes de grafos de **textos legislativos** del **código civil alemán** basándose en entidades nombradas. Cordellini et al. (2017) extraen las entidades nombradas a partir de un conjunto de textos formados por **juicios** de la Corte Europea de Derechos Humanos y la Wikipedia en **inglés** con el fin de enriquecer una ontología.

Los modelos de aprendizaje automático clásico y las aproximaciones híbridas siguen dominando el panorama en el tratamiento de textos legales, Chalkidis et al., 2017 aplican clasificadores lineales (*Logistic Regression* y

SVM) junto a reglas para el reconocimiento de entidades en **contratos en inglés**. Asimismo, Glaser et al. (2018) integran diferentes aproximaciones para el reconocimiento y la desambiguación de entidades nombradas en **contratos** en **alemán**. Por último, Andrew y Tannier (2018) combinan un modelo estadístico (CRF) y reglas lingüísticas para identificar entidades nombradas y enlazarlas en textos de **registros legales en francés**.

El paso de los modelos estadísticos y el aprendizaje automático clásico al aprendizaje profundo fue inminente en el panorama del PLN. Por eso, siguiendo esta línea, Chalkidis and Androutsopoulos (2017) aplican modelos de aprendizaje profundo utilizando un modelo de BiLSTM (*Bidirectional Short Term Memory*) para identificar y clasificar 11 tipos de elementos en **contratos** en **inglés** obteniendo resultados que superan los estudios anteriores que aplicaban SVM. Los modelos de aprendizaje profundo ya se consolidan y superan los resultados de los modelos clásicos. Leitner et al. (2019) proponen una metodología para el reconocimiento de un total de diecinueve tipos de entidades nombradas agrupadas en siete clases generales. La tarea NERC fue aplicada a un conjunto de **sentencias alemanas** con *Conditional Random Fields* (CRFs) y BiLSTMS. Los modelos de BiLSTM han alcanzado mejores resultados con un f-score de 0,9546 para el conjunto de 19 tipos y 0.9595 para el conjunto de las 7 clases generales.

2.2 NERC de textos legales en español

Son pocos los estudios que han abordado la tarea NERC en el dominio legal en español. Sin embrago, uno de los trabajos pioneros es el estudio de Martínez-González et al. (2005) que tiene como objetivo automatizar la extracción de referencias en textos legales, su resolución y su indexación mediante reglas y gramáticas para mejorar la recuperación de la información en este dominio. El estudio se ha realizado sobre colecciones de documentos de una editorial de textos legales, pero no queda claro qué tipo de documentos son si son legislativos o de otra categoría.

Pasan muchos años hasta que empiecen a aparecer otros estudios como Badji (2018), en el que se presenta una aproximación basada en reglas y patrones para reconocer y enlazar entidades legales (referencias a leyes, decretos, sentencias, cortes o fases judiciales) en fuentes legislativas españolas y fuentes no oficiales como las noticias y redes sociales donde aparecen con otras denominaciones populares (Ej. El caso de la "Ley Celaá").

En la misma línea, Rodríguez-Doncel et al. (2018) han transformado un conjunto de la legislación española en "Linked Data" en formato RDF basándose en identificar entidades (leyes, organismos/empresas y lugares) en los textos del Boletín Oficial del Estado (BOE) y enlazarlas con fuentes externas como la Wikipedia, Wikidata, el Registro Europeo de Autoridades, el Directorio Común de Unidades Orgánicas y Oficinas españolas o la base terminológica europea Eurovoc.

Por otro lado, Haag (2019) adopta una metodología híbrida combinando modelos estadísticos y reglas para el reconocimiento de las entidades legales en las fuentes legislativas argentinas utilizando el corpus de legislación argentina InfoLEG, en el que se recogen los textos del Boletín Oficial de la República Argentina.

Por último, Navas-Loro (2020) y (Navas-Loro y Rodríguez-Doncel, 2019), basándose en reglas y gramáticas, han abordado las expresiones temporales en textos legales desarrollando "TimeLex" para la detección, la normalización en TimeML y la resolución de las expresiones temporales en textos legales.

Tras este análisis del estado del arte, se puede observar la relevancia de la tarea NERC en el dominio legal y la evolución de las técnicas empleadas a lo largo de los años. Sin embargo, los esfuerzos y los recursos desarrollados en español se centran en tipos concretos de entidades nombradas como las referencias a leyes o las expresiones temporales desde una perspectiva vertical enfocada en unas entidades específicas. La única excepción, a este respecto, es el estudio sobre la conversión de la legislación española a "Linked Data" (Rodríguez-Doncel et al., 2018) donde se ha tratado varios tipos de entidades en los textos legislativos. No obstante, las entidades nombradas no constituyen el objetivo principal, dado que el enfoque se centra en enlazar los datos.

Partiendo de estas observaciones, se echa en falta una aproximación transversal para las entidades nombradas en el dominio legal español que aborda la tarea a gran escala. Este planteamiento integral es imprescindible como

punto de partida para establecer un *baseline* de la tarea NERC en el español jurídico. Además, ayudará a analizar aspectos principales como: los tipos de entidades más frecuentes, su distribución a través de los distintos tipos de texto legal, por ejemplo, entre textos legislativos y textos jurisprudenciales, etc.

De ahí, el presente trabajo pretende dar este primer paso hacia la tarea NERC en el español jurídico. No obstante, para conseguir un *baseline* objetivo es necesario unir esfuerzos y contrastar aproximaciones mediante tareas de evaluación (nuestro próximo objetivo) o iniciativas y proyectos interdisciplinarios e inter-institucionales como el proyecto europeo LYNX (Rehm et al., 2019) o los recursos desarrollados por Stanford Codex "*Stanford Center for Legal Informatics*" (Waltl y Vogl, 2018) (Rios, 2019).

3 Alcance y metodología

El objetivo de este estudio es anotar cinco tipos básicos de entidades nombradas en los textos legislativos en español peninsular. Los cinco tipos de entidades nombradas son:

- Personas
- Organizaciones
- Lugares
- Referencias a leyes, decretos, órdenes, normativas y artículos.
- Fechas absolutas entendidas como expresiones temporales referentes a fechas y días concretos como por ejemplo "el 3 de mayo de 2021".

Como primer acercamiento a la anotación de NE en el ámbito legal, se ha decantado por las categorías básicas de NE (Organizaciones, Personas y Lugares) junto a dos categorías relevantes del dominio legal; las fechas y las referencias a leyes. La selección de las dos últimas categorías ha teniendo en cuenta tres criterios: 1) La utilidad de cara a futuras aplicaciones (Ej. Enlazar las leyes, las líneas temporales, detección de periodos de vigencia, etc.); 2) La alta frecuencia de estos tipos de entidades en el dominio legal como se demostrará en sección 5 y, por último, 3) La naturaleza de los textos legales donde las referencias a leyes, artículos, etc. es el rasgo distintivo, por excelencia del dominio objeto de estudio. No se han ampliado las categorías a cantidades u otras categorías porque este estudio se considera como un primer paso y para futuros trabajos, sí se valora incluir nuevos tipos de entidades.

El previo análisis del estado del arte demuestra que los mejores resultados en cuanto a acierto y cobertura, se han logrado combinando diferentes estrategias para afrontar la tarea de anotación, especialmente si se trata de distintos tipos de entidades. Por este motivo, hemos optado por una metodología híbrida que reúne tres técnicas principales y que emplea las últimas técnicas en PLN y en el tratamiento de textos legales:

- Patrones de expresiones regulares
- Listas de fuentes externas
- Entrenamiento de tres modelos NERC utilizando la librería abierta spaCy, v3

El trabajo se estructura en seis fases principales: Extracción de datos, pre-anotación, validación parcial, entrenamiento, evaluación y visualización

Figura 1:Estructura del trabajo.

4 Datos

Para el presente estudio, se ha utilizado el conjunto de la legislación del BOE incluido junto a otros conjuntos en el meta-corpus Legal-ES (Samy et al., 2020). Cabe señalar que el portal del BOE incluye más conjuntos como anuncios o código electrónicos, etc. Sin embargo, aquí nos limitamos al conjunto de Legislación.

4.1 Estructura del conjunto

El conjunto consiste en un total de 215870 ficheros legislativos desde el año 1661 hasta septiembre de 2019. Los ficheros están en formato XML según el estándar de ELI *(European Lesgislation Identifier*). Al analizar la estructura y la nomenclatura del conjunto, se han identificado las siguientes categorías:

Categoría	Nº ficheros	Explicación
BOE-A	144599	Leyes-Decretos
BOE-T	1135	Sentencias constitucionales
DOUE	69448	Diario Oficial de la Unión Europea-Legislación, Decisiones, Recomendaciones y otros
Algunos boletines de Comunidades Autónomas	688	Algunos conjuntos de diarios oficiales de Comunidades Autónomas

Tabla 1. Categorías de documentos en el BOE legislativo.

Para esta fase, se ha decidido limitarse a los subconjuntos BOE-A y DOUE porque son los mayores conjuntos y, por tanto, son más representativos. Se han descartado los subconjuntos de boletines oficiales de comunidades autónomas porque pueden contener lenguas oficiales que no están en el alcance del presente estudio. Los conjuntos del DOUE emplean diferentes normas a la hora de referenciar leyes o directivas, lo cual implica incluir distintos patrones para el reconocimiento de estas entidades. A continuación, se muestran algunos ejemplos de referencias a leyes en el BOE y en el DOUE teniendo en cuenta las variaciones en el BOE a lo largo de los años

Ejemplos de referencia a leyes en el BOE

- Ley 13/2016, de 28 de julio
- Ley EDU/1234/2020 → En leyes recientes se añaden a veces tres letras para indicar el campo temático/ministerios en cuestión como Educación, Fomento.

-EL DECRETO MIL QUINIENTOS SESENTA/MIL NOVECIENTOS SETENTA Y CUATRO, DE TREINTA Y UNO DE MAYO, → En textos de los 60 y los 70, se mencionaban las fechas de forma alfabética.

Ejemplos de referencia a una Directiva Europea

- Directiva 2006/123/CE, de 12 de diciembre, del Parlamento Europeo y del Consejo
- Directiva 2006/112/CE del Consejo, de 28 de noviembre de 2006

Directiva 2006/112/CE

4.2 Extracción de los datos

Para cada fichero XML del conjunto, se ha procedido a la extracción automática del título a partir de los metadatos además del contenido y se ha transformado el contenido extraído a ficheros en formato txt para facilitar su posterior tratamiento.

Gracias a la nomenclatura, se ha podido organizar el proceso de extracción y transformación en grupos divididos por ventanas temporales correspondientes a las diferentes décadas. Por ejemplo, se han agrupado los ficheros de BOE-A de la década de los 70, los 80, los 90, etc. Esta agrupación también nos permite analizar la evolución de las formas de referenciar las leyes y las fechas. Por ejemplo, en los textos legislativos de los años 70, los números en leyes y fechas se oscilaba entre las referencias numéricas y las referencias alfabéticas. Al finalizar el proceso, el recuento final del texto extraído de BOE-A se asciende a: **370860624** *tokens* y DOUE **201840806** *tokens*.

Aunque se ha pre-anotado todo el conjunto, los experimentos de este trabajo se centran en el BOE-A, puesto que manejar este volumen de datos es imposible por la inviabilidad de validar esta cantidad y las limitaciones de infraestructura y capacidad de cómputo. Por estos motivos y de cara al entrenamiento del modelo, se han creado de forma aleatoria tres conjuntos de datos del BOE-A:

- Un conjunto de entrenamiento (training) → 1272254 tokens (21116 oraciones).
- Un conjunto de desarrollo (develop/validation) → 151600 tokens.
- Un conjunto de evaluación (test) →200438 tokens.

Al crear estos conjuntos, se ha tenido en cuenta que sean textos relativamente modernos para garantizar la utilidad del modelo en el contexto actual, ya que un enfoque diacrónico queda fuera del alcance del presente estudio.

Siguiendo este criterio, los tres conjuntos se han creado a partir de los textos del BOE de las tres últimas décadas: los años 90, la década de 2000-2010 y de 2010-2019.

5 Pre-anotación

Para realizar la pre-anotación, es imprescindible decidir: 1) ¿Qué se va a anotar? 2) ¿Cómo se va a realizar esta pre-anotación?

Las respuestas a estas preguntas suelen recogerse en las guías de anotación. Establecer

unos criterios claros y unívocos garantiza la consistencia de la anotación y, por consiguiente, la calidad del proceso. Para el presente trabajo, se han desarrollado unas guías de anotación internas básicas como documentos preparativos para una tarea de evaluación (IberLegal2020@Iberlef) que al final, no se ha celebrado (PlanTL, 2020). Para estas guías, se ha partido de las guías de referencia empleadas en tareas de evaluación de entidades nombradas en Iberlef (Porta-Zamorano y Espinosa-Anke, 2020).

Es importante señalar que la decisión acerca de la tipología de entidades surge de las características propias del corpus y el dominio. Los componentes NERC genéricos suelen incluir tipos básicos como Personas, Organizaciones, Lugares y Miscelánea u Otros. Se han realizado dos pruebas iniciales con los componentes NER de las librerías de spaCy y Stanza, pero los resultados no fueron satisfactorios, ya que no se adapten al dominio en cuestión. Por eso, hemos estimado necesario crear un modelo nuevo con una tipología que refleje la naturaleza del dominio y que sea de utilidad.

Por otro lado, en los textos legales, son comunes las entidades anidadas, es decir, compuestas. Por ejemplo, son comunes las menciones a leyes que incluyen una fecha como "La ley 1234/2010, de 12 de mayo de 2010". Ante estos casos, se ha optado por una aproximación simplificada considerando cada tipo de entidad de forma independiente. Eso resulta en anotar la mención "Ley 1234/2010" como una entidad legal y el segmento "12 de mayo de 2010" como fecha.

En cuanto a la metodología de anotar, se ha recurrido a las tres estrategias mencionadas en la sección 3 y se han integrado diferentes recursos dependiendo de cada tipo de entidad.

- **Leyes, referencias a leyes, decretos, normativas, órdenes, artículos**. Se ha desarrollado una serie de patrones de expresiones regulares para identificar referencias como "Ley 27/2014". Además, se ha recopilado una lista con los nombres oficiales completos de todas las leyes aprobadas desde el año 1977 disponibles en la página del Senado. (Ej. Ley 27/2014, de 27 de noviembre, del Impuesto sobre Sociedades).
- **Fechas absolutas.** Para este tipo de entidades se han desarrollado patrones de expresiones regulares que cubren menciones alfabéticas y numéricas.
- **Organismos.** Se han obtenido diferentes listas del Directorio Común de Unidades Orgánicas y Oficinas con un total de 16 mil entradas. Sin embargo, fue imprescindible un proceso de depuración para evitar duplicados, normalizar formatos y corregir faltas de ortografía, etc. Se ha añadido una lista adicional que incluye todos los nombres de ministerios en todas las legislaturas.
- **Lugares.** Las listas del Directorio Común incluyen 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".
- **Personas.** Para esta categoría, hemos optado por utilizar el componente NER de spaCy, dado que los resultados son aceptables, aunque hay un margen de mejora. Asimismo, se ha enriquecido la pre-anotación de este tipo con una lista de cargos y puestos.

La salida de la pre-anotación es un texto enriquecido con las entidades marcadas en formato de "offsets", es decir, incluyendo las posiciones de inicio y fin de cada entidad junto a su tipo. Se ha optado por este formato de salida porque es uno de los que admite spaCy para el entrenamiento. En el siguiente ejemplo, se anotan dos entidades: a) Referencia a la Orden INT/985/2005 empezando en posición 77 y termina en 95. b) Fecha "7 de abril" que empieza en el carácter 100 y termina en el carácter 110.

```
("Uno. Se introducen las siguientes modificaciones en el apartado Cuarto de la Orden INT/985/2005, de 7 de abril:", {"entities": [(100, 110, 'TIME'), (77, 95, 'LEGAL')]})
```

Se ha pre-anotado el total del conjunto, pero para gestionar esta cantidad de texto, se ha realizado la pre-anotación en agrupaciones divididas por las ventanas temporales indicadas en la sección 4.1. El resultado es un total de **10424216 entidades nombradas pre-**

anotadas en el BOE[1]. La cifra indicada refleja la alta frecuencia de entidades en los textos legislativos, lo cual confirma la relevancia de la tarea NERC en este dominio. Conviene señala que esta cifra es antes del proceso de validación y puede contener falsos positivos, así como entidades anotadas dos veces por dos categorías. No obstante, estas cifras nos pueden ofrecer algunos indicadores generales acerca de la distribución de los distintos tipos de entidades según la tabla siguiente:

Tipo	Porcentaje[2] %
• Leyes-nombres completos. • Referencias a leyes, etc.	1-2% 25-40%
Fechas	11-15%
Organizaciones	35-40%
Personas	6-7%
Lugares	10-11%

Tabla 2. Rangos de distribución de los tipos de entidades nombradas.

6 Validación

Una vez terminada la fase de pre-anotación, se procede a una validación parcial de los conjuntos creados para el entrenamiento. Esta validación pretende revisar de forma manual las anotaciones ambiguas para asegurar un conjunto de entrenamiento de calidad. Se han observado dos casos comunes de ambigüedad:

- **Persona vs. Lugar.** Algunos apellidos coinciden con nombres de lugares. Por ejemplo, "Segovia" que aparece como apellido y como ciudad.
- **Organización vs. Lugar.** Las entidades como "Comunidad de Madrid" puede referirse a un lugar o a una institución. Esa distinción depende del contexto y requiere de un proceso de desambiguación que queda fuera del alcance de este estudio.

Por otro lado, la anotación de las menciones a **leyes** en su forma completa constituye un reto por la longitud y la complejidad sintáctica de la entidad. Asimismo, a veces aparece completa y a veces aparece de forma parcial. De este modo, a la misma ley se puede referir de tres maneras: el título completo, el título parcial o la referencia con el número, el año y la fecha.

Respecto a la categoría de **Persona**, esta categoría incluye tanto las menciones a nombres propios como a puestos o cargos[3]. La preanotación de esta categoría depende, en parte, de la anotación automática de spaCy y, en otra parte de listas de cargos. Por eso, requiere de una revisión manual para evitar introducir al modelo ejemplos erróneos, sobre todo, en lo que se refiere a los nombres propios anotados de forma automática.

7 Entrenamiento del modelo

Se ha elegido la arquitectura de spaCy por su flexibilidad y su eficiencia. Además, ofrece una forma relativamente sencilla de manejar el entrenamiento de modelos permitiendo aplicar nuevas técnicas de aprendizaje profundo a la tarea de NERC de una forma flexible y sencilla.

El entrenamiento de modelos en spaCy se basa en la arquitectura de aprendizaje automático "*thinc"* que implementa redes neuronales convolucionales profundas (*Deep CNN*) integrando los Bloom embeddings (Honnibal y Montani, 2017).

Cabe señalar que, aunque spaCy 3.0 ha incluido modelos de lenguaje de *Transformers* para el español, pero hasta la fecha, el componente NER no está disponible para el modelo de *Transformers* español. Así que, se han utilizado los vectores del modelo grande de spaCy para el español (es_core_news_lg).

Hemos entrenado tres modelos de NERC para comparar los resultados y valorar las diferentes opciones de entrenamiento que ofrece la arquitectura de spaCy3. Se han utilizado los mismos conjuntos de datos de entrenamiento, desarrollo y evaluación en el entrenamiento de los tres modelos. En cuanto a la arquitectura del modelo, se ha utilizado los parámetros de entrenamiento de spaCy por defecto, ya que el objetivo del estudio se centra en el recurso y no en la arquitectura en sí.

- NERC-Legal-1. Se trata de una actualización sobre el modelo original de spaCy *(model_update*).
- NERC-Legal-2. Se entrena un modelo nuevo desde cero (*blank-model*) basándose en la arquitectura del modelo NER de

[1] Ejemplo del conjunto disponible en: https://github.com/dosamy/NERC-Legal-ES-Example-

[2] Se incluyen rangos porque la distribución varía entre los subconjuntos de las distintas ventanas temporales.

[3] El presente estudio no aplica sub-clases.

spaCy, pero desde cero sin tener en cuenta el modelo de NER ofrecido por defecto.

- NERC-Legal-3. Es básicamente el mismo que el modelo anterior, pero utilizando la arquitectura optimizada que ofrece la última versión de spaCy. Es un modelo entrenado desde cero. El proceso se ha realizado mediante el fichero de configuración de spaCy.

Model	NERC-Legal-1	NERC-Legal-2	NERC-Legal-3
Iteraciones	20	20	10
Drop-out	0.1	0.1	0.1
Batch_size	256	256	1000

Tabla 3. Parámetros de entrenamiento.

8 Evaluación

Para la evaluación de los tres modelos se ha utilizado el mismo conjunto de evaluación. Se ha realizado la evaluación mediante el *Scorer* de spaCy que aplica una evaluación estricta.

Los resultados obtenidos demuestran que entrenar un modelo de NERC en el dominio legal alcanza resultados comparables con el estado de la cuestión. Otro aspecto a resaltar, es que los altos valores de precisión y cobertura se deben a la alta frecuencia de entidades y al uso formal y normalizado del lenguaje en los textos jurídico, lo cual ayuda al modelo a aprender estas estructuras y generalizarlas.

Modelo	Precisión	*Recall*	f-score
NERC-Legal-1	**0.9403**	**0.9230**	**0.9316**
NERC-Legal-2	0.8942	0.8984	0.8963
NERC-Legal-3	0.8636	0.8449	0.8541

Tabla 4. Evaluación de los tres modelos.

A continuación, se presentan los resultados obtenidos por cada tipo de entidad nombrada. Las fechas y las entidades legales han obtenido los mejores resultados, mientras que los tipos de Persona y Organizaciones han tenido valores inferiores al resto de las categorías.

Entidades legales			
	Precisión	*Recall*	f-score
NERC-Legal-1	**0.9862**	**0.9862**	**0.9862**
NERC-Legal-2	0.8855	0.8883	0.8869
NERC-Legal-3	0.9415	0.9538	0.9476
Fechas			
	Precisión	*Recall*	f-score
NERC-Legal-1	0.9782	0.9782	0.9782
NERC-Legal-2	0.9865	**0.9782**	**0.9836**
NERC-Legal-3	**0.9870**	0.9261	0.9556
Organizaciones			
	Precisión	*Recall*	f-score
NERC-Legal-1	**0.9327**	**0.9505**	**0.9415**
NERC-Legal-2	0.8914	0.9290	0.9098
NERC-Legal-3	0.8382	0.8695	0.8536
Lugares			
	Precisión	*Recall*	f-score
NERC-Legal-1	**0.9496**	**0.8650**	**0.9053**
NERC-Legal-2	0.8821	0.8639	0.8728
NERC-Legal-3	0.8762	0.6008	0.7128
Personas			
	Precisión	*Recall*	f-score
NERC-Legal-1	0.8426	0.7436	0.7900
NERC-Legal-2	**0.8475**	**0.7883**	**0.8169**
NERC-Legal-3	0.7151	0.7116	0.7133

Tabla 5. Resultados de los tres modelos por cada tipo de entidad.

En cuanto a la categoría **Persona**, la razón detrás de los bajos valores en precisión y cobertura es la escasez de ejemplos en el conjunto de datos de entrenamiento, dada su poca frecuencia en comparación con otros tipos en los textos legislativos. Además, la pre-anotación de este tipo de entidades se llevó a cabo de forma automática salvo la anotación de puestos y cargos que se han anotado a partir de una lista. Todo esto influye en la calidad de los ejemplos y por tanto afecta negativamente al aprendizaje del modelo.

Por otro lado, en este tipo de texto, las entidades de tipo Organización son frecuentes, pero las listas empleadas en la pre-anotación solo incluyen entidades públicas españolas. No incluye entidades internacionales ni privadas. Además, la lista del Directorio Común es poco consistente, lo cual influye el proceso de la pre-anotación y por tanto los resultados del entrenamiento.

Por último, se ha llevado a cabo una evaluación de la anotación automática por reglas y listas comparándola con los resultados del modelo NERC-Legal-1. La Tabla 6 presenta la comparativa entre la anotación por reglas y la anotación por el modelo NERC-1.

	Precisión	*Recall*	f-score
Reglas y listas	**0.9666**	0.8710	0.9185
Modelo	0.9403	**0.9230**	**0.9316**

Tabla 6. Anotación-reglas vs. Modelo.

El uso de modelo supone una pequeña mejora (teniendo en cuenta que hay poco margen de mejora dados los altos valores de la

pre-anotación). De ahí, surge una cuestión: ¿Es viable optar por modelos cuyo coste de entrenamiento es alto, cuando se puede obtener resultados parecidos con técnicas menos costosas como las reglas y las listas? Sin embargo, la respuesta es sí, es más viable a medio y largo plazo porque una vez entrenado, el modelo puede generalizar y, por tanto, permite mayor cobertura y flexibilidad en comparación con un anotador basado en reglas y listas que requieren un alto coste de mantenimiento pese a su precisión.

La ventaja más destacada es que, el modelo ofrece mayor eficiencia en cuanto a tiempos de anotación, lo cual permite mejor integración en soluciones que requieren tiempos de ejecución reducidos y, por consiguiente, permite una mejor escalabilidad. Además, al ser entrenado con spaCy, permite beneficiarse del abanico de posibilidades que ofrece esta librería como integrar el componente en pipelines adaptados al dominio legal español como este ejemplo de pipeline del inglés jurídico[4]. Por otro lado, este modelo puede adaptarse para subdominios como la jurisprudencia o los textos administrativos.

9 Visualización

Por último, para la visualización de los resultados del modelo, se ha utilizado "displacy" que ofrece spaCy. En Figura 2, presentamos el resultado del texto de evaluación anotado con el Modelo-NERC-Legal-1.

10 Conclusiones y trabajo futuro

El presente trabajo ha abordado la importancia de la tarea NERC en el dominio legal destacando los retos en cuanto a los recursos y herramientas de PLN para el español legal.

En la parte práctica, se ha presentado una metodología para la anotación de 5 tipos básicos de entidades mediante diferentes técnicas. La anotación no pretende resolver todos los retos de los diferentes tipos de entidades, sino que se trata de una aproximación transversal básica y como un primer paso en un camino que requiere más esfuerzo y trabajo. Se han entrenado tres modelos y se han presentado los resultados de los modelos entrenados con spaCy. Los resultados son muy satisfactorios, ya que son comparables con los resultados de modelos del NERC español en dominio general (Agerri y Rigau, 2020). La alta frecuencia y el uso normalizado de las menciones a leyes y fechas, etc. son factores que ayudan a obtener altos valores de precisión y cobertura.

Partiendo de estos resultados esperanzadores, se abre camino para un abanico de posibilidades para líneas futuras como las entidades anidadas o tipologías jerárquicas donde se contemplen subtipos de entidades como artículos dentro de una ley, etc. Asimismo, se plantea abordar los textos legales hispanoamericanos, otros sub-dominios como las sentencias, los contratos. Por último, otra línea es el tratamiento más completo de las expresiones temporales para incluir expresiones deícticas o abordar los acrónimos y abreviaturas.

Por otro lado, se plantea organizar una tarea de evaluación con la finalidad de contrastar aproximaciones y asentar criterios en el tratamiento de los textos legales en español.

En cuanto a las conclusiones generales, reiteramos que el dominio legal es un ámbito que ofrece numerosas oportunidades para la Inteligencia Artificial, en general, y el PLN, en particular.

Por último, destacar la administración pública como otro sector relacionado con el dominio legal donde el PLN puede desempeñar un papel relevante en el tratamiento de documentos y en el desarrollo de aplicaciones que asistan en optimizar procesos internos y agilizar los servicios públicos de cara a la ciudadanía.

Figura 2: Visualización de la anotación Modelo-NERC.

Agradecimientos

Este estudio se ha hecho realidad gracias al apoyo continuo del Coordinador del PlanTL, David Pérez-Fernández. Asimismo, agradezco a Prof. Amal Shawer y a Óscar Redondo-Carrasco por su apoyo en todo momento.

[4] https://spacy.io/universe/project/blackstone

Bibliografía

Agerri, R. y G. Rigau. 2020. Projecting Heterogeneous Annotations for Named Entity Recognition. En *Proceedings of Iberlef Workshop*. Co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020). Málaga, Spain, September 2020. Disponible en: http://ceur-ws.org/Vol-2664/capitel_paper2.pdf

Andrew, J. y X. Tannier. 2018. Automatic Extraction of Entities and Relation from Legal Documents. En *Proceedings of the Seventh Named Entities Workshop*, Association for Computational Linguistics. pages 1–8. Melbourne, Australia, July 20, 2018.

Badji, I. 2018. Legal entity extraction with NER Systems. Tesis (Master), E.T.S. de Ingenieros Informáticos (UPM).

Cardellino, C., M. Teruel, L. Alemany, y S. Villata. 2017. A low-cost, high-coverage legal named entity recognizer, classifier and linker. En *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law*.

Chalkidis, I., I. Androutsopoulos, y A. Michos. 2017 Extracting contract elements. In *Proceedings of the 16th Int. Conf. on Artificial Intelligence and Law*, pages 19–28, London, UK, 2017.

Chalkidis I. e I. Androutsopoulos. 2017. A deep learning approach to contract element extraction. En *Proceedings of the 30th International Conference on Legal Knowledge and Information Systems*, Luxembourg, pp 155–164.

Cormack, G., M. R. Grossman, B. Hedin., y D. Oard. 2010. Overview of the TREC 2010 Legal Track. TREC.

Dozier, C., R. Kondadadi, M. Light, A. Vachher, S. Veeramachaneni, y R. Wudali. 2010. Named entity recognition and resolution in legal text. En Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (eds.) *Semantic Processing of Legal Texts*. LNCS (LNAI), vol. 6036, pp. 27–43. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12837-02.

Francesconi, E., S. Montemagni, W. Peters, y D. Tiscornia. 2010. Semantic Processing of Legal Texts: where the language of law meets the law of language (Lecture notes in computer science: lecture notes in artificial intelligence, Vol 6036).

Glaser, I., B. Waltl, y F. Matthes. 2018. Named entity recognition, extraction and linking in German legal contracts. En: *IRIS: Internationales Rechtsinformatik Symposium*, pp. 325–334.

Honnibal, M. y I. Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Landthaler, J., B. Waltl, y F. Matthes. 2016. Unveiling references in legal texts – implicit versus explicit network structures. En *IRIS: Internationales Rechtsinformatik Symposium*, pp. 71–78 (2016).

Leitner, E., G. Rehm, y J. Moreno-Schneider. 2019. Fine-grained Named Entity Recognition in Legal Documents. En Maribel Acosta, et al., (eds.), *Semantic Systems. The Power of AI and Knowledge Graphs. Proceedings of the 15th International Conference (SEMANTiCS2019)*, number 11702. Lecture Notes in Computer Science, pages 272–287, Karlsruhe, Germany, 9. Springer. 10/11 September 2019.

Martínez-González, M., P. de la Fuente, y D.J. Vicente. 2005. Reference extraction and resolution for legal texts. En *International Conference on Pattern Recognition and Machine Intelligence*, pages 218-221. Springer. Nadeau, D., y S. Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*, 3-26.

Navas-Loro, M. 2017. Mining, Representation and Reasoning with Temporal Expressions in the Legal Domain. *Proceedings of the Doctoral Consortium, Challenge, Industry Track, Tutorials and Posters*.

Navas-Loro, M. y V. Rodríguez-Doncel. 2020. Annotador: a Temporal Tagger for Spanish, Journal of Intelligent and Fuzzy Systems, Vol. 39 (2020)

PlanTL-IberLegal. 2019. Recursos y aplicaciones de tecnologías del lenguaje

para el dominio legal en lenguas de la Península Ibérica. Disponible en: https://plantl.mineco.gob.es/tecnologias-lenguaje/comunicacion-formacion/eventos/Paginas/iberlegal-2019.aspx

PlanTL-IberLegal. 2020. Tarea de evaluación de Entidades Nombradas en textos legales (Cancelada). Disponible en: https://temu.bsc.es/iberlegal/

Porta-Zamorano, J. y L. Espinosa-Anke. 2020. Overview of CAPITEL Shared Tasks at IberLEF 2020: Named Entity Recognition and Universal Dependencies Parsing. *IberLEF@SEPLN*. Disponible en: https://arxiv.org/pdf/2011.05932.pdf

Qi, P., Y. Zhang, Y. Zhang, J. Bolton, y C.D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. En *Association for Computational Linguistics (ACL) System Demonstrations*. 2020.

Quaresma, P. y T. Gonçalves. 2010. Using Linguistic Information and Machine Learning Techniques to Identify Entities from Juridical Documents. *Semantic Processing of Legal Texts*.

Rehm, G., J. Moreno-Schneider, J. Gracia, A. Revenko, V. Mireles, M. Khvalchik, I. Kernerman, A. Lagzdins, M. Pinnis, A. Vasilevskis, E. Leitner, J. Milde, y P. Weißenhorn. 2019. Developing and Orchestrating a Portfolio of Natural Legal Language Processing and Document Curation Services. En: Aletras, N., et al. (eds.) *Proceedings of Workshop on Natural Legal Language Processing* (NLLP 2019), co-located with NAACL 2019, Minneapolis, USA, 7 June 2019, pp. 55–66.

Rios, S. 2015. Lead Generation for BigLaw? The Business and Ethics of Providing Free Legal Tools and Information Online, 2015. Working paper. Disponible en: https://law.stanford.edu/publications/lead-generation-for-biglaw-the-business-and-ethics-of-providing-free-legal-tools-and-information-online/

Rodríguez-Doncel, V., M. Navas-Loro, E. Montiel-Ponsoda, y P. Casanovas. 2018. Spanish Legislation as Linked Data. *TERECOM@JURIX*.

Roy, A. 2021. Recent Trends in Named Entity Recognition (NER). *ArXiv, abs/2101.11420*.

Samy, D., J. Arenas-García, y D. Pérez-Fernández. 2020. Legal-ES: A Set of Large Scale Resources for Spanish Legal Text Processing. En Samy, D. et al. (eds.) *Proceedings of Workshop on Language Technologies in Government and Public Administration* (*LT4Gov 2020), co-located with LREC 2020, Marseille, France*.

Sekine, S. 2004. Named Entity: History and Future. Disponible en: http://cs. nyu. edu/sekine/papers/

Waltl, B. y R. Vogl. 2018. Explainable Artificial Intelligence – the New Frontier in Legal Informatics. En Jusletter IT 22. February 2018.