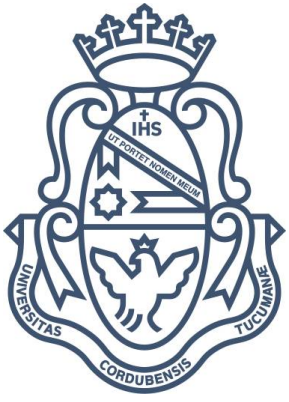
Facultad de Matemática, Astronomía,
Física y Computación

Universidad Nacional de Córdoba



Reconocimiento de entidades nombradas en texto de dominio legal

Tesis para obtener el título de

Licenciada en Ciencias de la Computación

Autora: Karen Haag

Director: Cristian Cardellino

Córdoba, Argentina 2019

Agradecimientos

Agradezco de manera muy especial a todas las personas que han sido parte de este camino de aprendizaje; a Cristian, mi director que dedico mucho tiempo y esfuerzo para ser el mejor guía en el desarrollo de este trabajo. A Laura Alonso, una mujer fuerte e inspiradora que me guió en mis primeros pasos en el aprendizaje automático. A mi familia que me apoyo y creyó en mi, a mis amigos y compañeros que estuvieron siempre de manera incondicional. A Bitlogic, por su confianza y por concederme muchas horas de trabajo para que finalice mis estudios. Y finalmente a Diego Piloni por ser siempre un gran apoyo y estar siempre ahí, siendo parte de mi aprendizaje en esta carrera.

Resumen

En la práctica legal, el acceso inteligente a la documentación es un importante capital. En particular, en el sistema jurídico argentino, la aplicación de métodos de inteligencia artificial para el acceso a la documentación legal es escasa o nula. Por este motivo, este trabajo se centra en la detección, clasificación y anotación de entidades nombradas (como Leyes, Resoluciones o Decretos, entre otros) para el corpus de InfoLEG, una base de datos que contiene los documentos de todas las leyes de la República Argentina. El problema de reconocimiento y clasificación de entidades en esta tesis se trabajó desde dos frentes. En primera instancia se hizo reconocimiento mediante patrones definidos por expresiones regulares.

Luego, se entrenó y evaluó un modelo basado en aprendizaje automático para tratar entidades que no eran regulares y así poder ampliar la cantidad de instancias capturadas. Por último, se realizó una aproximación utilizando anotación semántica para cada entidad y obtener así el acceso a la fuente de información correspondiente.

Índice general

1. Introducción y motivación **1**
1.1. Introducción 1
1.2. Motivación 2
1.3. Esquema de la tesis 4

2. Trabajo previo **7**
2.1. Conceptos básicos 8
2.1.1. Procesamiento de lenguaje natural 8
2.1.2. Expresiones regulares 8
2.1.3. Aprendizaje automático 8
2.2. Trabajo relacionado 9
2.2.1. A Low-cost, High-coverage Legal Named Entity Recognizer, Classifier and Linker 9
2.2.2. Named Entity Recognition and Resolution in Legal Text 10
2.2.3. Using machine learning algorithms to identify named entities in legal documents: a preliminary approach 11

3. Descripción y análisis de InfoLEG **13**
3.1. Descripción del corpus 13
3.2. Obtención del corpus 14
3.2.1. Problemas con el corpus obtenido 15
3.2.2. Estadísticas del corpus 15

3.3. Preproceso del corpus 17
3.4. Análisis y anotación del corpus 18
3.5. Tipología de entidades 19
3.5.1. Documentos por tipo de entidad 20
3.5.2. Ocurrencia de entidades en el corpus 21

4. Metodología de experimentación 23
4.1. Expresiones regulares para NERC 23
4.1.1. Expresiones regulares utilizadas 25
4.2. NERC mediante aprendizaje automático 27
4.2.1. Entrenamiento de un clasificador sobre el InfoLEG 28
4.3. Anotación semántica 32
4.3.1. Anotación semántica de entidades reconocidas . 33
4.3.2. Problemas con la anotación semántica 34

5. Análisis de resultados 37
5.1. Expresiones regulares 37
5.2. NERC con Stanford NER-CRF 39
5.2.1. Parámetros del sistema propuesto 39
5.2.2. Modelos de clasificación de entidades 42
5.3. Análisis de resultados 45
5.3.1. Evaluación de los clasificadores automáticos . . 45
5.3.2. Comparación de los modelos 48
5.3.3. Anotación semántica 50

6. Conclusiones y trabajo futuro 51
6.1. Conclusiones 51
6.2. Trabajo futuro 52

Bibliografía 55

Capítulo 1

Introducción y motivación

1.1. Introducción

El **reconocimiento de entidades nombradas** (NER por sus siglas en inglés), también conocido como extracción de entidades, es una tarea de extracción de información que busca localizar y clasificar en categorías predefinidas como personas, organizaciones, lugares, expresiones de tiempo y cantidades, las entidades nombradas encontradas en un texto.

El reconocimiento de entidades nombradas a menudo se divide conceptualmente en dos problemas distintos: detección de nombres, y clasificación de los nombres según el tipo de entidad al que hacen referencia. Es por eso que muchas veces en la literatura se lo conoce como **reconocimiento y clasificación de entidades nombradas** (NERC por sus siglas en inglés).

Una tercera fase que se desprende del reconocimiento y clasificación de entidades nombradas se conoce como **anotación semántica** (*entity linking* en inglés) donde se anota una entidad con una referencia a algún link de una base de conocimiento que contenga una definición semántica de la entidad (Carreras et al., 2003).

La primera fase generalmente se reduce a un problema de segmentación: los nombres son una secuencia contigua de tokens, sin solapa-

miento ni anidamiento, de modo que *Banco de la Nación Argentina* es un nombre único, a pesar del hecho de que dentro de este nombre aparezca la subcadena *Argentina* que es a su vez el nombre de un país. La segunda fase se trata de asignar una categoría, de entre un conjunto predeterminado, a cada una de las entidades previamente reconocidas en la fase uno.

El reconocimiento y clasificación de entidades nombradas se puede aprovechar de varias maneras, incluyendo el suministro de enlaces de hipertexto a la información almacenada sobre por ejemplo un artículo en particular. Por ejemplo, una mención del "Banco de la Nación Argentina" podría resolverse en un link a la página de Wikipedia que contenga un artículo sobre esta entidad.

1.2. Motivación

Las investigaciones realizadas indican que incluso los sistemas de reconocimiento de entidades más avanzadas son frágiles, dado que los sistemas desarrollados para un dominio no suelen comportarse bien en otros dominios. La puesta a punto de un sistema para un nuevo dominio conlleva un esfuerzo considerable. Esto es cierto para modelos basados en reglas y para sistemas estadísticos.

En la práctica legal, el acceso inteligente a la documentación de todo tipo (leyes, jurisprudencia, etc.) es un importante capital. Un acceso rápido, preciso y que garantice exhaustividad puede hacer una gran diferencia entre un éxito o un fracaso judicial. Además reduce significativamente el esfuerzo requerido para obtener toda la información relevante para llevar adelante una acción.

En los últimos años han crecido en número y capacidades las iniciativas comerciales que proveen acceso inteligente a la información o incluso servicios más proactivos, como automatismos para tratar casos fuertemente tipificados. Sin embargo, la mayor parte de estas iniciativas se han desarrollado para el entorno de Estados Unidos o la Unión Europea, mientras que los desarrollos para Argentina y su región son

mucho más superficiales. Por otro lado, la mayor parte de estas iniciativas son privadas, con un elevado coste, lo cual causa desigualdad en el acceso a la justicia.

El sistema jurídico argentino se beneficiaría fuertemente de la aplicación de métodos de inteligencia artificial para el acceso a la documentación legal. Efectivamente, el sistema argentino se caracteriza por una amplia dispersión normativa. Esto se evidencia en el elevado número de leyes que se han ido promulgando (si bien no todas están vigentes), situación que se ve aumentada por las constantes modificaciones, ampliaciones o complementaciones que muchas de esas leyes experimentan. A esta circunstancia se suman los innumerables decretos y resoluciones administrativas que día a día son emitidas por los organismos estatales, que definen o le otorgan mayor precisión a la aplicación de aquellas leyes, y que por ende deben ser tenidos en cuenta para lograr un entendimiento cabal de cómo operan los derechos y obligaciones, así como los procesos y operaciones de diversa índole contemplados en los cuerpos normativos.

Sin embargo, aquella dispersión normativa no es el único obstáculo para el acceso a la información en nuestro sistema legal. A esta situación se suman los recurrentes reenvíos que un cuerpo normativo, un decreto o una resolución administrativa realizan a otras disposiciones legales, lo cual muchas veces dificulta o torna engorroso el análisis o entendimiento correcto de la norma en cuestión. Esto se agrave aún más si la normativa a la cual se está reenviando no es de rápido y fácil acceso para su lectura, por ejemplo, porque el artículo o el cuerpo normativo al cual se está reenviando no se encuentra hipervinculado en formato electrónico.

De ahí la importancia de poder aplicar herramientas de Inteligencia Artificial tales como las propuestas en este documento, que ayuden a la reducción del tiempo en la búsqueda y análisis de la información. Gracias a lo cual se puede mejorar el estudio y análisis de los cuerpos legales, así como la toma de decisiones a la hora de resolver un caso; elevando la rentabilidad y eficacia del trabajo efectuado.

El procesamiento automático del lenguaje humano o lenguaje na-

tural (PLN) ha producido grandes avances en el acceso inteligente a la información. Gracias a estos avances se responden con gran precisión preguntas sobre la ubicación de una dirección en un mapa, las horas en las que se desarrolla un espectáculo o el autor de un libro. En algunos dominios restringidos, como por ejemplo el de los artículos científicos sobre temas de ciencias de la vida, se pueden contestar preguntas sobre interacciones entre entidades complejas como por ejemplo si un gen se expresa en la presencia de un determinado compuesto químico.

En este trabajo de tesis se hizo uso de la base de datos **InfoLEG** (InfoLEG, 2018), que se describe en detalle en el capítulo 3. Este recurso es un compendio de documentos legislativos digitales de la República Argentina. Si bien la existencia de esta herramienta es extremadamente útil a la hora de facilitar el acceso a la información de manera abierta, también presenta algunas limitaciones respecto a su naturaleza que serán objeto de estudio en este trabajo y sobre el cuál se buscará expandir.

En esta tesis se utilizó el corpus de InfoLEG para la detección y clasificación de entidades nombradas y su posterior vinculación a sus hipervínculos correspondientes.

Del desarrollo de esta tesis surgió la publicación "Mejora del acceso a InfoLEG mediante técnicas de procesamiento automático del lenguaje" (Cardellino et al., 2018).

1.3. Esquema de la tesis

En el capítulo 2 se dará una breve introducción a tres trabajos de investigación que tratan temas relacionados a los abordados en esta tesis. Esto servirá para poner en contexto el trabajo sobre el que esta tesis está basado. En particular, se hablará del los trabajos *A Low-cost, High-coverage Legal Named Entity Recognizer, Classifier and Linker* (Cardellino et al., 2017), *Named Entity Recognition and Resolution in Legal Text* (Dozier et al., 2010) y *Using machine learning algorithms to identify named entities in legal documents: a preliminary approach*

(Poudyal et al., 2019).

En el capítulo 3 se detalla información y estadísticas del corpus de InfoLEG, que fue el recurso principal sobre el cuál se trabajó y experimentó a lo largo de esta tesis de grado. Además, se hará un análisis de como está compuesto dicho corpus, cuáles son sus problemas y qué se buscó para lograr sobreponerse a estos.

En el capítulo 4 se introducirá los diferentes experimentos y métricas que se realizaron en las tres técnicas implementadas para el reconocimiento y vinculación de entidades nombradas: Expresiones regulares, el clasificador de entidades nombradas y enlazamiento de las entidades reconocidas mediante un trabajo de scraping en la web.

En el capítulo 5 se muestran y analizan los resultados obtenidos de los experimentos descritos en el capítulo 4 además de extraer conclusiones sobre dichos resultados, en busca de ir sobreponiéndose con cada nuevo experimento y metodología a las falencias de cada uno de los métodos trabajados.

Por último en el capítulo 6 se realiza una conclusión de los mejores métodos encontrados en la investigación de esta tesis y se detallan futuros trabajos que se pueden realizar a partir de los resultados de los experimentos.

Capítulo 2

Trabajo previo

El presente capítulo hace una breve introducción a los conceptos básicos abarcados en este trabajo de tesis, y además describe tres artículos relacionados a los temas que abordan este trabajo y que fueron guía para encaminar el desarrollo del mismo.

El capítulo empieza con una introducción sencilla a los temas fuertemente trabajados en este documento: procesamiento de lenguaje natural, expresiones regulares y aprendizaje automático.

Se sigue con la descripción de los tres artículos relacionados a este trabajo. El primero es una aproximación para mejorar la extracción de información en textos de dominio legal mediante la creación de un reconocedor, clasificador y vinculador de enlaces de hipertexto. El segundo se trata de un trabajo de reconocimiento y clasificación de entidades nombradas en documentos legales para la jurisprudencia de los Estados Unidos. El último es un trabajo de investigación sobre algoritmos de aprendizaje automáticos para identificar entidades nombradas en documentos legales.

2.1. Conceptos básicos

2.1.1. Procesamiento de lenguaje natural

El **procesamiento de lenguaje natural**, comúnmente abreviado como PLN (o NLP por su acrónimo en inglés), es una rama de la Inteligencia Artificial cuyo objetivo es facilitar la interacción entre humano y máquina a través de lenguajes naturales (e.g. inglés, español, francés, etc.), es decir, la comunicación mediante lenguajes no formales, en contraste con la interacción mediante lenguajes formales como lógica, matemática o lenguajes de programación.

2.1.2. Expresiones regulares

Una expresión regular (también conocida como *regexp* o *regex*, por su acrónimo en inglés) es un patrón que describe una secuencia de caracteres. Una "coincidencia" con la expresión regular, es el fragmento de texto o secuencia de bytes o caracteres a los que el software de procesamiento de expresiones regulares encontró que el patrón corresponde.

2.1.3. Aprendizaje automático

El aprendizaje automático supervisado es una técnica para deducir una función a partir de datos de entrenamiento. Estos datos consisten de pares de objetos: una componente del par son los datos de entrada y el otro, los resultados deseados (Russell and Norvig, 2009).

La salida de la función puede ser un valor numérico (como en los problemas de regresión) o una etiqueta de clase (como en los de clasificación). El objetivo del aprendizaje supervisado es el de crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber visto una serie de ejemplos, los datos de entrenamiento. Para ello, tiene que generalizar a partir de los datos presentados a las situaciones no vistas previamente.

Campos aleatorios condicionales

Una técnica de aprendizaje supervisado que ha sido muy útil en la tarea de clasificación de entidades nombradas son los campos aleatorios condicionales.

Un campo aleatorio condicional (Conditional Random Field o CRF en inglés) es un modelo estocástico utilizado habitualmente para etiquetar y segmentar secuencias de datos o extraer información de documentos. En algunos contextos también se lo denomina campo aleatorio de Márkov (Lafferty et al., 2001).

Estos modelos necesitan ser entrenados con N muestras; cada una contiene un conjunto de observaciones así como las etiquetas asociadas a esas observaciones. El modelo extrae un conjunto de características, que representan las dependencias existentes entre diferentes estados y entre estos y la secuencia de observaciones.

Cada estado puede depender de varias observaciones al mismo tiempo, incluso de la secuencia completa si fuese necesario. En el entrenamiento del modelo éste asigna unos pesos a cada una de esas características, indicando su relativa importancia según el caso.

2.2. Trabajo relacionado

2.2.1. A Low-cost, High-coverage Legal Named Entity Recognizer, Classifier and Linker

En este trabajo de investigación (Cardellino et al., 2017) se intenta mejorar la extracción de información en textos de dominio legal mediante la creación de un reconocedor, clasificador y vinculador de enlaces de hipertexto. Esta herramienta es utilizada para reconocer partes significativas de un texto de dominio legal y conectarla a una base de conocimientos, la ontología LKIF (Hoekstra et al., 2009).

Esta herramienta ha sido desarrollada primero mediante una correspondencia realizada manualmente entre las ontologías LKIF, de

dominio Legal, y YAGO, de dominio general. A través de esta alineación, se ha delimitado el dominio de las entidades legales y se ha obtenido todas las menciones de esas entidades en Wikipedia. Estas menciones se utilizan como ejemplos anotados manualmente para entrenar al reconocedor de entidades nombradas, al clasificador y al vinculador.

Luego se evaluó la herramienta en textos de Wikipedia y también en una pequeña muestra de sentencias del Tribunal Europeo de Derechos Humanos, que ha dado como resultado un rendimiento muy bueno, alrededor del 80 % de F1-score para diferentes niveles de granularidad.

2.2.2. Named Entity Recognition and Resolution in Legal Text

Este documento analiza el reconocimiento y la clasificación de entidades nombradas en documentos legales para la jurisprudencia de los Estados Unidos, declaraciones y alegatos y otros documentos de prueba (Dozier et al., 2010).

Los tipos de entidades en los que se han clasificado son jueces, abogados, empresas, jurisdicciones y tribunales. El trabajo incluye la descripción y análisis tres métodos para el reconocimiento: Búsqueda, reglas de patrones de texto y modelos estadísticos. Luego el trabajo describe un sistema real para encontrar entidades nombradas en un texto legal y evalúan su exactitud.

De manera similar, para la clasificación, se analizan técnicas de bloqueo, resolución de características y técnicas de aprendizaje automático supervisadas y semi-supervisadas.

En este trabajo de tesis se realizaron tareas complementarias a este trabajo utilizando expresiones regulares y un clasificador supervisado basado en CRF para obtener una mayor cobertura en textos en español.

2.2.3. Using machine learning algorithms to identify named entities in legal documents: a preliminary approach

Este documento trata sobre la precisión y el desempeño de diversos algoritmos de aprendizaje automático para el reconocimiento y extracción de diferentes tipos de entidades nombradas, como fechas, organizaciones, leyes de regulación y personas (Poudyal et al., 2019).

El experimento se basa en 20 documentos de decisión judicial del sitio European Lex [1], equivalente a InfoLEG para toda la comunidad europea.

Se utilizó el framework opensource Minorthird[2] que es una herramienta de aprendizaje automático desarrollado en Java, para la anotación, aprendizaje y categorización de entidades en texto.

En total, se aplicaron 8 algoritmos para la identificación y extracción, enumerados a continuación:

- Maximum entropy Markov model (MEMM)
- Conditional random fields (CRF)
- Semi-conditional random fields (SemiCRF)
- Support vector machine conditional Markov model (SVMCMM)
- Voted perceptron semi-Markov model (VPSMM)
- Voted perceptron conditional Markov model(VPCMM)
- Voted perceptron hidden Markov model (VPHMM)
- Voted perceptron semi-Markov model 2 (VPSMM2)

[1] `https://eur-lex.europa.eu/homepage.html?locale=es`
[2] `http://minorthird.sourceforge.net/old/doc/`

Como se puede observar, todos estos algoritmos están pensados para lidiar con secuencias (en este caso secuencias de palabras que forman el texto a anotar). Para analizar el rendimiento de los algoritmos, los datos obtenidos de las entidades etiquetadas fueron comparados con trabajo manual como referencia. Los que llegaban a mejor performance en las distintas clasificación de entidades fueron los algoritmos Hidden Semi Markov Model, Support Vector Machine conditional Markov model y Conditional Random Fields. Se llegó a la conclusión que los algoritmos de aprendizaje automático pueden ser un buen enfoque para resolver este tipo de problemas. En base a esto se eligió para trabajar un algoritmo de CRF para el clasificador de base en el sistema.

Capítulo 3

Descripción y análisis de InfoLEG

Este Capítulo vamos a dar una breve descripción del corpus de InfoLEG, recurso principal en el trabajo realizado para esta tesis de grado. Se busca hacer un resumen general de como se obtuvo el corpus, que técnicas de trabajo se utilizaron para procesarlo y estadísticas sobre su composición. Además, se introducirán brevemente las fases de desarrollo llevadas a cabo sobre el corpus durante esta tesis.

3.1. Descripción del corpus

InfoLEG es una base de datos legislativos del Ministerio de Justicia y Derechos Humanos de la Nación, Ministerio que administra además el Sistema Argentino de Información Jurídica (SAIJ) (InfoLEG, 2018). InfoLEG está conformada por documentos digitales tales como leyes, decretos, decisiones administrativas, resoluciones, disposiciones y todo acto que en sí mismo establezca su publicación obligatoria en la primera sección del Boletín Oficial de la República Argentina[1].

[1] `http://www.infoleg.gob.ar/?page_id=310`

El objetivo de esta base de datos es lograr la recopilación de los cuerpos normativos que integran el sistema jurídico argentino, privilegiando "el acceso gratuito, oportuno, rápido, y sencillo a la información, como así también a todos los otros servicios que se prestan tales como: consulta y asistencia documental, búsquedas especializadas en bases de datos legislativas y extranjeras a través de Internet, enlaces a bases de datos externas, capacitación y asistencia técnica, préstamos, reprografía, etc.".

Si bien esta herramienta constituye un recurso muy útil tanto para el ejercicio profesional como para el análisis y estudio de la normativa nacional, también presenta algunas limitaciones.

Una de las limitaciones más evidentes es que si bien en el texto algunas de las entidades mencionadas se encuentran hipervinculadas, las mismas están contempladas por fuera del texto legal como una nota al pie o un comentario, dejando mayoritariamente de lado a las referencias o reenvíos que en el mismo texto se efectúan.

Es por ello que consideramos que agilizaría mucho la navegación de los documentos hipervincular todas las entidades que se encuentran en el texto, teniendo en cuenta no sólo entidades como leyes o decretos, sino también otro tipo de entidades, tales como: resoluciones ministeriales, tratados internacionales, códigos, resoluciones administrativas, notas emitidas por entidades públicas, etc.

3.2. Obtención del corpus

Para la obtención del corpus de InfoLEG con el cual se realizaron todos los experimentos e investigaciones de este trabajo, se utilizó la herramienta Scrapy[2] de Python que permite obtener sitios web de manera automática (haciendo *scrapping*). Si el documento tenía alguna revisión, se obtenía siempre la última revisión del mismo, en caso contrario, se obtenía el documento original.

[2] `https://scrapy.org/`

Muchos de los documentos hacen referencia en su texto a otros documentos (e.g. decretos que citan leyes, o leyes que citan artículos, etc.). En particular, algunos de los documentos revisados (e.g. leyes revisadas, o cambiadas) tienen además un hipervínculo que dirige a la referencia. Los hipervínculos son tratados como menciones a entidades nombradas.

A partir de estos datos se generó el corpus sobre el cuál realizar la tarea de reconocimiento y clasificación de entidades nombradas que se explora y sobre el que se realizaron los experimentos de este trabajo.

3.2.1. Problemas con el corpus obtenido

La principal dificultad que se presentó al tratar con el corpus de InfoLEG es que las anotaciones están incompletas.

No todos los documentos están completamente anotados, con todas sus referencias debidamente hipervinculadas. Un documento puede o no tener una o más referencias hipervinculadas, y no todas las referencias a otros documentos están necesariamente hipervinculadas. Muchas veces puede pasar que una referencia esté hipervinculada en un párrafo pero no en el siguiente.

Más allá del problema de completitud, el corpus también carece de una anotación exacta del tipo de la entidad. Ya que lo que se denotó como una entidad es simplemente un hipervínculo en el recurso original. No había ninguna manera sencilla de saber si dicha entidad era una Ley, Decreto, Resolución, etc. Esto requirió un primer análisis manual, con ayuda de un experto de dominio, para detectar cuáles eran los tipos de entidades que conformaban el corpus, algo que se detalla en la Sección 3.5.

3.2.2. Estadísticas del corpus

La totalidad del corpus conseguido fue de 121.136 documentos (entre leyes, decretos, resoluciones, etc.). Del total de documentos, sin

embargo, solamente 8.603 tienen al menos una referencia hipervinculada, es decir sólo 7 % del total.

Como se mencionó previamente, los hipervínculos son tratados como menciones de entidades nombradas en el texto. Hay un total de 47.662 referencias hipervinculadas en estos 8603 documentos, que vienen a ser las entidades a reconocer y clasificar en el corpus.

A partir de estos 8.603 documentos y 47.662 entidades se procedió a generar una primera versión etiquetada de corpus que sólo serviría para reconocimiento en una instancia de exploración (es decir, las entidades sólo estaban marcadas como entidades, pero no poseían una clase definida).

La utilidad de este primer corpus es encontrar cuáles son los tipos de instancias que existen, o al menos cuáles tienen alguna referencia hipervinculada y a partir de dicho análisis avanzar con el mejor curso de acción en relación a dicha exploración. A partir de ello, con la ayuda del experto de dominio, se avanzó en el paso de anotación que se detalla en la Sección 3.4.

A su vez, estos documentos anotados fueron divididos en 3 subconjuntos: 60 % de los documentos fueron utilizados para entrenamiento del algoritmo de clasificación, 20 % fue utilizado como conjunto de validación y finalmente otro 20 % fue utilizado como conjunto de evaluación.

En el Cuadro 3.1 se detalla la cantidad de documentos que quedaron para entrenamiento, validación y evaluación, y a su vez la cantidad de referencias hipervinculadas que existen en cada subconjunto de datos.

	Entrenamiento	**Validación**	**Evaluación**
Documentos	5132	1684	1787
Hipervínculos	27461	9413	10784

Cuadro 3.1: Cantidad de documentos con hipervínculos y cantidad de hipervínculos totales, dividido en conjuntos de entrenamiento, validación y evaluación.

3.3. Preproceso del corpus

El corpus extraído directamente desde la página web de InfoLEG estaba en formato HTML, del que se tuvieron que limpiar las "etiquetas" para obtener una versión reducida sin agregados de formato (títulos, negritas, cursivas, etc.). Sin embargo, se dejaron intactas las etiquetas HTML que hacían referencia a hipervínculos a otros documentos (i.e. etiquetas del estilo `<a></a>`).

Una vez limpios los documentos HTML, se pasó a crear documentos XML con el texto limpio y las anotaciones correspondientes a los hipervínculos encontrados.

Sobre el texto limpio, luego de realizar un análisis con un experto de dominio que ayudara a identificar los tipos de entidades presentes en el texto (que se detallan en la Sección 3.5), se pasó a aplicar expresiones regulares como se detalla en la Sección 4.1. Cada documento se leyó línea por línea, y se aplicó en dicha línea una serie de expresiones regulares para detectar algún patrón que siguiera el lineamiento de las entidades que se buscaban detectar.

Para los experimentos utilizando el clasificador, que se detallan en la Sección 4.2, se necesitaba transformar los documentos en formato de columnas, donde cada línea representaba una palabra, y cada columna representaba la clase de la palabra definiendo si era una entidad o no.

Dicho documento en formato de columnas es el tipo de documento que requiere el clasificador para poder entrenarse y evaluarse. Por lo tanto se procesó el corpus para poner en cada fila del documento una palabra (se tomó como palabras a una secuencia de caracteres separado por espacios en el texto). El documento separa las distintas oraciones y distintos documentos mediante líneas en blanco.

Una limitación fue que las oraciones no podían superar las 50 palabras, en caso contrario el clasificador de Stanford no podía procesarlo, por lo que en dicho caso se hacía una nueva oración a partir de la palabra 51.

3.4. Análisis y anotación del corpus

Si bien el corpus obtenido fue procesado para tratar los hipervínculos como anotaciones de entidades, a la hora de clasificarlas en distintas categorías, esto no podía hacerse simplemente en base a los hipervínculos. Estos solo servían para generar un corpus con el cual se podía realizar reconocimiento de entidades nombradas, pero no así clasificación.

Para definir las clases posibles se requirió de un experto del dominio legal, que ayudara con la identificación de de los tipos de entidades presentes en el texto. Las clases encontradas se detallan en la Sección 3.5.

Sin embargo, no se pudieron anotar todas las instancias manualmente puesto que no se disponía del tiempo para ello. En cambio se decidió utilizar una combinación de anotaciones manuales, utilización de expresiones regulares, y el uso de un clasificador de entidades nombradas, entrenado sobre las anotaciones manuales originales.

En particular, las entidades detectadas por el experto de dominio, fueron anotadas mediante el formato BIO en primera instancia, luego solo con su clasificación para ser luego utilizadas en el entrenamiento del modelo de clasificación.

El formato BIO de anotación subdivide una etiqueta de clasificación X en B_X (inicio de entidad) y I_X (continuación de la entidad). Sirve para poder detectar cuando comienza y cuando termina una entidad. Esto nos resulta útil para diferenciar dos entidades que se encuentran consecutivas en un texto.

En el Cuadro 3.2 se refleja la diferencia de utilizar el método BIO de anotación y utilizar su clasificación.

Palabra	Anotación BIO	Clasificación
Las	O	O
facultades	O	O
conferidas	O	O
por	O	O
la	O	O
Ley	B_LEY	LEY
Nº	I_LEY	LEY
22.091,	I_LEY	LEY
artículo	B_ARTICULO	ARTICULO
33	I_ARTICULO	ARTICULO

Cuadro 3.2: Anotación de entidades

3.5. Tipología de entidades

A partir del análisis exploratorio y la ayuda del experto de dominio en la detección de entidades en el corpus, se buscaron detectar clases candidatas a ser trabajadas en una primera aproximación a realizar la tarea de clasificación de entidades nombradas.

A partir de los datos, se detectaron siete entidades principales para trabajar, que eran aquellas que más frecuencia tenían y aparecían de manera más uniforme. Estas se describen a continuación.

Artículo Es cada una de las disposiciones, generalmente enumeradas de forma consecutiva, que conforman un cuerpo legal, como un tratado, una ley o un reglamento.

Decisión Está relacionada con el dictamen o resolución emitida por el poder judicial para resolver un caso determinado.

Decreto Es un tipo de acto administrativo emanado habitualmente del poder ejecutivo y que, generalmente, posee un contenido normativo reglamentario, por lo que su rango es jerárquicamente inferior a las leyes.

Disposición se utiliza con un sentido más estricto, para designar uno de los enunciados lingüísticos en el sentido en el cual se articula el texto de un acto jurídico.

Expediente Es un instrumento público, que resulta de la agregación de las distintas actuaciones, de las partes y del órgano judicial, en forma de legajo.

Ley Es una norma jurídica dictada por el legislador, es decir, un precepto establecido por la autoridad competente, en que se manda o prohíbe algo en consonancia con la justicia cuyo incumplimiento conlleva a una sanción.

Resolución Se refiere a medidas que no se han convertido en leyes. La resolución es a menudo usada para expresar la aprobación o desaprobación del cuerpo de algo que no pueden votar de otra manera.

Si bien también había otras entidades de menor ocurrencia, se decidió no trabajarlas en esta primera instancia, y dejarlas como trabajo futuro, ya que las ocurrencias eran mucho menores y no contribuían tanto como las siete que se listaron.

3.5.1. Documentos por tipo de entidad

De los siete tipos de entidades identificados, algunos son documentos que pueden ser vinculados directamente: leyes, decretos, disposiciones, resoluciones, decisiones y expedientes son representados por documentos. Los artículos no lo son puesto que son secciones dentro de un documento (e.g. al artículo de una ley).

Como una de las tareas que se desarrolló en este trabajo fue la anotación semántica (también conocido como *entity linking*), que se detalla en la Sección 4.3, se buscó hacer un análisis para clasificar automáticamente el tipo de documento, dentro de estas siete entidades

(seis si no se cuenta artículo, que no puede ser un documento), a partir del título del documento.

Como la clasificación de los tipos de los documentos se hizo de manera automática a partir de los títulos, no es exacto, sin embargo da una idea general de cuál es la distribución aproximada de los tipos de documentos que existen en el corpus. En el cuadro 3.3 se muestra el resultado de esta clasificación de los documentos sobre los documentos que sí pudieron clasificarse.

Tipo de documento	Cantidad de documentos
Decisión	3631
Decreto	17916
Disposición	7872
Expediente	209
Ley	7336
Resolución	31341

Cuadro 3.3: Cantidad de documentos identificados por tipo de documento.

3.5.2. Ocurrencia de entidades en el corpus

Una de las tareas que se exploró en este trabajo de tesis fue el entrenamiento de un modelo supervisado mediante el clasificador automático de Stanford NER-CRF (Finkel et al., 2005). Para ello se necesitó que el corpus tuviera anotadas las clases de cada una de las referencias (que estaban dadas por los hipervínculos encontrados). Esto, sumado a las anotaciones hechas por el experto de dominio sirvió para conformar el corpus inicial con el que se entrenaría el clasificador inicialmente.

Como se dijo anteriormente, el corpus no estaba debidamente anotado, y los hipervínculos sirven como una referencia que nos indica la presencia de una entidad en el corpus dentro de cierta extensión de texto. Algunas de estas referencias fueron clasificadas por el experto

de dominio, pero no así los más de 47 mil hipervínculos que se encontraron en el corpus, puesto que no se disponía del tiempo para dicha tarea.

Mediante algunas heurísticas se realizó un proceso de anotación automática que se detalla en la Sección 4.2 del Capítulo siguiente.

Las anotaciones conseguidas sirvieron para generar un corpus para una primera iteración de experimentos entrenando algoritmos de aprendizaje automático supervisado como se detallará en el próximo capítulo. En el Cuadro 3.4 se detalla cuántas entidades de cada tipo (contadas como menciones en el texto) se utilizaron para la primera tanda de experimentos con el clasificador de Stanford.

Tipo de entidad	**Cantidad de menciones**
Artículo	3387
Decisión	680
Decreto	1527
Disposición	139
Expediente	30
Ley	6066
Resolución	1855
Otros (MISC)	754

Cuadro 3.4: Cantidad de entidades (como menciones en el corpus) utilizadas para el entrenamiento del clasificador automático.

Capítulo 4

Metodología de experimentación

El presente Capítulo desarrolla la metodología de trabajo que se utilizó durante este trabajo de tesis. Las siguientes secciones expandirán sobre los métodos utilizados durante la experimentación, los problemas encontrados y las formas que fueron abarcados para solucionarlos. Como se mencionó anteriormente, esta tesis tuvo tres grandes fases de trabajo. En primer lugar, se utilizaron expresiones regulares para la detección y clasificación de entidades. Se siguió con el entrenamiento de un modelo de clasificación basado en aprendizaje automático. Finalmente se realizó un proceso de anotación semántica sobre las menciones encontradas por los dos métodos anteriores.

4.1. Expresiones regulares para NERC

Cómo se mencionó en la Sección 3.4 del Capítulo anterior, se realizó un análisis exploratorio de los tipos de entidades a detectar con ayuda de un experto de dominio.

Fue en este análisis que se observó la regularidad con que muchas de las entidades aparecen en el corpus. Esto es algo natural y esperable

por la naturaleza misma del dominio trabajado, ya que los textos legales son muy estructurados.

A partir de estas observaciones, se concluyó que un trabajo con el uso de expresiones regulares para detectar patrones donde el texto de un documento (ley, decreto, resolución, etc.) hiciera referencia a otro documento, se podrían obtener buenos resultados que servirían como una primera instancia para expandir las anotaciones del corpus original.

Las expresiones regulares tienen la ventaja de poseer precisión absoluta sobre el dominio de patrones con los que se está trabajando. Esto quiere decir que, si una instancia es marcada como una entidad nombrada por una expresión regular, se tiene la certeza de que la instancia marcada es tal (suponiendo que la expresión regular esté correctamente definida). Más aún, el patrón es exhaustivo, lo que quiere decir que se puede estar seguro de que capturamos todas las entidades que coinciden con cierto patrón.

Otra gran ventaja de las expresiones regulares es que pueden realizar las tareas de reconocimiento y clasificación de entidades sin mayores cambios (i.e. no es necesario re-entrenar un modelo de reconocimiento para que actúe como modelo de clasificación), ya que la expresión misma codifica en su patrón el tipo (o clase) de la entidad reconocida.

El mayor problema con las expresiones regulares, es que su exhaustividad está limitada a los patrones definidos. Esto quiere decir que no encontrarán ninguna entidad que no coincida con alguna de las expresiones existentes. Y esto es un problema porque no es escalable hacer una expresión regular para cada patrón, puesto que sería equivalente a la anotación manual completa del corpus.

No obstante, los experimentos demostraron que con relativamente pocas expresiones regulares se podían encontrar una gran cantidad de entidades en el corpus.

4.1.1. Expresiones regulares utilizadas

En base a un análisis exploratorio del corpus, los patrones más comunes eran de la forma:

$$< \text{Tipo Entidad} > [\text{N}^\circ] < \text{Número} > [/ < \text{Año} >]$$

Dónde el "Tipo Entidad" es algo de una de las categorías nombradas en la Sección 3.5 (e.g. Ley, Decreto, Resolución, etc.), el "Número" es el identificador de la entidad y el "Año" (que es opcional, puesto que no todos los tipos de entidades lo tienen) es el año en que se creó o modificó dicha entidad.

Ejemplos de entidades capturadas por una expresión regular de las definidas en este trabajo serían: "Ley N° 23.611", ó "Decreto N° 1.567/2003".

El patrón genérico "< Tipo Entidad > N° < Número >" se utilizó en un primer momento para detectar los distintos tipos de entidades sobre los cuáles construir las anotaciones del corpus.

Partiendo de dicha base, y sumando la ayuda de las anotaciones hechas por el experto de dominio, se fueron descubriendo los distintos tipos de entidades, que se detallan en la Sección 3.5, además de algunos que se encontraron pero se decidieron no trabajar en esta tesis porque su ocurrencia era bastante escasa en comparación a las siete entidades principales.

Con las distintas clases de entidades nombradas encontradas, se fueron generando patrones más específicos. En particular se comenzaron a utilizar expresiones regulares tipificadas de acuerdo a la clase de entidad.

Esta tarea de especificar los patrones permitió reconocer nuevas entidades. Por ejemplo se identificó que algunas entidades podían aparecer sin la palabra "N°" (e.g. "Ley 23.456"), o, incluso si esta aparecía podía tener el símbolo de grado de la forma °, como de la forma º.

Además también se pudo observar en las entidades no capturadas que algunas aparecían con el año de su creación o reforma adjuntado.

Entidad	Expresión Regular
Artículo	art(í|i)culo[s] [N$^{(o\|º)}$] <Número> [/<Año>]
Decición	decisi(ó|o)n[es] [N$^{(o\|º)}$] <Número> [/<Año>]
Decreto	decreto[s] [N$^{(o\|º)}$] <Número> [/<Año>]
Disposición	disposici(ó|o)n[es] [N$^{(o\|º)}$] <Número> [/<Año>]
Expediente	expediente[s] [N$^{(o\|º)}$] <Número> [/<Año>]
Ley	ley[es] [N$^{(o\|º)}$] <Número>
Resolución	resoluci(ó|o)n[es] [N$^{(o\|º)}$] <Número> [/<Año>]

Cuadro 4.1: Expresiones regulares utilizadas para reconocimiento y clasificación de entidades nombradas.

Por ejemplo, el caso de los decretos que pueden aparecer como "Decreto N° 1.567" o como "Decreto N° 1.567/2003", por lo tanto fueron agregadas al patrón de búsqueda también.

Por último se detectó que algunas de las entidades que tenían tilde no eran capturadas ya que en los documentos había ocurrencias de ellas sin tilde. Es el caso de las entidades Artículo, Disposición, Resolución y Decisión. Por lo tanto se agregó la letra con tilde como opcional.

En base a las tipologías encontradas y descriptas en la Sección 3.5, el Cuadro 4.1 define la lista de expresiones regulares para capturar las distintas entidades.

Todo lo que se encuentre entre corchetes es opcional, y coincidirá con expresiones que tengan o no la parte que se encuentre entre corchetes. E.g. el patrón "Ley[es]" marca una coincidencia con las palabras "Ley" y "Leyes".

Todo lo que se encuentra entre paréntesis, dividido por una barra vertical "|", es condicional y coincidirá si tiene una de las dos opciones separadas por la barra. E.g. el patrón "resoluci(o|ó)n"marca una coincidencia con la palabra "resolución" y "resolucion"

Por supuesto, estas expresiones regulares no captarán todas las entidades por las limitaciones mismas que tienen, luego, no hay manera de capturar entidades de la forma:

- Ley Nacional N° 12.345
- Ley de la Nación N° 12.345

Además, si bien las expresiones regulares capturan plurales (e.g. Ley/Leyes), en el caso de que haya más de una entidad listada, las reglas sólo podrán capturar la primera de la lista. Ejemplo serían casos de la forma:

- Leyes N° 12.345 y N° 34.567
- Decretos N° 12.345/2003 y N° 34.567/2005

En dónde las expresiones regulares sólo pueden capturar la "Ley N° 12.345" y el "Decreto N° 12.345/2003".

Dada esta limitación de las expresiones regulares, se decidió avanzar con algoritmo de clasificación supervisado utilizando la suite del Stanford NER-CRF (Finkel et al., 2005).

4.2. NERC mediante aprendizaje automático

Las expresiones regulares, si bien probaron ser de gran utilidad y establecieron una base sólida de trabajo sobre la cual avanzar, tienen una limitación muy clave como se habló en la sección previa respecto a su posibilidad de encontrar todas las posibles formas en las que una mención de una entidad aparece en el corpus.

Muchas de las entidades nombradas aparecen en contextos muy similares, tienen las mismas palabras alrededor, la misma estructura sintáctica y, para el ojo humano, pueden parecer iguales. Sin embargo, pequeñas variaciones (desde cosas sencillas como un espacio de más, hasta errores humanos como faltas de ortografía) hacen que un patrón de una expresión regular no sirva para captar la mención de la entidad. Las expresiones regulares son estáticas en ese sentido.

La manera de trabajar con estos datos es mediante el uso de técnicas de procesamiento de lenguaje natural basadas en información estadística y modelos de aprendizaje automático. Se busca entrenar un clasificador supervisado de entidades nombradas que pueda capturar aquellas entidades que no son englobadas por algunos de los patrones de las expresiones regulares.

4.2.1. Entrenamiento de un clasificador sobre el InfoLEG

Armado del corpus de entrenamiento

Partiendo del corpus de InfoLEG descargado y procesado como se detalló en el capítulo anterior, se establecieron a los hipervínculos como las "anotaciones" necesarias para la construcción del corpus inicial sobre el cual entrenar el algoritmo de clasificación.

Como se aclara en la Sección 3.4, del Capítulo anterior, el corpus fue analizado y, con la ayuda de un experto de dominio legal, se establecieron algunas de las clases a buscar, las cuales se detallan en la Sección 3.5.

No obstante, sólo algunas de las referencias encontradas en los más de 47 mil hipervínculos fueron anotadas por el experto de dominio, puesto que la anotación manual total de estas era una tarea que excedía el alcance de este trabajo de tesis.

Para sobreponerse a esta limitación en las anotaciones pero si poder aprovechar la mayor parte de lo encontrado automáticamente mediante el uso de hipervínculos, una vez que el experto de dominio dio una idea general de que tipo de entidades esperar, se optó por una heurística muy sencilla basada en ver si el hipervínculo contenía entre sus palabras alguna de los identificadores de entidades (y no más de 1). En cuyo caso se consideraba a dicho hipervínculo como una mención de esa entidad.

En caso de que el hipervínculo no tuviera un identificador o bien tuviera más de un identificador posible, se lo clasificó como una en-

tidad sin clase específica asignada (parecida a la entidad "MISC" en la literatura más clásica). La idea de hacer esto es no perder la información de dicha entidad y en todo caso que la desambiguación de la clase en la entidad sea un proceso posterior, dentro del análisis de error.

Con las anotaciones encontradas, se dispuso ya de un corpus para el entrenamiento y evaluación de un detector y clasificador de entidades nombradas con ayuda de la suite de Stanford CFR-NER (Finkel et al., 2005), una implementación de CRFs para su uso en detección de entidades nombradas.

En este caso el par de objetos de entrenamiento serían las *palabras* de los documentos de InfoLEG y su correspondiente *clasificación* que puede corresponder a un tipo de entidad de las listadas en la Sección 3.5, o bien al tipo "MISC" (que engloba aquellas entidades que son reconocidas como tales, por tener un hipervínculo en el corpus original, pero que no se pudieron anotar por diversos motivos), o al tipo "*O*" cuando es una palabra que no pertenezca a ninguna de los tipos de entidades que se definieron en el dominio de este trabajo. En el Cuadro 3.2 del capítulo anterior se puede observar un ejemplo de clasificación del dominio del corpus.

Cabe destacar que los hipervínculos no son exhaustivos, por lo que el entrenamiento se realizó con datos ruidosos, pero aún así se pudieron obtener buenos resultados capturando aquellas entidades que no se podrían haber capturado mediante las expresiones regulares.

Con el corpus anotado que se generó en este paso, y ayuda del clasificador del Stanford NER-CRF que se describirá más abajo, se comenzó a experimentar para entrenar los modelos de clasificación para poder reconocer nuevas entidades no capturadas por las expresiones regulares que se describieron en la sección previa.

En las distintas iteraciones de experimentación, que se detallan en el capítulo siguiente, se utilizaron distintas variantes de este corpus inicial, desde un clasificador entrenado con la totalidad del corpus inicial hasta un clasificador particular que solo se entrenara con aquellas instancias que no lograron ser capturadas por las expresiones regula-

res.

En particular, una de las iteraciones que probó ser de particular utilidad a la hora de entrenar un modelo más completo, que fuera capaz de encontrar más instancias, se hizo a partir de un corpus híbrido entre el que se describió en esta sección, sumado a todas aquellas menciones de entidades que no fueran hipervínculos en el InfoLEG original, pero hayan sido detectados por una expresión regular. Es decir, aprovechando la propiedad de precisión que poseen las expresiones regulares, se extendió el corpus original con nuevas anotaciones para el paso de entrenamiento del clasificador supervisado, logrando así una búsqueda más exhaustiva de las entidades en el corpus.

Esta sinergia entre el clasificador automático y las expresiones regulares fue exitosa y requerida para la anotación semántica utilizada posteriormente, la cual es descripta en la Sección 4.3.

Stanford NER-CRF

Stanford NER-CRF (Finkel et al., 2005) es una implementación en Java de un reconocedor de Entidades Nombradas.

Viene con la implementación de un extractor de atributos diseñados y probados para el Reconocimiento de Entidades Nombradas, y muchas opciones para definir que características utilizar.

Un extractor de características se encarga de transformar una serie de datos de entrada, a una representación interna adecuada o un vector de características, los cuales serán utilizados por algún sistema de aprendizaje (en este caso el Standford NER), para detectar o clasificar patrones de dichos datos.

El Stanford NER-CRF incluye en la descarga reconocedores de entidades nombradas para inglés, particularmente para 3 clases (PERSONA, ORGANIZACIÓN, UBICACIÓN), y también está a disposición en la página oficial varios otros modelos para diferentes idiomas. No obstante, para este trabajo, los modelos pre-entrenados no servían puesto que los tipos de entidades nombradas buscados eran muy diferentes (empezando por las clases).

El clasificador permite la configuración del extractor de características para poder extraer solo aquellas que se consideren adecuadas para cada caso de uso. En particular, en el caso de este trabajo, se seleccionó la clase de la palabra que se está analizando, de la palabra previa, y la siguiente, bi-gramas de caracteres, y la forma de la palabra y sus secuencias. En la Sección 5.2.1 se profundiza sobre las características seleccionadas, ejemplificando cada una de las que se utilizaron para los experimentos.

Cabe mencionar que el clasificador tiene algunas limitaciones también. En la web la documentación sobre su utilización es bastante escasa, mucha información sobre configuración y utilización de la herramienta tuvo que ser recolectada desde páginas no oficiales, como foros de ayuda o similar. Esto provocó que trabajar con el clasificador al principio se volviera engorroso, en ciertos momentos teniendo que recurrir a ensayo y error para entender que es lo que el clasificador hace. Una vez superada la etapa, habiendo logrado lidiar con la falta de documentación, se pudo trabajar continuamente sobre los diferentes experimentos.

Otra limitación importante del clasificador del Stanford NER-CRF, tiene que ver con el consumo de recursos, en particular de memoria RAM. La utilización de muchas características para representar las instancias de entrenamiento crece con el tamaño del vocabulario y la cantidad de documentos, ya sea para las tareas de entrenamiento o evaluación del clasificador.

Por esta razón, se tuvo que reducir el número de atributos configurados desde un principio y también se redujo notablemente la cantidad de documentos seleccionados para entrenar, además de tener que truncar las oraciones como se mencionó anteriormente a una cierta cantidad de palabras.

Para anotar nuevas entidades, luego de ya haber obtenido un modelo entrenado, se requirió del procesamiento de documentos a anotar en grupos limitados, logrando evitar de esta forma las limitaciones de memoria del clasificador. Este método permitió anotar una mayor cantidad de documentos.

4.3. Anotación semántica

Una vez que las expresiones regulares fueron definidas, y el modelo de reconocimiento del Stanford NER-CRF fue entrenado, se lograron capturar suficientes expresiones en el corpus para pasar a la última parte de este proceso, que fue el procedimiento de anotación semántica de documentos.

La anotación semántica de una entidad es la referencia de dicha entidad no a una clase general (e.g. Ley, Resolución, Decreto, etc.), sino a una instancia con información sobre dicha entidad dentro de alguna base de conocimientos (e.g. una ontología) o algo similar. Es por eso que, necesariamente, la tarea de anotación semántica requiere de la existencia de una base de conocimiento que cuente con información unívoca respecto a la entidad anotada.

La base de conocimiento donde se puede encontrar toda la información de cada entidad reconocida, en este trabajo, es la misma página web del InfoLEG, donde se detalla la información de la mayoría de las entidades que fueron reconocidas mediante los distintos experimentos.

Por lo tanto, en este trabajo se intentó aproximar cada mención de una entidad que aparecía en los documentos del InfoLEG, convirtiéndolas a hipervínculos que hagan referencia al documento de dicha entidad (que es, a su vez, una página del InfoLEG). La idea entonces es tomar el corpus original que está parcialmente anotado con los hipervínculos de solo algunas entidades y completarlo en su totalidad mediante la anotación semántica de cada mención de un documento dentro del corpus.

Un problema de esta aproximación es que, en primera instancia, no se tenía un mapeo claro entre las menciones de las entidades y su identificador dentro de la página del InfoLEG.

Por lo tanto se decidió utilizar la herramienta Scrapy para obtener desde la página de InfoLEG tanto las URLs como el título de cada uno de los documentos y así poder generar los hipervínculos para las entidades capturadas por las expresiones regulares y el Stanford NER-CRF, a su respectiva referencia en la página web del InfoLEG.

Clasificación	Cantidad
Ley	4381
Decreto	11455
Disposición	5473
Resolución	35438
Expediente	48
Decisión	2039

Cuadro 4.2: Cantidad de documentos capturados para anotación semántica.

Las URLs donde se encontraba cada uno de los documentos era estándar y tenía el siguiente patrón:

```
http://servicios.infoleg.gob.ar/infolegInternet/verNorma.do?id=<id-doc>
```

Utilizando este método sencillo, se lograron identificar 58.834 documentos con sus respectivos identificadores dentro del InfoLEG. En el Cuadro 4.2 se puede ver la cantidad de documentos por cada entidad de las detalladas en la Sección 3.5 que se identificaron.

En particular, recordar que la entidad "Artículo" no tiene documentos asignados, sino que es una sección dentro de otro documento por lo que queda fuera del alcance de esta tesis por necesitar de un paso previo de segmentación de documentos que no se trabajó aquí.

4.3.1. Anotación semántica de entidades reconocidas

Luego de haber obtenido los títulos de cada documento publicado en InfoLEG y sus respectivos identificadores con los cuales se podía construir el hipervínculo en donde se encuentra el documento, se prosiguió con la generación de un documento HTML con los textos normativos, agregando a cada entidad reconocida por el Stanford NER-CRF y las expresiones regulares su vínculo correspondiente.

Para este procedimiento, se utilizó la salida de los documentos ya clasificados, recorriendo el texto y escribiéndolo dentro del cuerpo del documento HTML (i.e. dentro de las etiquetas <body><\body>).

Cuando se encontraba en el texto alguna entidad y la entidad encontrada tenía una correspondencia entre los identificadores obtenidos durante el proceso de recolección de datos entonces se generaba el vínculo correspondiente. Para ello se introduce dicho vínculo, dentro del alcance de la mención de la entidad, con las etiquetas correspondientes (i.e. <a><\a>)

Si, por el contrario, la entidad no tenía correspondencia con ninguno de los identificadores, se las escribe en el documento con un énfasis para subrayarlas y resaltarlas en el texto como una entidad reconocida pero no anotada semánticamente (i.e. entre etiquetas <u><\u>).

A modo de ejemplo supongamos que en los resultados obtenidos por el Stanford NER-CRF y las expresiones regulares se reconoce como entidad la *Ley 19640* y ésta se encuentra dentro de los documento identificados con el número 28185. Entonces, agregaremos al documento HTML la siguiente referencia donde aparezca nombrada dicha ley:

```
<a href="http://servicios.infoleg.gob.ar/infolegInternet/verNorma.do?id=28185">
    Ley 19.640
</a>
```

4.3.2. Problemas con la anotación semántica

Uno de los principales problemas con la anotación semántica fue precisamente la identificación de documentos, tanto para determinar unívocamente su identificador dentro del sitio del InfoLEG, como para discriminar las menciones de la entidad en un documento. Esto es porque al comparar el título de los documentos, que era la principal fuente de identificación, con las entidades reconocidas se encontraron diferentes formas de nombramiento que dificultaron la tarea de anotar semánticamente las entidades reconocidas. Un ejemplo de esto es el caso:

- Ley N° 23.697

Donde se puede observar que aparecía nombrada con la palabra "N°" y los dígitos de su número de ley contienen un punto. Por el contrario, en la página de InfoLEG su título es:

- Ley 23697 HONORABLE CONGRESO DE LA NACION ARGENTINA

El cual cuenta con muchos espacios entre las palabras de su título, con un nombre más extendido que especifica un poco más de lo que la ley puede tratar, sin la palabra "N°" y sin puntos en sus dígitos.

También se encontraron dificultades para anotar documentos en los que figura su año de creación o modificación ya que se podía encontrar variaciones en la forma de mencionar los años. Por ejemplo en los documentos se puede encontrar un decreto nombrado de la siguiente manera:

- Decreto 281/97

Sin embargo, su título de los documentos tiene la siguiente forma:

- Decreto 281/1997 PODER EJECUTIVO NACIONAL (P.E.N.)

Donde se puede notar a simple vista que se refiere al mismo Decreto, pero en la página de InfoLEG además de aparecer con un título más extenso, el año aparece escrito con cuatro dígitos y en la entidad reconocida solo lo denota con los dos últimos dígitos.

Para contraponerse a este problema, se optó por estandarizar tanto los nombres de las entidades reconocidas, como el título con el que aparecía en la web para poder asociarlas. Para esto se utilizaron las siguientes normalizaciones tanto en las entidades reconocidas como en los nombres de InfoLEG:

- Llevar títulos y entidades a minúsculas.

- Eliminar los puntos de los números.
- En los títulos capturados desde InfoLEG tomamos solo los nombres hasta la especificación del número de su norma, ignorando el resto de especificaciones.
- Eliminar la palabra Nº de las entidades reconocidas.
- Normalizar los años a solo sus últimos dos dígitos
- Normalizar los títulos separando sus palabras solo con un espacio.

Esta estandarización permitió enlazar muchas más entidades y darle más valor a la anotación semántica dentro de este trabajo.

Capítulo 5

Análisis de resultados

Este capítulo realiza un análisis de los resultados de los experimentos de las distintas fases de trabajo de esta tesis, los cuales se describen en el Capítulo 4. Además, los resultados de los distintos experimentos serán comparados entre si para ver como los distintos métodos de detección de entidades (expresiones regulares y aprendizaje automático) se complementan.

5.1. Expresiones regulares

A partir del corpus que se separó para entrenamiento, como se describió en la Sección 3.2.2, constituido por 72.679 documentos, que contaban con un total de 90.750.998 palabras, se aplicaron las expresiones regulares que se definieron en la Sección 4.1, para detectar las menciones de entidades definidas en la Sección 3.5.

Las reglas detectaron automáticamente 723.074 entidades. Dichas entidades pueden observarse en el Cuadro 5.1, donde se establece la cantidad de entidades reconocidas por cada tipo.

Como podemos observar por el Cuadro 5.1, la mayoría de las entidades capturadas son del tipo "Artículo", seguido muy de cerca por la entidad del tipo "Ley" y la entidad de tipo "Decreto".

Tipo de entidad	Cantidad de menciones
Artículo	268699
Decreto	148395
Disposición	13236
Expediente	17515
Ley	205798
Resolución	69431

Cuadro 5.1: Menciones de entidades reconocidas mediante expresiones regulares.

El hecho de que se capturen más artículos viene dado a que un artículo puede estar asociado a un decreto, una ley, u otro tipo de documentos (como los detallados en el Cuadro 3.3). Además un documento puede listar más de un artículo en su texto. Esta asociación entre la entidad "Artículo" y el documento del que forma parte es algo que no se tratará en el alcance de esta tesis.

La naturaleza de las expresiones regulares nos asegura que la precisión de las menciones encontradas en el Cuadro 5.1 es 100 % (o muy cercana al 100 % puesto que los documentos cuentan con cierto error por el factor humano). Sin embargo, sabemos que no se pueden capturar todas las menciones de entidades por razones que se explicaron en detalle en la Sección 4.1 respecto a las variaciones en las menciones.

Una particularidad que se puede notar de este Cuadro es que tiene una distribución que cambia ligeramente en comparación a la del Cuadro 3.4. Esto da idea de que hay ciertas menciones que, en el corpus original, tienen mayor tendencia a ser hipervinculadas (e.g. "Ley" por sobre "Artículo"), siempre teniendo en cuenta que ciertas menciones pueden no haber sido descubiertas por las expresiones regulares.

Por otro lado, en comparación con el Cuadro 3.3, vemos que hay ciertas entidades que, por más que tengan muchos documentos asociados, no tienen tantas referencias en otros documentos. Este es el caso particular de la entidad "Resolución" que tiene muchos documentos (en comparación a, por ejemplo "Ley" o "Decreto"), pero no tantas

menciones encontradas. En contraste, la entidad "Ley" no tiene tantos documentos encontrados, pero muchas más menciones. Esto es de esperar puesto que claramente es más natural hacer referencia a otras leyes que a, por ejemplo, resoluciones.

5.2. NERC con Stanford NER-CRF

5.2.1. Parámetros del sistema propuesto

Como se explicó en la Sección 4.2, del Capítulo anterior, se utilizó la herramienta "Stanford NER-CRF" para entrenar el modelo de clasificación automática de entidades nombradas.

La herramienta cuenta con un extractor de características, que vienen a representar las instancias, del conjunto de datos. Las características, en la herramienta, se definen mediante un archivo de configuración que se pasa al momento de entrenar el modelo.

El Stanford NER-CRF ofrece varias características posibles a la hora de representar las entidades, estas se pueden encontrar en la documentación correspondiente[1].

Para el entrenamiento del clasificador, se decidió utilizar las características recomendadas en el trabajo de Finkel (Finkel et al., 2005). Dichas características se describen a continuación:

useWord Utiliza la palabra actual (la que se busca clasificar).

usePrev Utiliza la palabra previa y su clase.

useNext Utiliza la palabra posterior y su clase.

useNGrams Utiliza n-gramas de caracteres (i.e. información de subpalabras), de la palabra a clasificar, como atributos. En particular, se usaron n-gramas de tamaño 2 debido a limitaciones de memoria con una mayor cantidad de caracteres.

[1] `https://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/ie/NERFeatureFactory.html`

wordShape Utiliza la forma de la palabra como atributo.

Tener en cuenta que un atributo que se utiliza implícitamente en el uso de CRFs es la clase de las palabras de la secuencia. Esto es así porque la optimización del algoritmo hace uso de toda la información disponible para encontrar los mejores parámetros.

El Cuadro 5.2 (página siguiente) muestra un ejemplo de que atributos son extraídos de una instancia de entrenamiento.

Instancia	Clase	useWord	usePrev	useNext	useNGrams	wordShape
Las	O	Las	-	facultades	(L, a); (a, s)	Xxx
facultades	O	facultades	Las	conferidas	(f,a);(a,c);(c,u); (u,l);(l,t);(t,a); (a;d);(d,e)(es)	xxxxxxxxxx
conferidas	O	conferidas	facultades	por	(c,o);(o,n); (n,f); (f,e);(e,r);(r,i); (i,d);(d,a);(a,s)	xxxxxxxxxx
por	O	por	conferidas	la	(p,o);(o,r)	xxx
la	LEY	la	por	Ley	(l,a)	xx
Ley	LEY	Ley	la	Nº	(L,e);(e,y)	Xxx
Nº	LEY	Nº	Ley	22.091	(N,º)	Xº
22.091	ART	22.091	Nº	,	(2,2);(2,.);(.,0); (0,9);(9,1);	xx.xxx
,	O	,	22.091	artículo	-	,
artículo	ART	artículo	,	33	(a,r);(r,t);(t,í); (í,c);(c,u);(u,l);(l,o)	xxxxxxxx
33	ART	33	artículo	-	(3,3)	dd

Cuadro 5.2: Ejemplo de atributos de una instancia.

5.2.2. Modelos de clasificación de entidades

Se experimentó creando varios modelos distintos que utilizaran distinta información a partir del conjunto de datos supervisados que se obtuvo a partir del corpus del InfoLEG. A continuación describiremos algunos de los modelos junto con los resultados nos brindaron.

Modelo entrenado con corpus de hipervínculos

El primer modelo del Standford NER-CRF se realizó utilizando como datos de entrenamiento el corpus obtenido a partir de las entidades anotadas mediante los hipervínculos existentes en el recurso original, que fueron identificados mediante las heurísticas explicadas en la Sección 4.2.

La Figura 5.1 refleja las métricas de precisión y exhaustividad (o

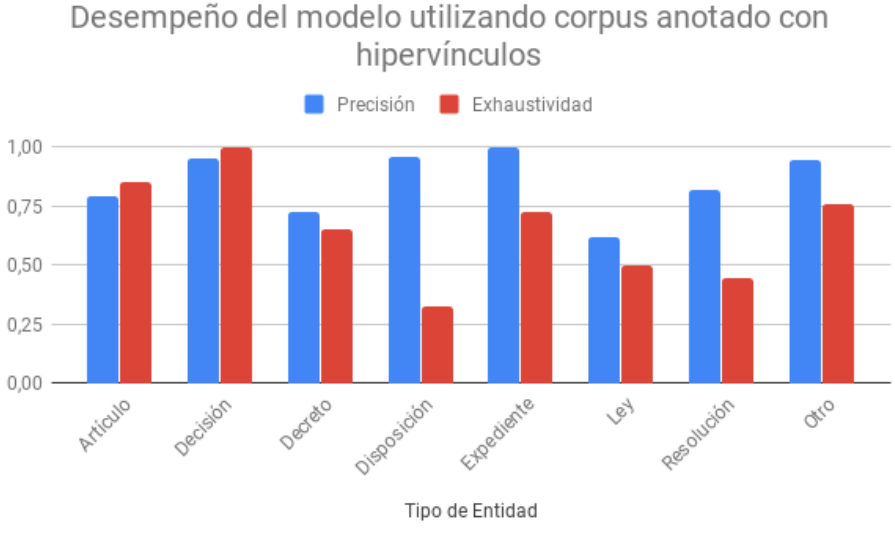


Figura 5.1: Desempeño de modelo entrenado con el corpus obtenido a partir de los hipervínculos

recall en inglés) del modelo sobre el conjunto de datos de evaluación del corpus. Se puede observar que el Cuadro es un gráfico de barras agrupadas. Hay dos barras por cada entidad de las definidas en la Sección 3.5. Las barras azules representan el valor de precisión y las barras rojas representan el valor de exhaustividad.

Modelo específico para entidades no regulares

Dado que las expresiones regulares hacían un trabajo bastante bueno en capturar muchas expresiones, como demuestra el Cuadro 5.1, y en vista de que lo que se necesitaba hacer era precisamente complementar aquello que estos patrones ya capturaban, se buscó desarrollar un modelo que se encargara de clasificar específicamente aquellas entidades que no podían ser capturadas por las reglas manuales.

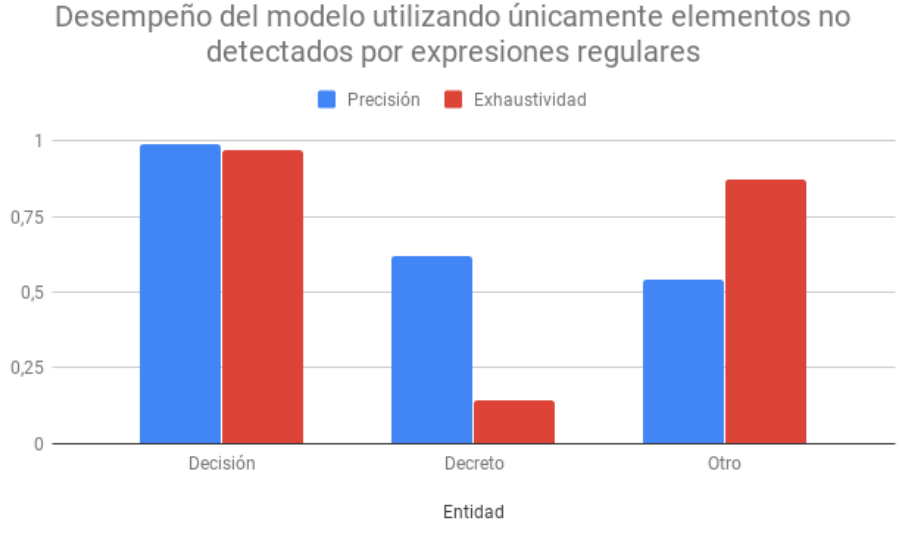


Figura 5.2: Desempeño del modelo específico para entidades no capturadas mediante expresiones regulares

A razón de esto, se optó por sacar las entidades que ya podían ser capturadas del corpus de hipervínculos que se utilizó en el entrenamiento del modelo que se describió anteriormente. Así fue que se entrenó y clasificó con menos clases ya que se encontró que la gran mayoría de instancias recuperada a partir de los hipervínculos (utilizando las heurísticas descriptas en la Sección 4.2) contenían algún patrón establecido por las expresiones regulares.

La Figura 5.2 muestra los resultados de evaluar el modelo entrenado sobre los datos de evaluación (nuevamente obviando todas aquellos casos de entidades detectadas por expresiones regulares). El gráfico es similar en estructura al de la Figura 5.1 que se detalló en la sección inmediata anterior. Como se puede observar, sólo quedaron para este caso algunos tipos de entidades. Esto es porque, como se dijo antes, gran cantidad de las menciones del corpus original ya son capturadas por expresiones regulares existentes. Esto habla del alcance de las expresiones regulares, dado que en su gran mayoría están cubriendo las entidades que se pudieron reconocer automáticamente a partir de los hipervínculos en el corpus.

Modelo aumentado con expresiones regulares

En el último modelo entrenado con el Stanford NER-CRF se decidió aumentar el corpus original de hipervínculos mediante la extensión con todas aquellas entidades detectadas mediante expresiones regulares (las encontradas en la Sección anterior, cuya información de ocurrencia se encuentra en el Cuadro 5.1). Es decir, hacer el trabajo inverso que el descripto para el modelo previo.

La Figura 5.3 muestra el desempeño del clasificador entrenado con el corpus aumentado. En este caso se pueden volver a observar todos los diferentes tipos de clases, similar al caso de la Figura 5.1. El gráfico mostrado en la Figura tiene las mismas características que se describen para el gráfico de la Figura 5.1, detallado previamente.

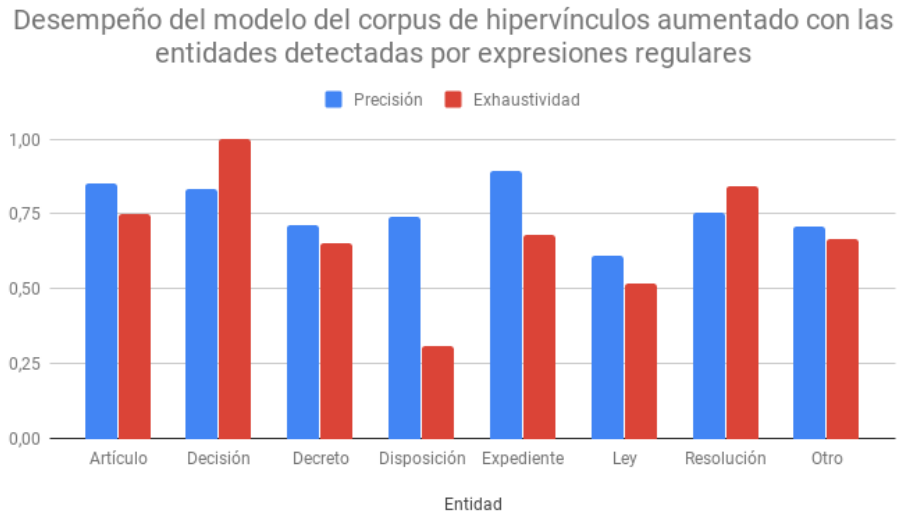


Figura 5.3: Desempeño del modelo utilizando corpus original aumentado mediante expresiones regulares

5.3. Análisis de resultados

5.3.1. Evaluación de los clasificadores automáticos

Los resultados de los clasificadores automáticos que se muestran en las Figuras 5.1, 5.2 y 5.3 fueron evaluados con métricas estándar de precisión y exhaustividad (*recall*), comparándolos con las anotaciones del corpus de evaluación. Esto es en contraste con las expresiones regulares que tienen una precisión casi absoluta, por su naturaleza.

Lamentablemente, la evaluación que se pudo mostrar, no es exacta. Esto es debido a que el corpus de evaluación utilizado, al igual que el de entrenamiento, no es exhaustivo. Es decir, no todas las menciones de entidades en el corpus de evaluación están debidamente anotadas con su clase correspondiente (y en muchos casos ni siquiera están

anotadas como una entidad).

A causa de esto, el clasificador del Stanford NER-CRF, aplicado sobre el corpus de evaluación, anotaba como entidades a ciertas menciones que no estaba anotadas en el corpus original que está incompleto. Luego tenemos un caso de un "verdadero positivo" que es marcado como "falso positivo" por no coincidir con la falta de etiqueta en el corpus de evaluación. Al incrementar el número de "falsos positivos" y decrementar el número de "verdaderos positivos", métricas como "precisión" y "recall" se ven afectadas negativamente.

Para poder revisar cuál era el verdadero impacto de este problema y que tanto se podía confiar en las clasificaciones hechas por el Stanford NER-CRF, se realizó un examen exploratorio sobre una muestra de aproximadamente 1500 entidades clasificadas automáticamente por el Stanford NER-CRF sobre los datos de evaluación. El examen consistió en la revisión manual, una a una, de dichas entidades.

Esta exploración arrojó resultados muy contundentes, con una exactitud casi absoluta por parte del Stanford NER-CRF a la hora de clasificar entidades en el conjunto de datos. Muy pocos de las menciones anotadas por el clasificador eran erróneas.

Al analizar los gráficos de las Figuras 5.1 y 5.2 la primera impresión que dan es que empeoró el modelo para esas clases que en principio se intentó mejorar (i.e. aquellas no cubiertas por expresiones regulares). Pero estos datos no reflejan con claridad los resultados reales por el problema de exhaustividad en el corpus de evaluación.

Se puede notar que en el caso de la entidad "Decisión" el modelo sin expresiones regulares mejora levemente y da un buen resultado. Lo que contrasta con la entidad "Decreto", que parece empeorar de un modelo a otro. La razón de esto, en primera instancia, tendría que ver con que, al eliminar aquellas instancias capturadas por expresiones regulares, se estarían eliminando gran cantidad de instancias de la clase decreto, lo que lleva al clasificador a ser entrenado con muchos menos ejemplos de dicha clase.

Sin embargo, cuando se realizó el examen exploratorio, se pudo observar que se capturaron muchos decretos que no seguían un patrón

predefinido por las expresiones regulares y que si estaba bien su reconocimiento por parte del clasificador. Esto es algo bueno puesto que mejora la cobertura del modelo a nuevos elementos no vistos y no posibles de capturar previamente.

Por otro lado también pareciera que el reconocimiento de la entidad de tipo "Otro" empeora levemente. No obstante, sucede también algo similar que para el caso de la entidad "Decreto" al reconocer muchas nuevas entidades no cubiertas por hipervínculos en el corpus original.

El caso particular del modelo mostrado en la Figura 5.3 es particular, puesto que el hecho de haber aumentado el corpus inicial con tantas entidades ya reconocidas por expresiones regulares, fue más pernicioso que lo que se muestra en un principio. Esto es porque, si bien uno esperaría que modelos automáticos funcionaran mejor con mayor cantidad de datos anotados, a la hora de revisar los resultados del modelo de forma manual, se encontraron mayor cantidad de errores, sobre todo en comparación a los otros modelos.

Una hipótesis que surgió a este respecto es que la gran cantidad de anotaciones agregadas mediante las expresiones regulares, desbalancea mucho más la distribución de clases del corpus, y en particular sólo agrega información de una cantidad selecta de ejemplos (que son los que ocurren con los patrones que pueden capturar las expresiones regulares). Esto hace que los modelos automáticos no logren mejor generalización de resultados y que simplemente aprendan a capturar aquellas menciones que las expresiones regulares ya trabajan bien.

Dicha hipótesis se pudo confirmar observando los datos etiquetados por el clasificador entrenado con el corpus aumentado por las expresiones regulares. Efectivamente se pudo notar, al hacer la revisión manual de los resultados, que la mayoría de las entidades de tipo "Ley" o "Artículo" (dos de las entidades con mayor ocurrencia según lo que muestra el Cuadro 3.4) que capturó el clasificador, pero no fueron capturadas por expresiones regulares, estaban mal etiquetadas.

El modelo aumentado con expresiones regulares sí pudo aprender mejor y logró capturar nuevas entidades (no cubiertas por los patro-

nes) de las entidades de tipo "Resolución", en particular casos donde las expresiones regulares no servían simplemente por pequeños errores de ortografía.

5.3.2. Comparación de los modelos

Se busca analizar de una manera mas exhaustiva el algoritmo de clasificación del Stanford NER-CRF, comparándolo con la cantidad de entidades que fueron reconocidas por las expresiones regulares, con el objetivo de observar que alcance tiene el clasificador.

Los resultados que se muestran en esta sección son aproximados ya que no se puede saber, a priori, cuantas de las entidades reconocidas por el clasificador forman la misma entidad, esto es porque el clasificador se entrenó utilizando el método I/O, que marca determinadas palabras como parte o no de una entidad, pero si hay más de una entidad ubicadas de manera continua en el texto, y son del mismo tipo, no se puede saber donde termina una y comienza la otra. Se decidió entrenar un clasificador mediante este método porque hacía más sencilla la evaluación en primera instancia. Una opción a esto hubiese sido el entrenamiento mediante anotación del tipo BIO, pero queda como trabajo futuro.

El Cuadro 5.3.2 muestra los resultados comparando las expresiones regulares con el modelo de clasificación entrenado sin considerar las expresiones regulares (i.e. el modelo entrenado que se muestra en la Figura 5.2). Las entidades capturadas son en el conjunto de datos de evaluación. Sólo se muestra la cantidad de entidades que no encontraron ambos modelos, es decir, si una entidad fue encontrada por el clasificador y las expresiones regulares, esta no se contabilizó.

Lo primero a observar en el Cuadro, es el limitado número de tipos de entidades que trabajó el modelo, esto es reflejo de lo que se mostró previamente en la Figura 5.2. Y se da porque muchas de las entidades que efectivamente podían discriminarse y clasificarse de manera automática mediante los hipervínculos del corpus, eran reconocidas por algún tipo de expresión regular.

Entidad	Exp. Reg.	Clasificador	TOTAL
Artículo	2657	-	2657
Decreto	1467	57	1524
Decisión	-	249	249
Disposición	233	-	233
Expediente	23	-	23
Ley	2780	-	2780
Resolución	685	-	685
Otro	-	2483	2483
TOTAL	7845	2800	10645

Cuadro 5.3: Cantidad de entidades capturadas por cada una de los métodos

En particular, varios tipos de entidades directamente no tuvieron contrapartida en el clasificador simplemente porque, al ocurrir de manera tan regular en el corpus ya eran capturadas por expresiones regulares y eliminadas completamente del clasificador que se muestra en el Cuadro 5.3.2.

De las entidades que sí trabajó el clasificador, que son "Decreto", "Decisión" y "Otros", llama particularmente la atención de que el clasificador no captura muchas entidades nuevas más allá de las que sí agarran las expresiones regulares, en particular por la uniformidad en la que aparecen menciones de este tipo de entidad. Por otra parte, la entidad del tipo "Decisión" que sí tuvo mayor impacto lo encontrado por el clasificador y complementó mucho más al haber reconocido nuevas entidades que las expresiones regulares no pudieron. Sin embargo, el punto de comparación más importante en este caso es quizás el caso de la entidad "Otros", por el sencillo hecho de que directamente, al no poder definirse una expresión regular para este tipo de entidad (debería ser demasiado genérica y muy difícil de especificar), y únicamente en base a los valores obtenidos en el corpus de hipervínculos, se lograron reconocer muchas entidades en el corpus de evaluación.

5.3.3. Anotación semántica

Para el caso de la anotación semántica de las entidades, lamentablemente no se pudo hacer análisis automático al respecto de si fueron correctamente realizadas las anotaciones (i.e. los hipervínculos del documento anotado). La manera de evaluarlo, nuevamente, fue a través de un análisis de error a partir de observaciones sobre una muestra de los documentos generados con las anotaciones.

A priori, las observaciones no mostraron errores en cuanto a la anotación de las entidades mediante su nombre y su asociación con el título del documento legal al que hacen referencia. Es decir, si el título de la Ley, Decreto, etcétera, es igual al de la entidad, en principio la anotación estaría bien hecha. Puede ocurrir, no obstante, que se esté vinculando la mención a una revisión de incorrecta del documento (recordar que los documentos pueden tener revisiones más recientes si hubo cambios en la ley o similar).

Capítulo 6

Conclusiones y trabajo futuro

6.1. Conclusiones

Utilizar reglas manuales resultó muy eficiente para capturar entidades por la naturaleza del dominio trabajado. Muchas de las entidades buscadas seguían un patrón muy regular.

Estas aportan gran valor a este trabajo dado que, por su naturaleza, nos permiten asegurar con gran certeza, que no tienen precisión absoluta en el etiquetado, lo que es de gran provecho cuando no se tiene un corpus de evaluación suficientemente completo como es el caso del corpus de InfoLEG.

Por otro lado, a pesar de las limitaciones del clasificador, respecto a los recursos necesarios para entrenarse con la totalidad de los datos, resultó muy útil para capturar entidades no regulares, que de otra manera serían prácticamente imposibles de encontrar (definitivamente no era una solución utilizar expresiones regulares para este caso).

Para darle más aporte a este trabajo, y enriquecer la variedad de entidades reconocidas, pudimos observar que el clasificador automático entrenado con el Stanford NER-CRF tuvo muy buen desempeño

cuando fue entrenado con entidades que no seguían los patrones ya capturados por las expresiones regulares. Esto permitió que se capturen muchas entidades no accesibles desde las expresiones regulares. En particular, varias de estas no se pudieron clasificar en primera instancia (e.g. las entidades de la categoría "Otros"), pero aún así el hecho de capturarlas da pautas para seguir mejorando el sistema.

Por último utilizar estandarización de las palabras hizo que podamos generar hipervínculos para anotar semánticamente gran cantidad de entidades. Si bien el paso de anotación semántica requiere de un trabajo más profundo para mejorarse, el aporte inicial fue muy bueno.

6.2. Trabajo futuro

En el trabajo futuro a realizar, una de las primeras etapas sería migrar el clasificador a un etiquetado de tipo BIO, que sirva para poder diferenciar entre dos entidades que ocurran continuas en una predicción hecha. Esto es de especial importancia a la hora de realizar la parte de anotación semántica, que depende en primera instancia de una clasificación y diferenciación correcta de las entidades.

Otra área de trabajo muy importante está dada por la extensión en las posibles clases para las entidades. Como se estableció en el Capítulo 3, la anotación fue limitada a 7 categorías que eran más comunes para poder comenzar a trabajar con una primera aproximación. Un análisis más detallado de los documentos, también en conjunto con un anotador experto. La idea es poder diferenciar nuevas clases, y trabajar todas aquellas entidades que dentro de este trabajo se clasificaron como "Otros".

Asimismo, dentro de las categorías definidas, una opción es desglosar la categoría "Artículo" y diferenciar las menciones de entidades de un artículo de una "Ley" de un artículo de algún otro documento, como un "Decreto".

Realizar análisis de error más profundo para tratar de mejorar el modelo de Stanford NER a partir del corpus de validación. La idea

es explorar más de cerca en que se están equivocando los modelos y eventualmente trabajar en mejorar el corpus de validación para poder tener mejor juicio de valor a partir de las métricas. Otra opción es la exploración de otras métricas que no se vean tan afectadas por los errores clásicos que descubrimos en este corpus obtenido automáticamente.

Además, otra línea de acción es el entrenamiento de otros modelos de aprendizaje automático. Si bien el Stanford NER es muy estándar y, por lo que muestran los resultados, bastante bueno, sería útil también explorar sistemas más cercanos al estado del arte como son los sistemas de aprendizaje profundo, especialmente las redes recurrentes bidireccionales (Chiu and Nichols, 2016).

Finalmente, el trabajo de anotación semántica requiere mayor dedicación, haciendolo de manera menos automática en la instancia inicial. La idea es eventualmente poder integrar esto de manera directa al sitio del InfoLEG, por lo que buscamos realizar un enlazado que, además de ser más certero, siga el estilo de la página, de manera tal que pueda reemplazarse eventualmente.

Bibliografía

Cardellino, C., Teruel, M., Alonso Alemany, L., and Villata, S. (2017). A Low-cost, High-coverage Legal Named Entity Recognizer, Classifier and Linker. In *ICAIL-2017 - 16th International Conference on Artificial Intelligence and Law*, page 22, Londres, United Kingdom.

Cardellino, F., Cardellino, C., Haag, K., Alonso i Alemany, L., Soto, A., Teruel, M., and Villata, S. (2018). Mejora del acceso a infoleg mediante técnicas de procesamiento automático del lenguaje. In *XVIII Simposio Argentino de Informática y Derecho (SID)-JAIIO 47 (Buenos Aires, 2018).*

Carreras, X., Màrquez, L., and Padró, L. (2003). A simple named entity extractor using adaboost. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 152–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chiu, J. P. and Nichols, E. (2016). Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., and Wudali, R. (2010). Named entity recognition and resolution in legal text. pages 27–43.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hoekstra, R., Breuker, J., Di Bello, M., and Boer, A. (2009). Lkif core: Principled ontology development for the legal domain. In *Proceedings of the 2009 Conference on Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood*, pages 21–52, Amsterdam, The Netherlands, The Netherlands. IOS Press.

InfoLEG (2018). InfoLEG: Información legislativa y documental. `http://www.infoleg.gob.ar/?page_id=310`. Accessed: 2018-06-28.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Poudyal, P., Borrego, L., and Quaresma, P. (2019). Using machine learning algorithms to identify named entities in legal documents: a preliminary approach.

Russell, S. and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition.

Los abajo firmantes, miembros del Tribunal de evaluación de tesis, damos fe que el presente ejemplar impreso se corresponde con el aprobado por este Tribunal.