Mejora del acceso a Infoleg mediante técnicas de procesamiento automático del lenguaje

Fernando Cardellino†, Cristian Cardellino‡, Karen Haag‡, Axel Soto*⊙, Milagro Teruel‡, Laura Alonso Alemany‡, Serena Villata**

† Facultad de Derecho, Universidad Nacional de Córdoba, Argentina
‡ Facultad de Matemática, Astronomía Física y Computación, Universidad Nacional de Córdoba, Argentina
* Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur, Argentina
⊙ Insituto de Ciencias e Ingeniería de la Computación, CONICET–UNS, Argentina
**INRIA Sophia Antipolis, Francia

Resumen En este artículo[1] presentamos una aproximación para la detección Automática de Entidades en textos legales, y su aplicación al corpus InfoLeg. La aproximación se basa en diversas técnicas de Extracción de Información, entre ellas Aprendizaje Automático a partir de ejemplos y reglas creadas manualmente.
Presentamos un análisis de los diferentes tipos de entidades que se encuentran en los textos, sus realizaciones lingüísticas y sus dificultades para el análisis automático. El diseño de la solución se basa en las dificultades propias de estas entidades.
En la fase actual de desarrollo de la aproximación hemos anotado manualmente una porción del corpus y hemos desarrollado reglas para anotar casos sencillos automáticamente. Hemos entrenado y evaluado una aproximación basada en aprendizaje automático para casos sencillos, con resultados muy prometedores.

1. Introducción y Motivación

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 de reducir 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

[1] Parte del trabajo publicado en este artículo fue financiado por el programa de investigación e innovación Horizonte 2020 de la Unión Europea, en el marco del proyecto MIREL: MIning and REasoning with Legal texts, No 690974 del programa RISE - Marie Skodowska-Curie.

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 una 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 coadyuvan 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 natural (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.

El dominio legal se puede beneficiar también de estos avances en las técnicas de PLN para mejorar la exhaustividad y precisión en el acceso a la información. Algunas de las utilidades que se podrían implementar son:

- Tratamiento automático y exhaustivo de una gran cantidad de documentos legales.
- Generación automática de hipervínculos entre documentos.

- Integración de textos de diferente tipo, por ejemplo, normas y sentencias.
- Identificación automática de entidades relevantes para indexación y recuperación de documentos.
- Facilitar la lectura de los textos identificando secciones, entidades relevantes, etc.
- Aumento de la búsqueda por palabras clave mediante palabras relacionadas como sinónimos, hiperónimos, temas, etc.
- Recomendación de textos relacionados, no solamente por relaciones de reenvío sino también por semejanza semántica.

Sin embargo, hay que tener en cuenta que las técnicas automáticas de PLN conllevan una tasa de error debida a la ambigüedad propia del lenguaje humano. Aunque en los textos de dominio legal esta ambigüedad es más acotada que en textos más informales, sigue existiendo un nivel de vaguedad que es propio del lenguaje humano. Por esta razón es necesario delimitar bien los objetivos de las aplicaciones, conocer la casuística de los textos a tratar, diseñar soluciones que tengan en cuenta los fenómenos textuales concretos y las capacidades de las técnicas de PLN, evaluar las diferentes propuestas e implementar soluciones con diferentes grados de fiabilidad.

En este artículo presentamos una propuesta para aplicar técnicas de reconocimiento y clasificación de entidades nombradas (NERC, por sus siglas en inglés) al corpus InfoLeg [1]. El corpus InfoLeg es una base de documentos digitales normativos nacionales: leyes, decisiones administrativas, decretos, resoluciones, disposiciones, acordadas y todo acto administrativo publicado en la primera sección del Boletín Oficial de la República Argentina desde mayo 1997, más la normativa referenciada. Se encuentra disponible públicamente[2], y actualmente cuenta con un sistema de búsqueda mediante el número y tipo de norma o bien por el texto literal de los documentos.

El objetivo primario de esta propuesta es desarrollar una herramienta para identificar automáticamente en este corpus entidades como leyes, principios, sentencias, modificatorias, etc. Dependiendo del tipo de entidad, se podrán aplicar acciones como establecer automáticamente hipervínculos desde el texto donde se menciona una ley al documento de la ley. El mismo procedimiento se puede aplicar para modificatorias, resoluciones, sentencias previas, jurisprudencia de todo tipo, etc. Para entidades del tipo institución, se pueden establecer hipervínculos a la página de la institución. Para conceptos más abstractos se puede hipervincular la página correspondiente de la Wikipedia. También se pueden usar las entidades nombradas para recomendar textos con las mismas entidades, o bien para expandir la búsqueda por palabras clave usando ontologías a las que se asocian las entidades.

Esta propuesta se encuentra en desarrollo, habiéndose completado ya una primera fase de análisis de los fenómenos lingüísticos y una segunda fase de anotación preliminar de textos para realizar la prueba de concepto. Después de esta primera fase, aplicaremos y evaluaremos diferentes aproximaciones para obtener una herramienta automática para identificar entidades nombradas.

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

El resto de este artículo se organiza como sigue: en la siguiente sección presentamos trabajo relacionado al que estamos desarrollando. Después seguimos con una descripción del corpus InfoLeg y una delimitación de la casuística de entidades que pretendemos tratar, con sus dificultades y potenciales. En la sección 5 presentamos la solución propuesta, y seguimos con una descripción de la anotación manual que se ha llevado a cabo hasta el momento. Finalizamos con el trabajo que estamos realizando en estos momentos y las fases de desarrollo posterior.

2. Trabajo Relacionado

Las posibilidades ofrecidas por los métodos de procesamiento automático del lenguaje natural en el contexto del ámbito legal han sido investigadas en diversos trabajos previos. En particular, se destacan las técnicas de reconocimiento de entidades realizadas sobre legislación gubernamental, resoluciones judiciales o patentes. Por ejemplo, Dozier et al. [2] emplean una combinación de métodos (diccionarios, reglas y métodos estadísticos) para identificar nombres de jueces, abogados, compañías, jurisdicciones y juzgados. En los casos de Nanda et al. [3] y Surdeanu et al. [4] se emplean campos aleatorios condicionales (CRF por sus siglas en inglés: *conditional random fields*) para reconocer entidades, con la salvedad que en el caso de Surdeanu et al, se apunta a reconocer menciones de entidades que se encuentran anidadas dentro de entidades de otros tipos.

El paso siguiente al reconocimiento de entidades, aunque no siempre presente, es la vinculación de la entidad reconocida con un elemento o concepto en una ontología o base de conocimiento, tal como se lleva a cabo dentro del contexto legal en [3,2,5]. Si bien existen ontologías dentro del dominio legal, tales como DOLCE+ [6], LRI-Core [7], LKIF [8], ELTS [9], o LegalRuleML [10], las mismas sólo comprenden conjuntos pequeños de conceptos legales o sólo poseen descripciones de conceptos a niveles abstractos sin contar con la necesaria instanciación en conceptos más especificos. Dado el alto costo humano que requiere el diseño y modificación de una ontología de dominio específico, se ha visto la necesidad de enriquecer las ontologías a través de enfoques computacionales [11,12,13,14]. Estos trabajos emplean una combinación de métodos estadísticos y aprendizaje supervisado, apoyándose en ciertos casos en otras ontologías o bases de conocimiento de propósito general como YAGO y Wikipedia.

Aún cuando la cantidad de herramientas de software abierto y trabajos de investigación sobre aplicaciones no es tan amplia dentro del dominio legal como sí lo es en otros dominios, advertimos los beneficios de las técnicas de procesamiento de lenguaje para dicho dominio. En particular, el reconocimiento de entidades en texto facilita el desarrollo de aplicaciones que hacen uso de los conceptos extraídos. Uno de ellos son los motores de búsqueda semánticos, los cuales uno de los objetivos es lograr máxima cobertura (*recall*) en el resultado de las búsquedas [15,16,17,18]. Otro ejemplo es la simplificación o resumen automático de grandes porciones de texto, en donde la identificación de conceptos importantes facilita la selección de las porciones de texto con mayor relevancia [19,20]. Los sistemas

de recomendación también permitirían hacer uso de conceptos reconocidos para la sugerencia de otros documentos o conceptos relacionados [21]. Finalmente, la exploración interactiva de grandes volúmenes de datos puede facilitarse mediante la extracción de conceptos de relevancia y su posterior visualización tal como se realiza por ejemplo en Collins et al. [22] o en Ekstrom y Lau [23].

3. El corpus InfoLeg

Según la información brindada desde la página web de InfoLeg, el mismo "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)"[1]. En dicha base de datos es posible encontrar en forma digitalizada diversos tipos de documentaciones legales como: leyes, decretos, resoluciones administrativas, disposiciones "y todo acto que en sí mismo establezca su publicación obligatoria en la primera sección del Boletín Oficial de la República Argentina". 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 encuentras hipervinculadas, sobretodo cuando se trata de leyes modificatorias o complementarias, 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 todos los reenvíos que presenta 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.

En la versión de InfoLeg que nos hemos descargado desde el servicio de Datos Abiertos de la Nación[3] se encuentran 121.136 documentos, con un total de casi 159 millones de palabras y 47.662 hipervínculos. De estos documentos, un 30 % son resoluciones, un 10 % son decretos y un 4 % son leyes.

4. Delimitación de Entidades a Tratar

Como dijimos, la delimitación del problema a tratar es crucial para un abordaje óptimo. En esta sección presentamos una descripción de los diferentes tipos

[3] http://www.datos.gob.ar/ca/dataset/base-infoleg-normativa-nacional

de entidades que se encuentran en InfoLeg, su caracterización lingüística y las dificultades que presentan su tratamiento automático. En nuestro análisis, hemos ordenado los diferentes tipos de entidades de menor a mayor proporción de error en el análisis humano y automático, para desarrollar una herramienta en la que se pueda configurar el nivel de fiabilidad y cobertura de las anotaciones resultantes.

Existen muy diversos tipos de entidades en los textos. En esta primera fase nos hemos centrado en las entidades que son más fácilmente reconocibles por anotadores humanos y métodos automáticos. De esta forma, conseguiremos tratar en primer lugar entidades con alta fiabilidad (bajo error), y así tener siempre disponible una versión del sistema que puede dejar algunas entidades afuera pero que identifica con alta fiabilidad las que sí identifica.

Este tipo de entidades más fiables son principalmente documentos: leyes, resoluciones, decretos, tratados, etc. Por ejemplo, en la Nota Externa 59/2009 de la Administración Federal de Ingresos, Dirección General de Aduanas del 23 de Junio de 2009, es posible encontrar las siguientes entidades, que se muestran subrayadas en el ejemplo:

> Bs. As., 23/6/2009
> VISTO la Resolución del MINISTERIO DE ECONOMIA Y PRODUCCION N° 127 del 10 de marzo de 2008, modificatoria de la Resolución N° 534 del 14 de julio de 2006, por la que se instruyó a la DIRECCION GENERAL DE ADUANAS de la ADMINISTRACION FEDERAL DE INGRESOS PUBLICOS, para que aplique como base de valoración de las exportaciones de gas natural el precio fijado para esta mercadería por el CONVENIO MARCO ENTRE LA REPUBLICA ARGENTINA Y LA REPUBLICA DE BOLIVIA PARA LA VENTA DE GAS NATURAL Y LA REALIZACION DE PROYECTOS DE INTEGRACION ENERGETICA, suscripto el 29 de junio de 2006.

En una segunda fase pasaremos a anotar entidades que se expresan en el texto con más vaguedad, como por ejemplo personas jurídicas, roles, etc., así como entidades nombradas no específicas de dominio legal, como fechas, lugares, cantidades, instituciones, etc. Aparecen subrayadas en el siguiente ejemplo:

> Bs. As., 23/6/2009
> VISTO la Resolución del MINISTERIO DE ECONOMIA Y PRODUCCION N° 127 del 10 de marzo de 2008, modificatoria de la Resolución N° 534 del 14 de julio de 2006, por la que se instruyó a la DIRECCION GENERAL DE ADUANAS de la ADMINISTRACION FEDERAL DE INGRESOS PUBLICOS, para que aplique como base de valoración de las exportaciones de gas natural el precio fijado para esta mercadería por el CONVENIO MARCO ENTRE LA REPUBLICA ARGENTINA Y LA REPUBLICA DE BOLIVIA PARA LA VENTA DE GAS NATURAL Y LA REALIZACION DE PROYECTOS DE INTEGRACION ENERGETICA, suscripto el 29 de junio de 2006.

Para esta tarea evaluaremos el desempeño de un identificador de entidades nombradas genérico, ya entrenado para identificar este tipo de entidades también genéricas, como por ejemplo el de Freeling [24].

En una tercera fase anotaremos conceptos abstractos, como los que se ven en el siguiente ejemplo:

> Bs. As., 23/6/2009
> VISTO la Resolución del MINISTERIO DE ECONOMIA Y PRODUCCION N°
> 127 del 10 de marzo de 2008, modificatoria de la Resolución N°
> 534 del 14 de julio de 2006, por la que se instruyó a la DIRECCION
> GENERAL DE ADUANAS de la ADMINISTRACION FEDERAL DE INGRESOS PUBLICOS,
> para que aplique como base de valoración de las exportaciones
> de gas natural el precio fijado para esta mercadería por el CONVENIO
> MARCO ENTRE LA REPUBLICA ARGENTINA Y LA REPUBLICA DE BOLIVIA
> PARA LA VENTA DE GAS NATURAL Y LA REALIZACION DE PROYECTOS DE
> INTEGRACION ENERGETICA, suscripto el 29 de junio de 2006.

En nuestra aproximación, estos diferentes tipos de entidades serán tratados de forma incremental, tanto por lo que respecta al desarrollo de la anotación, desarrollo de las herramientas específicas que los traten y aplicación de las herramientas que los identifican en texto.

Los conceptos más abstractos tienen más diversidad en su realización lingüística, por lo tanto esperamos que los métodos de aprendizaje basados en características superficiales tengan un rendimiento más bajo. En cambio, esperamos que los recientes métodos proyectivos de aproximación a la semántica, como los *word embeddings* [25], mejoren el rendimiento en el reconocimiento de este tipo de entidades.

Como referencia para determinar si una expresión es una entidad nombrada o no, hemos usado el diccionario Eurovoc[4], un tesauro multilingüe usado para indexación de documentación en la Unión Europea.

5. Descripción de la Aproximación

Nuestra propuesta para la identificación automática de entidades nombradas puede verse resumida en el gráfico de la Figura 1. Se trata de una aproximación híbrida, que combina técnicas de aprendizaje automático con reglas manuales, aplicando la técnica más adecuada al tipo de entidad, en una arquitectura de cascada: los tipos de entidad más fiables se tratan primero, aplicando mecanismos específicos, y los tipos menos fiables se van tratando por nivel de fiabilidad descendente, aplicando mecanismos cada vez más genéricos, basados principalmente en aprendizaje automático a partir de corpus anotados.

Los principios básicos del sistema son la modularidad y el diagnóstico de fiabilidad. De esta forma, el usuario podrá elegir obtener mayor fiabilidad en los resultados o mayor cobertura de tipos de entidad, o directamente seleccionar el tipo de entidades a tratar.

[4] http://eurovoc.europa.eu/

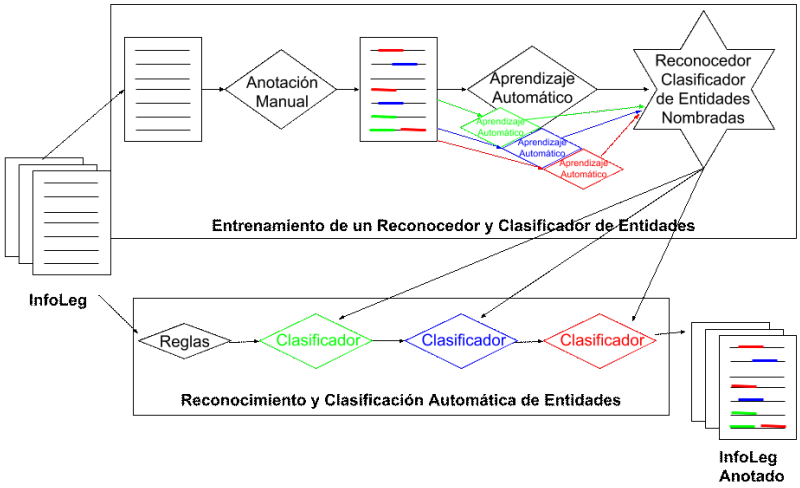

Figura 1. Arquitectura de la propuesta para el tratamiento automático de entidades nombradas en InfoLeg.

6. Evaluación

Hemos implementado y evaluado una primer versión del clasificador de entidades nombradas basado en aprendizaje automático.

A partir del corpus etiquetado, se entrenó y evaluó un detector de entidades nombradas de referencia, el Stanford CFG-NERC [26]. Se entrenó con dos corpus etiquetados diferentes: un corpus etiquetado mediante reglas desarrolladas manualmente y un corpus etiquetado totalmente manualmente. Describimos ambos a continuación, y finalmente describimos los resultados del clasificador entrenado con cada uno de ellos.

6.1. Reglas manuales

Las reglas manuales tratan solamente las entidades que se expresan de forma más predecible en el texto, y solamente algunas de sus menciones, las que son inambiguas, con las siguientes expresiones regulares, donde el paréntesis indica la opcionalidad de `N°`:

```
Ley (N°) <dígito>
Decreto (N°) <dígito>/<año>
Expediente (N°) <dígito>/<año>
Resolución (N°) <dígito>/<año>
```

```
Resolución (N°) <dígito>
Disposición (N°) <dígito>/<año>
Artículo <dígito>
```

Estas reglas se implementan mediante expresiones regulares y se aplican mediante correspondencia de patrones. No tienen error porque solamente se aplican a casos inambiguos, pero sí pueden tener problemas de cobertura, ya que pueden no identificar entidades que se nombran con patrones distintos.

Se aplicaron estas reglas sobre el corpus completo de InfoLeg, y se reconocieron más de 120.000 entidades. Luego este corpus se dividió en dos mitades de más o menos el mismo tamaño, una para entrenamiento y otra para evaluación.

6.2. Anotación manual de entidades nombradas para aprendizaje automático

Se han anotado manualmente 11 documentos con un total de 56.424 palabras, detectando 495 entidades nombradas, 315 de las cuales eran únicas y el resto repeticiones, con un total de 1.994 palabras, distribuidas como se ve en la Tabla 1.

Tipo	Número de instancias
Ley	140
Resolución	98
Nota	52
Decreto	52
Sin clasificar	153

Cuadro 1. Tipos de entidades anotadas manualmente en el corpus Infoleg.

En los procesos de anotación manual es muy importante mantener la consistencia en las anotaciones, ya que eso garantiza un mejor rendimiento de los algoritmos de aprendizaje automático, que pueden aprender de las sistematicidades pero no de las inconsistencias.

A los fines de lograr una anotación óptima de las entidades mencionadas en el texto normativo, se han establecido unas directrices para guiar el proceso de anotación manual por parte de los expertos y aumentar su consistencia. Entre los criterios que se han desarrollado, destacamos los siguientes:

1. Cuando se reenvía a un artículo (y eventualmente a un inciso o punto) determinado de una Ley o del mismo texto, se anotan por separado la ley y el artículo o inciso:
   ```
   ...según lo dispuesto en el segundo párrafo del artículo 43
   de la Ley de Impuesto al Valor Agregado, texto ordenado en
   1997 y sus modificaciones...
   ```
2. Se anotan por separado la fecha de publicación y la entidad publicada

... la Resolución SECRETARIA DE ENERGIA Nº 659/2004 del 17 de junio de 2004...

3. Cuando dos entidades mencionadas están unidas por la "y", se anotan como entidades separadas:
   ... los beneficios establecidos por las leyes 24.043 y 24.411...
4. Cuando junto a la entidad mencionada se aclara la existencia de otras entidades que la complementan y/o modifican, se anotan como entidades separadas:
   .... los beneficios establecidos por las leyes 24.043 y 24.411, sus complementarias y modificatorias...

Para las anotaciones manuales se usó la interfaz de anotación Brat [27], que se muestra en la Figura 2.

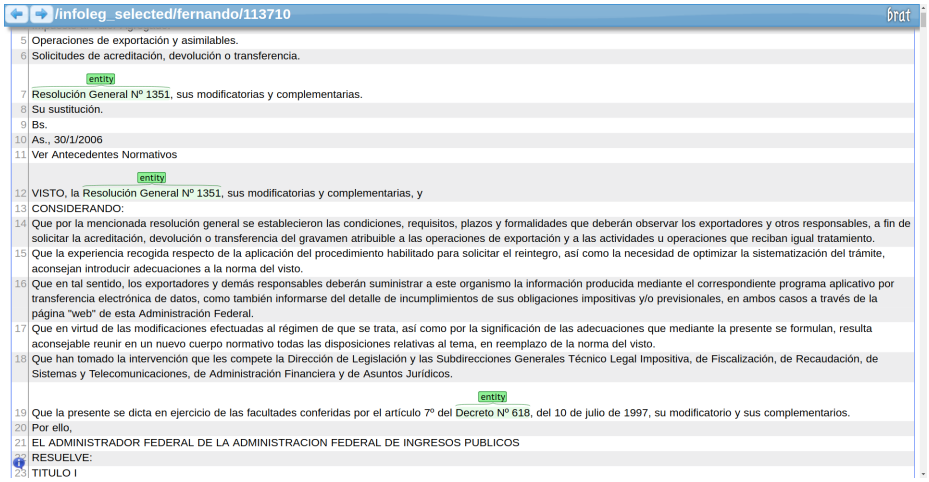

Figura 2. Ejemplo de uso de la interfaz de anotación brat [27] para anotar Entidades Nombradas en InfoLeg.

Para entrenamiento y evaluación, este corpus se dividió de forma que el 80 % se usó para entrenar (370.000 palabras, 324 entidades) y el 20 % para evaluar (81.000 palabras, 173 entidades).

6.3. Resultados

En el Cuadro 2 se muestran los resultados obtenidos aplicando el Stanford NERC al corpus etiquetado con reglas o etiquetado manual, entrenando con el corpus etiquetado con reglas o manualmente. Se muestran resultados de precisión (porcentaje de aciertos entre los casos clasificados como entidades nombradas), cobertura (porcentaje de entidades nombradas que son clasificadas como tales) y su media harmónica, la medida-F. También se muestran números absolutos

de cantidad de errores (número de casos clasificados como entidades nombradas que no lo son) y cantidad de silencios (número de entidades nombradas que no son clasificadas como tales).

Podemos observar que cuando entrenamos y evaluamos con el corpus etiquetado con reglas, las cifras de precisión y cobertura son casi del 100 %. Esto es debido a que hay muchas entidades nombradas que no se consideran como tales tampoco en la evaluación, por lo tanto, no son consideradas ni como errores ni como silencios.

En cambio, al evaluar con el corpus etiquetado a mano, donde se reconocen muchas más entidades nombradas, observamos que aunque el número de errores es relativamente bajo, lo cual mantiene una precisión alta, el número de silencios es muy alto, lo cual afecta directamente a la cobertura. Para paliar este problema, necesitamos un mayor número de ejemplos etiquetados e incorporar técnicas de generalización de los ejemplos como *word embeddings*.

	precisión	cobertura	medida-F	N. de errores	N. de silencios
anotación con reglas	99 %	99 %	99 %	178	95
anotación manual	84 %	56 %	67 %	38	157

Cuadro 2. Resultados (en porcentaje) de aplicar el StanfordNERC al corpus etiquetado con reglas o etiquetado manual, entrenando con el corpus etiquetado con reglas o manualmente, respectivamente. El error son los falsos positivos y el silencio los falsos negativos.

7. Conclusiones y Trabajo Futuro

En este artículo hemos presentado una aproximación para el etiquetado automático de Entidades Nombradas en el corpus InfoLeg. Se trata de una aproximación híbrida, que combina reglas manuales con aprendizaje automático. En este momento hemos finalizado la fase de análisis de tipos de entidades, hemos diseñado una aproximación gradual, en cascada, por fiabilidad del tipo de entidad, hemos desarrollado reglas manuales para detectar casos sencillos automáticamente, hemos anotado una pequeña porción del corpus y hemos entrenado y evaluado un aprendedor automático con el corpus anotado manualmente y con el corpus anotado mediante reglas, obteniendo resultados preliminares prometedores. Para mejorar la cobertura del sistema, necesitamos aumentar el número de anotaciones manuales, para lo cual requeriremos la colaboración de la comunidad.

La herramienta resultante quedará a disposición de la comunidad. También se aplicará al corpus InfoLeg para aumentar los hipervínculos de que dispone el corpus, mejorar la recuperación de información basada en palabras clave y mejorar otras aplicaciones como recomendación de textos relacionados. La porción del

corpus anotado manualmente y todo el corpus anotado automáticamente también estarán a disposición de la comunidad, así como las diferentes aplicaciones resultantes.

Referencias

1. InfoLEG información legislativa y documental. http://www.infoleg.gob.ar/?page_id=310. Accessed: 2018-06-28.
2. Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. Named entity recognition and resolution in legal text. In Semantic Processing of Legal Texts, pages 27–43. Springer, 2010.
3. Rohan Nanda, Giovanni Siragusa, Martin Theobald, Guido Boella, Livio Robaldo, Francesco Costamagna, et al. Concept recognition in European and national law. Frontiers in Artificial Intelligence and Applications, Volume 302: Legal Knowledge and Information Systems, 2017.
4. Mihai Surdeanu, Ramesh Nallapati, and Christopher Manning. Legal claim identification: Information extraction with hierarchically labeled data. In Workshop Programme, page 22, 2010.
5. Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. A low-cost, high-coverage legal named entity recognizer, classifier and linker. In Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law, pages 9–18. ACM, 2017.
6. Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Luc Schneider. Sweetening ontologies with DOLCE. In International Conference on Knowledge Engineering and Knowledge Management, pages 166–181. Springer, 2002.
7. JAPJ Breukers and RJ Hoekstra. Epistemology and ontology in core ontologies: FOLaw and LRI-Core, two. 2004.
8. Rinke Hoekstra, Joost Breuker, Marcello Di Bello, Alexander Boer, et al. The LKIF core ontology of basic legal concepts. LOAIT, 321:43–63, 2007.
9. Gianmaria Ajani, Guido Boella, Luigi Di Caro, Livio Robaldo, Llio Humphreys, Sabrina Praduroux, Piercarlo Rossi, and Andrea Violato. The European taxonomy syllabus: A multi-lingual, multi-level ontology framework to untangle the web of European legal terminology. Applied Ontology, 11(4):325–375, 2016.
10. Tara Athan, Guido Governatori, Monica Palmirani, Adrian Paschke, and Adam Wyner. LegalRuleML: Design principles and foundations. In Reasoning Web International Summer School, pages 151–188. Springer, 2015.
11. Llio Humphreys, Guido Boella, Livio Robaldo, Luigi Di Caro, Loredana Cupi, Sepideh Ghanavati, Robert Kevin Muthuri Kiriinya, and Leon van der Torre. Classifying and extracting elements of norms for ontology population using semantic role labelling. In The 15th International Conference on Artificial Intelligence & Law—San Diego, June 8-12, 2015, 2015.
12. Mírian Bruckschen, Caio Northfleet, DM Silva, Paulo Bridi, Roger Granada, Renata Vieira, Prasad Rao, and Tomas Sander. Named entity recognition in the legal domain for ontology population. In Workshop Programme, page 16, 2010.
13. Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, and Giulia Venturi. Ontology learning from Italian legal texts. Law, Ontologies and the Semantic Web, 188:75–94, 2009.

14. Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. Ontology population and alignment for the legal domain: YAGO, Wikipedia and LKIF. In Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with the 16th International Semantic Web Conference (ISWC 2017), 2017.
15. Gordon V Cormack, Maura R Grossman, Bruce Hedin, and Douglas W Oard. Overview of the TREC 2010 legal track. In Proc. 19th Text REtrieval Conference, page 1, 2010.
16. Maura R Grossman and Gordon V Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. Rich. JL & Tech., 17:1, 2010.
17. Erich Schweighofer. Semantic indexing of legal documents. In Semantic Processing of Legal Texts, pages 157–169. Springer, 2010.
18. Maura R Grossman, Gordon V Cormack, and Adam Roegiest. TREC 2016 Total Recall Track Overview. In TREC, 2016.
19. M Saravanan and Balaraman Ravindran. Identification of rhetorical roles for segmentation and summarization of a legal judgment. Artificial Intelligence and Law, 18(1):45–76, 2010.
20. Ben Hachey and Claire Grover. Extractive summarisation of legal texts. Artificial Intelligence and Law, 14(4):305–345, 2006.
21. Rachoud Winkels, Alexander Boer, Bart Vredebregt, and Alexander von Someren. Towards a legal recommender system. In Proc. of the 27th Int'l Conf. on Legal Knowledge and Information Systems, 2014.
22. Christopher Collins, Fernanda B Viegas, and Martin Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on, pages 91–98. IEEE, 2009.
23. Julia A Ekstrom and Gloria T Lau. Exploratory text mining of ocean law to measure overlapping agency and jurisdictional authority. In Proceedings of the 2008 international conference on Digital government research, pages 53–62. Digital Government Society of North America, 2008.
24. Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, May 2012. ELRA.
25. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. January 2013.
26. Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 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, 2005. Association for Computational Linguistics.
27. Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. BRAT: A web-based tool for NLP-assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL '12, pages 102–107, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.