elDial.com
Contenidos Jurídicos

La inteligencia artificial entiende el lenguaje "talcahuanensis"

(Impacto de los lenguajes de procesamiento de lenguaje natural (NLP) en la recuperación de información legal)

Por el Dr. Horacio R. Granero ***[1]***

Se augura que dentro de poco no hará falta leerse una resolución judicial para extraer conclusiones: qué se pedía, qué argumentos se utilizaron y qué sentenció el tribunal, en la medida que el **procesamiento de lenguaje natural (NLP, *Natural Language Processing*) se perfila como un avance en el proceso de extracción automatizada de conocimiento de textos jurídicos**, cualse pondría en manos de los investigadores **una potente herramienta de "*minería de argumentos*" aplicada a la argumentación jurídica**, así como para analizar –tanto desde una perspectiva académica como sociológica– las resoluciones judiciales de los tribunales argentinos.

Hoy podemos saber, con una certeza bastante aproximada a la realidad, cuál es la tendencia de los Tribunales en la resolución de un tema, y la probabilidad de un resultado favorable o desfavorable y en cuestión de segundos informarle al cliente las perspectivas de su petición.

Ciencia ficción, no, realidad gracias a la inteligencia artificial. Intentaremos ver cómo ello es posible.

A través de diversas técnicas, entre las que se incluyen la *tokenización* y *vectorización del texto*, el desarrollo de herramientas para el análisis sintáctico y *semántico* (y, eventualmente, el *pragmático* o de significado pretendido) y la utilización de herramientas de Machine Learning y Deep Learning, se han desarrollado diferentes modelos de procesamiento (y comprensión) de texto que resultan útiles en el campo de la investigación jurídica.

Aunque los algoritmos de "*deep learning*" (aprendizaje profundo) han evolucionado mucho en los últimos años, la comprensión de un lenguaje tan complejo, en forma y fondo, como el jurídico es todo un reto**.** Una herramienta de tecnología informática aplicada al Derecho debe contemplar necesariamente el procesamiento de lenguaje natural, "*machine learning*" (aprendizaje automático) y redes neuronales profundas, todo ello enmarcadas en la idea de IA o Inteligencia Artificial.Ahora bien, ¿qué significan estos términos y qué

implican?Comenzando con el procesamiento de lenguaje natural o NLP se puede decir que es el campo que estudia la comprensión y manipulación del lenguaje natural humano, es decir tal y como nos expresamos por escrito o de viva voz, por parte de un ordenador. Por ello trabaja áreas como el entendimiento por parte de una máquina del lenguaje humano, su percepción o generación. Por ejemplo, un software de traducción aplica NLP, siendo una de sus tareas entender que *«lawyer»* es una palabra inglesa que en castellano se traduce como *«abogado»*.

Todos los modelos de NLP requieren un entrenamiento con datos de acuerdo al objetivo al que apuntan y serán tan buenos como la calidad, cantidad y completitud de los datos con que fue entrenado y el poder del algoritmo seleccionado.

Los desarrollos de modelos de NLP incluyen (entre otros) herramientas para:

Categorización de textos: Ordena el texto en diferentes taxonomías definidas por el usuario

Agrupamiento de textos: Agrupa documentos de acuerdo a similitudes en su contenido

Extracción de información: identifica contenido valioso en texto no estructurado

Resolución nombre-entidad: extrae nombres que clasifica en etiquetas predeterminadas

Identificación de relaciones: identifica relación semántica entre entidades

Análisis de sentimientos: descifra el significado emocional detrás del lenguaje

Lo cierto es que elNLP, en mayor o menor grado de complejidad, se aplica a múltiples tareas de nuestro día a día, y desde hace bastantes años como para la clasificación de texto para detectar *spam* en el correo electrónico o la indicación de errores gramaticales mientras se escribe un texto. Ahora bien, ¿Cómo intenta un ordenador analizar nuestras palabras y sus múltiples significados y variantes? Hay dos grandes corrientes: la simbólica y la estadística. La aproximación simbólica consiste en un sistema de reglas del estilo *«Si ocurre esto, haz eso»* o "*if-then*" en programación- Puede llegar a generar árboles de reglas realmente complejos, y para un humano resulta más sencillo de entender y predecir su comportamiento. Es la técnica utilizada generalmente con los "contratos inteligentes" o *"smart contracts"* [2]

La estadística es más moderna y explotó con la aparición de las técnicas de *machine learning* o aprendizaje automático y el acceso a bases de datos voluminosas. Esencialmente consiste en anotar y estructurar una serie de textos relacionados con la materia que nos interesa (divorcio, por ejemplo). Es decir, textos que el software puede identificar sin dificultad. A partir de esos primeros datos anotados, se crea un modelo estadístico al que se le comenzarán a proporcionar datos no anotados ni estructurados, que el algoritmo por su cuenta deberá comenzar a estructurar y clasificar de acuerdo a la información inicialmente suministrada. [3]

Así el modelo estadístico generado por el programa puede "predecir estadísticamente" el tipo de palabras que se usarán en una frase. De modo que con relativamente menos trabajo, y muchos datos, se puede avanzar de forma más rápida y eficiente en una tarea como el NLP. Al momento existen tareas que NLP ya entiende como resueltas, y hay otras en las que todavía hay trabajo por hacer o las que aún son complejas para una máquina. Por ejemplo, son tareas casi resueltas la detección de spam, saber si lo leído es un verbo o un adjetivo, o el reconocimiento de nombres propios. Se progresa en conocer el sentido positivo, negativo o neutro de un texto, en deducir el significado de las palabras, la relación entre términos (por ejemplo que un *«él»* hace referencia a *«Juan»*), su traducción o la extracción de información.

Se debe reconocer que todavía es difícil para una máquina responder una pregunta con toda la exactitud deseada (los abogados lo hacemos siempre?...) o mantener un diálogo en tiempo real.

NLP intenta hacer comprensible el lenguaje humano para una máquina pero su principal enemigo es la *ambigüedad*, algo de lo que el lenguaje humano está repleto. Ya sea por el uso de la ironía, el sarcasmo, los registros informales, los errores de pronunciación o escritura, la mezcla de idiomas y tantas otras variantes que afectan al lenguaje humano escrito y hablado. Por tanto, que una máquina sea capaz de comprendernos y dar respuesta no es tarea fácil todavía. Hay progresos obviamente importantes (como el caso de asistentes inteligentes como Siri o Alexa), pero aún existe un camino largo por recorrer, especialmente cuando el software debe entender más allá de áreas o dominios muy concretos. [4]

Convengamos, igualmente, que el lenguaje humano legal -o "*talcahuanensis*" como se lo conoce cariñosamente en la jerga tribunalicia por la sede física del Palacio de Justicia en Buenos Aires ubicado en la calle Talcahuano- tiene complejidades bastante particulares**:** la sintaxis legal es enrevesada y poco natural, se usan frases mucho más largas (entre 20 y 25 palabras más de media que en un periódico, por ejemplo), hay un mayor número de preposiciones pero un menor número de verbos o adverbios, se usan múltiples complementos encadenados y frases subordinadas, lo que hace que existan vínculos entre palabras o frases muy separadas. Además, muchos de los términos legales usados solo tienen sentido en el ámbito jurídico, sin existir una correlación en el lenguaje natural común. [5]

Aplicar NLP a textos legales implica unos esfuerzos concretos, no es una tarea tan sencilla como aplicar los métodos del lenguaje natural común al "*talcahuanensis*", pero en cualquier caso no es una barrera insuperable. De hecho, en la actualidad la mayoría de las aplicaciones utiliza NLP, generalmente para recuperar información o para extraer información, que no es lo mismo.

La recuperación de información, a diferencia de la recuperación tradicional (como en el caso de la denominada "*lógica de Boole*", como por ejemplo "*hijos **Y** alimentos*") en el

caso del NLP ayuda a hacerla mediante conceptos, de modo que, si se buscan documentos que incluyan el término "*hijo*", el software muestra los que mencionan esa palabra expresamente pero también los que incluyen conceptos relacionados como "descendientes", etc. aunque no incluya el término "*hijo*" expresamente, dado que NLP entiende que aunque esos documentos no incluyen al término "hijo", están vinculados al mismo. La extracción de información, por su parte, consiste en la tarea de extraer automáticamente "*información estructurada"*, el descubrimiento de patrones ocultos, agrupación según taxonomías predeterminadas, de "*documentos desestructurados*" o "semi-estructurados". Es decir, sacar datos útiles y ordenados de textos en principio no preparados para ser "entendidos" por una máquina, y ésta es la función que se utiliza para analizar jurisprudencia, para extraer información relevante para plantear una estrategia procesal. Luego esta información se usa para generar informes, crear visualizaciones de un gran conjunto de documentos o ayudar en la preparación de un asunto, y es lo que en definitiva ayuda al profesional a tomar mejores decisiones. [6]

En elDial.com surgió en 2016 la idea de desarrollar un programa que, aproveche la experiencia de la empresa en el manejo de bases de datos on line y las nuevas tecnologías de tratamiento de la información en inteligencia artificial, particularmente en el campo del NLP y así nació Sherlock-Legal creado para ser un asistente del abogado, a través de una interfaz gráfica dinámica, intuitiva y sencilla donde se efectúan las preguntas en "lenguaje natural" (por ejemplo, "En el caso de una separación matrimonial los hijos deben quedar con la madre?") no con "*descriptores*" o "*lógica boolena*" mencionada anteriormente. Mediante algoritmos generados al efecto se analizan sintácticamente y se interpretan los precedentes judiciales con el fin de encontrar los fragmentos relacionados con la pregunta formulada que el programa considere más relevantes. Posteriormente *Sherlock* despliega un grupo de las distintas respuestas que considera pertinentes, generándose gráficos que indican los porcentajes de aceptación o rechazo con la pregunta efectuada, dando, finalmente, su opinión en forma automática sobre la probabilidad que ésta sea afirmativa o negativa con relación a la consulta efectuada. El sistema si bien en esta primera etapa fue diseñado para ser aplicable a jurisprudencia, el desarrollo puede ser utilizado con cualquier base de datos estructurada y aún no estructurada. [7]

El problema a resolver por el equipo de desarrollo consistió principalmente en generar una herramienta que responda relacionando preguntas realizadas en lenguaje natural conparte de los textos jurídicos de la base de datos, razón por lo que se optó por un modelo de *Question Answering* (QA)para un dominio cerrado. El programa, en definitiva, se resume a dos módulos, buscando hallar sumarios pertinentes con la pregunta realizada (problema de la pertinencia) y luego hallar el/los párrafos dentro de los *sumarios pertinentes* que directa o indirectamentemejor respondanse busca interpretar en el texto una *respuesta* a la pregunta realizada, analizando si ésta es por "*si*" o por "*no*". En el primer Módulo, con relación al problema de la pertinencia, se realiza en primer lugar un análisis sintáctico de la pregunta realizada y luego el programa se queda con los lemas y entidades, posteriormente se obtienen las raíces de la palabras, se quitan los *stopwords* (palabras que agregan volumen pero no información) y luego se busca en la base de fallos utilizando un modelo de *Bag of words* (Bag of N-grams words) o sea vectores de ocurrencia de las palabras/N-gramas de

dimensión n que forman matrices para todos los sumarios y TF-IDF*(Term frequency – Inverse document frequency)*, técnica de recuperación de información que pesa la frecuencia de un término (TF) y su frecuencia de documento inversa (IDF). Cada palabra o término tiene su respectivo puntaje TF e IDF y el producto de los puntajes TF e IDF de un término se lo considera el peso TF * IDF de ese término. Posteriormente se utilizan criterios de similitud para encontrar documentos similares como el *criterio del coseno* (utilizando los vectores creados para los documentos se genera su producto escalar y se aplica el teorema del coseno resultando una nueva matriz) y se generan distintos escenarios para la pregunta. Los resultados se verifican contra un modelo entrenado por *Naive Bayes* (modelo probabilístico) si la lista de documentos resultante es la más pertinente. En el segundo Módulo, con respecto al problema de la respuesta,se recibe en primer lugar la pregunta formulada junto con los *ids* de los documentos que fueron previamente pre-procesados, luego se analizan buscando el/los fragmento/s que más se acerquen a la respuesta y en el caso de preguntas fácticas si el/los fragmentos seleccionados responden por "si" o por "no". Para una aproximación sintáctica (resolver oraciones en voz pasiva y oraciones subordinadas) se pre procesa el texto y se lo *divide en párrafos* (que permite resolución de anáforas) para luego dividirlo en sentencias.

A partir de 2018 se produjo el desarrollo de modelos basados en el concepto *"transformer"* que significó un salto enorme, ya que permite, a través de su modelo Codificador-Decodificador, procesar e interpretar significados (semántica) en textos extensos. El modelo desarrollado por Google, BERT (*Bidirectional Encoder Representation from Transformers*), fue primero. A partir de su estructura y concepto se desarrollaron modelos como XLNet (también de Google/CMU), RoBERTa (de Facebook), ALBERT (de Google) y Megatron (de NVIDIA) y ya se tiene conocimiento de BETO desarrollado por investigadores de la Universidad de Chile para textos en castellano. Estos sistemas, si bien no "*comprenden*", responden en base a las estructuras en las que se basó su entrenamiento. BERT, básicamente implica entrenar una red neuronal para aprender " *lenguaje* ", y luego esta red se usa como una columna vertebral para realizar diversas tareas de PNL. Lo que hizo revolucionario su uso fue el hecho de que se erigió en la primera representación de lenguaje profundamente *bidireccional*, lo que significa que miró las palabras después de una palabra dada, no solo las palabras anteriores, y su arquitectura está basada en *transformadores*, lo que le ayuda a ponderar el contexto en el que aparece una palabra en una oración. Todos estos factores combinados explican su éxito de alguna manera. En una investigación reciente dio excelentes resultados analizando, por ejemplo, opiniones de películas [8]

En conclusión, el procesamiento de lenguaje natural o NLP consigue, como en el caso de *Sherlock-Legal*, que las máquinas puedan entender, manipular y generar textos a partir del lenguaje humano, escrito o verbal y dado que el Derecho es mayoritariamente texto -incluso con las particularidades apuntadas- se aumentan las capacidades del profesional en la solución de problemas consultados, ahorrándole tiempo y ganando en eficiencia, si bien eso le obliga a centrarse cada vez más en tareas de mayor valor, imposibles de suplantar por la máquina, como es la necesaria empatía que los profesionales deben contar con el cliente para entender el problema planteado, o la Justicia y Equidad que deben regir al magistrado

en las resoluciones que dicta, dejando al software que se encargue de las tareas más rutinarias y comunes (y que la inteligencia artificial muy probablemente haga mejor…)

[1] Presidente de elDial.com, CEO del proyecto Sherlock-Legal y Profesor Emérito de la UCA.**[2]**GRANERO, Horacio R. "Los contratos inteligentes y la tecnología "blockchain" (su encuadre en el Código Civil y Comercial de la Nación)", publicado el 7 de marzo 2018 en elDial.com, Citar DC24BB

[3] Ver antecedentes en "Informática Jurídica: Acercando la Informática al Derecho en las puertas del siglo XXI" por Daniel Pastor en file:///C:/Users/USUARIO/Downloads/Dialnet-InformaticaJuridica-5109931.pdf

[4]https://legaltechies.es/2017/08/11/que-es-el-procesamiento-de-lenguaje-natural-y-como-afecta-en-tareas-legales/**[5]**Valga como ejemplo los Considerandos *"...Premito que asiste razón a la recurrente en su mohín.- Es que debe diferenciarse la obligación que los condóminos tenían, ya sea con las empresas de servicios públicos, el fisco o el consorcio, de la que ahora reclama la actora en concepto de reembolso.."* autos L. 74.104/2012/CA1 - "B, B. M. c/ De B. S. s/ Cobro de sumas de dinero" – CNCIV – SALA G - 18/11/2015Publicado el 29/01/2016Citar: elDial.com - AA9444

[6]https://legaltechies.es/2017/08/11/que-es-el-procesamiento-de-lenguaje-natural-y-como-afecta-en-tareas-legales/

[7]https://www.cronista.com/legales/Inteligencia-juridica-artificial-20180606-0006.html

[8]https://towardsdatascience.com/how-does-bert-reason-54feb363211

Citar: elDial DC2991

Publicado el: 04/03/2020

copyright © 1997 - 2020 Editorial Albrematica S.A. - Tucumán 1440 (CP 1050) - Ciudad Autónoma de Buenos Aires – Argentina