Hacia un sistema de extracción de definiciones en textos jurídicos

Alexy J. Sánchez H. y Melva J. Márquez R.
Universidad de Los Andes, Postgrado en Modelado y Simulación de Sistemas, Centro de Investigación y Proyectos en Simulación y Modelos
Mérida, Venezuela
alexy@ula.ve, melva@ula.ve

La necesidad de procurarnos información de forma selectiva y eficiente es una realidad innegable que surge como consecuencia directa del cada vez mayor volumen de información. En este sentido, los Sistemas de Extracción de Información (SEI) representan herramientas de gran valor para el aprovechamiento eficiente del conocimiento presente en los textos. Sin embargo, los conocimientos presentes en los textos varían de acuerdo a la naturaleza de éstos; de allí que los SEI deben ser desarrollados para tipos de textos específicos y con el fin de extraer información específica. En este sentido, los cuerpos de leyes son creados con la finalidad de establecer reglas sociales obligatorias, lo cual requiere crear y definir de forma unívoca muchos conceptos que dejen bien en claro cuando se comete o no una falta. Además, los cuerpos de leyes son por lo general muy extensos (algunos constan de más de 600 artículos), y todas las personas que residen en un país tienen necesidad revisarlas con algún propósito específico en algún momento. Es así pues, que en el presente estudio presentaremos un SEI diseñado para la extracción de definiciones en textos jurídicos, y más concretamente en los cuerpos de leyes venezolanas.

Palabras clave: Sistemas de Extracción de Información, Definiciones, Inteligencia Artificial, Procesamiento del Lenguaje Natural,

I. Introducción

La necesidad de procurarnos información de forma selectiva y eficiente es una realidad innegable que surge como consecuencia directa del cada vez mayor volumen de información. En este sentido, los Sistemas de Extracción de Información (SEI) representan herramientas de gran valor para el aprovechamiento eficiente del conocimiento presente en los textos. La extracción de información es una técnica de la Inteligencia Artificial (IA) cuyo fin es el de obtener hechos a partir de una serie de documentos. Estos hechos son pues obtenidos a partir de documentos seleccionados y estructurados luego en una base de datos. Sin embargo, los conocimientos presentes en los textos varían de acuerdo a la naturaleza de éstos; de allí que los SEI deben ser desarrollados para tipos de textos específicos y con el fin de extraer información específica. A estos textos específicos a los que sirve un SEI se le conoce como *dominio de trabajo*. Es así pues, que en el presente estudio presentaremos un SEI diseñado para la extracción de definiciones en textos jurídicos, y más concretamente en los cuerpos de leyes venezolanas. Nuestro SEI extraerá definiciones presentadas particularmente el la Ley Orgánica del Trabajo, y los almacenará en una base de datos. La manipulación del SEI se hará por medio de una Interfaz Gráfica de Usuario (IGU), la cual tanto permitirá manipular los archivos a ser leídos como el acceso a la base de datos donde se almacenan los definens con sus definiciones. Tanto la base de datos como la IGU fueron diseñadas utilizando el lenguaje de

programación JAVA. Confiamos en que nuestro SEI será de gran utilidad dado el valor inherente que tienen los SEIs para la terminología, y por la necesidad real de extraer información selectiva que la naturaleza del tipo de texto con los que trabajamos impone; los cuerpos de leyes son creados con la finalidad de establecer reglas sociales obligatorias, lo cual requiere crear y definir de forma unívoca muchos conceptos que dejen bien en claro cuando se comete o no una falta. Además, los cuerpos de leyes son por lo general muy extensos (algunos constan de más de 600 artículos), y todas las personas que residen en un país tienen necesidad revisarlas con algún propósito específico en algún momento. Huelga decir los beneficios que el sistema representa para los profesionales del derecho. Es así pues, que es el objetivo del presente estudio, proponer un sistema de extracción de definiciones en textos jurídicos, y específicamente en La ley Orgánica del trabajo. En la siguiente sección presentaremos algunos aspectos teóricos de los textos jurídicos relevantes para este estudio. Seguidamente, expondremos las cuestiones lingüísticas con las que tuvimos que lidiar en la realización de nuestro SEI y los seudo algoritmos de las reglas empleadas en la detección y extracción de las definiciones. En la sección cuatro presentaremos las conclusiones del estudio.

II. Cuerpos de leyes: aspectos teóricos relevantes para el estudio.

En el campo jurídico, y específicamente en el ámbito de las leyes, observamos que éstas, además de ser numerosas, utilizan tanto un vocabulario como un conjunto de estructuras gramaticales sui generis que no son usualmente utilizados en el español corriente, lo que conlleva a que la interpretación de las leyes por parte del ciudadano común no sea una tarea fácil y que deba ser desarrollado por expertos. Estas estructuras gramaticales sui generis se caracterizan, por una parte, por una cierta rigidez en su estructura sintáctica y, por otra, por la utilización de un vocabulario conservador, lleno de tecnicismos y estable. Ambos componentes gramaticales, el sintáctico y el léxico, hacen del lenguaje jurídico una especie de lenguaje semiformal en el que no hay cabida para la creatividad. Esto se debe, como lo señalan algunos autores, a que el lenguaje jurídico ha sido concebido como un instrumento para connotar ideas con sentido unívoco. Elena de Miguel, profesora de la Universidad Autónoma de Madrid, muestra cuáles son esos rasgos léxico-sintácticos, de entre los cuales presentamos a continuación los más característicos con ejemplos tomados de la constitución (de Miguel, 2000):

a) Una marcada tendencia a construcciones nominales: la cual se hace patente al comparar el número de sustantivos y adjetivos en relación con el número de verbos utilizados.

[Con el fin supremo de refundar la República para establecer una sociedad democrática, participativa y protagónica, multiétnica y pluricultural en un Estado de justicia, federal y descentralizado, que consolide los valores de la libertad, la independencia, la paz, la solidaridad, el bien común, la integridad territorial, la

convivencia y el imperio…] (tomado del preámbulo de la constitución)

b) La abundante presencia de formas no personales del verbo; v.gr., infinitivos, participios presentes y pasados.

[Ninguna persona puede ser arrestada o detenida sino en virtud de una orden judicial, a menos que **sea** sorprendida in fraganti. En este caso, será llevada ante una autoridad judicial en un tiempo no mayor de cuarenta y ocho horas a partir del momento de la detención. Será juzgada en libertad, excepto por las razones determinadas por la ley y apreciadas por el juez o jueza en cada caso.] (apartado 1, Artículo 44)

c) Exceso de subordinación, abundancia de incisos y gran extensión de los párrafos

[El pueblo de Venezuela, en ejercicio de sus poderes creadores e invocando la protección de Dios, el ejemplo histórico de nuestro Libertador Simón Bolívar y el heroísmo y sacrificio de nuestros antepasados aborígenes y de los precursores y forjadores de una patria libre y soberana; con el fin supremo de refundar la República para establecer una sociedad democrática, participativa y protagónica, multiétnica y pluricultural en un Estado de justicia, federal y descentralizado, que consolide los valores de la libertad, la independencia, la paz, la solidaridad, el bien común, la integridad territorial, la convivencia y el imperio de la ley para esta y las futuras generaciones; asegure el derecho a la vida, al trabajo, a la cultura, a la educación, a la justicia social y a la igualdad sin discriminación ni subordinación alguna; promueva la cooperación pacífica entre las naciones e impulse y consolide la integración latinoamericana de acuerdo con el principio de no intervención y autodeterminación de los pueblos, la garantía universal e indivisible de los derechos humanos, la democratización de la sociedad internacional, el desarme nuclear, el equilibrio ecológico y los bienes jurídicos ambientales como patrimonio común e irrenunciable de la humanidad; en ejercicio de su poder originario representado por la Asamblea Nacional Constituyente mediante el voto libre y en referendo democrático, decreta la siguiente…] (preámbulo de la constitución)

d) Uso excesivo de construcciones pasivas, tanto perifrásticas como reflejas.

[Las leyes de procedimiento se aplicarán desde el momento mismo de entrar en vigencia, aun en los procesos que se hallaren en curso; pero en los procesos penales, las pruebas ya evacuadas se estimarán en cuanto beneficien al reo o a la rea, conforme a la ley vigente para la fecha en que se promovieron.](Artículo 24).

e) Abundancia de oraciones impersonales con “se”.

[Se garantiza, así mismo, la independencia y la autonomía de las iglesias y confesiones religiosas, sin más limitaciones que las derivadas de esta Constitución y de la ley. El padre y la madre tienen derecho a que sus hijos o hijas reciban la educación religiosa que esté de acuerdo con sus convicciones.](Artículo 59)

f) Abundancia de locuciones prepositivas; v.gr.; en el supuesto de, de conformidad con, etc.

[El Ejecutivo Nacional, para mantener y restablecer el orden público, proteger a los ciudadanos y ciudadanas, hogares y familias, apoyar las decisiones de las autoridades competentes y asegurar el pacífico disfrute de las garantías y derechos constitucionales, de conformidad con la ley, organizará:…](Artículo 332).

Por otra parte, Los cuerpos de leyes presentan una organización bien estructurada; e.g., las leyes están compuestas de títulos, estos últimos están compuestos de capítulos y los capítulos de artículos, que siguen una secuencia numerada de principio a fin. Además, existe una organización jerárquica de los conceptos enunciados; e.g., los conceptos más generales son presentados antes de los más específicos.

III. Cuestiones lingüísticas en la detección y extracción de definiciones

En los SEI es común extraer patrones heurísticos a partir de los textos de interés; esto es, del dominio de trabajo y que son específicos del mismo, lo que hace inviable la aplicación de tales patrones a otros dominios. Estos patrones heurísticos permiten crear reglas para la extracción de información pertinente. En el caso de los cuerpos de leyes, y específicamente de la extracción de definiciones en este tipo de documentos, estos patrones no son muy difíciles de extraer dado que, por una parte, como ya mencionamos en la sección anterior, los textos jurídicos presentan cierta rigidez en su estructura, tanto léxica y sintáctica como organizativa. Por otra parte, los marcadores de definición presentes en los cuerpos de leyes estudiados son muy limitados. En el siguiente cuadro se muestran los marcadores de definición presentes en la Ley Orgánica del Trabajo.

Forma Básica	Formas presentes
Entenderse	Se entiende por
	Se entenderá por
	Se entienden como
	Se entenderá que
Ser	Es
	Son
	Será
	Serán
Ser considerado	Será considerado
	Será considerada
	Serán considerados
	Serán consideradas

Cuadro No. 1. marcadores de definición presentes en la LOT.

No obstante, para los fines del SEI se trabajo solamente con las formas presentes de *entenderse*. Esto es así porque tanto las formas de *ser* como las de *considerarse* eran un tanto más complejas, y en nuestro caso estamos realizando un estudio preliminar

En la creación de patrones, nuestro mayor esfuerzo estuvo pues orientado en dos cuestiones fundamentales: a) la detección de marcadores metalingüísticos de la forma básica *entenderse que* nos dieran pista de la presencia de definiciones y b) el análisis sintáctico de las oraciones donde se hubiesen detectado definiciones para poder establecer los límites tanto de las palabras o frases definitorias como de las palabras o frases que se definían. En la mayoría de los SEI se incluyen diccionarios que permiten realizar análisis semánticos, pero este no fue nuestro caso. Dado a que este sistema es una humilde propuesta para un SEI para extracción de definiciones específicamente en las leyes, y que surge como resultado de un curso de maestría, quisimos mantenerlo lo más simple posible. Es así pues, como aunque un componente de análisis semántico, encontrar las palabras o frases que estaban siendo definidas, fue tomado en cuenta, el componente sintáctico jugo un papel crucial en la consecución del sistema extractor de definiciones. Esto se debe a que los marcadores metalingüísticos para establecer definiciones presentes en el corpus de las leyes estudiadas como ya dijimos son limitados y su uso está bastante restringido. No obstante, al momento de establecer los límites de las cláusulas definitorias y de las palabras o frases que ellas definían, fue necesario realizar un análisis sintáctico parcial; esto es, descubrir las relaciones de los distintos elementos de las oraciones que nos permitieran saber dónde terminaba un definen y comenzaba su definición. Esta forma de abordar el problema hizo posible que no fuera necesaria crear una gramática libre de contexto (GLC), común en los SEI. Los pasos seguidos para la extracción de definiciones se muestran a continuación en seudo código.

```
Se carga la ley que se desea analizar al programa
    Mientras no sea fin del documento
        Se leen las oraciones
        Si en oración está presente un marcador de definición entonces
            Se etiquetan las partes de la oración
            Se extrae el definens
            Se extrae su definición
            Se almacenan el definens con su definición.
```

Como se pudo apreciar en el seudo algoritmo anterior, una vez que se determina que existe una posible definición en una oración por medio de la identificación de los marcadores de definición, se realiza un análisis sintáctico parcial de la oración en cuestión. En dicho análisis se marcan los elementos constitutivos de la oración siguiendo el siguiente patrón.

1) Se etiquetan todas las clases de palabras constituidos por un conjunto de elementos finitos; e.g. artículos definidos e indefinidos,

preposiciones, cuantificadores, adjetivos y pronombres demostrativos, adjetivos y pronombres posesivos, conjunciones, pronombres relativos, conectores discursivos y reformuladores. Se etiquetan además algunas marcas que consideramos marcas de definición como las comas y los dos puntos.

2) En el caso de que la palabra a etiquetar no pertenezca a ninguna de las categorías anteriores; e.g. verbos, sustantivos y adjetivos, entonces la palabra es etiquetada como "vacío".

Según estas reglas de etiquetado, el siguiente artículo de la ley orgánica del trabajo

Artículo 42. Se entiende por empleado de dirección el que interviene en la toma de decisiones u orientaciones de la empresa, así como el que tiene el carácter de representante del patrono frente a otros trabajadores o terceros y puede sustituirlo, en todo o en parte, en sus funciones.

Quedaría etiquetado, después del marcador de definición "se entiende por", así:

vacío, preposición, vacío, artículo definido, conjunción, vacío, preposición, artículo definido, vació, preposición, vacío, conjunción, …

Ahora bien, la extracción del definen y de su definición dependen del marcador metalingüístico presente. En el ejemplo anterior, observamos que el marcador de definición "se entiende por" impone que el definen "empleado de dirección" se encuentre inmediatamente después de él y que la definición se encuentre inmediatamente después del definen. En este caso particular el definen está separado de su definición por medio del artículo definido "el". Pero esta no es la única forma en que el definen puede estar separado de su definición para este marcador de definición. En algunas ocasiones como en el caso del tercer párrafo del artículo 16

Se entiende por explotación, toda combinación de factores de la producción sin personería jurídica propia ni organización permanente, que busca satisfacer necesidades y cuyas operaciones se refieren a un mismo centro de actividad económica.

El definen está separado de su definición por medio de una coma seguida de un adjetivo demostrativo. Basados entonces en todas las formas en que el definen podría estar separado de su definición en la ley orgánica del trabajo cuando el marcador de definición era "se entiende por" se crearon las reglas. A continuación mostramos todos los casos como el definen podría estar separado de su definición para el caso del marcador "se entiende por".

- un artículo (definido o indefinido) después de un vacío
- una coma después de un vacío y un artículo después de una coma
- una coma después de un vacío y un adjetivo demostrativo después de la coma
- una coma después de un vacío y un cuantificador después de la coma
- un pronombre demostrativo después de un vacío
- un adjetivo demostrativo después de un vacío
- dos puntos después de un vacío
- la preposición “a” después de un vacío
- un cuantificador después de un vacío

Así pues, en seudo código, nuestra regla para encontrar el definen y la definición después de encontrar el marcador de definición “se entiende por” es el siguiente

Si en oración presente “se entiende por” entonces
    define los límites del definen
    extrae definen
    define los límites de la definición
    extrae definición

donde los límites del definen van desde inmediatamente después del marcador de definición hasta alguno de los patrones anteriormente mostrados. Para la extracción de la definición los límites estarían determinados por el límite del definen hasta el final de la oración.

Por ser un prototipo, en nuestro sistema de extracción de información no nos preocupamos por identificar co-referencias; esto es, no implementamos ningún mecanismo para saber cuando una oración independiente de la oración donde tanto un definen con su definición se encuentran deba ser unida a la anterior.
El criterio para extender la definición hasta el final de la oración obedece a que la oración debe tener sentido completo; aunque con ello dejamos por fuera oraciones complementarias que puedan arrojar más detalles sobre el definen en cuestión. Este fue el caso de 16 de las 39 definiciones extraídas por nuestro SEI. En la mayoría de los casos, las ideas que complementaban una definición se presentaban en parágrafos únicos que complementaban la idea expresada en un artículo anterior. Esto demuestra la importancia de tomar en cuenta co-referencias en un estudio posterior.

IV. Resultados y conclusiones

La manera habitual de evaluar los Sistemas de Recuperación de Información (SRI) es por medio de dos criterios conocidos como precisión y cobertura. Estos dos criterios permiten medir la efectividad del proceso de recuperación de información; esto es, la capacidad del SRI para extraer información relevante. En nuestro estudio hemos decidido adoptar estos dos criterios para la evaluación de nuestro

SEI. La precisión, adaptada a nuestro SEI podría ser definida como el porcentaje de definiciones, de entre las que devuelve el sistema, que son correctas. Una precisión del 100% indica que todas las definiciones extraídas fueron correctas, aunque no se asegura que se hayan extraído todas las definiciones. La cobertura, también adaptada a nuestro SEI, podría ser definida como el porcentaje de definiciones correctas que fueron extraídos por el sistema, aunque no dice nada sobre cuántas definiciones incorrectas también fueron extraídas. Si se tienen pues, una cobertura del 100% y una precisión del 100% significa que se extrajeron todas las definiciones correctas y sólo las definiciones correctas. Ahora bien, para poder determinar tanto el número como la fidelidad de las definiciones extraídas fue necesario extraer manualmente todas las definiciones que el marcador de definición *entenderse* presentaba en el texto que conformaba nuestro corpus. A continuación se muestra un cuadro en el que se compara el número de definiciones extraídas por el SEI con el número de definiciones presentes en el corpus.

Número de definiciones organizado por orden alfabético																											
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	Ñ	O	P	Q	R	S	T	U	V	W	X	Y	Z	Ot
SEI	2	2	2	1	6	1	0	1	2	1	0	1	2	0	0	2	1	0	1	4	6	0	0	0	0	0	0	3
	total			39																								
LEY	2	2	2	1	6	1	0	1	1	1	0	1	2	0	0	2	1	0	1	4	6	0	0	0	0	0	0	3
				38																								

Cuadro No. 2 Número de definiciones organizado por orden alfabético

A partir del cuadro anterior podemos hacer el cálculo para medir la cobertura de nuestro SEI utilizando las siguientes formulas

Precisión = DCE/DE (38/39)*100 = 97.44%

Cobertura= DCE/DC (38/38)*100 = 100%

donde

DE = Definiciones Extraídas = 39
DC = Definiciones Correctas = 38
DCE = Definiciones Correctas Extraídas = 38

Así pues, nuestro SEI presenta una cobertura del 100%; esto es, se extrajeron todas las definiciones que el corpus presentaba para el marcador *entenderse*. Por otra parte, nuestro SEI arrojó una precisión del 97%; esto es, se extrajo una definición más de las que se esperaban. Estos cálculos se realizaron sin tomar en cuenta el hecho de que en 16 de las 39 definiciones extraídas existían oraciones que complementaban las definiciones. Si se tomara ese hecho como un criterio para esgrimir que las definiciones no eran correctas entonces la cobertura sería de un 61% y la precisión de 59%, cifras que aun siguen siendo alentadoras. Sin

embargo, la gran diferencia entre los pares de cifras nos alertan sobre la importancia de tomar en cuenta las co-referencias en el diseño de los SEI.

Referencias

Arntz, R. y Pitch H. (1995) Introducción a la terminología. [Traducción castellana: A. de Irazabal el al.] Madrid: Fundación Germán Sáchez Ruipérez (Obra original publicada en 1989).

[De Miguel, 2000]. De Miguel, E. El texto Jurídico-Administrativo: Analisis de una Orden Ministerial. Circulo de Lingüística Aplicada a la Comunicación, CLAC, No. 4 noviembre de 2000

Gonzalo Arroyo, J. (2003). Recuperación y extracción de información. En Martí, Ma. A. (Coord.) Las tecnologías del lenguaje. (pp. 157-192). Barcelona(España): Editorial UOC.

Pérez Hernandes, M. (2002). *Explotación de los corpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento*. Estudios de Lingüística Española. Vol. 18. Fecha de acceso julio de 2005. Disponible en: http://elies.rediris.es/elies18/index.html

Corpus

Ley organica del trabajo (Venezuela). Fecha de acceso julio de 2005. Disponible en : http://www.analitica.com/bitblioteca/congreso_venezuela/ley_del_trabajo.asp