Extracción de Datos Enlazados desde Información No Estructurada Aplicando Técnicas de PLN y Ontologías

Aramís Rodríguez-Blanco[1], Alfredo Simón-Cuevas[1], Wenny Hojas-Mazo[1], José M. Perea-Ortega[2]

[1]Universidad Tecnológica de La Habana José Antonio Echeverría, Cujae, La Habana, Cuba
[2]Universidad de Extremadura, Badajoz, España
{aridriguezb, asimon, whojas}@ceis.cujae.edu.cu
jmperea@unex.es

Resumen En este trabajo se presenta un método para la extracción automática de datos enlazados a partir de información textual no estructurada en idioma español e inglés. El método está basado en la extracción de una conceptualización del texto en forma de mapa conceptual, la cual es posteriormente transformada en un modelo de dados RDF. En la extracción de la conceptualización se aplican diversas técnicas de PNL y se brinda la posibilidad de utilizar una ontología, para incrementar las capacidades de extracción de información del texto. Se realizaron pruebas usando tres colecciones de textos en español e inglés para evaluar la propuesta, con resultados prometedores en la extracción de conceptos y relaciones entre ellos, y mostrando también los beneficios del uso de ontologías como recurso de conocimiento externo.

Palabras claves: extracción de datos enlazados; PLN; mapas conceptuales; y ontologías

Abstract In this work a method for the automatic extraction of linked data from unstructured textual information in Spanish and English language is presented. The method is based on the extraction of a conceptualization from the text in form of concept map, which is transformed later in a RDF data model. In the conceptualization extraction several NLP techniques are applied and the possibility to use an ontology is offered, for increasing the capacities of information extraction from the text. Several tests using three collections of texts in Spanish and English were carried out for evaluating the proposal, with promising results in the concepts and relationships extraction and showing the benefits of the use of ontologies as external knowledge resource.

Keywords: linked data extraction; NLP; concept map; and ontologies

1. Introducción

El surgimiento de aplicaciones inteligentes, tales como la búsqueda semántica, y la popularidad de las tecnologías de la Web Semántica están requiriendo cada vez más que el contenido en la Web sea reorganizado a través de datos semánticos [1]. El paradigma de los Datos Enlazados (DE) (*linked data* en inglés) ha evolucionado como un poderoso facilitador en la transición de la actual Web (orientada a documentos) a una Web de datos interconectados [2]. En este sentido, los DE se han convertido en un activo de gran valor para la búsqueda, el análisis, la ingeniería del conocimiento, la integración y la recuperación de información. La extracción de información en fuentes no estructuradas, semi-estructuradas o estructuradas y su descripción en un modelo de datos RDF son tareas claves en el ciclo de vida de los DE [2]. Los DE son generalmente construidos a partir del minado de fuentes de información semi-estructuradas, como Wikipedia donde *DBpedia* [3] es el referente principal, estructuradas como las bases de datos [4], y en menor medida a partir de contenido textual no estructurado. La extracción de DE a partir de fuentes textuales se ha abordado en [5][6][7], pero estas soluciones son aplicables a contenido en inglés y presentan algunas limitaciones en cuanto a la cantidad de información que puede ser extraída, así como en los mecanismos diseñados para ello. La ausencia de propuestas para extraer DE desde textos en español, no posibilita aprovechar el volumen de información actualmente disponible en ese idioma, siendo esta situación una de las motivaciones de este trabajo. Un ejemplo de esta necesidad y también motivación, lo constituye *DBpedia-LatAm*[1], como iniciativa que se une al esfuerzo colaborativo impulsado desde *DBpedia* para extraer DE desde artículos de Wikipedia[2], pero con énfasis en los documentos escritos en español.

En este trabajo se presenta un método para extraer DE a partir de contenido textual no estructurado escrito en español e inglés, basado en la extracción automática de una conceptualización del texto, representada en forma de Mapa Conceptual (MC) [8], la cual es posteriormente transformada en un modelo de datos RDF. Los MC son un tipo de grafo de conocimiento no formalizado semánticamente, donde el conocimiento se representa en lenguaje natural y se estructura a través de nodos que representan conceptos y relaciones entre ellos formando proposiciones. Los conceptos pueden representar entidades, eventos, objetos o regularidad percibida de ellos, y las proposiciones son declaraciones (unidades semánticas) constituidas por dos o más conceptos interconectados mediante una relación dirigida y etiquetada por una frase-enlace que define el tipo de relación [8]. Estos elementos indican la existencia de similitudes estructurales entre el MC y el grafo subyacente al modelo de datos RDF, ya que este último se basa en tripletas (*sujeto*, *predicado*, *objeto*) y los MC en proposiciones (*concepto*, *frase-enlace*, *concepto*), lo que sirve de fundamento para diseñar una propuesta de extracción de DE bajo la perspectiva de los MC. El método consta de tres fases: pre-procesamiento, extracción de la conceptualización y genera-

[1] http://es-la.dbpedia.org/home/
[2] http://es.wikipedia.org/wiki/Wikipedia

ción del modelo RDF. En las dos primeras se aplican un conjunto de técnicas de Procesamiento de Lenguaje Natural (PLN), tales como: análisis sintáctico superficial y de dependencias, reconocimiento de entidades, patrones lingísticos, entre otras, algunas de ellas incluidas en propuestas similares [9][10][11][12][13], pero no integradas en una misma solución. También brinda la posibilidad de procesar textos en idioma inglés y español, a diferencia de las propuestas reportadas en [9][10][11][12][13], dirigidas solo a textos en inglés, y de usar una ontología OWL como recurso de conocimiento externo para apoyar la extracción de conceptos y relaciones del texto, siendo esta otra de las contribuciones de la propuesta. Este método no solo posibilita extraer datos que representen entidades, como principalmente se realiza en [5][6][7], sino que también permite extraer otros conceptos que pueden ser útiles para ampliar la información asociada a las entidades y facilitar la identificación de vínculos entre ellas. Por otra parte, también incluye mecanismos para identificar relaciones semánticas, ya sean taxonómicas y no taxonómicas, significando una ventaja respecto a trabajos similares [5][6][7][9][10][11][12][13]. El método fue evaluado mediante pruebas realizadas con tres colecciones formadas por resúmenes en español e inglés, y textos más extensos en inglés sobre temas de Inteligencia Artificial (IA). Los resultados obtenidos se muestran alentadores, alcanzándose en la mayoría de las colecciones valores de precisión superiores al 90 % en la extracción de conceptos y al 50 % en la extracción de proposiciones. En las pruebas realizadas también se constataron los beneficios del uso de una ontología en cuanto al aumento de la cantidad de información extraída y a la calidad de su obtención.

El resto del trabajo se organiza según se describe a continuación. En la Sección 2 se describen y analizan trabajos relacionados con la propuesta. En la Sección 3 se describe la solución propuesta y las fases que la componen. En la Sección 4 se presentan y analizan los resultados de las pruebas realizadas. Las conclusiones se exponen en la Sección 5.

2. Trabajos Relacionados

La extracción automática de DE a partir de contenido textual no estructurados ha sido un tema abordado con anterioridad en [5][6][7], pero estas soluciones son aplicables mayoritariamente a textos en idioma inglés, no identificándose propuestas para textos en español. En [6] se reporta una solución para extraer DE a partir de textos del dominio de la ciberseguridad, en la cual la extracción de términos y conceptos se realiza aplicando un reconocedor de entidades adaptado para reconocer clases de entidades asociadas a ese dominio. Los vínculos entre las entidades se identifican mediante relaciones existentes entre las clases representadas en una ontología, estableciendo un proceso de mapeado entre dichas clases y las entidades identificadas en el texto, y no a partir del análisis sintáctico del contenido. En este sentido, las relaciones posibles a extraer están limitadas a los tipos de relaciones representadas en la ontología, las cuales se enmarcan en el ámbito de la ciberseguridad. En [5] se presenta una arquitectura para extraer DE con independencia del dominio, en la que se combina un reconocedor de enti-

dades nombradas, con una estructura de representación del discurso que modela el significado de los textos para extraer los datos y sus vínculos. En esta arquitectura se representan otras entidades relevantes del texto, identificadas a partir de frases sustantivas o eventos, y sus relaciones, así como información sintáctica y gramatical. Las relaciones binarias entre esas entidades se establecen a través de proposiciones y roles verbales, siendo este último aspecto su limitación fundamental. En [7] se realiza un procesamiento del texto aplicando diferentes técnicas de PLN que concluye con un árbol de dependencias de las sentencias del texto, sobre el cual se realiza la extracción de entidades y relaciones entre ellas. En esta propuesta se utiliza una Base de Datos de Entidades, construida manualmente, donde se almacenan las posibles entidades a extraer. Las relaciones son identificadas a partir de consultas almacenadas en una Base de Datos de Consultas, igualmente construida de forma manual y analizando los vínculos entre las entidades identificadas en el árbol de dependencia. Estos elementos sugieren que con esta propuesta se pueden alcanzar altos índices de precisión en la extracción de la información, pero una baja cobertura del texto, ya que la información a extraer está limitada a lo que se pueda identificar de lo almacenado en las bases de datos que son utilizadas.

La construcción de MC a partir de textos ha sido tratado en [9][10][11][12][13], aunque igualmente todas estas propuestas están concebidas para textos en inglés. Entre las técnicas de PLN empleadas en el pre-procesamiento del texto se encuentran: etiquetado POS [10][12], el análisis sintáctico superficial [11] y de dependencia [9][13], el reconocimiento de entidades [10] y la desambiguación del sentido de las palabras [12][13]. Los conceptos han sido identificados fundamentalmente a partir de frases sustantivas [9][10][11][12][13], en algunos casos empleando patrones léxico-sintácticos predefinidos [9][10], y también a partir de entidades reconocidas [10]. En la mayor parte de los trabajos las relaciones entre conceptos (proposiciones) se identifican a partir de relaciones verbales [9][11][12][13], en muy pocas propuestas se aprovechan los beneficios que ofrece el análisis de dependencias [9][13] para la identificación de relaciones entre conceptos, y solo son identificadas relaciones taxonómicas en [10].

A partir de este análisis, se decidió diseñar una nueva propuesta para extraer DE aplicable principalmente a textos en español, aspecto no considerado en [5][6][7], y también en inglés, independiente del dominio, a diferencia de [6], y que posibilite extraer una mayor cantidad de conceptos (datos) y relaciones, respecto a trabajos similares [5][6][7]. Asociado a este último objetivo, se proponen algunas contribuciones: (1) el uso de patrones lingísticos, en combinación con una ontología de referencia, para extraer conceptos; y (2) la identificación de relaciones semánticas (taxonómicas y no taxonómicas) entre los conceptos, usando para ello también patrones lingísticos y la ontología; (3) uso intensivo del árbol de dependencia para identificar relaciones entre los conceptos, principalmente relaciones no taxonómicas.

3. Método de Extracción de Datos Enlazados

El método propuesto fue diseñado en varias fases, como se muestra en la Fig. 1. En la obtención de la conceptualización se combinan varias técnicas de PLN, y el uso de una ontología OWL de referencia, para facilitar y ampliar la cobertura del texto en la identificación de conceptos y relaciones entre ellos. Una de las ventajas del diseño del método propuesto es su flexibilidad, ya que permite utilizar cualquier ontología, aunque es recomendable que el conocimiento que represente sea del mismo dominio (o cercano) al del contenido de los textos, para lograr un mayor aprovechamiento de ese recurso de conocimiento.

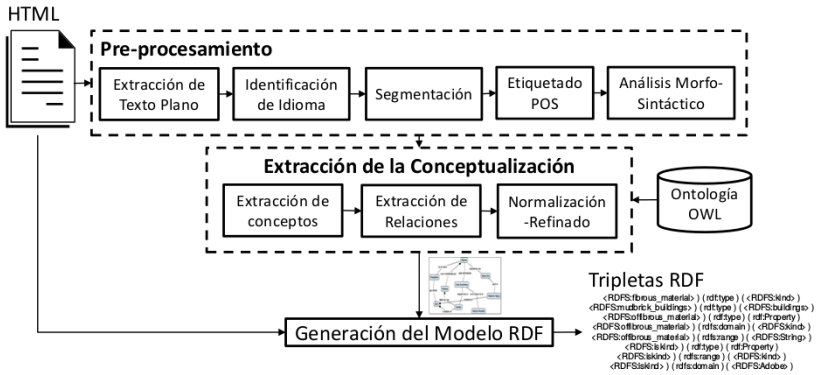

Figura 1. Fases y componentes del método propuesto para la extracción de DE.

3.1. Pre-procesamiento

En esta fase, el contenido del fichero de entrada es segmentado y caracterizado sintácticamente, partiendo de la extracción del texto plano usando una biblioteca de clases de Java desarrolladas para este propósito, que brinda soporte para ficheros html, htm, pdf, docx, doc, rtf y txt. Luego, se identifica si el texto ha sido escrito en español o inglés aplicando la solución reportada en [14]. La identificación del idioma constituye una tarea importante en este método, pues a partir del resultado de esta tarea es que se decide que patrones utilizar para la extracción de conceptos y relaciones, dado que los mismos están definidos para textos en español o inglés. La segmentación se realiza en oraciones (unidad de análisis), considerando principalmente el punto final como elemento que delimita los segmentos. En este proceso se ignoran los puntos presentes en números, en siglas y abreviaturas, y los puntos suspensivos. Sobre cada una de las oraciones, se identifican los *tokens* (ej. palabras, números, signos de puntuación, etc.), y se ejecuta el análisis morfosintáctico, el cual incluye el análisis sintáctico superficial y de dependencias. Este análisis del texto se realiza usando el analizador

sintáctico *Freeling* [15], teniendo en cuenta que es uno de los pocos analizadores que brinda soporte para textos en español, además de inglés. Como resultado del análisis morfológico cada uno de los *tokens* se etiqueta con su raíz morfológica, categoría gramatical y otros datos.

El análisis sintáctico superficial está dirigido principalmente a la identificación de conceptos, y consiste en agrupar los *tokens* de la oración en *chunks* o constituyentes que representan estructuras gramaticales categorizadas como: sintagmas nominales, adjetivales, preposicionales, grupos verbales, entre otros, a partir de los cuales se obtiene un árbol sintáctico. A través del análisis de dependencias se establecen las relaciones de dependencias existentes entre las estructuras gramaticales, las cuales son representadas en un árbol de dependencias. Los resultados de este análisis no solo son muy útiles para identificar vínculos entre los conceptos presentes en las diferentes estructuras gramaticales, sino también para identificar o construir las frases-enlace a utilizar en el etiquetado de esas relaciones.

3.2. Extracción de la Conceptualización

En esta fase, se construye automáticamente un MC del texto, como representación de la conceptualización, a partir de la identificación de conceptos y relaciones entre ellos (proposiciones). La extracción de conceptos se basa en la identificación de términos (constituidos por una o varias palabras) dentro de los sintagmas nominales o adjetivales, cuya composición se corresponda con alguno de los patrones lingísticos definidos para este propósito, así como en la identificación de conceptos representados en la ontología que estén presentes en el texto. La extracción de relaciones entre conceptos se basa también en el uso de patrones lingísticos, en la identificación de vínculos entre conceptos representados en la ontología, y en el análisis de las dependencias existentes entre las estructuras gramaticales del texto.

La extracción de conceptos se inicia con la identificación en el texto de términos representados como clases o instancias en la ontología, los cuales son tratados como entidades. En este proceso de mapeado se evalúa la similitud sintáctica existente entre los términos del texto y los conceptos de la ontología usando la métrica de Levenshtein. Seguidamente, se analiza el árbol sintáctico de cada oración para identificar otros términos que cumplan con alguno de los patrones lingísticos definidos, los cuales se extraen también como conceptos. Estos patrones han sido formalizados sobre la base del etiquetado gramatical que realiza *Freeling* y constituyen los patrones más frecuentes a partir de los cuales están formados los conceptos representados en la ontología del proyecto *DBpedia*. Mediante estos patrones no solo se identifican como conceptos las entidades nombradas (ej. NP) sino también sustantivos, adjetivos y frases que expresan combinaciones de ambos. En la Tabla 1 se muestra una selección de los patrones definidos para textos en español, los cuales son equivalentes a los identificados en análisis de los conceptos de la ontología, y que se utilizan para textos en inglés.

El método propuesto brinda la posibilidad de extraer relaciones explícitas e implícitas entre los conceptos, siendo las primeras las identificadas en la oración

Cuadro 1. Patrones lingísticos para identificar de conceptos en textos en español.

Patrones	Ejemplos	Patrones	Ejemplos
NC	procesos	Z NC	cinco archivos
NP	San Cristóbal	AQ NC	gran personalidad
NC AQ	sistema informático	NC AQ AQ	procesador lógico operativo
NC NP	director Juan	AQ	lejos

Leyenda: NC: sustantivo común; NP: sustantivo propio; Z: número o numeral; AQ: adjetivo

y las segundas las identificadas a través de la ontología, las cuales permiten relacionar conceptos presentes en diferentes partes del texto a partir de vínculos semánticos. En este proceso intervienen dos tareas: identificación del vínculo entre los conceptos y la obtención de la frase-enlace. Entre los tipos de relaciones a identificar se encuentran las relaciones semánticas del tipo taxonómicas y no taxonómicas (ej. verbales). Las primeras se identifican mediante los patrones lingísticos que se muestra en Tabla 2, definidos sobre la base de lo reportado en [16][17][18], y a partir de la ontología, la cual es consultada mediante SPARQL y utilizando Jena.

Cuadro 2. Patrones para identificar relaciones taxonómicas en textos en español.

Patrones	Ejemplos	Patrones	Ejemplos
SN_0 'tales como' $\{SN_1, SN_2..., (y \mid o)\}$ SN_n	...animales tales como perro y ave.	SN {,} 'incluyendo' {SN,}+{$y \mid o$} NP	...países de leyes comunes, incluyendo Canadá y Reino Unido
'tales' SN 'como' {NP,}+{$(o \mid y)$} NP	...tales autores como Hearst, y Cimiano.	SN{,} 'especialmente' {NP,}+{$y \mid o$} NP	...los países europeos, especialmente Francia, Reino Unido, y España
SN {, SN} + {,} 'u otros' SN	Quemaduras, heridas, u otros daños...	SN_1 'es un tipo de' SN_0	el mapa conceptual es un tipo de grafo de conocimiento
SN {, sn}+{,} 'y otros' SN	...templos, viviendas y otras edificaciones		

Leyenda: SN: sintagma nominal; NP: sustantivo propio; +: concatenación de elementos; |: disyunción de elementos; : elementos opcionales; ( ): grupo de elementos

A través de la ontología, se identificaría una relación taxonómica entre dos conceptos C_0 y C_1 si ambos constituyen clases en la ontología y están vinculados mediante una relación de sub-clase, siendo especificada la frase-enlace como 'subClaseDe' (o '*subClassOf*' en inglés). Se identifican otras relaciones semánticas si: C_0 y C_1 son clases en la ontología que están vinculadas a través de una relación de Propiedad de Objeto, siendo especificada la frase-enlace por el identificador

de ese tipo de relación en la ontología; o C_0 constituye una clase y C_1 una de sus instancias, o viceversa, siendo especificada la frase-enlace como 'instanciaDe' (o '*instanceOf*' en inglés).

En un segundo paso, se identifican otros tipos de relaciones a partir del análisis el árbol de dependencia generado de cada oración. Este proceso se basa en el análisis de las estructuras asociadas a diferentes tipos de nodos representados en ese árbol, tales como: preposicionales, conceptuales, verbales, subordinantes y de coordinación, a partir de los cuales se establecen los vínculos de dependencia entre las estructuras gramaticales de la oración. Por ejemplo, las relaciones verbales se pueden extraer identificando conexiones existentes entre un concepto identificado en el sujeto de la oración (sintagma nominal) y otros identificados en los complementos que pertenecen a su ámbito de dependencia, a través de un nodo verbal. En este caso, la proposición se construye vinculando el concepto incluido en el sujeto con cada uno de los conceptos identificados en los complementos y la frase-enlace se construye a partir del nodo verbal y sus modificadores. Las frases-enlace que etiquetan las relaciones identificadas en este análisis son identificadas (o construidas) a partir de los siguientes patrones lingísticos: VM, VM+SP, VM+CS+VM, VM+CS, PR+VM, CS+VM+SP, VM+VM, y AQ+SP, siendo CS: conjunción subordinada; SP: preposición; AQ: adjetivo; VM: verbo principal; PR: pronombre relativo.

El proceso de normalización-refinado se realiza luego de haber extraídos los conceptos y proposiciones del texto, y tiene como objetivo reducir redundancias o incoherencia. Se lleva a cabo a partir de la eliminación de las proposiciones repetidas (enlazan conceptos iguales), así como la unificación de conceptos y frases-enlace sintácticamente similares, usando para ello también la métrica de Levenshtein.

3.3. Generación del Modelo RDF

En este proceso, la conceptualización extraída automáticamente del texto es transformada en un modelo de datos RDF, donde cada una de las proposiciones se codifica como una tripleta RDF. En esta transformación se sigue como convención que: (1) el concepto origen de la proposición se codifica como el *sujeto*; (2) el concepto destino se codifica como el *objeto*; y (3) la frase-enlace se codifica como el *predicado*. Las URIs que identifican los elementos de las tripletas RDF se construyen a partir de las direcciones URL obtenidas del fichero HTML procesado, específicamente, la asociada al propio fichero y las correspondientes a los hipervínculos asociados a términos representados como conceptos, y las etiquetas de los conceptos. Si un concepto no posee un hipervínculo, la dirección URL de referencia sería la del propio fichero origen, y en otro caso sería la del hipervínculo.

4. Resultados y Discusión

La obtención de forma automática de la conceptualización del texto constituye el elemento que mayor influencia tiene en la calidad de los resultados a

obtener por el método propuesto, por lo que las pruebas realizadas se centraron en evaluar este aspecto. Hasta el momento, no se han identificado marcos de evaluación de referencia, ni corpus de textos reconocidos para probar este tipo de soluciones, por lo que su evaluación se hace bastante compleja. No obstante, se decidió realizar las pruebas sobre textos en español e inglés usando tres colecciones: *DBpedia_ES*, *DBpedia_EN* e IA. Las dos primeras fueron construidas con resúmenes en español e inglés, respectivamente, tomados de un *dataset* de resúmenes cortos disponible en *DBpedia*[3], y la segunda con textos en inglés sobre temas de IA[4]. En la construcción de las colecciones se tuvo en cuenta que los textos debían estar incluidos en corpus disponibles en Internet, tuvieran diferentes tamaños, y que existiera alguna ontología que tuviera algún vínculo con el contenido de los mismos. La caracterización de las colecciones utilizadas se muestra en la Tabla 3.

Cuadro 3. Caracterización de las colecciones de prueba.

Características	**DBpedia_ES**	**DBpedia_EN**	**IA**
Idioma	Español	Inglés	Inglés
Documentos	50	50	6
Prom. de palabras	75.9	75.2	1111.83
Prom. de oraciones	3.0	3.4	46.83

Inicialmente se ejecutó el método sobre cada una de las colecciones sin usar la ontología y luego usando la ontología. Esta última ejecución se realizó solo sobre las colecciones *DBpedia_EN* e IA, ya que eran de las se disponía de una ontología de referencia. En las pruebas realizadas con la colección *DBpedia_EN* se utilizó la ontología de *DBpedia* y en las realizadas con la colección IA se utilizó una ontología del dominio de IA, tomada de la misma fuente de los textos seleccionados. En los experimentos se evaluaron aspectos tales como: cantidad de información extraída (conceptos y relaciones), nivel de contribución de la ontología, y precisión en la extracción de conceptos y proposiciones. Considerando que no se tenía información sobre los resultados correctos a obtener para cada texto, se decidió evaluar la calidad de la extracción de conceptos y proposiciones a través de evaluadores humanos (3 profesores de la carrera de Ingeniería Informática). Los evaluadores debían clasificar cada uno de los conceptos y proporciones extraídas en correctos o incorrectos, sobre la base del contenido de los textos y considerando lo siguiente:

- concepto correcto: sustantivos o adjetivos que tuvieran un significado importante en el texto, nombres de entidades, o frases que tuvieran sentido;
- proposición correcta: si puede ser interpretada con un sentido propio (ej. cuando ambos conceptos son correctos y la frase-enlace está bien definida).

[3] http://wiki.dbpedia.org/Datasets

[4] Corpus de textos y ontología asociada disponibles en http://azouaq.athabascua.ca/goldstandards.htm

A partir de la información emitida por los evaluadores se calculó la precisión como la razón entre los elementos extraídos (conceptos o proposiciones) correctamente y el total, cuyos valores se muestran en la Tabla 4. La precisión se calculó a partir de los resultados coincidentes de los evaluadores, con un índice de acuerdo de 92 %.

Según se aprecia, la mayor precisión se obtuvo en la extracción de conceptos, sin una diferencia significativa respecto al idioma cuando no se usaron ontologías, demostrándose la calidad y utilidad de los patrones lingísticos definidos para este propósito. Sin embargo, se observa una menor precisión en la extracción de proposiciones. Este resultado está dado en buena medida por la alta complejidad que tiene la identificación de forma automática de relaciones entre conceptos en un texto, así como por algunos resultados no favorables del análisis sintáctico realizado con *Freeling*. Esto último también incide en la identificación de conceptos, pero repercute más en la extracción de las proposiciones porque en su evaluación no solo se mide si los conceptos vinculados son correctos o no, sino que también que la frase-enlace esté bien formada. La identificación del vínculo entre los conceptos, así como la obtención de la frase-enlace dependen en gran medida del análisis que se realiza sobre el árbol de dependencias, y por tanto las deficiencias que posea repercuten en los resultados.

Cuadro 4. Resultados experimentales en la extracción de la conceptualización.

		DBpedia_EN		**IA**	
Aspectos	**DBpedia_ES**	**NoOnt**	**SiOnt**	**NoOnt**	**SiOnt**
CC	7,88	10,06	10,2	112,4	120,2
CE	2,2	4,86	5,08	29,2	36,2
CR	6,10	8,38	8,48	103	114,8
CCOnt.	-	-	3,72	-	37,2
CROnt.	-	-	0	-	8,6
PC	93,66	89,21	92,58	96,81	98,57
PR	53,92	51,26	53,97	66,67	83,10

Leyenda: CC: Prom. cant. conceptos; CE: Prom. cant. de entidades; CR: Prom. cant. relaciones; CCOnt.: Prom. cant. conceptos obtenidos de la ontología; CROnt.: Prom. cant. relaciones obtenidasde la ontología; PC: Precisión en la identificación de conceptos; PR: Precisión en la identificación de relaciones; NoOnt: sin usar ontología; SiOnt: usando ontología.

Un resultado importante es el incremento de la precisión (aunque discreto) cuando se utiliza la ontología, así como el incremento de la cantidad de información extraída, apreciado en mayor medida en la colección IA. Esto demuestra los beneficios que se pueden obtener cuando es usada una ontología de referencia en la extracción automática de conceptos y relaciones desde textos no estructurados.

5. Conclusiones

En el trabajo se ha presentado un nuevo método para extracción de los DE a partir de textos no estructurados, aplicable a textos escritos en el idioma español e inglés y no dependiente del dominio. En esta nueva propuesta se combinan un conjunto de técnicas de PLN, tanto para el pre-procesamiento de los textos como para la identificación de conceptos y relaciones, que permiten aumentar las capacidades para extraer una mayor cantidad de información desde los textos. El uso de patrones lingísticos para identificar conceptos y relaciones semánticas entre ellos, la posibilidad de utilizar de manera flexible una ontología de referencia, así como el análisis de las dependencias entre los conceptos en el texto, constituyen elementos que han propiciado alcanzar una mayor cobertura del texto a la hora de extraer la información. La extracción de conceptos, que no necesariamente representen entidades, así como la identificación de vínculos semánticos entre ellos, propicia que se obtenga una mayor riqueza en los DE que se obtienen con el método propuesto. Las pruebas realizadas mostraron resultados prometedores, alcanzándose valores de precisión en la mayoría de los casos superiores al 90 % en la extracción de conceptos y al 50 % en la extracción de relaciones. También se constataron los beneficios de usar una ontología, referentes al incremento de la cantidad de información extraída de los textos, y al aumento de la precisión en la extracción de conceptos y relaciones.

Agradecimientos. Este trabajo ha sido parcialmente financiado por el Ministerio de Economía y Competitividad del Gobierno de España, en el marco del proyecto REDES (TIN2015-65136-C2-1-R).

Referencias

1. Wang, P., y Zhang, X.: Finding, Extracting, and Building Academic Linked Data. In Li, J. et al. (Eds): Semantic Web and Web Science, Springer & Business Media, 25–39 (2013)
2. Auer, S., Lehmann, J., Ngonga A. C., y Zaveri, A.: Introduction to Linked Data and its Lifecycle on the Web. Proc. of Reasoning Web 2013, LNCS, Vol. 8067, 1–90 (2013)
3. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., y Bizer, C.: DBpedia A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web, 1, 1–27 (2012)
4. Dimitrios-Emmanuel, S., Stavrou, P., y Mitrou, N.: Bringing relational databases into the semantic web: A survey. Semantic Web, 3(2), 169–209 (2012)
5. Augenstein, I., Padó, S., y Rudolph, S.: LODifier: Generating Linked Data from Unstructured Text. In Proc. of ESWC 2012, LNSC, Vol. 7295, 210–224 (2012)
6. Joshi, A., Lal, R., Finin, T., y Joshi, A.: Extracting Cybersecurity Related Linked Data from Text, IEEE 7th Int. Conf. on Semantic Computing (ICSC), 252–259 (2013)

7. Krí, V., y Hladká, B.: RExtractor: a Robust Information Extractor. Proc. of the 2015 Conf. of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 21–25 (2015)
8. Novak, J. D., y Cañas,A. J.: The Theory Underlying Concept Maps and How to Construct Them, Technical Report IHMC CmapTools, USA (2006)
9. Kowata, J. H., Cury, D., y Silva, M. C.: Concept maps core elements candidates recognition from text. In Proc. of CMC 2010, 120–127 (2010)
10. Raymond, M., Song, D., Yuefeng, L. I., Terence, Ch., I., y Hao, J.-X.: Towards A Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning. Knowledge and Data Engineering, 21(6), 800–813 (2009)
11. Valerio, A., Leake, D.,y Cañas, A. J.: Using Automatically Generated Concept Maps for Document Understanding: A Human Subject Experiment. In Proc. of CMC 2012, 438–445 (2012)
12. Wang, W. M., Cheung, C. F.,Lee, W. B. y Kwok, S. K.: Mining knowledge from natural language texts using fuzzy associated concept mapping. Information Processing and Management, 44(5), 1707–1719 (2008)
13. Attia S. S., Arafa, W. M. y Eldin,A. S.: A Framework of Proposed Intelligent Tool for Constructing Concept Map, In Proc. of ICL 2010, 524–533 (2010)
14. Amine, A., Elberrichi, Z., y Simonet, M.: Automatic language identification: An alternative unsupervised approach using a new hybrid algorithm. Int. J. on Computational Science & Applications, 7(1), 94–107 (2010)
15. Padró, L. y Stanilovsky, E.: FreeLing 3.0: Towards Wider Multilinguality. International Conference on Language Resources and Evaluation, In Proc. of LREC 2012, 2473–2479 (2012)
16. Hearst, M. Automatic acquisition of hyponyms from large text corpora. In Proc. of the 14th Int. Conf. on Computational Linguistics, 539–545 (1992)
17. Cimiano P. y Vlker, J.: Text2Onto, In Proc. of the 10th Int. Conf. on Applications of Natural Language to Information Systems, 227–238 (2005)
18. Jiang, X., y Tan, A. H.: CRCTOL: A semanticbased domain ontology learning system. J. of the American Society for Information Science and Technology, 61(1), 150–168 (2010)