Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 133-140 recibido 11-10-2020 revisado 24-01-2021 aceptado 01-02-2021

Mejoras aplicadas a la extracción de relaciones semánticas para la Web en español

Improvements applied to Open Information Extraction in Spanish

Juan M. Rodríguez[1], Hernán D. Merlino[2], Patricia Pesado[1]
[1] Facultad de Informática. Universidad Nacional de La Plata
[2]Departamento de Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús
jmrodriguez1982@gmail.com, hmerlino@gmail.com, ppesado@lidi.info.unlp.edu.ar

Resumen: En este trabajo de investigación se presenta un método novedoso de extracción de relaciones semánticas para la Web en español llamado ECMes. Este método es comparado con otros métodos similares en el *estado-del-arte* utilizando para ello dos conjuntos de prueba diferentes: uno conocido y ya utilizado en trabajos anteriores relacionados y otro construido específicamente para este artículo utilizando fuentes heterogéneas de datos. En el primero de los conjuntos de datos ECMes obtuvo la mejor precisión, la mejor exhaustividad (*recall*) y por lo tanto la medida F1 más alta en relación a los otros métodos evaluados. En el segundo de los conjuntos de datos obtuvo la mejor exhaustividad (*recall*) y la medida F1 más alta.
Palabras clave: Extracción de conocimiento, extracción de relaciones semánticas, procesamiento de lenguaje natural.

Abstract: This paper introduces a novel method of Open Information Extraction in Spanish called ECMes. This method is compared to other similar methods in the *state-of-the-art* using two different testing datasets: one is a well known dataset, already used in previous related works and the second dataset was constructed specifically for this article using heterogeneous data sources. In the first dataset ECMes obtained the highest precision, recall and F1 measurement. In the second dataset, ECMes obtained the highest recall and the highest F1 measurement.
Keywords: Knowledge extraction, semantic relationship extraction, natural language processing, Open Information Extraction.

1 Introducción

En este documento presentamos un método novedoso de extracción de relaciones semánticas para la Web en idioma español llamado ECMes (Extractor de Conocimiento Mejorado en Español). ECMes está construido sobre la arquitectura base de TP-OIE-ES (Rodríguez y Merlino, 2020) pero implementa una serie de mejoras sobre el algoritmo original que le permiten incrementar considerablemente su precisión, elevar su exhaustividad (*recall*) y por lo tanto mejorar la medida de rendimiento F (*F-measure*).

En las siguientes secciones se detallarán cuáles fueron las mejoras introducidas en el algoritmo, cómo se construyó el conjunto de datos de prueba y cuáles fueron los criterios utilizados para evaluar los diferentes métodos.

La extracción de relaciones semánticas para la Web, u otros grandes corpus de datos, (en inglés *Open Information Extraction*) es un paradigma de extracción de información presentado por primera vez en (Banko et al., 2007). En dicho trabajo se presenta al método TEXTRUNNER, un sistema informático capaz de devolver una tupla de la forma: *(argumento 1, relación, argumento 2)* por cada relación semántica existente en una oración dada como entrada. Típicamente información fáctica del tipo: "Quién hizo qué a quién y cómo" (Rodríguez et al., 2018).

ISSN 1989-7553. DOI 10.26342/2021-66-11

Un método de OIE (*Open Information Extraction)* debe cumplir las siguientes condiciones según Banko et al. (2007).

- Hacer una sola pasada por el corpus garantizando la escalabilidad, independientemente del tamaño de este.
- Ser independiente del dominio.
- Tener un corpus como única entrada (*input*). Su salida (*output*) debe ser un conjunto de relaciones extraídas.
- Ser no supervisado.
- Extraer cualquier relación existente y no depender de relaciones previamente establecidas por el desarrollador.

2 Trabajos relacionados

Investigaciones documentales (Glauber y Barreiro, 2018; Rodríguez, Merlino y García-Martínez, 2015) muestran que fueron desarrollados diversos métodos de OIE, aunque no todos están disponibles de forma pública. De todos los métodos presentados solo algunos fueron comparados de forma experimental con métodos ya existentes para determinar su desempeño relativo. La mayoría de ellos fueron creados para el idioma inglés exclusivamente. Los más destacados en cantidad y calidad de extracciones son:

- ClausIE (Del Corro y Gemulla, 2103).
- ReVerb (Fader, Soderland y Etzioni, 2011).
- OLLIE (Schmitz et al., 2012).
- MinIE (Gashteovski, Gemulla y Del Corro, 2017).
- ArgOE (Gamallo y Garcia, 2015).
- Stanford OpenIE (Angeli, Premkumar y Manning, 2015).
- DepOE (Gamallo y Garcia, 2012).
- ExtrHech (Zhila y Gelbukh, 2013).

De la lista anterior solo trabajan con textos en idioma español ExtrHech, ArgOE y DepOE. ArgOE y DepOE fueron diseñados para soportar múltiples lenguajes (Glauber y Barreiro, 2018). ExtrHech soporta español e inglés (Zhila y Gelbukh, 2013).

Para los métodos en español: DepOE (2012), ExtrHech (2014) y ArgOE (2015), se muestra la precisión en la Tabla 1. Corresponde a las pruebas realizadas por los autores de cada método, desafortunadamente no se cuenta con el valor de exhaustividad.

Medidas	**ExtrHech**	**ArgOE**	**DepOE**
Precisión	0.55	0.55	**0.68**

Tabla 1: Resultados obtenidos de los trabajos: (Zhila y Gelbukh, 2013; Gamallo y Garcia, 2015; Gamallo y Garcia, 2012).

En (Rodríguez y Merlino, 2020) se presenta un nuevo método de *OIE* en idioma español llamado TP-OIE-ES. En el mismo trabajo se realizan pruebas para evaluar la precisión, la exhaustividad (*recall*) y la medida F1 de dicho método y de los métodos ArgOE y DepOE. Las pruebas se realizaron con un conjunto de datos compuesto por 69 oraciones extraídas de Wikipedia y propuestas por Gamallo y García (2015). En la Tabla 2 se muestran los resultados obtenidos:

Medidas	**TP-OIE-ES**	**DepOE**	**ArgOE**
Precisión	0.62	**0.89**	0.67
Exhaustividad	**0.36**	0.29	0.29
Medida F1	**0.46**	0.44	0.40

Tabla 2: Resultados obtenidos en (Rodríguez y Merlino, 2020) con el mismo conjunto de prueba utilizado en Gamallo y García (2015).

3 Problemas abiertos

Aunque TP-OIE-ES logra mejorar la exhaustividad y con ello una medida F1 más alta respecto a los otros métodos evaluados, en la prueba realizada por Rodríguez y Merlino en (2020) quedan al menos tres problemas importantes por ser resueltos. Los mismos se explican en las siguientes subsecciones.

3.1 Mejorar precisión

Si bien la medida F1, que evalúa conjuntamente la precisión y la exhaustividad con igual ponderación es más alta para TP-OIE-ES que para los otros métodos, la diferencia es de apenas 0.02 puntos con respecto a DepOE. Y la precisión de TP-OIE-ES es la más baja de los tres métodos. La precisión se convierte en el punto más débil de este método y es un problema que debe ser resuelto si se pretende mejorar al algoritmo.

3.2 Mejorar la evidencia disponible

Para soportar con mayor evidencia el desempeño de los métodos y sus medidas de rendimiento es necesario realizar por lo menos una segunda evaluación contra un conjunto de datos de prueba que no sea tan uniforme como el presentado por Gamallo y García en (2015), el cual está compuesto solo por sentencias extraídas de Wikipedia.

Además uno de los principios de los métodos de OIE es que estos son independientes del dominio. Es decir que deberían funcionar de forma similar con cualquier conjunto de textos de entrada para un mismo idioma (Banko et al., 2007).

3.3 Mejorar la informatividad de las extracciones realizadas

Muchas de las extracciones semánticas realizadas por el método TP-OIE-ES son correctas pero poco informativas, por ejemplo en la oración siguiente:

(1) *La bibliografía se estructura con los datos de las fichas bibliográficas de esos textos.*

Extrajo la siguiente tupla:

(2) *(La bibliografía; se estructura; con los datos de las fichas).*

La relación semántica es correcta ya que corresponde a lo expresado en la oración, sin embargo sería más informativa en este caso particular, si el argumento segundo contuviese al resto de las palabras. Por ejemplo:

(3) *(La bibliografía; se estructura; con los datos de las fichas bibliográficas de esos textos).*

La cantidad de información que debe contener una tupla para expresar correctamente la relación semántica expresada en una oración es una cuestión subjetiva. Dependerá en parte del problema que se esté intentando resolver y del procesamiento posterior que reciban las tuplas extraídas. Considérese la oración del ejemplo 4 a continuación:

(4) *Albert Einstein fue galardonado con el Premio Nobel en Suecia en 1921.*

Y las siguientes tuplas posibles:

1. *(Albert Einstein; fue galardonado con; el Premio Nobel).*
2. *(Albert Einstein; fue galardonado con; el Premio Nobel en Suecia en 1921).*
3. *(Albert Einstein; fue galardonado en; 1921).*
4. *(Albert Einstein; fue galardonado en; Suecia).*

Todas las tuplas expresan relaciones semánticas correctas. Sin embargo la tupla 2 es la más informativa, aunque no necesariamente es preferible por sobre las demás. Si el problema a resolver fuese, por ejemplo: "Ganadores del Premio Nobel por año" la tupla 3 sería más conveniente. No obstante, si un método devuelve una tupla como la 3, es esperable que genere al menos dos tuplas más como la 4 y la 1 para que toda la información existente en la oración original quede reflejada en las relaciones semánticas extraídas. Pero si las posibilidades del método consisten en que devuelva solo la 3 o solo la 2, la 2 es en este caso preferible para propósitos generales, ya que no hay pérdida de información.

Teniendo en cuenta lo anterior, en las pruebas realizadas por Rodríguez y Merlino en (2020) de las 49 extracciones semánticas marcadas como correctas, 19 fueron marcadas como "poco informativas". Es decir que casi un 39 % de las extracciones pueden ser mejoradas o están expresando la idea principal de la oración de forma pobre. Por lo cual mejorar la informatividad de las relaciones semánticas extraídas es un punto importante para mejorar el desempeño global del algoritmo.

4 ECMes

ECMes es una versión mejorada del método TP-OIE-ES. El algoritmo principal de TP-OIE-ES intenta identificar relaciones semánticas en una oración dada utilizando una lista de patrones. Estos patrones son generados automáticamente por el mismo método a partir de un conjunto de datos etiquetados, entiéndase en este caso una lista de oraciones con sus respectivas relaciones semánticas, expresadas estas en forma de tupla. Una vez entrenado, utiliza estos patrones para encontrar coincidencias en el árbol de dependencias sintácticas de la oración (enriquecido con información adicional: la categoría gramatical y los nombres de las entida-

des halladas). TP-OIE-ES buscará exhaustivamente todas las coincidencias existentes.

Estas coincidencias permiten identificar la relación propiamente dicha y el argumento primero (también llamado sujeto o primera entidad). El argumento segundo es obtenido mediante una serie de reglas para buscar la frase nominal más próxima, tal y como lo hace ReVerb (Fader, Soderland y Etzioni, 2011) en idioma inglés. Sin embargo, si encuentra una relación candidata y un argumento segundo candidato, pero no encuentra un argumento primero usando la lista de patrones, intentará encontrar este argumento inicial buscando la frase nominal más próxima a la izquierda de la relación.

La versión original de TP-OIE-ES fue entrenada con oraciones etiquetadas en idioma inglés. Si bien el árbol de dependencias sintácticas que se puede construir para idioma inglés es válido también en español ya que el *parser* utilizado es *depparse* de la biblioteca *Stanford CoreNLP* (Chen y Manning, 2014) el cual es universal. Esto significa que las aristas que conectan las palabras son siempre las mismas independientemente del idioma (Buchholz y Marsi, 2006). Sin embargo los ejemplos provistos en idioma inglés podrían no ser representativos de las oraciones más comunes del idioma español. Es decir, los patrones generados son válidos pero no necesariamente útiles, porque quizás son patrones que permiten identificar relaciones semánticas en oraciones válidas pero poco frecuentes, como podrían ser oraciones en idioma español con una estructura sintáctica similar a la de una frase en idioma inglés.

Por otro lado las categorías gramaticales utilizadas en la versión original de TP-OIE-ES son conocidas como *Penn Treebank POS tags* (Ratnaparkhi, 1996) y son exclusivas del idioma inglés, mientras que para español tanto TP-OIE-ES como su versión mejorada ECMes utilizan las categorías gramaticales *Universal POS tags* (Petrov, 2011). Por ende para buscar las coincidencias de los patrones, las categorías gramaticales que aparecían en estos debían ser traducidas en TP-OIE-ES de un sistema al otro (Rodríguez y Merlino, 2020). Como no existe una correlación unívoca entre estos dos sistemas, esta traducción es susceptible a introducir algún error.

4.1 Mejorar precisión y exhaustividad

Para intentar mejorar las medidas de rendimiento se implementaron tres mejoras en el algoritmo original.

4.1.1 Regeneración de los patrones de búsqueda

La principal medida que se tomó para mejorar la precisión y exhaustividad del método original fue la de reentrenar al mismo con un nuevo conjunto de ejemplos, esta vez en idioma español. Esto implicó convertir los métodos nativos que trabajaban con las categorías gramaticales en idioma inglés en el formato de *Penn Treebank POS tags* al formato *Universal POS tags*. No solo para el sistema de búsqueda de coincidencias de patrones, sino también para el sistema de puntajes, el cual utiliza información de las oraciones como la categoría gramatical para asignarle un puntaje a una relación semántica extraída. Solo aquellas que logran cierto puntaje son devueltas.

La base de datos de entrenamiento se construyó con un total de 110 oraciones en idioma español. De dichas oraciones se extrajo un total de 209 relaciones semánticas de forma manual. El resumen se puede apreciar en la Tabla 3.

Fuente	Oraciones	Relaciones
es.wikipedia.org	33	62
tweets Covid-19	13	13
tweets municipalidad	9	15
libros	23	54
periódicos de noticias	32	65

Tabla 3: Cantidad de oraciones por origen, junto a sus relaciones semánticas en el conjunto de entrenamiento.

El detalle de las oraciones y sus respectivas relaciones semánticas puede hallarse en la siguiente URL: https://bit.ly/3a5VgUT. Los tweets y las frases de periódicos de noticias fueron obtenidas al azar de tres conjuntos de datos separados disponibles públicamente en el sitio https://www.kaggle.com. Las frases de libros fueron extraídas manualmente de diversos libros.

A este conjunto de datos de entrenamiento se sumó uno adicional con 12 oraciones creadas especialmente para propósitos de prueba durante la fase de desarrollo del método. La mayoría de estas frases son traducciones de las frases

propuestas en (Del Corro y Gemulla, 2013) como ejemplos de los diferentes tipos de oraciones para el idioma inglés.

4.1.2 Mejora en el sistema de puntaje

La otra tarea que se adicionó para mejorar la precisión fue mejorar el sistema de puntaje. El sistema de puntaje asigna valores positivos o negativos a las extracciones encontradas para determinar qué tan correctas son y finalmente determinar si serán devueltas o no. Este sistema está basado en el que propuso Fader y Etzioni en (2011). Contiene las mismas reglas básicas (menos una que fue eliminada) más 7 reglas adicionales. La regla eliminada asignaba un puntaje negativo a oraciones con más de 20 palabras. Las reglas adicionadas se listan en la Tabla 4.

Regla	Puntaje
Si **e2** igual **e1**	-100
Si **e1** contiene a **r**	-100
Si **e1** termina con la misma palabra con que empieza **r**	-50
Si **e2** es un determinante	-1000
Si *longitud*(**r**) = 1 palabra y **r** es determinante	-200
Si *longitud*(**r**) = 1 palabra y **r** es verbo	10

Tabla 4: Reglas adicionales utilizadas para puntuar las relaciones semánticas extraídas, teniendo cada una la forma: (e1, r, e2).

4.1.3 Detección de sujetos tácitos

Un problema adicional del idioma español es el sujeto tácito, si bien este existe en idioma inglés es poco frecuente y en general solo se omite el sujeto si ya fue nombrado antes en la misma oración. Esta particularidad del idioma español provocaba que el algoritmo no detectase relaciones semánticas en muchas oraciones, ya que no encontraba dentro de la misma un sujeto candidato para el primer argumento. Supóngase la oración del ejemplo 5.

(5) *Jugábamos al fútbol.*

La relación semántica en este caso, indica que "*nosotros"* es el argumento primero, el sujeto que está relacionado con "*fútbol*" mediante la relación "*jugar al*". La tupla debería quedar de la siguiente forma:

(6) *(Nosotros; jugábamos; al fútbol).*

Sin embargo, el algoritmo original no encontrará nunca una palabra en la oración de entrada que pueda asociar al argumento primero, simplemente porque esa palabra no está presente. Para resolver este problema, cuando el algoritmo no es capaz de hallar un argumento primero, añadirá al comienzo de la oración una palabra comodín, que será analizada como un pronombre personal cualquiera. Si con esta nueva palabra encuentra una relación semántica tal que el comodín coincide con el argumento primero, el algoritmo devolverá una tupla con el argumento primero vacío. Siguiendo con el ejemplo anterior:

(7) *(; jugábamos; al fútbol).*

El espacio inicial vacío indica que hay un sujeto tácito en la relación semántica devuelta. Esta mejora busca no solo aumentar la precisión sino también la exhaustividad (*recall*) del método.

4.2 Añadir evidencia

Se construyó un conjunto de datos de prueba usando las mismas fuentes utilizadas para la construcción del conjunto de datos de entrenamiento. Este conjunto de datos consta de 55 oraciones diferentes y un total de 120 relaciones semánticas extraídas de forma manual, Tabla 5.

Fuente	Oraciones	Relaciones
es.wikipedia.org	16	36
tweets Covid-19	6	10
tweets municipalidad	5	9
Libros	12	29
periódicos de noticias	16	36

Tabla 5: Cantidad de oraciones por origen, junto a sus relaciones semánticas en el conjunto de pruebas.

El detalle de las oraciones y sus respectivas relaciones semánticas puede hallarse en la siguiente URL: https://bit.ly/3a5VgUT

4.3 Mejorar la calidad de las extracciones realizadas

Para mejorar la calidad de las extracciones realizadas, se implementaron tres mejoras al algoritmo original, las cuales se detallan en las secciones siguientes.

4.3.1 Expandir la relación

El algoritmo original está pensado para construir la relación propiamente dicha según patrones de coincidencia en el árbol de dependencias sintácticas. Esto implica que puede tomar palabras no consecutivas dentro de la oración para formar la relación. Teniendo en cuenta el ejemplo 1, podría construir la relación: *fue galardonado en,* aunque las palabras *galardonado* y *en* no son consecutivas. Si bien en este ejemplo esto es correcto, hay muchos casos en los cuales se generan relaciones poco informativas o bien el algoritmo no logra encontrar un argumento segundo para la relación armada. Para estos casos se decidió ampliar la relación y que esta contenga todas las palabras existentes entre su palabra inicial y final. Para el ejemplo 1, quedaría: *fue galardonado con el Premio Nobel en.*

4.3.2 Agregar nombres de entidades

Existen casos en donde una oración contiene un nombre de entidad (NER) y sin embargo esta no aparece en la extracción realizada. Por ejemplo, TP-OIE-ES para la siguiente oración:

(8) *El tío Juan nos escribió una carta.*

Extrajo la siguiente tupla:

(9) (El tío; nos escribió; una carta).

Esto se debe a que el *parser* superficial que busca un posible argumento primero, no detecta a la entidad (en este caso Juan) como parte de la frase nominal. Se agregó en este caso una corrección al algoritmo para que no ignore entidades detectadas que están junto a frases nominales candidatas.

4.3.3 Tener en cuenta múltiples verbos

El método original fallaba al extraer la relación en oraciones donde aparecen dos o más verbos seguidos, ya que, por lo general, solo extrae uno solo de los verbos, respetando las coincidencias del patrón de extracción. Para ilustrar este punto, supóngase que una oración como la del ejemplo 10, (cuyo árbol de dependencias sintáctico se muestra en la Figura 1), fue utilizada para entrenar al algoritmo.

(10) La ciencia mejoró la sociedad.

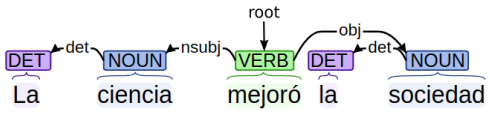

Figura 1: Árbol de dependencias sintáctico y categorías gramaticales. Oración del ejemplo 10.

La relación que en este caso es: *mejoró,* es el verbo raíz en el árbol de dependencias sintácticas. Este ejemplo generará un patrón que servirá para identificar a cualquier verbo raíz como una posible *relación* candidata para la tupla. Con lo cual, una oración como la del ejemplo 11, cuyo árbol de dependencias sintácticas se muestra en la Figura 2, detectará como *relación* candidata la palabra: *permitido* e ignorará el verbo *mejorar*.

(11) La ciencia ha permitido mejorar la sociedad.

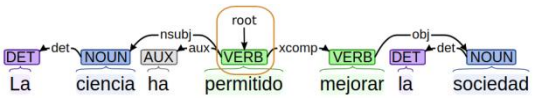

Figura 2: Árbol de dependencias sintáctico y categorías gramaticales, oración del ejemplo 11. En naranja se muestra la coincidencia del patrón utilizado.

En este ejemplo el patrón tampoco está considerando al verbo auxiliar: *ha*, aunque este podría ser detectado por un patrón diferente.

La mejora que se introdujo al método original implica mantener unidos los verbos que están juntos en la oración (incluidos los auxiliares). Esto ha permitido mejorar la informatividad en las relaciones extraídas.

5 Resultados

Para medir la precisión, la exhaustividad (*recall*) y la medida F1 se utilizaron los dos conjuntos de pruebas mencionados en las secciones 2.1 y 4.2. En el primer conjunto se comparó ECMes con otros tres métodos de Open IE en español: ArgOE (Gamallo y Garcia, 2015), DepOE (Gamallo y Garcia, 2012) y TP-OIE-ES (Rodríguez y Merlino, 2020). Y en el segundo conjunto se lo comparó solo contra ArgOE y DepOE.

Se detallan a continuación las fórmulas utilizadas para calcular la precisión, la exhaustividad y la medida F:

$$\text{Precisión} = \frac{\text{extracciones correctas}}{\text{total de extracciones}} \quad (1)$$

$$\text{Exhaustividad} = \frac{\text{extracciones correctas}}{\text{total extracciones manuales}} \quad (2)$$

$$F_\beta = \frac{(1+\beta^2) \cdot \text{Precisión} \cdot \text{Exhaustividad}}{(\beta^2 \cdot \text{Precisión}) + \text{Exhaustividad}} \quad (3)$$

En fórmula (2) para la variable: *total extracciones manuales* se asume que las relaciones semánticas extraídas de forma manual conforman la totalidad de las existentes. Para el conjunto de datos de prueba de Gamallo y García es 137 según se indica en el trabajo de (2015). Para el nuevo conjunto de datos, descripto en la sección 4.2 es de 122.

En la fórmula (3) el parámetro β se estableció igual a 1, para que la precisión y la exhaustividad tuviesen el mismo peso en la fórmula. Por ello nos referimos a la medida F directamente como *medida F1* o simplemente *F1*.

Los resultados obtenidos se resumen en las tablas 6 y 7.

Medidas	**DepOE**	**ArgOE**	**TP-OIE-ES**	**ECMes**
Precisión	0.89	0.67	0.62	**0.92**
Exhaustividad	0.29	0.29	0.36	**0.42**
Medida F1	0.44	0.40	0.46	**0.57**

Tabla 6: Resultados obtenidos sobre el conjunto de datos de prueba de Gamallo y García (2015).

Medidas	**DepOE**	**ArgOE**	**ECMes**
Precisión	**0.81**	0.68	0.68
Exhaustividad	0.18	0.22	**0.34**
Medida F1	0.30	0.33	**0.45**

Tabla 7: Resultados obtenidos sobre el conjunto de datos de prueba presentado en la sección 4.2.

Como puede observarse en los resultados mostrados en ambas tablas ECMes supera a los otros métodos con un margen de al menos 10 puntos porcentuales para la medida F1. Y en el primero de los conjuntos de prueba supera incluso a DepOE en precisión. Sin embargo la precisión disminuye bastante en el segundo conjunto de pruebas, el cual tiene oraciones mucho menos uniformes, aunque en compensación logra mantener una exhaustividad relativamente alta en relación a los otros dos métodos.

Respecto a la informatividad de las relaciones semánticas extraídas, en el conjunto de pruebas de la tabla 6, solo 4 relaciones semánticas fueron identificadas como poco informativas a diferencia del método original TP-OIE-ES que tenía un total de 19 relaciones semánticas poco informativas. Además como extrajo mayor cantidad de relaciones semánticas el porcentaje de relaciones poco informativa cayó de 39% a solo 7%. Para el conjunto de datos presentado en la tabla 7 el número de extracciones semánticas poco informativas es de 13.

El código fuente del método junto a la totalidad de las pruebas realizadas pueden encontrase en *Github* en la siguiente URL:

https://github.com/juanma1982/ECMes

6 Conclusiones

A partir de los resultados presentados en la sección 5 podemos concluir que el método propuesto en este artículo: ECMes supera al método original sobre el cual está construido: TP-OIE-ES y que puede ubicarse entre los métodos de *Open Information Extraction* en idioma español en el *estado-del-arte*.

Si bien ECMes está construido como una serie de mejoras sobre TP-OIE-ES, al haber reentrenado al algoritmo desde cero utilizando datos en español, y al haber adaptado todo el algoritmo de búsqueda de coincidencias por patrones a idioma español nos hemos alejado un poco de la propuesta original de TP-OIE que pretendía construir patrones universales capaces de ser utilizados en diferentes idiomas con solo algunas pequeñas adaptaciones. TP-OIE-ES es la adaptación a español de TP-OIE. Sin embargo, ECMes muestra que la focalización exclusiva en idioma español arroja mejores resultados.

Bibliografía

Angeli, G., M. J. Premkumar, y C. D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint*

Conference on Natural Language Processing, 1, págs. 344--354.

Banko, M., M. J. Cafarella, S. Soderland, M. Broadhead, y O. Etzioni. 2007. Open information extraction for the web. *IJCAI , 7*, 2670-2676.

Buchholz, S. y E. Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. *In Proceedings of the tenth conference on computational natural language learning*, (págs. 149-164).

Chen, D. y C. Manning. 2014. A fast and accurate dependency parser using neural networks. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, (págs. 740--750).

Del Corro, L. y R. Gemulla. 2103. ClausIE: clause-based open information extraction. *22nd international conference on World Wide Web*, (págs. 355-366).

Fader, A., S. Soderland, y O. Etzioni. 2011. Identifying relations for open information extraction. *Association for Computational Linguistics*, (págs. 1535-1545).

Gamallo, P. y M. Garcia. 2012. Dependency-based open information extraction. En A. f. Linguistics (Ed.), *Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP*, (págs. 10--18).

Gamallo, P. y M. Garcia. 2015. Multilingual open information extraction. En Springer (Ed.), *Portuguese Conference on Artificial Intelligence*, (págs. 711--722).

Gashteovski, K., R. Gemulla, y L. Del Corro. 2017. Minie: minimizing facts in open information extraction. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (págs. 2630--2640).

Glauber, R. y D. Barreiro Claro. 2018. A systematic mapping study on open information extraction. *Expert Systems with Applications* (págs. 372--387). Elsevier.

Petrov, S. D. 2011. A universal part-of-speech tagset.

Ratnaparkhi, A. 1996. A maximum entropy model for part-of-speech tagging. *In Conference on Empirical Methods in Natural Language Processing*.

Rodríguez, J. M. y H. D. Merlino. 2020. TP-OIE-ES: Método autónomo de extracción de relaciones semánticas para la Web en Español. *Conferencia Iberoamericana de Complejidad, Informática y Cibernética: CICIC 2020*. Orlando, Florida, USA. Manuscript submitted for publication.

Rodríguez, J. M., H. D. Merlino, P. Pesado, y R. García-Martínez. 2018. Evaluation of open information extraction methods using Reuters-21578 database. En ACM (Ed.), *2nd International Conference on Machine Learning and Soft Computing (ICMLSC '18)*, (págs. 87--92).

Rodríguez, J. M., H. D. Merlino, y R. García-Martínez. 2015. Revisión Sistemática Comparativa de Evolución de Métodos de Extracción de Conocimiento para la Web. *XXI Congreso Argentino de Ciencias de la Computación (CACIC 2015)*. Buenos Aires, Argentina.

Schmitz, M., R. Bart, S. Soderland, y O. Etzioni. 2012. Open language learning for information extraction. *2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (págs. 523-534).

Zhila, A. y A. Gelbukh. 2013. Comparison of open information extraction for English and Spanish. *Computational Linguistics and Intelligent Technologies*, *12, number 19*, págs. 714--722.