Revisión Sistemática Comparativa de Evolución de Métodos de Extracción de Conocimiento para la Web

Juan Manuel Rodríguez[1,2], Hernán D. Merlino [1,2], Ramón García-Martínez[2]

[1] Cátedra de Sistemas de Soporte para Celdas de Producción Flexible.
Departamento de Computación. Facultad de Ingeniería, Universidad de Buenos Aires.
[2] Laboratorio de Investigación y Desarrollo en Sistemas de Inteligencia Artificial
Grupo Investigación en Sistemas de Información (LIDSIA-GISI).
Depto. Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús
http://sistemas.unla.edu.ar/sistemas/gisi/LIDSIA.htm
jmrodriguez1982@gmail.com, hmerlino@gmail.com, rgm1960@yahoo.com

Resumen. En esta comunicación se revisa la evolución los métodos de extracción de conocimiento, con foco en la extracción de relaciones semánticas desde el año 2005 hasta la fecha, teniendo como hilo conductor aquellos métodos creados para trabajar en la Web, centrados en grandes volúmenes de información no estructurada en lenguaje natural y sin dominio especifico. Se resumen distintas comparaciones realizadas entre los métodos relevados.

Palabras Clave. Extracción de conocimiento, extracción de relaciones semánticas, métodos de extracción auto-supervisados, *open information extraction*, procesamiento de lenguaje natural.

1. Introducción

Desde el cambio de milenio, la Web se ha convertido en un repositorio emergente de conocimiento embebido que crece exponencial de manera continua. La necesidad de explotar estos conocimientos ha servido para recuperar la tradición de amplio campos de la Inteligencia Artificial en algo que la comunidad llama en la actualidad computación cognitiva (*cognitive computing*) [Modha et al., 2011] pero en el nuevo contexto de grandes datos. En este nuevo contexto aparecen necesidades específicas dentro de las técnicas de extracción de conocimiento como lo son las técnicas de extracción de relaciones semánticas para grandes volúmenes de datos no estructurados, en particular en lenguaje natural.

Utilizando el método de revisiones sistemáticas [Argimón, 2004] se ha realizado una investigación documental sobre métodos de extracción de conocimiento para la web y se han comparado la calidad de los resultados.

En este contexto, se introducen conceptos de extracción de conocimiento (sección 2), se revisan las distintas técnicas de extracción de relaciones semánticas (sección 3) focalizándose en los llamados métodos basados en conocimiento (sección 3.1), los métodos de tipo supervisados (sección 3.2), y los métodos auto-supervisados (sección 3.3). Se formula una revisión exhaustiva de los distintos métodos de extracción de relaciones semánticas auto-supervisados creados para la Web (sección 4), detallando

las distintas formulas utilizadas para comparar le efectividad de los métodos (sección 4.1), y revisando en dichas comparaciones (sección 4.2). Por último, se presenta una comparación global formulada por los autores sobre la calidad de respuesta de unos métodos frente a otros con detalle de la fuente de cada comparación (sección 5), se indican las diferencias observadas entre ellas y se da cuenta de las comparaciones que quedan por hacerse. Finalmente se presentaln conclusiones parciales y se señalan las futura lineas de trabajo (sección 6).

2. Introducción a la Extracción de Conocimiento

La extracción de conocimiento es cualquier técnica mediante la cual un proceso automatizable es capaz de analizar fuentes de información no estructurada, como por ejemplo textos escritos en lenguaje natural y extraer el conocimiento allí embebido para representarlo de una manera estructurada, manipulable en procesos de razonamiento automático, como por ejemplo: una regla de producción o un subgrafo en una red semántica. A la información obtenida como salida de este tipo de procesos se la llama: pieza de conocimiento [García-Martínez & Britos, 2004; Gómez et al., 1997] Si se piensa a la extracción de conocimiento como una transformación algebraica podría plantearse:

```
extracción_de_conocimiento(estructuras_de_información) = piezas_de_conocimiento.
```
(1)

El desafío de la extracción de conocimientos comienza a fines de la década de 1970 como es señalado en [Cowie & Lehnert, 1996]. Más tarde en los años 90 la investigación fue alentada y financiada por la Agencia de Proyectos Avanzados de Defensa (DARPA) [Konstantinova, 2014].

Los métodos de extracción de conocimiento comenzaron trabajando en la detección y clasificación de nombres propios, utilizando como entrada fuentes de información no estructurada, este tipo de extracción de conocimiento es llamado Reconocimiento de Nombres de Entidades (NER según sus siglas en inglés). En general estos sistemas de extracción de conocimiento buscan nombres de personas, compañías, organizaciones y lugares geográficos [Konstantinova, 2014]. El siguiente paso que dieron los métodos de extracción de conocimiento fue el de resolver correferencias y el de extraer relaciones entre nombres de entidades [Jurafsky & Martin, 2000].

Hacía fines de la década de 2000 los métodos de extracción de conocimiento se habían diversificado y especializado. En [Jurafsky & Martin, 2000] se reconocen distintos tipos de piezas de conocimiento susceptibles de ser extraídas: nombres de entidades, expresiones temporales, valores numéricos, relaciones entre entidades y expresiones previamente identificadas, eventos, entre otras.

La extracción de conocimiento tradicionalmente ha requerido de participación humana en la forma de reglas de extracción o bien de ejemplos de entrenamiento etiquetados de forma manual. En particular para los casos de extracción de relaciones entre entidades, es el usuario quien debe explícitamente especificar cada relación que le interese, tarea ardua, sobre todo cuando se trabaja con fuentes heterogéneas de

información no estructurada y con volúmenes de datos demasiado grandes, como podría ser la Web. Debido a ello en general los sistemas de extracción de conocimiento fueron utilizados sobre fuentes de información no estructurada más bien pequeñas y homogéneas [Banko et al., 2007].

3. Extracción de Relaciones Semánticas

Una subtarea comprendida dentro del conjunto de métodos de extracción de conocimiento es la de extraer de relaciones semánticas. En [Culotta et al., 2006] se define a la extracción de relaciones semánticas como: "la tarea de descubrir conexiones semánticas entre entidades". Y agrega que es de uso común realizar esta tarea utilizando como entrada textos en lenguaje natural en los cuales se suele identificar primeramente grandes cantidades de pares de entidades por documento para luego determinar si existe una relación entre estas utilizando pistas basadas en las características del lenguaje analizado.

En [Etzioni et al., 2008] se clasifican los métodos de extracción de relaciones semánticas en tres clases:

– métodos basados en conocimiento (*knowledge-based methods*)

– métodos supervisados (*supervised methods*)

– métodos auto-supervisados (*self-supervised methods*)

3.1. Métodos Basados en Conocimiento

Los primeros sistemas de extracción de relaciones eran específicos para un dominio, por ejemplo en 1991 DARPA desafió a la comunidad que estaba trabajando en procesamiento de lenguaje natural a "construir sistemas robustos capaces de llenar plantillas con piezas de conocimiento sobre el terrorismo en América Latina", los campos requeridos eran: fechas, ubicaciones, perpetradores, armas, víctimas y objetivos físicos. Más adelante los dominios fueron cambiando y se centraron en *joint ventures*, microelectrónica y planes para la sucesión de gestiones empresariales.

Este tipo de sistemas estaban basados en reglas de coincidencia de patrones (*pattern-matching)* creadas a mano para cada dominio. Estos sistemas tenían la desventaja de no ser escalables ni portables entre dominios diferentes [Etzioni et al., 2008].

3.2. Métodos Supervisados

Este tipo de métodos trabaja con un conjunto de datos de entrenamiento en donde ciertos ejemplos específicos, para un dominio de interés, son previamente etiquetados. Para luego, utilizando dichos ejemplos, entrenar un extractor de forma automática. La principal contra de este tipo de métodos radica en el tiempo y el esfuerzo que se requiere para construir el conjunto de datos de entrenamiento [Konstantinova, 2014].

3.3. Métodos Auto-supervisados

En 2005 Oren Etzioni en [Etzioni et al., 2005] presenta un método de extracción de relaciones semánticas llamadas KnowItAll, el cual es capaz de aprender a etiquetar sus propios ejemplos de entrenamiento utilizando solo un conjunto pequeño de patrones de extracción, independientes de cualquier dominio. Este fue el primer sistema publicado capaz de encarar la extracción de conocimiento de páginas Web ya que era no supervisado, independiente del dominio y escalable [Etzioni et al., 2005; Etzioni et al., 2008]. Los métodos estudiados en el presente trabajo pertenecen a esta última categoría.

4. Métodos de Extracción de Conocimiento Creados para la Web

En el año 2007 Michele Banko introduce un nuevo concepto en materia de extracción de conocimiento, al que llama en inglés: *Open Information Extraction* (OIE). Se trata de un paradigma de extracción de conocimiento en donde un sistema informático realiza una sola pasada sobre el total de las fuentes de información no estructurada en formato de lenguaje natural (llamado *corpus* de documentos), dadas como entrada y extrae un gran conjunto de tuplas relacionales sin requerir ningún tipo de participación humana. Cabe aclarar que este paradigma de extracción de conocimiento pertenece a la clase de métodos auto-supervisados. En el mismo trabajo Banko presenta un método llamado TEXT RUNNER, el cual es el primer método que trabaja dentro de este nuevo paradigma [Banko et al., 2007].

A partir de este trabajo se propusieron otros métodos de extracción de conocimiento bajo el paradigma que Banko llamó Open IE y que podríamos identificar de forma más concreta como métodos de extracción de conocimiento para la Web.

4.1. Comparaciones entre Distintos Métodos de Extracción de Conocimiento Creados para Web

En el punto 3.2 se presentan distintos métodos de extracción de conocimiento para la Web y comparaciones entre ellos. Dichas comparaciones fueron obtenidas del relevamiento de distintos artículos y es por ello que utilizan distintas formulas para evaluar de forma cuantitativa la calidad y cantidad de las piezas de conocimiento extraídas por los diferentes métodos de extracción de conocimiento.

La formula más comúnmente utilizada es la precisión, la cual se calcula como los casos de éxito sobre las extracciones totales, o más específicamente en este caso:

$$\text{Precisión} = \frac{\text{cantidad\_de\_piezas\_de\_ conocimiento\_extraídas\_ correctamente}}{\text{cantidad\_de\_piezas\_de\_ conocimiento\_extraidas}} \quad (2)$$

La segunda fórmula más utilizada, pero prácticamente en conjunto con la precisión fue la exactitud (*recall* en inglés), la cual se calcula como la cantidad de casos de

éxito sobre la cantidad de casos relevantes totales, o más específicamente en este caso:

$$\text{Exactitud} = \frac{\text{cantidad\_de\_piezas\_de\_ conocimiento\_extraídas\_ correctamente}}{\text{cantidad\_de\_piezas\_de\_conocimiento\_totales\_en\_el\_documento}} \quad (3)$$

En la mayoría de los casos la cantidad de piezas totales de conocimiento fueron etiquetadas a mano, en ocasiones por más de una persona.

Otra fórmula utilizada fue la Medida-F, la cual se calcula utilizando las dos medidas anteriores más un parámetro β, que indica a cuál de las dos se le da una ponderación mayor. Su fórmula es la siguiente:

$$F_\beta = \frac{(1+\beta^2)2 * \text{Precisión} * \text{Exactitud}}{(\beta^2 * \text{Precisión}) + \text{Exactitud}} \quad (4)$$

En todos los artículos relevados, siempre se utiliza esta medida con $\beta=1$, para simplificar en este trabajo se referirá a la Medida-F con $\beta=1$, como Medida-F1 o simplemente F1.

Por último en algunos artículos se utiliza el área bajo la curva *Receiver Operating Characteristic* (ROC) como medida de la calidad de las piezas de conocimiento extraídas. Esta medida se basa en una representación grafica de la tasa de verdaderos positivos contra la tasa de falsos positivos, y el área que encierra dicha curva es una medida de la calidad. Un área de 1 representa una calidad perfecta, en este caso significaría que el método extrajo correctamente todas las piezas de conocimiento sin extraer ninguna de más ni de menos. Un área de 0,5 representa una calidad nula, en este caso significaría que el método no logró extraer ninguna pieza de conocimiento correctamente [Bradley, 1997].

4.2. Evolución de los Métodos de Extracción de Conocimiento Creados para Web

El siguiente trabajo en el que se presentó un método de extracción de conocimiento para la Web fue el de Wu y Weld en 2010 [Wu & Weld, 2010], allí se presentan dos métodos WOE-parse y WOE-pos, el primero WOE-parse utiliza un enfoque ligeramente distinto, trabaja con un árbol de dependencias, realizando un análisis sintáctico en cada oración para extraer las relaciones, y si bien logra un mayor número de extracciones que TEXT RUNNER (1.42 tuplas por oración frente a 0.75), es 30 veces más lento que su predecesor. WOE-pos por el contrario es igual de rápido que TEXT RUNNER y ligeramente mejor (1.05 tuplas extraídas por oración). Si bien WOE-parse y WOE-pos, son métodos de propósito general su base de entrenamiento fue Wikipedia.

En [Mesquita et al., 2010] se presenta un método de extracción de conocimiento llamado SONEX pensado para extraer relaciones de redes sociales y de la *blogosfera*. Si bien el trabajo realizado es de interés, no se realizan comparaciones con otros métodos que permitan discernir si aporta alguna mejora considerable.

En [Christensen et al., 2011] se presenta un nuevo enfoque bajo las mismas consignas de extracción de conocimiento planteadas por Banko, se busca utilizar la técnica de etiquetamiento secuencial, basado en la función semántica (en inglés

Semantic Role Labeling) para la extracción de relaciones entre entidades. Se crean dos métodos nuevos: SRL-Lund y SRL-UIUC, se los compara con TEXT RUNNER en dos conjuntos de piezas de información no estructurada, uno pequeño y otro grande. Ambos demuestran ser más precisos que TEXT RUNNER, SRL-Lund obtiene una precisión de 0.7 y una medida F1 de 0.59, SRL-UIUC obtuvo una precisión de 0.63 y una medida F1 de 0.68 mientras que TEXT RUNNER obtuvo una precisión de 0.55 y una medida F1 de 0.35. Sin embargo, al trabajar con el conjunto más grande de datos de entrada, TEXT RUNNER demuestra tener una ventaja adicional, es 20 veces más rápido que SRL-LUND y 500 veces más rápido que SRL-UIUC.

En [Fader et al., 2011] Fader propone un método de extracción de conocimiento, el cual logra un área bajo la curva ROC mayor que WOE-parse, WOE-pos y que TEXT RUNNER, se trata de ReVerb. ReVerb fue puesto a prueba utilizando un conjunto de información no estructurada que constaba de 500 millones de sentencias web, demostró ser más rápido incluso que TEXT RUNNER. En un subconjunto de 100 000 sentencias se obtuvieron los siguientes tiempos: WOE-parse tardó 11 horas, WOE-pos y TEXT RUNNER tardaron 21 minutos cada uno y ReVerb 16 minutos. La mejora introducida por Fader consistió en agregar restricciones a TEXT RUNNER y centrarlo en la extracción de relaciones basadas en verbos.

La restricción de ReVerb no le permite encontrar relaciones basadas en otro tipo de categorías gramaticales que no sean verbos, es por eso que algunos autores proponen diversos métodos para extraer relaciones basadas en otro tipo de categorías gramaticales, en particular sustantivos, es el caso de OLLIE [Schmitz, 2012], de ReNoun [Yahya et al., 2014] y de TRIPLEX [Mirrezaei et al., 2015]. OLLIE es planteado directamente como una mejora a ReVerb, siendo su objetivo encontrar relaciones basadas no solo en verbos, sino también en sustantivos y adjetivos. Además analiza la posibilidad de hacer un análisis del contexto para encontrar relaciones no explicitas. OLLIE logra obtener 2.7 veces más área sobre la curva ROC que ReVerb y 1.9 veces más área bajo la curva ROC que WOE-parse, además OLLIE encuentra 4.4 veces más extracciones correctas que ReVerb y 4.8 veces más que WOE-parse [Schmitz, 2012]. El enfoque de TRIPLEX es ligeramente distinto ya que funciona como un complemento a ReVerb o a OLLIE, en el estudio realizado en [Mirrezaei et al., 2015], TRIPLEX por sí solo no logra superar a OLLIE (se compara utilizando la medida F1 en este caso) y es el uso conjunto de OLLIE más TRIPLEX el que arroja un mejor resultado, aunque no muy lejano al que arroja OLLIE por sí solo.

Para concluir en [Del Corro & Gemulla, 2013] se presenta un método de extracción de conocimiento llamado ClauseIE (respetando el paradigma propuesto por Banko); en dicho trabajo ClauseIE es comparado contra ReVerb, OLLIE, TEXT RUNNER y WOE-pos utilizando distintas fuentes de información no estructurada en formato de lenguaje natural: 500 oraciones extraídas del conjunto de datos de prueba utilizado con ReVerb en [Fader et al., 2011], 200 oraciones aleatorias extraídas de Wikipedia y 200 oraciones aleatorias extraídas del New York Times. El resultado en todos los casos fue favorable a ClauseIE quien obtuvo una mejor precisión que los demás métodos.

En la figura 1 se ilustra la evolución de los distintos métodos presentados. Figura 1 se puede ver de forma cualitativa como fue evolucionando el desempeño de los distintos métodos a lo largo del tiempo. Las comparaciones fueron obtenidas al analizar los distintos trabajos y al recoger la información allí presentada. Dicha

información consiste en datos relacionados con calidad y cantidad de las piezas de conocimiento extraídas. Como se mencionó, estos datos pueden estar representado de diversas formas: precisión, exactitud, medida-F1 o área bajo la curva ROC.

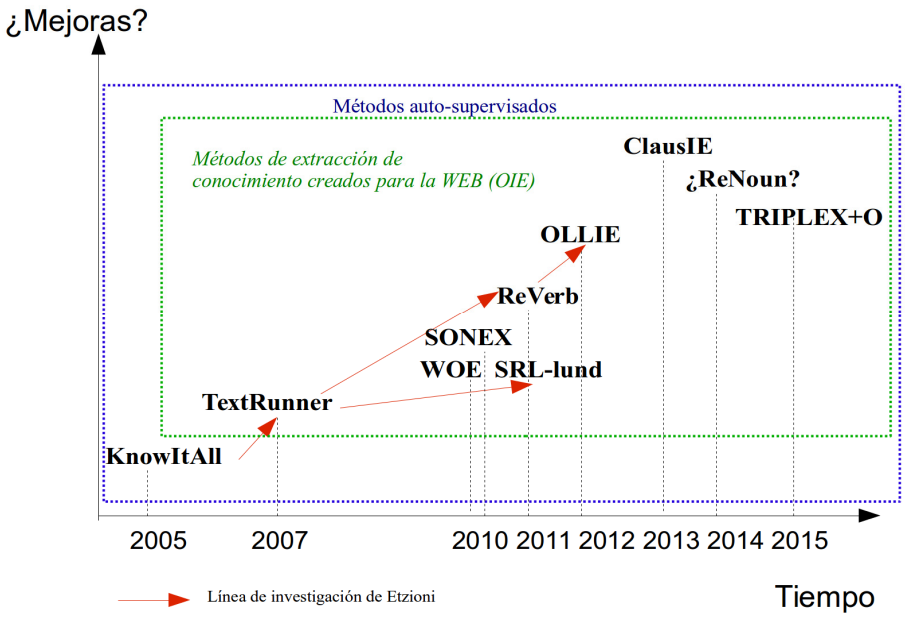

Fig. 1. En el siguiente gráfico se muestra la mejora supuesta entre los distintos métodos de extracción de conocimiento creados para la Web *versus* el tiempo, partiendo desde el trabajo de Etzioni de 2005 en donde presenta a KnowItAll.

Hay que tener en cuenta también que los métodos no siempre son comparados utilizando los mismos conjuntos de datos de entrada. Las publicaciones en donde se presentan los métodos: KnowItAll, TextRunner, ReVerb, SRL-Lund y OLLIE tienen como autor a Oren Etzioni; es decir que en el desarrollo de las comparaciones entre métodos estuvo involucrada al menos una misma persona, esta clase de continuidad sobre una línea de investigación es un fuerte indicio a favor de los datos presentados.

5. Consideraciones

En relación a la valides de las comparaciones anteriores, hay que tener en cuenta que métodos fueron comparados entre sí en un mismo ambiente controlado y cuáles no. En la Tabla 1 se muestra que método fue comparado contra que otro, indicando cual resultó mejor en dicha comparación y la publicación de referencia.

La tabla 1 es una tabla de doble entrada, en donde cada celda debe entenderse como una comparación hecha entre el método indicado en la columna contra el método indicado en la fila. En la celda se indica de manera genérica qué método logró una mayor calidad y cantidad de piezas de conocimiento extraídas, independiente-

mente de la medida utilizada en el artículo. Se indica también la referencia al artículo o los artículos de donde fue relevada la comparación.

Tabla 1. Resumen de comparaciones relevadas entre métodos

MÉTODOS	TextRunner	WOE	SRL-Lund	ReVerb	OLLIE	ClausIE	ReNoun	TRIPLEX
KnowItAll	TextRunner [1]							
TextRunner		WOE [2, 5, 9]	SRL-Lund [4]	ReVerb [5, 9]		ClausIE [9]		
WOE				ReVerb [5, 9]	OLLIE [6, 9]	ClausIE [9]		
SRL-Lund					SRL-Lund [6]			
ReVerb					OLLIE [6, 8], ReVerb [9]	ClausIE [9]		TRIPLEX, TRIPLEX + ReVerb [8]
OLLIE						ClausIE [9]		OLLIE, TRIPLEX + OLLIE [8]
ClausIE								
ReNoun								
TRIPLEX								

Referencias: 1. [Banko et al., 2007], 2. [Wu & Weld, 2010], 3. [Mesquita et al., 2010], 4. [Christensen et al., 2011], 5. [Fader et al., 2011], 6. [Schmitz et al., 2012], 7. [Yahya et al., 2014], 8. [Mirrezaei et al., 2015], 9. [Del Corro & Gemulla, 2013]

Para el caso de TRIPLEX, se indica además la comparación realizada tomando de forma conjunta TRIPLEX más el segundo método, ya que cómo se indicó TRIPLEX se crea como un método suplementario a ReVerb u OLLIE.

Se puede ver que algunos métodos fueron ampliamente comparados, es el caso de TextRunner con 8 comparaciones totales realizadas, de WOE con 8 comparaciones totales realizadas, de ReVerb con 9 comparaciones totales realizadas y de OLLIE con 8 comparaciones totales realizadas. El caso opuesto es el de ReNoun el cual no fue comparado con ningún otro método. Es interesante observar el caso de la comparación de OLLIE con ReVerb hay publicaciones [Schmitz et al., 2012; Mirrezaei et al., 2015] que concluyen que OLLIE es superior a ReVerb y otra [Del Corro & Gemulla, 2013] que concluye que ReVerb obtiene mejores resultados que OLLIE. Por último las comparaciones que se realizaron contra ClauseIE pertenecen todas a la publicación de [Del Corro & Gemulla, 2013]

6. Conclusiones

Del análisis presentado e esta comunicación sobre distintas distintos métodos de extracción de conocimiento creados para la Web (es decir para grandes *corpus* de documentos sin un dominio definido) se puede formular las siguientes conclusiones preliminares:

[i] El mejor de los métodos estudiados en términos de cantidad y calidad de piezas de conocimiento extraído es ClauseIE. No se registran comparaciones con ReNoun ni contra TRIPLEX.

[ii] Dado que TRIPLEX cuando es utilizado en combinación con OLLIE es apenas un poco mejor que OLLIE por sí solo, sería esperable que ClauseIE lo superase en precisión.

[iii] Respecto ReNoun no hay evidencia suficiente para deducir si su desempeño estará por encima o no de ClauseIE.

[iv] A continuación de ClauseIE se ubicarían los métodos: OLLIE, ReVerb y WOE, en ese orden (según su calidad), aunque según el caso de prueba utilizado alguno podría superar a otro.

Queda como trabajo futuro realizar una comparación de todos los métodos relevados para poder completar las comparaciones faltantes, unificando además los criterios para la evaluación del desempeño y la calidad de los métodos. En dicho trabajo, se tendrán en cuenta los distintos conjuntos de datos utilizados como entrada, ya que estos generarán diferentes casos de usos, capaces de aportar evidencia sobre las diferencias observadas en el desempeño de los métodos, como la mencionada entre OLLIE y ReVerb.

Financiamiento

Las investigaciones que se reportan en este artículo han sido financiadas parcialmente por los Proyectos UNLa-33B145, UNLa-33B133 y UNLa-33A205 de la Secretaria de Ciencia y Tecnología de la Universidad Nacional de Lanús.

Referencias

Argimón J. 2004. Métodos de Investigación Clínica y Epidemiológica. Elsevier España, S.A. ISBN 9788481747096.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007, January). Open information extraction for the web. In IJCAI (Vol. 7, pp. 2670-2676).

Bradley, Andrew P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 1997, vol. 30, no 7, p. 1145-1159.

Christensen, J., Soderland, S., & Etzioni, O. (2011, June). An analysis of open information extraction based on semantic role labeling. In Proceedings of the sixth international conference on Knowledge capture (pp. 113-120). ACM.

Cowie, J., & Lehnert, W. (1996). Information extraction. Communications of the ACM, 39(1), 80-91.

Culotta A, McCallum A, Betz J (2006) Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Association for Computational Linguistics, New York, New York, pp 296–303

Del Corro, L., & Gemulla, R. (2013, May). ClausIE: clause-based open information extraction. In Proceedings of the 22nd international conference on World Wide Web (pp. 355-366). International World Wide Web Conferences Steering Committee.

Etzioni O, Banko M, Soderland S, Weld DS (2008) Open information extraction from the web. Communications of the ACM 51:68–74

Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., ... & Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. Artificial intelligence, 165(1), 91-134.

Fader, A., Soderland, S., & Etzioni, O. (2011, July). Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1535-1545). Association for Computational Linguistics.

García-Martínez, R. & Britos, P. V. (2004). Ingeniería de sistemas expertos. Nueva Librería. ISBN 987-1104-15

Gómez, A., Juristo, N., Montes, C., & Pazos, J. (1997). Ingeniería del conocimiento. Editorial Centro de Estudios Ramón Areces. ISBN 84-8004-269-9.

Jurafsky, D., & Martin, J. H. (2000). Speech and language processing an introduction to natural language processing, computational linguistics, and speech recognition, 2nd edn. Prentice-Hall, Inc.

Konstantinova, N. (2014). Review of Relation Extraction Methods: What Is New Out There?. In Analysis of Images, Social Networks and Texts (pp. 15-28). Springer International Publishing.

Mesquita, F., Merhav, Y., & Barbosa, D. (2010). Extracting information networks from the blogosphere: State-of-the-art and challenges. In Proceedings of the Fourth AAAI Conference on Weblogs and Social Media (ICWSM), Data Challenge Workshop.

Mirrezaei, S. I., Martins, B., & Cruz, I. F. (2015). The Triplex Approach for Recognizing Semantic Relations from Noun Phrases, Appositions, and Adjectives. In The Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (Know@LOD) co-located with Extended Semantic Web Conference (ESWC), Portoroz, Slovenia.

Modha, D., Ananthanarayanan, R., Esser, S., Ndirango, A.., Sherbondy, A.., Singh, R. 2011. Cognitive Computing. Communications of the ACM, 54(8): 62-71.

Schmitz, M., Bart, R., Soderland, S., & Etzioni, O. (2012, July). Open language learning for information extraction. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 523-534). Association for Computational Linguistics.

Wu, F., & Weld, D. S. (2010, July). Open information extraction using Wikipedia. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 118-127). Association for Computational Linguistics.

Yahya, M., Whang, S. E., Gupta, R., & Halevy, A. (2014, October). Renoun: Fact extraction for nominal attributes. In Proc. 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar.