ojs.uv.es/index.php/qfilologia/index
Rebut: 18.07.2023. **Acceptat**: 15.09.2023
Per a citar aquest article: Periñán-Pascual, Carlos. 2023. "Un uso sostenible de WordNet en la inteligencia artificial". *Quaderns de Filologia: Estudis Lingüístics* XXVIII: 141-155.
doi: 10.7203/QF.28.26595

Un uso sostenible de WordNet en la inteligencia artificial

A Sustainable Application of Wordnet in Artificial Intelligence

Carlos Periñán-Pascual
Universitat Politècnica de València
jopepas3@upv.es

Resumen: La elaboración manual de extensos recursos lingüísticos para su uso en sistemas de inteligencia artificial es una tarea que requiere un gran esfuerzo económico y humano en un período largo de tiempo, razón por la cual se recomienda un uso sostenible de los recursos existentes. En este contexto, el objeto de investigación de este artículo es WordNet, una base de datos léxica diseñada originariamente para el estudio de las redes semánticas. En concreto, analizamos la sostenibilidad de este recurso lexicográfico en campos de la inteligencia artificial como el procesamiento del lenguaje natural y la ingeniería del conocimiento desde la perspectiva de tres atributos esenciales: extensibilidad, interoperabilidad y reusabilidad.
Palabras clave: sostenibilidad; WordNet; inteligencia artificial; procesamiento del lenguaje natural; ingeniería del conocimiento.

Abstract: The manual compilation of comprehensive language resources for artificial intelligence systems is a labour-intensive, time-consuming and economically costly task, which is why a sustainable use of existing resources is recommended. In this context, the research goal of this article is WordNet, a lexical database originally designed for the study of semantic networks. In particular, we analyse the sustainability of this lexicographical resource in the artificial intelligence fields of natural language processing and knowledge engineering from the perspective of three core attributes: extensibility, interoperability and reusability.
Keywords: sustainability; WordNet; artificial intelligence; natural language processing; knowledge engineering.

1. Introducción

La elaboración manual de extensos recursos lingüísticos (p. ej., bases de datos léxicas y ontologías, entre otros) para la inteligencia artificial es una tarea que exige mucho trabajo en un período largo de tiempo, razón por la cual se recomienda un uso sostenible de los recursos existentes, en línea con lo que

Colman (2016) denominó "lexicografía sostenible". En este sentido, la autora describió brevemente dos aspectos críticos que contribuyen a esta sostenibilidad: la optimización de materiales, productos y recursos económicos, y la automatización del propio proceso lexicográfico. No obstante, este concepto de sostenibilidad requiere una mayor exploración cuando se aplica a los recursos lingüísticos para la inteligencia artificial. Debemos recordar que la cuestión de la sostenibilidad se originó por una preocupación medioambientalista que llevó en 1948 a la creación de la Unión Internacional para la Conservación de la Naturaleza, cuya principal misión sigue siendo "contribuir a encontrar soluciones prácticas para los principales desafíos ambientales y de desarrollo que enfrentaba ya el planeta" (Rivera-Hernández *et al.*, 2017: 58). Por tanto, por analogía, podemos inferir que un uso sostenible de los recursos lingüísticos tiene como objetivo contribuir a encontrar soluciones prácticas para los principales desafíos de desarrollo a los que se enfrenta la inteligencia artificial.

Igualmente, es importante destacar que cualquier modelo sostenible debe adaptarse a la evolución del entorno. Por esta razón, es preciso explorar brevemente la evolución de la inteligencia artificial a través de dos ramas de interés para nuestro estudio: el procesamiento del lenguaje natural (PLN) y la ingeniería del conocimiento. Desde los orígenes del PLN en los años cincuenta hasta finales de la década de los ochenta, el paradigma simbólico dominó la investigación, caracterizada por la construcción de una representación formal del aducto textual y por una base de conocimiento que dota al sistema de la capacidad de explicar y razonar sobre los resultados generados, e incluso sobre los pasos intermedios del proceso. En cambio, el paradigma subsimbólico del PLN se vuelve gradualmente más popular a partir de la década de los noventa, donde se emplean modelos de aprendizaje automático (p. ej., clasificadores bayesianos ingenuos, árboles de decisión y máquinas de vectores de soporte, entre otros muchos) con extensos corpus textuales de entrenamiento. Además, es en la década de los noventa cuando comienza la investigación en ontologías, a través de las cuales los sistemas pueden intercambiar información sobre un dominio de interés. No obstante, y a diferencia del aprendizaje automático, el entusiasmo investigador en las ontologías disminuyó en la primera década del nuevo siglo, debido principalmente a la falta de madurez en las metodologías y las infraestructuras (Tudorache, 2020). En esta última década, el grueso de las investigaciones en el PLN sigue adoptando el paradigma subsimbólico, pero liderado por el aprendizaje profundo, que es sinónimo de redes neuronales, las cuales tratan de imitar el funcionamiento conexionista del cerebro humano. Actualmente, las investigaciones en el PLN apuestan

por un enfoque neurosimbólico, que combina ambos paradigmas con el fin de construir un sistema más eficaz. De acuerdo con Bader y Hitzler (2005), ambos paradigmas de la inteligencia artificial se pueden integrar a través de dos modelos diferentes: el sistema ejecuta los componentes simbólico y subsimbólico en paralelo para luego combinar los resultados (i. e., modelo híbrido) o el sistema consiste en un componente subsimbólico principal que utiliza conocimiento simbólico durante el procesamiento (i. e., modelo unificado). En estos últimos años, por otra parte, la ingeniería del conocimiento ha experimentado una revitalización del desarrollo ontológico en diversos campos (p. ej., biomedicina, finanzas e ingeniería, entre otros), debido en parte a la adopción de estándares ampliamente reconocidos.

Además, con el fin de adaptarse adecuadamente al entorno, destacamos tres atributos esenciales que favorecen la sostenibilidad de los recursos lingüísticos para la inteligencia artificial: la extensibilidad, la interoperabilidad y la reusabilidad. Esta afirmación se apoya en el hecho de que cualquier sistema informático se compone de una serie de instrucciones en forma de algoritmos que procesan un conjunto de datos. Tanto los algoritmos como los datos son componentes críticos durante el análisis, diseño, desarrollo, mantenimiento y evaluación de estos sistemas, tareas que forman parte de la disciplina de la ingeniería del software. Precisamente, desde esta disciplina, Venters *et al.* (2014) caracterizaron el concepto de "sostenibilidad de software" a través de un conjunto de cualidades básicas que es deseable que tengan los sistemas computacionales. Por consiguiente, esas cualidades también son deseables en los recursos lingüísticos que pueden ser aprovechados por las tecnologías del lenguaje, los cuales se dividen en cuatro grupos según Moreno *et al.* (2019): corpus, memorias de traducción (i. e., base de datos lingüística con segmentos de texto en parejas de lenguas), entidades nombradas (i. e., listas con nombres de personas, organizaciones, lugares, cantidades, etc., que se utilizan en tareas de recuperación de la información) y recursos léxicos (p. ej., glosarios, terminologías, tesauros y ontologías). Como se indicó anteriormente, nos centramos en tres de las cualidades que contribuyen a la sostenibilidad del software en general y que aplicamos a los recursos lingüísticos en particular.

Por una parte, los recursos lingüísticos deben ser extensibles, permitiendo la incorporación de nuevos contenidos para poder satisfacer las necesidades de los sistemas. Por ejemplo, si disponemos de un corpus de tuits, podemos anotar cada texto con una etiqueta de polaridad (p. ej., positivo, negativo y neutro) con el fin de utilizar el corpus extendido para el entrenamiento de un modelo orientado al análisis de los sentimientos. Por otra parte, los recursos

lingüísticos deben ser interoperables, facilitando así el intercambio de información con otros recursos dotados de conocimiento. Por ejemplo, si disponemos de varias ontologías que describen un mismo dominio temático (p. ej., la salud), podemos establecer relaciones de equivalencia entre los dos modelos ontológicos con el fin de automatizar el proceso de fusión y así obtener un recurso semánticamente más rico. Finalmente, los recursos lingüísticos deben ser reutilizables, favoreciendo el uso del recurso en aplicaciones para las cuales no fue inicialmente diseñado. Por ejemplo, si disponemos de un glosario de términos especializados, podemos procesar sus definiciones con el fin de generar un grafo conceptual que ayude a resolver los casos de ambigüedad semántica en un texto técnico.

En este contexto, el objeto de investigación de este artículo es la descripción del uso sostenible de WordNet en la inteligencia artificial, una base de datos léxica diseñada originariamente para el estudio de las redes semánticas. Por tanto, analizamos este recurso lexicográfico desde la perspectiva de su extensibilidad, interoperabilidad y reusabilidad para su adaptación a los avances en la inteligencia artificial. Con este fin, tras una breve descripción de WordNet (apartado 2), el artículo presenta algunos estudios destacados que (a) han logrado aumentar el conocimiento de WordNet de manera automática o semiautomática (apartado 3), (b) han permitido intercambiar los datos de WordNet con otros recursos en el ámbito de la ingeniería del conocimiento (apartado 4) y (c) han utilizado WordNet en diversas tareas del PLN (apartado 5). Finalmente, se resumen las principales conclusiones (apartado 6).

2. WordNet

En la década de los noventa, George A. Miller y sus colegas del Cognitive Science Laboratory en la Universidad de Princeton diseñaron WordNet (Miller *et al.*, 1990; Miller, 1995; Fellbaum, 1998a, 1998b) con el fin de construir manualmente una red semántica con palabras del inglés a partir de principios psicolingüísticos[1]. Con el transcurrir del tiempo, WordNet se convirtió en una extensa base de datos léxica que se ha utilizado como recurso de investigación por excelencia en el PLN, debido principalmente a su disponibilidad pública y

[1] Como es práctica habitual, se emplea el nombre WordNet (sin especificar la lengua) para referirnos al recurso original que se construyó para el inglés en la Universidad de Princeton. Siguiendo el modelo de esta base de datos, se han construido recursos similares para muchas otras lenguas, a los que nos referimos como WordNet del español, francés, italiano, etc.

gratuita y a su amplia cobertura. En WordNet, los nombres, verbos, adjetivos y adverbios del inglés se organizan en grupos de sinónimos denominados *synsets*, cada uno de los cuales representa un concepto lexicalizado. Las unidades léxicas que configuran los *synsets* se conectan por medio de diversas relaciones semántico-conceptuales (Miller, 1995): antonimia (p. ej., *wet-dry*), hiponimia/hiperonimia (p. ej., *plant-tree*), implicación (p. ej., *divorce-marry*), meronimia/holonimia (p. ej., *ship-fleet*), sinonimia (p. ej., *pipe-tube*) y troponimia (p. ej., *whisper-speak*). A continuación, describimos los significados de estas relaciones (Miller *et al.*, 1990; Fellbaum, 1998b), donde *x* e *y* representan dos unidades léxicas:

- *X* e *y* son sinónimos en un determinado contexto si la sustitución de una por la otra en ese contexto no altera el valor de la verdad. La sinonimia es la relación implícita que existe entre las palabras que pertenecen a un mismo *synset*.
- *X* es un antónimo de *y* si *x* tiene el significado opuesto de *y*. Cuando ambas son adjetivos cualitativos, *x* suele proporcionarse como respuesta a *y* en un test de asociación léxica.
- *X* es un hipónimo de *y* si los hablantes nativos de inglés aceptan la oración *An x is a kind of y*. La hiperonimia es la relación inversa de la hiponimia, es decir, si *x* es un hipónimo de *y*, entonces *y* es un hiperónimo de *x*.
- *X* es un tropónimo de *y* si los hablantes aceptan la oración *x is to y in some particular manner*.
- *X* es un merónimo de *y* si los hablantes aceptan la oración *An x is a part of y*. La holonimia es la relación inversa de la meronimia, es decir, si *x* es un merónimo de *y*, entonces *y* es un holónimo de *x*.
- *X* mantiene una relación de implicación con *y* si la verdad de la oración *Someone x* conlleva necesariamente la verdad de la oración *Someone y*.

Es importante destacar que, mientras que WordNet asigna las relaciones de antonimia y sinonimia entre unidades léxicas que pertenecen a determinados *synsets*, las relaciones de hiponimia, implicación, meronimia y troponimia se establecen directamente entre *synsets*.

En el caso de los nombres, la principal relación semántica es la hiponimia, la cual permite construir estructuras jerárquicas que pueden llegar a alcanzar hasta doce niveles, que van desde un concepto muy genérico hasta uno extremadamente específico, normalmente técnico. Otras de las principales relaciones entre los *synsets* nominales es la meronimia, que se limita a tres tipos

en WordNet (Miller, 1990): partes separables (p. ej., *blade-knife*), miembros de grupos (p. ej., *professor-faculty*) y sustancias (p. ej., *oxygen-air*).

Al igual que los nombres, los verbos se prestan a una organización jerárquica, pero en esta ocasión a través de la troponimia, donde un verbo especifica una determinada manera de llevar a cabo la acción referida por otro verbo (p. ej., *smack-hit*). Las estructuras arbóreas construidas en torno a los verbos poseen una altura más baja que las de los nombres[2], excediendo raramente de cuatro niveles. Los verbos también se relacionan a través de varias clases de implicación (p. ej., *snore-sleep*).

Por otra parte, WordNet diferencia dos tipos de adjetivos (i. e., cualitativos y relacionales). Los adjetivos cualitativos (p. ej., *heavy*, *light*, *tall*) se organizan en torno a la antonimia. Más concretamente, se organizan en grupos que se forman alrededor de dos adjetivos antónimos, que se consideran "antónimos directos" en WordNet. Las parejas de antónimos directos constituyen un grupo reducido de adjetivos. En cambio, existen muchos más adjetivos que WordNet clasifica como "antónimos indirectos" (p. ej., *huge* es un antónimo indirecto de *small*), los cuales son semánticamente similares a alguno de los antónimos directos. En cambio, los adjetivos relacionales (p. ej., *scientific*, *musical*) no se organizan como los adjetivos cualitativos, sino que se vinculan a los nombres a los que pertenecen.

Finalmente, la mayoría de los adverbios en WordNet se derivan de adjetivos por medio del sufijo *-ly*. Los adverbios también pueden vincularse a través de la antonimia, siguiendo la misma organización de los adjetivos a partir de los cuales se forman.

3. Extensibilidad de WordNet

Con respecto a la extensibilidad, concluimos que se han originado tres tipos de recursos a partir de WordNet. Por una parte, son constantes los esfuerzos por incorporar contenido multilingüe. Una de las primeras iniciativas fue EuroWordNet (Vossen, 1998), que se construyó como una base de datos léxica multilingüe (i. e., alemán, checo, español, estonio, francés, holandés e italiano) en forma de redes de palabras estructuradas de manera similar a WordNet, es decir, *synsets* conectados a través de relaciones semánticas. En concre-

[2] La altura de un árbol es la longitud máxima de la ruta que existe entre el nodo raíz (i. e., un *synset* que no tiene hiperónimo) y un nodo hoja (i. e., un *synset* que no tiene hipónimo), donde cada uno de los nodos (i. e., *synsets*) de la ruta determina un nivel.

to, cada lengua tiene su propia red de palabras, donde los *synsets* de cada red se vinculan a los *synsets* de WordNet, con lo cual las palabras de un determinado significado en una lengua pueden conectarse con las palabras del mismo significado en otra lengua. Esto es posible gracias a que EuroWordNet contiene el Índice Interlingüístico, una lista de registros que consisten básicamente en un *synset* y una glosa especificando el significado. Por tanto, las relaciones de equivalencia entre los conceptos de cada red de palabras y los *synsets* de WordNet se explicitan a través del Índice Interlingüístico. Igualmente, debemos destacar la iniciativa de Global WordNet Association[3], cuyo objetivo no es solo mantener y estandarizar las redes de palabras construidas para múltiples lenguas (actualmente, más de cincuenta) sino también integrarlas con WordNet y EuroWordNet.

Por otra parte, se ha incrementado la conectividad de los *synsets* en WordNet. A este respecto, Ponzetto y Navigli (2010) construyeron WordNet++, una versión extendida de WordNet que incluye millones de relaciones asociativas obtenidas de la Wikipedia. En concreto, este recurso se construyó en dos fases de manera automática. En primer lugar, las páginas de la Wikipedia se asociaron con los *synsets* de WordNet, para lo que se utilizaron contextos de desambiguación como (a) las etiquetas de los sentidos, los vínculos y las categorías que se encuentran en las páginas de la Wikipedia y (b) las relaciones de hiperonimia, hiponimia y cohiponimia de los *synsets*, además de sus glosas, que se encuentran en WordNet. En segundo lugar, todos los vínculos que interconectan las páginas de la Wikipedia se transfirieron a relaciones asociativas en WordNet.

Finalmente, se ha incorporado nuevo contenido lingüístico con el fin de facilitar el desarrollo de determinadas aplicaciones del PLN, como la clasificación de textos o el análisis de los sentimientos y las emociones. Por ejemplo, Magnini y Cavaglià (2000) crearon WordNet Domains, donde los *synsets* de WordNet se anotaron con campos temáticos (p. ej., ARCHITECTURE, MATHEMATICS, SPORT, etc.) por medio de un procedimiento semiautomático a partir de la propia estructura de WordNet. Los *synsets* que no pertenecen a un dominio específico se etiquetaron como FACTOTUM. Strapparava y Valitutti (2004) elaboraron WordNet-Affect, donde un subconjunto de *synsets* de WordNet se etiquetó a partir de un conjunto de categorías afectivas específicas (p. ej., *Alarm, Confusion, Cruelty, Envy, Hate, Love, Wrath,* etc.) que se organizaron jerárquicamente, donde conceptos generales como BEHAVIOUR,

[3] http://globalwordnet.org/resources/wordnets-in-the-world/

COGNITIVE STATE, EMOTION y MOOD, entre otros, se encuentran en los niveles superiores de esta taxonomía afectiva. Esuli y Sebastiani (2006) construyeron uno de los lexicones de polaridad más populares, SentiWordNet, que resultó de la anotación automática de todos los *synsets* de WordNet según sus grados de positividad, negatividad y objetividad. De este modo, cada sentido de una palabra puede tener puntuaciones diferentes en cada dimensión de polaridad, donde cada una de las dimensiones tiene asignado un valor de 0 a 1, siendo 1 la suma de las tres puntuaciones para cada *synset*.

4. Interoperabilidad de WordNet

La heterogeneidad semántica entre ontologías es el principal escollo para conseguir la interoperabilidad y así permitir el intercambio de información semántica a través de diferentes recursos dotados de conocimiento. En la ingeniería del conocimiento y la inteligencia artificial, una ontología es la especificación explícita de una conceptualización (Gruber, 1993). Por consiguiente, la ontología requiere un modelo donde se represente el vocabulario que describa el dominio de interés, para lo que deben definirse unas clases (i. e., conceptos) y unas relaciones entre las clases. Mediante las ontologías, se proporciona un medio para organizar el conocimiento, de tal forma que las personas puedan recuperar información eficazmente y los sistemas computacionales puedan automatizar el razonamiento sobre los datos. Establecer relaciones consistentes entre los elementos de las distintas ontologías no es una tarea trivial, ya que un mismo concepto puede estar representado de múltiples formas en diferentes recursos. Es importante destacar que WordNet no es una ontología formal *stricto sensu*, sino más bien una extensa red semántica. No obstante, muchos investigadores en el PLN, e incluso algunos en ingeniería del conocimiento, emplean WordNet como una ontología ligera sobre palabras, sentidos y sus relaciones, donde la relación de hiperonimia entre los *synsets* se trata como una relación de subsunción entre clases (i. e., en la lógica descriptiva, $C \sqsubseteq D$, siendo C y D clases). En este sentido, los intentos de contribuir a la interoperabilidad de WordNet han sido numerosos.

Por una parte, se ha empleado WordNet como base para el desarrollo de ontologías de dominio específico, asegurando así la interoperabilidad desde sus inicios. A este respecto, uno de los casos más conocidos es OntoLearn (Navigli & Velardi, 2002), una metodología para el aprendizaje automático de ontologías que se estructura en tres fases principales: (a) la extracción de una

terminología de dominio específico a partir de la compilación de un corpus de textos y el posterior filtrado de palabras basado en técnicas estadísticas, (b) la interpretación semántica de los términos y (c) la organización de los conceptos de acuerdo con sus relaciones taxonómicas. WordNet desempeña un papel determinante en las fases (b) y (c) de OntoLearn.

Por otra parte, WordNet se ha alineado con diversos modelos ontológicos de nivel superior y propósito general, por ejemplo, DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) o SUMO (Suggested Upper Merged Ontology), con el fin de encontrar correspondencias entre unidades léxicas y categorías ontológicas, lo cual resulta muy útil para la representación formal de los textos. Por ejemplo, Gangemi *et al*. (2002) presentaron los resultados de utilizar OntoClean como una herramienta para integrar la taxonomía de nivel superior de los nombres en WordNet con el nivel superior de DOLCE. Niles (2003) describió el método de proyección de los *synsets* de WordNet a los conceptos de SUMO.

5. Reusabilidad de WordNet

La reusabilidad de WordNet es evidente, ya que originalmente no se diseñó como un recurso computacional para el PLN. Además, dentro de este campo, WordNet se ha utilizado en diversas áreas, por ejemplo, en la categorización de documentos y la recuperación de información, entre otras. En este sentido, se ha empleado dentro del paradigma simbólico principalmente para la resolución de dos tipos de tareas: la expansión léxica y la desambiguación léxica. Por una parte, la expansión léxica permite enriquecer automáticamente los términos de las consultas en los sistemas de recuperación de documentos a partir de la relación de sinonimia (Moldovan & Mihalcea, 2000) o de las relaciones de sinonimia, hiperonimia e hiponimia (Gong *et al*., 2005) en WordNet. A modo de ilustración, consideremos la primera definición del nombre *plane* en WordNet (i. e., *an aircraft that has a fixed wing and is powered by propellers or jets*), cuyo *synset* contiene otras dos palabras como sinónimos: *airplane* y *aeroplane*. Con esta lista de tres palabras, podemos expandir la consulta en un motor de búsqueda de la siguiente manera: *plane OR airplane OR aeroplane*.

Por otra parte, la ambigüedad léxica, originada por la polisemia y homonimia de las palabras, es un problema crítico en el avance de los sistemas que requieran la comprensión del lenguaje natural. Por ejemplo, a partir del inventario de cinco significados que tiene el nombre *plane* en WordNet, el sis-

tema debería determinar que el sentido más adecuado de *plane* en la oración *Figure 2 shows how to find a tangent plane to a graph* corresponde a la segunda definición (i. e., *(mathematics) an unbounded two-dimensional shape*), para lo cual es preciso considerar el contexto que rodea a la palabra. A este respecto, se han aplicado dos estrategias diferentes en la resolución de la ambigüedad semántica de las palabras a través de WordNet: una basada en la similitud semántica y otra en la polisemia sistemática. Han sido numerosas las propuestas de medidas que permiten calcular la similitud semántica ente dos significados léxicos. Muchas de estas medidas recurren a la jerarquía taxonómica de WordNet, donde se diferencian tres enfoques. En primer lugar, se puede adoptar un enfoque basado en los nodos, es decir, en su contenido de información (Resnik, 1995; Lin, 1998). En segundo lugar, se puede adoptar un enfoque basado en la longitud de las rutas (i. e., el número de aristas o nodos), es decir, en su distancia conceptual (Wu & Palmer, 1994; Leacock & Chodorow, 1998). En tercer lugar, se puede adoptar un enfoque híbrido. Por ejemplo, Jiang & Conrath (1997) se apoyaron en la longitud de las rutas entre *synsets*, pero también añadieron el contenido de información definido por Resnik (1995) como un factor de decisión. Además, se han diseñado métodos basados en los vectores generados a partir del grafo de WordNet para calcular la similitud semántica (Agirre & Soroa, 2009; Goikoetxea *et al.*, 2015). De este modo, la asociación entre dos palabras se calcula a través de la similitud de coseno entre los vectores de sus correspondientes *synsets*.

Por otra parte, como indicó Palmer (1998), las distinciones de sentido de WordNet son demasiado finas, lo cual dificulta la desambiguación léxica. Diversos investigadores (Buitelaar, 2000; Peters & Peters, 2000) han mitigado este problema proponiendo métodos para la detección automática de la polisemia sistemática (o polisemia regular) entre los sentidos, es decir, un conjunto de sentidos léxicos que están relacionados de forma sistemática (p. ej., animal-comida, producto-productor o idioma-persona, entre otras relaciones). Desde esta premisa, cuando las palabras tienen sentidos relacionados y las relaciones son sistemáticas, estas relaciones pueden generalizarse sobre clases semánticas más amplias de palabras similares. De esta manera, estas clases reducen enormemente la cantidad de procesamiento, ya que el número de decisiones durante el proceso de desambiguación es mucho menor.

Igualmente, WordNet se ha empleado bajo el enfoque neurosimbólico en el desarrollo de aplicaciones del PLN, donde destacamos su uso en sistemas de generación de historias y de comprensión escrita. En la generación de historias, algunos investigadores demostraron que el conocimiento del sentido

común puede contribuir a construir textos más coherentes. En esta línea, Yang y Tiddi (2020) desarrollaron DICE, donde se inyecta conocimiento de ConceptNet, WordNet y DBpedia a un modelo GPT-2 (Radford *et al.*, 2019). En la comprensión escrita automática, Mihaylov y Frank (2018) emplearon WordNet y ConceptNet para enriquecer las representaciones textuales que aprendía una unidad recurrente cerrada bidireccional para inferir las palabras omitidas en un texto (p. ej., nombres comunes y entidades nombradas). Wang y Jiang (2019) propusieron Knowledge Aided Reader, un sistema que explota el conocimiento general extraído de pares pasaje-pregunta con la ayuda de WordNet para ayudar a los mecanismos de atención de una red neuronal bidireccional de memoria a largo y corto plazo. Yang *et al.* (2019) introdujeron KT-NET, que emplea un mecanismo de atención para seleccionar el conocimiento de WordNet y NELL (Carlson *et al.*, 2010) y luego inyectar dicho conocimiento a BERT (Devlin *et al.*, 2019) para realizar predicciones basadas en el contexto y en el conocimiento externo.

6. Conclusiones

Este artículo explora la sostenibilidad de un recurso lingüístico como WordNet en el campo de la inteligencia artificial, para lo cual se analizan tres atributos: extensibilidad, interoperabilidad y reusabilidad. Concluimos que WordNet tuvo un papel pivotal en el desarrollo de sistemas de PLN dentro del paradigma simbólico durante la década de los noventa y los primeros años del siglo xxi. Esto se reflejó en la creación de diversos recursos a partir de WordNet, los cuales lograron extender su multilingüismo, conectividad y contenido semántico, y en el uso de WordNet para la optimización de algunos componentes de estos sistemas, por ejemplo, en la clasificación de textos, la recuperación de documentos y el análisis de los sentimientos y las emociones. Igualmente, en los primeros años del siglo xxi, los ingenieros del conocimiento promovieron el alineamiento de WordNet con ontologías existentes de nivel superior y propósito general, además de recurrir a este recurso para el desarrollo de nuevas ontologías especializadas. En cambio, durante el transcurso del siglo xxi, la consolidación de los métodos tradicionales de aprendizaje automático redujo considerablemente la reusabilidad de WordNet, ya que estos métodos no suelen precisar recursos lexicográficos a gran escala.

Tras este auge y caída de la sostenibilidad de WordNet, esta última década es testigo de su resurgimiento, focalizado en el enriquecimiento de vectores

de palabras construidos a través de redes neuronales. De hecho, estos vectores de baja dimensionalidad contribuyen al desarrollo de modelos híbridos o unificados de sistemas neurosimbólicos, donde solo la sinergia de los enfoques simbólico y subsimbólico puede representar un verdadero avance en las aplicaciones de comprensión del lenguaje natural.

Agradecimientos

Esta publicación es parte del proyecto de I+D+i PID2020-112827GB-I00 financiado por MCIN/AEI/10.13039/501100011033 y del proyecto SMARTLAGOON [referencia 101017861] financiado por el Programa Horizonte 2020 de la Unión Europea.

Bibliografía

Agirre, Eneko & Soroa, Aitor. 2009. Personalizing PageRank for word sense disambiguation. En *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Atenas: Association for Computational Linguistics, 33-41. doi: https://doi.org/10.3115/1609067.1609070

Bader, Sebastian & Hitzler, Pascal. 2005. Dimensions of neuralsymbolic integration - A structured survey. En Artemov, Sergei; Barringer, Howard; d'Avila Garcez, Artur; Lamb, Luis C., & Woods, John (eds.) *We will show them: Essays in honour of Dov Gabbay*, vol. 1. Londres: King's College Publications, 167-194.

Buitelaar, Paul. 2000. Reducing lexical semantic complexity with systematic polysemous classes and underspecification. En Bagga, Amit; Pustejovsky, James, & Zadrozny, Wlodek (eds.) *Proceedings of the NAACL-ANLP 2000 Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems*. Seattle: Association for Computational Linguistics, 14-19. doi: https://doi.org/10.3115/1117543.1117546

Carlson, Andrew; Betteridge, Justin; Kisiel, Bryan; Settles, Burr; Hruschka, Estevam R., & Mitchell, Tom M. 2010. Toward an architecture for never-ending language learning. En *Proceedings of the 24th AAAI Conference on Artificial Intelligence*. AAAI Press: Palo Alto, 1306-1313. doi: https://doi.org/10.1609/aaai.v24i1.7519

Colman, Lut. 2016. Sustainable lexicography: Where to go from here with the ANW (Algemeen Nederlands Woordenboek, An Online General Language Dictionary of Contemporary Dutch)? *International Journal of Lexicography* 29(2): 139-155. doi: https://doi.org/10.1093/ijl/ecw008

Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton, & Toutanova, Kristina. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding.

En Burstein, Jill; Doran, Christy, & Solorio, Thamar (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1. Minneapolis: Association for Computational Linguistics, 4171-4186.

Esuli, Andrea & Sebastiani, Fabrizio. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. En Calzolari, Nicoletta; Choukri, Khalid; Gangemi, Aldo; Maegaard, Bente; Mariani, Joseph; Odijk, Jan, & Tapias, Daniel (eds.) *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. Génova: European Language Resources Association, 417-422.

Fellbaum, Christiane. 1998a. A semantic network of English: The mother of all wordnets. *Computers and the Humanities* 32(2-3): 209-220. doi: https://doi.org/10.1023/A:1001181927857

Fellbaum, Christiane (ed.). 1998b. *WordNet: An electronic lexical database*. Cambridge-Massachusetts: the MIT Press. doi: https://doi.org/10.7551/mitpress/7287.001.0001

Gangemi, Aldo; Guarino, Nicola; Masolo, Claudio; Oltramari, Alessandro & Schneider, Luc. 2002. Sweetening ontologies with DOLCE. En Gómez-Pérez, Asunción & Benjamins, V. Richard (eds.) *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*. Berlín / Heidelberg: Springer, 166-181. doi: https://doi.org/10.1007/3-540-45810-7_18

Goikoetxea, Josu; Soroa, Aitor, & Agirre, Eneko. 2015. Random walks and neural network language models on knowledge bases. En *Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver: Association for Computational Linguistics, 1434-1439. doi: https://doi.org/10.3115/v1/N15-1165

Gong, Zhiguo; Cheang, Chan Wa, & Leong Hou, U. 2005. Web query expansion by WordNet. En Andersen, Kim Viborg; Debenham, John, & Wagner, Roland (eds.) *Proceedings of the 16th International Conference on Database and Expert Systems Applications*. Berlín / Heidelberg: Springer, 166-175. doi: https://doi.org/10.1007/11546924_17

Gruber, Thomas R. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5(2), 199-220.

Jiang, Jay J. & Conrath, David W. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. En Chen, Keh-Jiann; Huang, Chu-Ren & Sproat, Richard (eds.) *Proceedings of the 10th Research on Computational Linguistics International Conference*. Taipei: Association for Computational Linguistics and Chinese Language Processing, 19-33.

Leacock, Claudia & Chodorow, Martin. 1998. Combining local context and WordNet similarity for word sense identification. En Fellbaum, Christiane (ed.) *WordNet: An electronic lexical database*. Cambridge / Massachusetts: the MIT Press, 265-283.

Lin, Dekang. 1998. An information-theoretic definition of similarity. En Shavlik, Jude W. (ed.) *Proceedings of the 15th International Conference on Machine Learning*. San Francisco: Morgan Kaufmann, 296-304.

Magnini, Bernardo & Cavaglià, Gabriela. 2000. Integrating subject field codes into WordNet. En *Proceedings of the Second International Conference on Language Resources and Evaluation*. Atenas: European Language Resources Association, 1413-1418.

Mihaylov, Todor & Frank, Anette. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. En Gurevych, Iryna & Miyao, Yusuke (eds.) *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Melbourne: Association for Computational Linguistics, 821-832. doi: https://doi.org/10.18653/v1/P18-1076

Miller, George A. 1990. Nouns in WordNet: A lexical inheritance system. *International Journal of Lexicography* 3(4), 245-264. doi: https://doi.org/10.1093/ijl/3.4.245

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38(11): 39-41. doi: https://doi.org/10.1145/219717.219748

Miller, George A.; Beckwith, Richard; Fellbaum, Christiane; Gross, Derek, & Miller, Katherine J. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography* 3(4): 235-244. doi: https://doi.org/10.1093/ijl/3.4.235

Moldovan, Dan I. & Mihalcea, Rada. 2000. Using WordNet and lexical operators to improve Internet searches. *IEEE Internet Computing* 4(1): 34-43. doi: https://doi.org/10.1109/4236.815847

Moreno, Antonio; Torre, Doroteo; Valverde, Ana, & Campillos, Leonardo. 2019. *Estudio sobre datos reutilizables como recursos lingüísticos*. Ministerio de Economía y Empresa. Gobierno de España. https://plantl.mineco.gob.es/tecnologias-lenguaje/actividades/estudios/Paginas/datos-reutilizables-como-recurso-linguitico.aspx

Navigli, Roberto & Velardi, Paola. 2002. Semantic interpretation of terminological strings. En *Proceedings of the 6th International Conference on Terminology and Knowledge Engineering*. Berlín: Springer, 95-100.

Niles, Ian. 2003. Mapping WordNet to the SUMO Ontology. En *Proceedings of the 2003 International Conference on Information and Knowledge Engineering*. Las Vegas: CSREA Press, 23-26.

Palmer, Martha. 1998. Are WordNet sense distinctions appropriate for computational lexicons? En *Proceedings of the Senseval Workshop on Word Sense Disambiguation (Siglex98)*. Brighton: Association for Computational Linguistics, 1-8.

Peters, Wim & Peters, Ivonne. 2000. Lexicalised systematic polysemy in WordNet. En *Proceedings of the Second International Conference on Language Resources and Evaluation*. Atenas: European Language Resources Association, 1-7.

Ponzetto, Simone Paolo & Navigli, Roberto. 2010. Knowledge-rich word sense disambiguation rivaling supervised systems. En Hajič, Jan; Carberry, Sandra; Clark, Stephen, & Nivre, Joakim (eds.) *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala: Association for Computational Linguistics, 1522-1531.

Radford, Alec; Wu, Jeffrey; Child, Rewon; Luan, David; Amodei, Dario, & Sutskever, Ilya. 2019. Language models are unsupervised multitask learners. *OpenAI*

blog 1(8): 9. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. En *Proceedings of the 14th International Joint Conference for Artificial Intelligence*. San Francisco: Morgan Kaufmann, 448-453.

Rivera-Hernández, Jaime Ernesto; Blanco-Orozco, Napoleón Vicente; Alcántara-Salinas, Graciela; Houbron, Eric Pascal, & Pérez-Sato, Juan Antonio. 2017. ¿Desarrollo sostenible o sustentable? La controversia de un concepto. *Posgrado y Sociedad - Revista Electrónica del Sistema de Estudios de Posgrado* 15(1): 57-67. doi: https://doi.org/10.22458/rpys.v15i1.1825

Strapparava, Carlo & Valitutti, Alessandro. 2004. WordNet Affect: An Affective Extension of WordNet. En Lino, Maria Teresa; Xavier, Maria Francisca; Ferreira, Fátima; Costa, Rute & Silva, Raquel (eds.) *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lisboa: European Language Resources Association, 1083-1086.

Tudorache, Tania. 2020. Ontology engineering: Current state, challenges, and future directions. *Semantic Web* 11(1): 125-138. doi: https://doi.org/10.3233/SW-190382

Venters, Colin C.; Lau, Lydia; Griffiths, Michael K.; Holmes, Violeta; Ward, Rupert R.; Jay, Caroline; Dibsdale, Charlie E., & Xu, Jie. 2014. The blind men and the elephant: Towards an empirical evaluation framework for software sustainability. *Journal of Open Research Software* 2(1): 1-6. doi: https://doi.org/10.5334/jors.ao

Vossen, Piek. 1998. Introduction to EuroWordNet. *Computers and the Humanities* 32(2-3): 73-89. doi: https://doi.org/10.1023/A:1001175424222

Wang, Chao & Jiang, Hui. 2019. Explicit utilization of general knowledge in machine reading comprehension. En Korhonen, Anna; Traum, David & Màrquez, Lluís (eds.) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florencia: Association for Computational Linguistics, 2263-2272. doi: https://doi.org/10.18653/v1/P19-1219

Wu, Zhibiao & Palmer, Martha. 1994. Verb semantics and lexical selection. En Pustejovsky, James (ed.) *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. Las Cruces: Association for Computational Linguistics, 133-138. doi: https://doi.org/10.3115/981732.981751

Yang, An; Wang, Quan; Liu, Jing; Liu, Kai; Lyu, Yajuan; Wu, Hua; She, Qiaoqiao, & Li, Sujian. 2019. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. En Korhonen, Anna; Traum, David, & Màrquez, Lluís (eds.) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florencia: Association for Computational Linguistics, 2346-2357. doi: https://doi.org/10.18653/v1/P19-1226

Yang, Xinran & Tiddi, Ilaria. 2020. Creative storytelling with language models and knowledge graphs. En Conrad, Stefan & Tiddi, Ilaria (eds.) *Proceedings of the CIKM 2020 Workshops co-located with 29th ACM International Conference on Information and Knowledge Management*. Galway: CEUR Workshop Proceedings, 1-9.